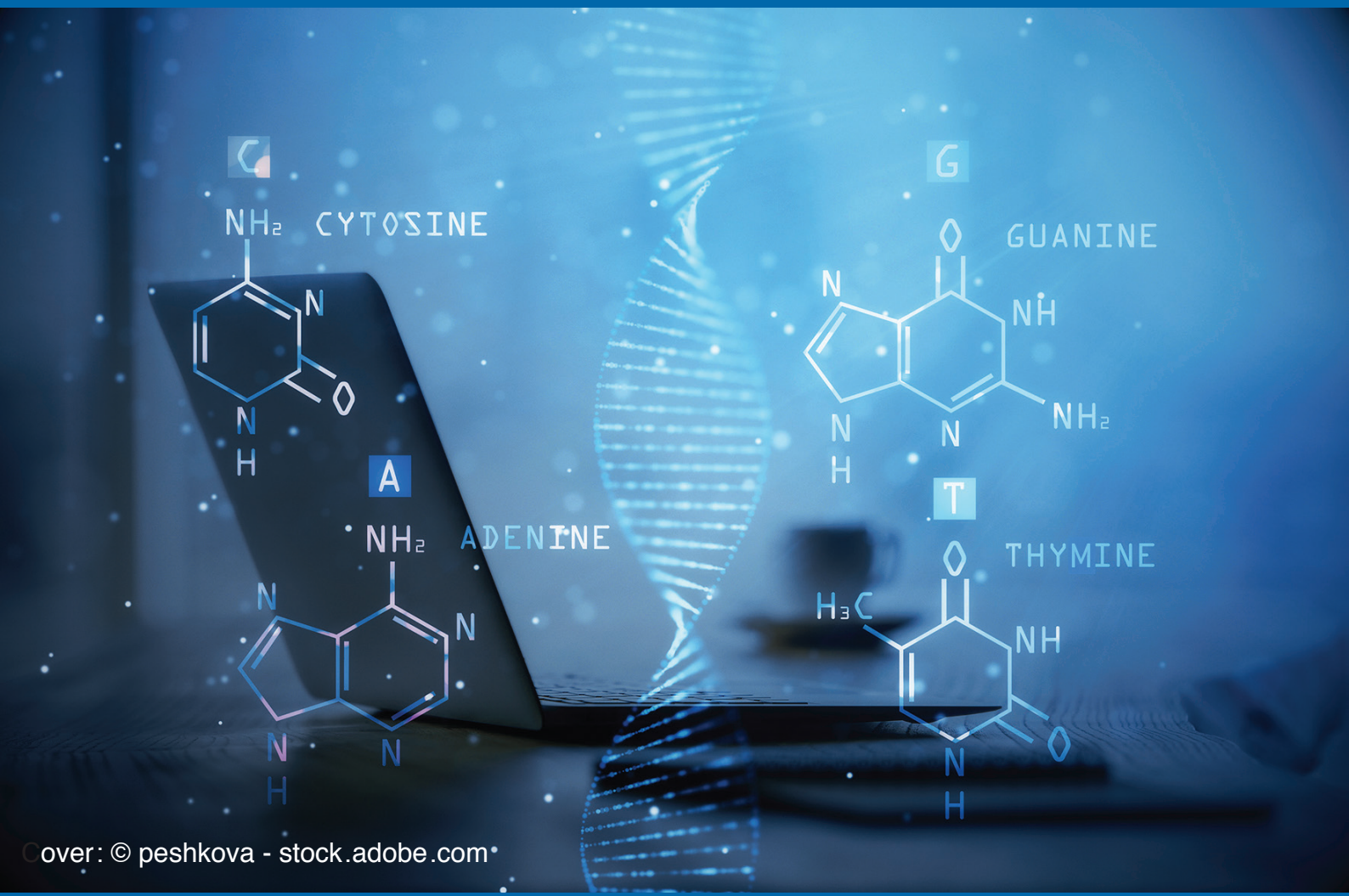
NATIONAL ACADEMY OF TECHNOLOGIES OF FRANCE

SHARING A REASONED, CHOSEN PROGRESS

Archiving big data beyond 2040: DNA as a candidate

DNA : reading , writing , storing information

A National Academy of Technologies of France report



Cover: © peshkova - stock.adobe.com

NATIONAL ACADEMY OF TECHNOLOGIES OF FRANCE

SHARING A REASONED, CHOSEN PROGRESS

Archiving big data beyond 2040: DNA as a candidate

—

"DNA: reading, writing, storing information"

—

A report by the National Academy of Technologies of France

September 2020

Académie des technologies
Le Ponant
19, rue Leblanc
Bâtiment A
75015 Paris
+33(0)1 53 85 44 44
secretariat@academie-technologies .fr
www .academie-technologies .fr

ISBN: 979-10-97579-19-7

Content

Abstract

The storage and archiving of digital big data by the current approach based on datacenters will not be sustainable beyond 2040. There is therefore an urgent need to focus sustained R&D efforts on the advent of alternative approaches, none of which are currently mature enough.

The Global DataSphere (GDS) was estimated in 2018 to be 33 thousand billion billion ($33x10^{21}$) of characters (bytes), which is of the same order as the estimated number of grains of sand on earth. This data comes not only from research and industry, but also from our personal and professional connections, books, videos and photos, medical information. It will be further increased in the near-future by autonomous cars, sensors, remote monitoring, virtual reality, remote diagnosis and surgery. The GDS is increasing by a factor of about a thousand every twenty years.

Most of this data is then stored in several million datacenters (including corporate and cloud datacenters), which operate within transmission networks. These centers and networks already consume about 2 % of the electricity in the developed world. Their construction and operating costs are globally in the order of one trillion euros. They cover one millionth of the world's land surface and, at the current rate, would cover one thousandth of it by 2040.

The storage technologies used by these centers are rapidly becoming obsolete in terms of format, read/write devices, and also of the storage medium itself, which requires making copies every five to seven years to ensure data integrity. They also pose increasing problems in the supply of scarce resources such as electronic-grade silicon.

An attracting alternative is offered by molecular carriers of information, such as DNA (used as a chemical rather than a biological agent) or certain, very promising, non-DNA heteropolymers. In principle, DNA should allow information densities ten million times higher than those of traditional memories: the entire current GDS would fit in a van. DNA is stable at ordinary temperature for several millennia without energy consumption. It can be easily multiplied or destroyed at will. Some calculations can be physically implemented with DNA fragments. Finally, DNA as a support will not become obsolete because it constitutes our hereditary material.

To archive and retrieve data in DNA, five steps must be followed: i) coding the binary data file using the DNA alphabet which has four letters, ii) writing, iii) storing, iv) reading that DNA, and finally v) decoding the information it contains. A prototype performing these steps has been in operation since March 2019 at Microsoft Corp. in the United States. The performance of such prototypes needs improving by several orders of magnitude to become economically viable: about a thousand-fold for the cost and speed of reading, and 100 million-fold for the

speed of writing. These factors may seem staggering, but this would be to ignore the speed of progress in DNA technologies which increases at close to a factor of one thousand every five years — much faster than in the electronic and computer fields.

The report ends with a short perspective that allows the reader to form an opinion on the technical and economic aspects of using DNA for archiving big data

Chapter I

Context and motivation

Humanity is accumulating data at an unprecedented and increasing rate. This data is that of our social and cultural life — our family, friendly and professional connections, our books, videos and photos, our medical data — and that of scientific research, industry, etc. We sometimes call them ***big data***[1]. And much more is to come: autonomous cars, sensors and other connected objects, remote surveillance, virtual reality, medical deserts compensated by consultation, diagnosis (telemedicine) and even surgery at a distance. In 2025, it is estimated that three quarters of humanity will be connected, and that we will each interact with data every 18 seconds on average[2]. Almost all this data is processed by computers, which means that it must be represented by long sequences of two elements, marked '0' and '1': in other words, binary data. These long sequences are often subdivided into groups of eight successive '0' or '1' elements, which are called "***bytes***" (B).

The Global DataSphere (GDS)

The set of digital data created by mankind, the "Global DataSphere" (GDS), contains about as many characters (bytes)[3] as the number of observable stars in the universe, or the estimated number of grains of sand on earth. This GDS was estimated in 2018 to be 33 Zettabytes (33 ZB; or 33 x 10^{21} characters[4]). It doubles every two to three years and will reach about 175 ZB in

1 The Académie des technologies (NATF) has looked at big data on several occasions, and two recent publications have resulted, one on the technological-strategical aspects ("***Big data: a paradigm shift may hide another one",*** EDP Sciences 2015), and the other on the ethical aspects ("***Big data - Ethical issues***", Académie des technologies 2019).

2 HiPEAC Vision 2015 (European Commission, FP7, 2015).

3 A "byte" is a sequence of **8** "bits". A bit takes only **2** values usually designated by the digits '0' and '1' (hence the term "binary" or "numerical" coding). So there are 256 (2^8) possible bytes. We consider here that a byte represents one character (a letter, a digit, or a symbol) out of the 256 that are possible. For example the byte '00100011' usually encodes the character '#'.

4 Prefixes of the International system of units, and corresponding numbers:

K	kilo	10^3	1 000
M	mega	10^6	1 000 000
G	giga	10^9	1 000 000 000
T	tera	10^{12}	1 000 000 000 000
P	peta	10^{15}	1 000 000 000 000 000
E	exa	10^{18}	1 000 000 000 000 000 000
Z	zetta	10^{21}	1 000 000 000 000 000 000 000
Y	yotta	10^{24}	1 000 000 000 000 000 000 000 000

2025[5]. For instance, approximately 400 hours of video (200 GB) are added every minute to the GDS. A different example, from the European Organisation for Nuclear Research (CERN), is the more than 100 PB of data that it has produced and that must be preserved for future generations of physicists. At the current rate, the GDS will reach more than 5,000 ZB in 2040. Another way of looking at such a high number is to say that it would take 50 million years to download this GDS with a medium-speed Internet connection.

Most of the data created by humanity is stored in long term storage installations. Centralized storage is rapidly increasing, especially in datacenters and now in the "cloud", relative to which, local storage on a computer or telephone is decreasing. The cloud allows individual users to take advantage of on-demand computing resources, and to increase the level of automation through server virtualization. By 2021, the cloud will store as much information as traditional datacenters. In the remainder of this report, we will consider all datacenters, including the cloud. Dedicated buildings for this centralized storage are constantly being constructed throughout the world, often in cold countries because this storage is a large consumer of electricity and requires significant cooling. The available storage capacity in datacenters grows slower than the GDS, and represents in 2020 only 40 % of the binary data created, a percentage that tends to decrease[6].

Datacenters

Consider a small datacenter of 300 m^2 built in 2008, and comprising only 2,000 computer servers for a total power of one megawatt (MW). Over its 20-year lifespan[7], it will have used 66 metric tons of copper, 15 metric tons of plastics, 33 metric tons of aluminum and 152 metric tons of steel. Each year, it will have been used 23 million liters of water. Including server cooling, it will have consumed 18 million kilowatt-hours (0.018 terawatt-hours or 0.018 TWh) of electricity[8]. In comparison, a large datacenter costs several billion euros in investment, occupies one million m^2, and comprises one million servers[9]. It consumes one gigawatt (GW) of electricity (of which about 40-50 % is needed for cooling), or about 10 TWh per year, which is more than the consumption of a city of 100,000 inhabitants. Of course, these centers are linked to the rest of the world by large networks of connections, which also consume various resources including electricity[10]. There was a total of 8.6 million datacenters, including those of companies, in the world in 2017,

5 Reinsel D, Gantz J, Rydning J (2018). The Digitization of the World - From Edge to Core (International Data Corporation & SeaGate).

6 International Data Corporation digital universe study - https://www.idc.com/

7 https://www.lemonde.fr/planete/article/2011/07/07/les-data-centers-de-vraies-usines-electriques_1546181_3244.html

8 Computer Guide (2008). https://www.guideinformatique.com/dossiers-actualites-informatiques/consommation-electrique-des-data-centers-29.html

9 Server density appears lower in the large 2018 center than in the small 2008 center. This reflects the evolution of servers, in terms of architecture, but also in terms of usage: emphasis on calculation and database management in 2008, rather than on internet services in 2018.

10 Davey J (2019). Powering the data revolution (HSBC Global Research).

they occupied a total surface area of more than 170 million m^2, the equivalent of 25,000 soccer fields[11]. This area represents about one millionth of the world's land mass, approximately 150 million km^2. If the current rate of doubling every two years were to continue, one thousandth of the land surface would be occupied by these centers before 2040. However, this is probably an overestimate because the efficiency of datacenters is continually increasing in terms of energy[12] and area. The consumption of electricity by the datacenters and their connection networks is roughly equivalent to that of the fifth largest electricity consuming country in the world, that is, somewhere between India and Japan. It was estimated that in 2007, datacenters and their associated connection networks consumed 623 TWh worldwide and generated 423 megatons of CO_2 equivalent emissions. In 2012, they were responsible for 2 % of the greenhouse gases produced globally. The annual global invesment to build new datacenters is in the order of several tens of billions of US dollars[13]. For example, Google alone has annually invested US$10 billion over the period 2015-2017.

Aerial photograph of a datacenter (Farrat Ireland).

Credit: The agency creative

Provisional Conclusion

The information given above, albeit limited, clearly shows that the growth of digital data, at the current rate and with the current technology, is not sustainable beyond about 2040.

Is there a technology that would allow us to archive all the big data in a few cubic meters

11 Reinsel D, Gantz J, Rydning J (2018). The Digitization of the World - From Edge to Core (International Data Corporation & SeaGate). http://hebergement-et-infrastructure.fr/actualites-et-innovations/8-6-million-de-datacenters-dans-le-monde-en-2017 https://www.businesswire.com/news/home/20141110005018/en/IDC-Finds-Growth-

12 https://www.datacenterknowledge.com/archives/2016/06/27/heres-how-much-energy-all-us-data-centers-consume

13 Cook G (2012). *How clean is your cloud?* (Greenpeace International) https://www.greenpeace.org/archive-international/Global/international/publications/climate/2012/iCoal/HowCleanisYourCloud.pdf

with virtually no energy expenditure? The conclusion of the report presented here is that indeed such a technology exists. This conclusion is based on the work of NATF cross-disciplinary working group (2018-2020) ***DNA: reading, writing, storing information***, which auditioned 26 world specialists.

Before presenting this technology, we summarize the state of the art in the field of binary data storage and archiving, highlighting its limitations and prospects.

Chapter II

State of the art in data storage and archiving

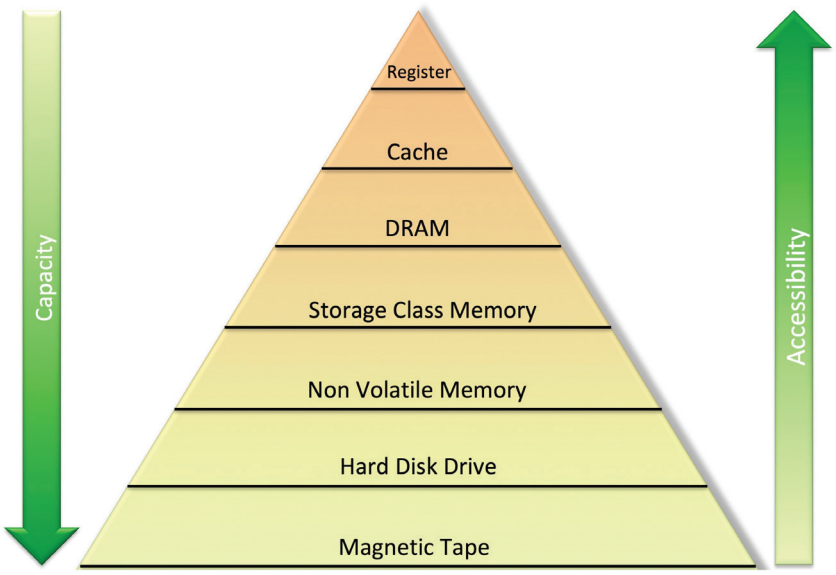
Issues

The issue of storing and archiving digital information is dependent on several parameters:

- the amount of information and its coding;
- the duration of short-term "storage" or "backup", or long-term "archiving", *i.e.* how long the information will be kept;
- the frequency of access to information;
- the monetary and environmental cost of producing, retaining and managing information.

Memory hierarchy in computer architecture

In modern computing, memory is an electronic device used to store information. It is organized hierarchically. This hierarchy can be represented as a pyramid composed of several levels defined by data access speed and storage capacity. As the capacity of the storage medium increases, so too does the time required to access the data.



Pyramid of memory types in computer systems.
Credit: François Képès and Carlo Reita.

At the top of the pyramid is the register, which is an internal memory in the processor. It is the fastest memory in a computer (0.1 nanosecond for data access) but has the highest manufacturing cost and is therefore reserved for a very small amount of data (a few thousand bytes).

Below the registry is the cache memory. This memory stores frequently accessed information for a short period of time. These memories are very fast (1 to 10 nanoseconds for data access) but also very expensive and reserved for a small amount of data (a few kilobytes - KB - to megabytes - MB).

Below the cache memory is the random-access memory, in which the information processed by the computer device is stored and then erased. This is the main storage space of the microprocessor, but its contents disappear when the computer is turned off. It is a relatively fast memory (10 to 1 000 nanoseconds for data access) and is reserved for a few gigabytes (GB) of data.

Finally, there is mass memory, which includes:

– hard disks and Flash memories, which store a large amount of information (several Terabytes - TB) over the long term;
– magnetic tapes used for very long-term archiving (> 10 years) of information. The cost of mass storage is relatively low, but its access speed is lower than other types of memory.

Information storage technologies in electronics

To meet the needs of digital information storage, a wide range of technologies is employed: physical, magnetic, optical, Flash memory or virtual storage. After a period when storage on optical media, such as CDs or DVDs, had found a place in this highly competitive market, today almost all storage in the field of information technology is based on magnetic or charge storage.

Magnetic storage technologies

Supports

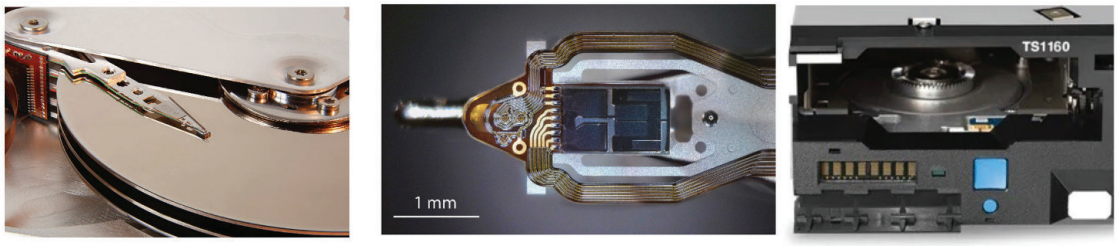
These media include magnetic tape, hard disk and MRAM memories[14]. They have a large storage capacity and an information retention time in the order of a decade. Thus, it is the technology of choice for data archiving. Tape has the highest storage density but intrinsically sequential data access limits read and write speeds. Disk has much faster random access but at the cost of lower density.

Principle

Writing on these media is done by magnetization and reading is magnetic. The current in a coil generates a magnetic field which induces dipoles in a substrate (disk or tape) coated with a magnetic material. Reading is done by measuring the current induced by the movement of the

14 MRAM: magneto-resistive random access memory

dipoles. Over time, this technology has improved due to the quality of the magnetic materials and the miniaturization of the read and write heads.



Magnetic storage media

Credits from left to right:
Eric Gaba - Wikimedia Commons user: Sting
Roman Starkov - Creative Commons Attribution-Share Alike 4.0
https://www.ibm.com/it-infrastructure/storage/tape/drives

Charge storage technologies

Media

This technology includes SRAM[15], DRAM[16] and Flash memory. Charge storage devices are easily integrated into complex electronic circuits. Performance requirements are:

- the speed of reading, writing and accessing information;
- storage capacity;
- the number of data read and write cycles during the life cycle.

Principle

The information stored on these media is represented by the state of charge of a capacity (loaded/not loaded). This charge can be read directly or affect the conduction of a transistor. Over time, this technology has become miniaturized.

The case of Flash memories

There are two classes of charge storage devices: i) volatile[17] (DRAM and SRAM) and ii) non-volatile (Flash) memories. Flash memory is the most widely used non-volatile memory today (static disks, memory cards for portable devices, USB keys).

The lifespan of a Flash memory is calculated by the number of data writes and erasures (typically 10 thousand to 1 million) that the media may undergo before degrading. Its particularly regular structure has allowed a gradual reduction in size, in order to obtain storage densities comparable to those of hard disks. In addition, microelectronic mass production processes have drastically reduced costs.

15 SRAM: static random access memory
16 DRAM: dynamic random access memory
17 Loss of information in the absence of power supply.

Currently, NAND 3D[18] Flash memory systems are being developed. Instead of being arranged on flat surfaces, the storage cells are arranged on folded surfaces, which allows for many more cells per unit volume (up to 72 stacking levels) and thus increases storage density.

Today, it is almost impossible to further reduce the size of Flash memory cells. Thus, researchers are working on new memories based on the change in resistance.

Comparison of storage technologies in electronics

Storage technologies and their evolution are compared in this table:

	Type of storage	Magnetic storage		Charge storage			
	Supports	Magnetic tapes	Magnetic Hard Disk Drives (HDD)	volatile memories		non-volatile memories	
				SRAM	DRAM	PCM[1]	Flash (NAND)
In 2008	Storage density (Gbits/cm2)	0.14	59	N/A	N/A	N/A	31
	Reading time (ns)	N/A	N/A	N/A	N/A	N/A	N/A
	Writing time (ns)	N/A	N/A	N/A	N/A	N/A	N/A
	Duration of storage	N/A	N/A	N/A	N/A	N/A	N/A
	Endurance (cycles)	N/A	N/A	N/A	N/A	N/A	N/A
	Production cost ($/GB)	0.091	0.272	N/A	N/A	N/A	3.33
	Revenues generated ($)	1	34	N/A	N/A	N/A	10.1
In 2016	Storage density (Gbits/cm2)	3.89	170	N/A	N/A	N/A	310
	Reading time (ns)	N/A	$5\text{-}8\text{x}10^6$	<10-50	10-50	20-70	25 000
	Writing time (ns)	N/A	$5\text{-}8\text{x}10^6$	<10-50	10-50	50-500	200 000
	Duration of storage	> 10 years	10 years	< second	< second	<10 years	10 years
	Endurance (cycles)	N/A	10^{15}	$>10^{17}$	10^{17}	$10^7\text{-}10^8$	$10^4\text{-}10^6$
	Production cost ($/GB)	0.016	0.039	$10^2\text{-}10^3$	10	1	0.32
	Revenues generated ($ billion)	0.65	26.8	N/A	N/A	N/A	38.7

1 PCM: a form of non-volatile random access memory.

Two conclusions can be drawn:

– Magnetic tape remains today the best compromise for long-term data archiving; indeed, the storage duration is the highest and the production cost the cheapest out of all the storage technologies; in addition, tape consumes less than 1 % of the total electricity in a datacenter;
– Magnetic hard disk drives (HDD) and solid-state (Flash) are the best trade-offs for mass storage as they have the highest storage density while providing fast access to stored data.

18 NAND: logic gate "NOT-AND"

Conclusion

The so-called *traditional* binary data storage and archiving technologies cited so far are all close to their theoretical optimum. In other words, future gains will be small in terms of density, access speed, longevity, durability and costs. Pursuing Moore's Law[19] is an increasingly difficult challenge. It should also be noted that the present production of electronic-grade silicon (an abundant component of hard disks) is 100-fold below anticipated needs.

Moreover, although the fundamental principles such as magnetism on which the traditional media are based remain unchanged, these media are rapidly becoming obsolete in three ways[20].

- **The storage format:** for example, the use of 3.5-inch floppy disks was gradually phased out between 2000 and 2010. Magnetic tape technology continues to evolve, which results in new generations being incompatible with the previous ones.
- **The read/write device:** to use the same examples, functional 3.5-inch floppy disk drives have become rare. Magnetic tape drives continue to change.
- **The medium itself:** because of their physical nature, all storage media have a limited lifespan, at most about ten years, with the risk thereafter of losing information. To protect against this, data must be regularly checked and recopied in order to save it. For example, due to the degradation of magnetic signals over time, it is standard practice to recopy tapes and disks every five to ten years.

Finally, it should be remembered that these traditional systems consume a lot of energy not only in order to function but also as a consequence of the three unavoidable forms of obsolescence described above.

In conclusion, the world is facing a serious data storage problem that cannot be solved by current technologies. These drawbacks are not shared by DNA information storage technology, which we discuss below.

19 *Moore's Law* was expressed in 1965 in Electronics magazine by Gordon E. Moore, one of the three founders of Intel. This empirical law states that the number of microprocessor transistors (rather than the more complex integrated circuits) on a silicon chip doubles every two years. Another empirical law seems to govern the evolution of storage capacities. Here too, so far, the doubling occurs approximately every two years.

20 Hourcade J-C, Laloë F, Spitz E (2010). *Longévité de l'information numérique*. Académie des technologies & Académie des sciences (EDP Sciences).

Chapter III

Storage of digital information in DNA

New methods of information storage on the molecular scale are being envisaged so as to increase storage capacity, reduce media size, and increase data retention time. Recent advances in DNA reading and writing technologies have led researchers to consider this natural polymer as a medium for digital archives. DNA can store **1 bit per about 50 atoms,** whereas magnetic storage requires about one million atoms.

Brief history of data storage in DNA

Richard Feynman in 1959, and Mikhail Neiman in 1964, were the first to propose that DNA could be used to store digital information. But it was only in 1977 that the first methods for reading DNA and in 1983 for writing DNA were developed.

In 1988, Joe Davis[21] designed and synthesized, for the first time, an 18-nucleotide DNA fragment containing a digitized message symbolizing the "MicroVenus" icon, which he then transferred to an intestinal bacterium, *Escherichia coli*.

In 2012, George M. Church's group (Harvard University, United States of America) stored 0.6 MB of information in the form of synthetic DNA fragments[22]. In 2013, Nick Goldman's group (European Bioinformatics Institute, United Kingdom) converted four computer files into DNA sequences totalling 0.7 MB[23]. The information was transcribed back without errors.

In 2018, Microsoft Corp. and the University of Washington in the United States of America stored 1 GB of information[24] from various types of files in DNA. They have the record ever since.

In 2024, it is planned to archive 1 TB (equivalent to about 1,000 films) in 24 hours at a cost of US$ 1,000[25].

21 https://en.wikipedia.org/wiki/Joe_Davis_(artist)

22 Church M. G, Gao Y, Kosuri S (2012). Next-Generation Digital Information Storage in DNA. *Science* 337:1628.

23 Goldman N, Bertone P, Chen S, Dessimoz C, Leproust EM (2013). Toward practical, high-capacity, low maintenance information storage in synthesized DNA. *Nature* 494:77-90.

24 Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA (2019). *Nat Rev Genet* 20: 456–466.

25 *Intelligence Advanced Research Projects Activity* [IARPA] (2020). https://www.dni.gov/index.php/newsroom/press-releases/item/2086-iarpa-announces-launch-of-the-molecular-information-storage-program

DNA: a high-performance storage medium

DNA

DNA is one of the carriers of hereditary information in our cells and, more generally, in the living world, as recalled in the following box.

DNA and biological information

The issue of biological information became a scientific subject in 1972 with the publication of Henri Atlan's book[1].

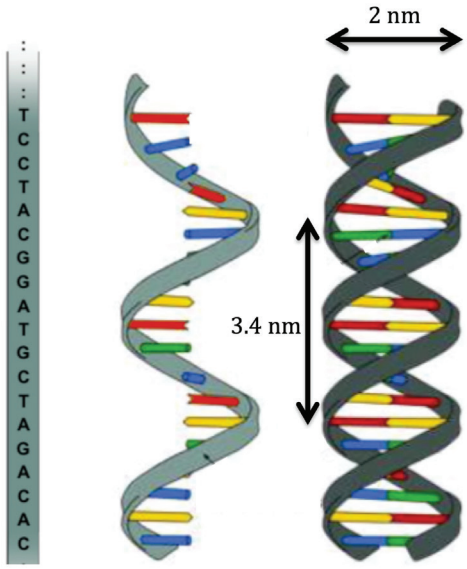
- Part of this information lies in the dynamic organization that the living cell actively maintains and passes on to its progeny.
- A second part, also inheritable, resides in all its epigenetic states (state of a switch or oscillator etc.).
- A third part is carried by its chromosomes, in the form of the DNA sequences they contain. This last part is by far the most easily describable.

Despite the relative ease of quantifying the capacities of DNA to carry information in an artificial system, it is currently impossible to do this for a living system or, indeed, for parts of that system. In an artificial system, the four letter alphabet (A, C, G, T) of DNA can be used directly as a binary information carrier encoding two bits, according to a principle similar to that of a two-letter (0 and 1) magnetic carrier encoding one bit. This principle is, of course, extremely reductive in relation to the complex, malleable and multiform role played by hereditary material in a living cell. Such complexity is not a problem when DNA is used as a carrier of binary — or, more exactly, quaternary — information, in a test tube outside the cell (*i.e.*, *in vitro*). In what follows, DNA will therefore be presented as a polymer with interesting properties that can be manipulated easily thanks to the remarkable tools that biologists have been making since the middle of the 20th century.

1 Atlan H. (1972) *L'organisation biologique et la théorie de l'information* (Seuil).

As a reminder, deoxyribonucleic acid (DNA) consists of two antiparallel strands wound around each other to form a double helix structure. Each DNA strand is a linear (unbranched) polymer[26] composed of an assembly of nucleotides. Each nucleotide is composed of one of the four nitrogenous bases, adenine (A), guanine (G), thymine (T), cytosine (C), linked to a deoxyribose sugar that is itself linked to a phosphate group. The nucleic bases of one DNA strand can interact with the nucleic bases of the other DNA strand through hydrogen bonds by respecting pairing rules. Adenine and thymine pair with one another by two hydrogen bonds, while guanine and cytosine pair by three hydrogen bonds.

26 A polymer is a large molecule made up of many repeating subunits, called monomers. If all the subunits are identical, it is called a ***homopolymer***. If the subunits are not all identical, it is called a ***heteropolymer*** or ***copolymer***. A copolymer is a polymer resulting from the copolymerization of at least two chemically different types of monomers.



DNA representations. On the right, a double-stranded DNA helix (showing AT and CG pairings); in the middle, a single-stranded DNA; on the left, the sequence of the same single strand, now unwound for a linear representation.

Credit: adapted by François Képès from the presentation by Nick Goldman (EBI).

Since the work of Friedrich Miescher in 1869, DNA can be manipulated outside cells, that is, *in vitro;* it is mainly *in vitro* that its use for storing digital data has been considered and where it has many advantages over traditional systems.

Advantages of storing binary information in DNA

Information density

The information density of DNA is about 10 million times greater than that of the best traditional systems. In principle, DNA can store half a ZB of information per gram (g). Thus, researchers estimate that the entire GDS could currently fit in less than 100 g of DNA, assuming that only a single DNA molecule was synthesized for each piece of information. In practice, however, many identical copies would be synthesized.

- In addition, parts of this DNA would have to carry quality control and indexing signals along with the data.
- Finally, the DNA must be preserved in macroscopic containers[27].

Taking these density losses into account, it can be estimated more realistically that the GDS, stored in DNA, would fit in a van.

27 Organick L, Ang SD, Chen YJ, Lopez R, Yekhanin S, Makarychev K, Racz MZ, Kamath G, Gopalan P, Nguyen B, Takahashi CN, Newman S, Parker HY, Rashtchian C, Stewart K, Gupta G, Carlson R, Mulligan J, Carmean D, Seelig G, Ceze L, Strauss K (2018). Random access in large-scale DNA data storage. *Nature Biotechnology* 36(3) :242-248.

Consumption

Storage of DNA at room temperatures does not involve any consumption of resources, and DNA operations are much less energy intensive than in electronics; a 1,000-fold gain has been mentioned[28].

Longevity

The longevity of DNA is about 10 thousand times that of traditional media. DNA molecules more than 560,000 years old have been analyzed from ancient samples[29]. In the laboratory, a half-life of 52,000 years has been demonstrated by artificially accelerating its ageing[30].

Obsolescence

The obsolescence of DNA as a carrier of information will not occur as long as humans have the technology to write and read DNA molecules, which is an integral part of modern medicine. The issue of coding and decoding will be resolved when a standard will emerge.

Copying or multiplication

The copying or replication of DNA, and therefore of the information it contains, is fast and inexpensive. This is because DNA is naturally replicated in cells before they divide. Such replication can be achieved *in vitro* by the "polymerase chain reaction" (PCR). PCR is performed using two primers, which are single-stranded DNA fragments of about twenty nucleotides that bind by complementarity to two specific zones flanking the region to be amplified. Thus, a single DNA fragment can be duplicated serially by benchtop thermocyclers, generating by this exponential process several billion copies in a few hours for a fraction of a euro[31]. This represents a considerable advantage over the cumbersome and costly duplication of data on traditional media.

Destruction at will

The destruction of DNA at will is easily and quickly achievable. Indeed, although this macromolecule is chemically not very reactive, cells possess protein catalysts (enzymes called DNases) that are extremely effective in reducing DNA to its nucleotide components. DNases are commercially available at a modest cost. More brutal but less sophisticated physical treatments, for example exposure to high temperature, destroy DNA in a fraction of a second, even when it is protected inside a macroscopic container.

28 https://www.iarpa.gov/index.php/research-programs/mist

29 See:

- Orlando L et al. (2006). Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence. *Current Biology* 16, R400-402.
- Orlando L et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* 499, 74–78.

30 Bonnet J, Colotte M, Coudy D, Couallier V, Portier J, Morin B, Tuffet S (2010). Chain and conformation stability of solid-state DNA: implications for room temperature storage. *Nucleic Acids Res.* 38(5):1531-46. http://www.imagene.fr/dnashell-rnashell/dnashell/

31 Each cycle duplicates the existing one. So 30 cycles produce 2^{30} copies, or more than a billion.

Computing with DNA

The physico-chemical properties of DNA lend themselves to the direct implementation of certain calculations. The principle of such a calculation (further developed in chapter IV) is to code a combinatorial problem using DNA strands that are custom-built, to simulate the operations needed to find the solution by manipulating these strands using the techniques of molecular biology, and then to read the solution by sequencing[32].

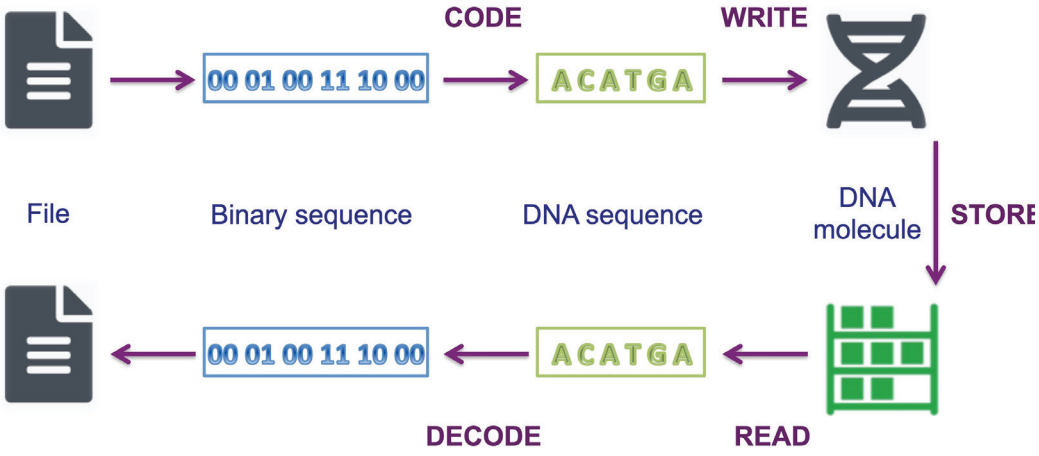
Technologies for information storage in DNA

The technology involves several steps: data coding, DNA writing, storage and reading, and decoding into data. Operations such as modification, amplification or destruction of the DNA can be grafted on these processes.

Coding of information

Principle

As a reminder, the binary system is the numbering system using base 2. The digits of the binary numeration are called "bits", from the English "binary digit". A bit can take two values, conventionally denoted '0' and '1'. Traditional media such as hard disks, USB keys or DVDs store digital data by changing the magnetic, electrical or optical properties of a material to store these 0's and 1's. A byte is a series of eight bits.



Steps in the process of storing big data in DNA. Here are shown for example 12 successive bits extracted from the digital file. These 12 bits are encoded in the form of six nucleotides which are written in succession in a DNA molecule. This DNA is stored on shelf, then read, and the sequence of nucleotides thus obtained is decoded to reconstitute the original digital file.

Credit: François Képès.

32 Adelman LM (1994). Molecular computation of solutions to combinatorial problems, *Science* 266, 5187, 1021-1024.

To store data in DNA, the concept is the same but the process is different. Rather than creating sequences of 0's and 1's, as with digital data, DNA data storage uses nucleotide sequences. The general idea is to assign numerical values to DNA nucleotides. For example, the bit pair 00 could be equivalent to nucleotide A, 01 to C, 10 to G and 11 to T. Thus, a new code is invented, where bits are converted into nucleotides to form a DNA fragment, which is then synthesized *in vitro*. However, more elaborate coding methods are beginning to appear (see an example in chapter IV).

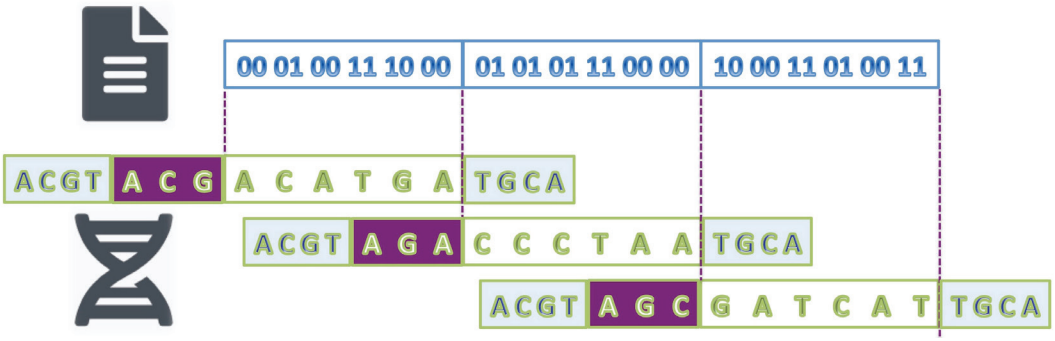
Current DNA synthesis technology is limited to fragments of the order of 200 nucleotides at most, which is very short compared to typical computer files. This is because the longer a strand of DNA, the more difficult it is to construct it chemically. We can therefore draw an analogy between "nucleotide packets" (the short strands of DNA) and byte packets that are sent during an Internet transaction such as sending an e-mail. In both cases, the complete message is correctly reconstructed on arrival from the packets using accompanying information about affiliation, indexing and addressing along with quality controls.

Splitting the computer file into fragments

The computer file is split into fragments of about 20 bytes. Each fragment has an identifier of a few bits at its end. The identifiers allow the fragments to be put in the same order as the information in the digital file.

Conversion of byte fragments to DNA nucleotides

Fragments of about 20 bytes are converted into DNA nucleotides. Each DNA segment has about 200 nucleotides and contains the payload and label. The latter allows the DNA fragments containing information from the same digital files to be grouped together and ordered. It is used for selective access to information, based on an indexing principle. Once the process of encoding digital information on DNA has been completed, the DNA fragments are synthesized.



Conversion of byte fragments into DNA nucleotides. The digital file is divided into segments of about 20 bytes (symbolized at the top by sequences of 12 bits). Each segment results in a DNA synthesis (bottom) containing the payload representative of the digital file (green). The other elements (blue and purple) allow indexing and error correction.

Credit: adapted by François Képès from the presentation by Karin Strauss (Microsoft Corp.).

Writing DNA

DNA synthesis makes it possible to write relatively short nucleic acid fragments with a defined sequence of nucleotides. There are two alternate routes of synthesis: chemical or enzymatic. If required, the sequence of the synthesized DNA can later be modified by mutagenesis.

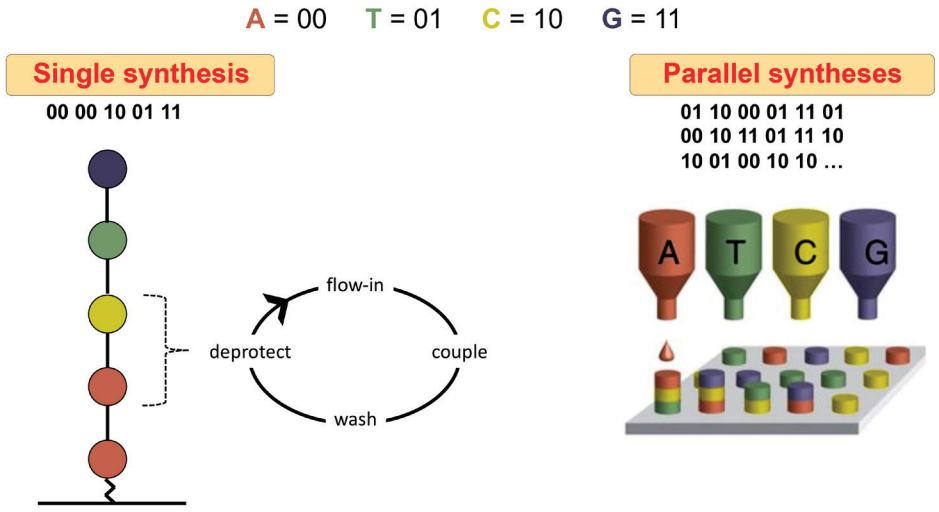
Principle of chemical synthesis

Since 1983, there has been a method for the synthesis of DNA by chemical means. It remains the only commercial method of writing DNA. This method, which has changed little over the years, is based on phosphoramidite chemistry. It involves adding successive nucleotides to the end of the DNA polymer being synthesized; the added nucleotide has a terminal group that blocks the addition of a second nucleotide. This allows only one nucleotide to be added to the sequence at a time. The excess free nucleotide is removed by washing, and the protective end group is then removed by chemical reaction. The following nucleotides are added one after the other in repeated cycles. Throughout its synthesis, the DNA remains attached to a resin. It is detached during the final "deprotection" step. During the synthesis of "a" DNA fragment, it is not actually a single molecule that is synthesized, but a population of identical copies — typically 10^{12} to 10^{15} copies — which can be reduced under controlled conditions to about ten copies in order to increase the density of information[33].

In the early days of chemical DNA synthesis, the process of adding nucleotides was manual. Automatic synthesizers were introduced in the 1990's. They have evolved and can have up to 200 columns, making it possible to synthesize 200 different DNA fragments simultaneously. Twist Bioscience (USA)[34], a company specializing in DNA synthesis, has miniaturized this process. It performs DNA synthesis in wells etched into a silicon microchip, simultaneously synthesizing 10,000 different DNA fragments comprising up to 200 nucleotides. It will quickly reach the million wells mark.The error rate for each added nucleotide is approximately 0.1 %, which explains why in practice the length of the usable fragment is limited to 200 nucleotides. To write much longer pieces of DNA, the usual method is to assemble many fragments of about 200 nucleotides end-to-end, for example by playing on the overlap of their ends, which are designed to be complementary to the preceding and following fragments so as to avoid any assembly error. Another disadvantage of chemical synthesis is that phosphoramidite chemistry is polluting.

33 Organick L, Chen YJ, Dumas Ang S, Lopez R, Liu X, Strauss K, Ceze L. Probing the physical limits of reliable DNA data retrieval. ***Nat Commun***. 2020 Jan 30;11(1):616.

34 https://www.twistbioscience.com/technology



Chemical synthesis of DNA on a solid support. On the left, a sequential synthesis; the addition of one nucleotide at a time takes place according to the cycle shown in the middle. On the right, this same process is carried out simultaneously at several sites.

Credit: adapted by François Képès from the presentation by Nick Gold (Catalog DNA).

Principle of enzymatic synthesis

Due to the above limitations, an alternative method of DNA synthesis by enzymatic means was invented in the early 2010's. This synthesis makes use of a special DNA polymerase present in immune cells, called Terminal Deoxynucleotidyl Transferase (TdT). *In vivo*, TdT adds nucleotides randomly to DNA[35], unlike most DNA polymerases which depend on a single-stranded template and extend the anti-parallel strand in complement to this template. In the process described here, the use of TdT allows the DNA to be extended with the desired single nucleotide by providing only that nucleotide at a given step. As in the chemical synthesis of DNA, the researchers add a chemical protecting group for each nucleotide, thus preventing TdT from adding more than one at a time. Once the desired nucleotide is added, its protection is removed, and the cycle is repeated.

As a result, this biological approach has several advantages over the traditional chemical route. TdT has a high rate of synthesis and miss nucleotide incorporation at a very low rate. The chemical group protecting the nucleotides is different; it preserves the solubility of the nucleotide in water, making this enzymatic approach less polluting than the chemical route using organic solvents.

Half a dozen companies have embarked on this new approach: Nuclera Nucleics, Ansa Biotechnology, Spindle Biotech Inc. (with emphasis on RNA), Molecular Assemblies, Merck, and DNA Script[36]. It is too early to estimate when this biological approach will become commercially attractive.

35 This enzyme is thus responsible for the variability of specific portions of genes encoding immunoglobulin protein chains (antibodies) being selected to adapt to a new antigen.

36 https://www.nuclera.com
https://www.ansabio.com
https://sosv.com/portfolio/spindle-biotech-inc/
https://molecularassemblies.com
https://www.merck.com
http://www.dnascript.com/

Post-Synthesis Sequence Modification

Having scalable access to data offers a flexibility which is valuable in the information technology world, as it allows rewriting part of a digital file without the costly need to rewrite everything. Similarly, it would be interesting to rewrite some parts of the DNA without touching other areas. Many rewriting methods have been developed since the beginning of genetic engineering in 1973[37].

Generally, we mean by "directed mutagenesis" those methods of changing an existing sequence into the desired sequence. Since 2009, techniques have been available for simultaneously introducing multiple mutations into a long double helix of DNA, including whole chromosomes[38].

As shown in the box, the molecular approach has a disadvantage in comparison with computer methods when it comes to handling changing data.

Directed mutagenesis

Broadly speaking, *in vitro* directed mutagenesis often involves PCR, where mutations are carried on the synthetic DNA strands that serve as the primer for the DNA polymerase. The latest method of *in vivo* directed mutagenesis is based on the use of CRISPR-Cas9[1] and its derivatives that, compared with previous methods, are just as precise but are faster and cheaper, even in the case of genomes like those of mammals. It consists of a "guide RNA", which targets a particular DNA sequence, associated with the protein-enzyme "Cas9" which, like "molecular scissors", cuts the DNA at the precise target site.

In the context of archiving big data in DNA, post-synthesis sequence modifications would be different depending on whether the DNA was stored *in vivo* or *in vitro*. *In vitro,* it would be necessary to:

- retrieve the DNA from its container;
- modify it locally by *in vitro* directed mutagenesis, notably mutagenic PCR[2];
- check by local sequencing that the desired mutations have been introduced;
- and finally put the modified and verified DNA back into storage.

This procedure is routine in molecular biology labs throughout the world. However, the choice of procedure depends on the circumstances; for instance, if the DNA to be modified is not exceptionally long, it would be more judicious to completely re-synthesize it as a new sequence. *In vivo*, various approaches are possible, including the use of "molecular scissors" derived from CRISPR-Cas9. This particular *in vivo* approach can be more delicate than the *in vitro* alternatives.

1 CRIPSR-Cas9 is a natural defence system of bacteria. which keeps the memory of exposure to a virus. The RNAs encoded by CRISPR bind to the protein-enzyme Cas9, which can then cut the virus' DNA in order to inactivate it. In 2012, the work of Emmanuelle Charpentier and Jennifer Doudna made it possible to derive from this natural system very efficient and precise genetic engineering tools: Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337(6096):816-21. Note that these two scientists were awarded the Nobel prize in chemistry in 2020.

2 As described above, PCR can be used to amplify a segment of DNA bounded by two short, synthetic single-stranded DNA "primers" chosen by the operator. In order to achieve directed mutagenesis, it is sufficient that one of the primers carries the desired mutations in the target zone. Assuming that the PCR amplifies the DNA by a factor of 1 million, all but the original (1 in 1 million) will carry the desired mutations.

Long-term storage of DNA

Once synthesized, the DNA is then stored by either a chemical or a physical method.

In the chemical storage system, developed by Robert Grass[39] (ETH Zurich, Switzerland), the synthesized DNA is encapsulated in silica nanobeads, then referenced and distributed in

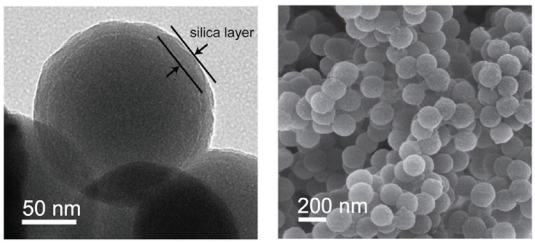
37 Esvelt KM, Wang HH. (2013) Genome-scale engineering for systems and synthetic biology. *Mol Syst Biol*. 9:641.

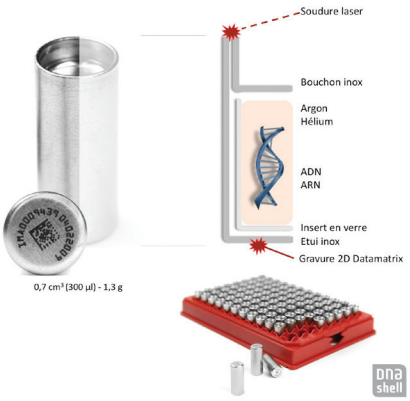
38 Niu D, Wei HJ, Lin L, George H, Wang T, Lee IH, Zhao HY, Wang Y, Kan Y, Shrock E, Lesha E, Wang G, Luo Y, Qing Y, Jiao D, Zhao H, Zhou X, Wang S, Wei H, Güell M, Church GM, Yang L. (2017) Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science* 357(6357):1303-1307.

39 Chen WD et al. (2019). Combining Data Longevity with High Storage Capacity—Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles. Advanced Functional Materials 1901672.

microwell plates. Before being read, the DNA stored in the nanobeads must be extracted by a chemical reagent capable of dissolving the silica while preserving the DNA.

Silica nanoparticles (Robert Grass)





Room temperature DNA storage technologies.

Upper panel: Silica nanoparticles. DNA encapsulated in silica nanoparticles is protected from potential environmental reagents (*e.g.*, oxygen and water) by an impermeable silica layer, which is only a few nanometers thick. This results in significantly increased DNA stability, and the information encoded in the DNA has an expected lifetime of several decades at room temperature.

Lower panel: Capsules from the company Imagene. The outside of these capsules is made of stainless steel, the glass inside contains up to 0.8 g of DNA. A plate (in red) can hold 96 of these capsules. The information encoded in the DNA has an expected lifetime of several tens of thousands of years at room temperature.

Credits: Robert Grass (ETH Zürich), top panel; and Sophie Tuffet (Imagene), bottom panel.

Twist Bioscience (USA)[40] uses physical storage capsules designed by the company Imagene[41] (France). These capsules are made of stainless steel on the outside, glass on the inside, and are the size of a button cell battery. Each capsule contains up to 0.8 g of DNA (potentially 1.4 EB of data when redundancy is taken into account). It is for single use only. Opening the capsule is an easy step that allows the DNA to be recovered. The half life of the DNA is estimated at 52,000 years at room temperature in this storage system where it is protected from water, oxygen and light[42].

40 https://www.twistbioscience.com/

41 http://www.imagene.fr/

42 Organick L et al. (2020). An empirical comparison of preservation methods for synthetic DNA data storage. *bioRxiv preprint* doi: https://doi.org/10.1101/2020.09.19.304014

For both approaches, information files are stored on the DNA in a structured manner. For example, DNA containing a large information file would be archived in its own capsule or nanobead whilst DNA fragments containing smaller information files would be grouped together in the same container, where they would be differentiated by their label sequences.

Reading DNA

Polymerase Chain Reaction (PCR)

To transcribe back the information stored on DNA, its nucleotide sequence must be read by a process known as sequencing that often starts with a PCR amplification.

The PCR technique is also used to selectively access information. This entails multiplying the desired DNA fragment, which may be mixed in with many other fragments, by choosing a primer that is only complementary to the label identifying this fragment. Reading some of this DNA does not destroy the information it contains because so many identical copies remain.

Sequencing

DNA sequencing consists of determining the sequence of nucleotides A, T, G and C within this linear polymer. Two approaches will be described below.

a) Principle of the Illumina technology

The Illumina company, leader in the DNA sequencing market, commercializes devices that can sequence up to 4 billion nucleotides per experiment (in comparison, the human genome contains around 3 billion nucleotides), the equivalent of 1 GB or a film. However, this long sequence is actually obtained in the form of a large number of shorter sequences corresponding to fragments, typically 300 nucleotides in length. The error rate of this sequencing is 1 %. This is the technology used by Microsoft Corp. for their 1 GB proof of concept described above. It is described in detail in the following box.

Sequencing Technology Using Synthesis

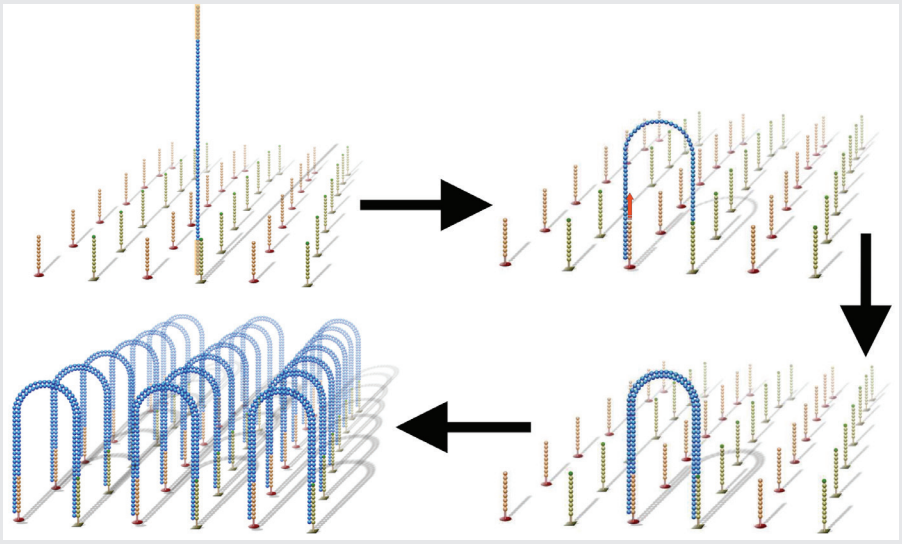
The first step in sequencing is to prepare a library of DNA fragments. To do this, DNA is extracted from the sample and then fragmented. The fragments are heated to separate their two strands from one other. Short, known, sequences of DNA known as "adaptors" are then added to the ends of these DNA fragments.

The liquid phase mixture of the single-stranded DNA thus marked by attached adaptors is then brought into contact with a standardized glass slide. This slide contains a billion single-stranded DNA deposits a few tens of nucleotides long, which are complementary to the adaptors added during the preparation of the DNA library. Both ends of the single-stranded DNA attach to the glass slide to form a bridge. An enzyme called DNA polymerase is added to this mixture to synthesize the DNA strand complementary to this bridge, one nucleotide at a time. The four nucleotides A, T, G and C, labeled with four different fluorophores[1], are added to the mixture sequentially. After incorporation of a nucleotide by the DNA polymerase, the glass slide is washed, thus removing excess nucleotides. Subsequently, a photograph of the glass slide is taken. Then the fluorescence of the nucleotide

►►

1 Chemical substance capable of emitting fluorescent light after excitation by light of a shorter wavelength.

is removed by chemical reaction. This allows a new labeled nucleotide to be incorporated into the next cycle. This process is performed 300 times in a row. For each photograph, the fluorescence is quantified at each point. The ordered set of the 300 photographs constitutes a reading of the sequence of 300 successive nucleotides for each DNA fragment that has been bridged on the glass slide.



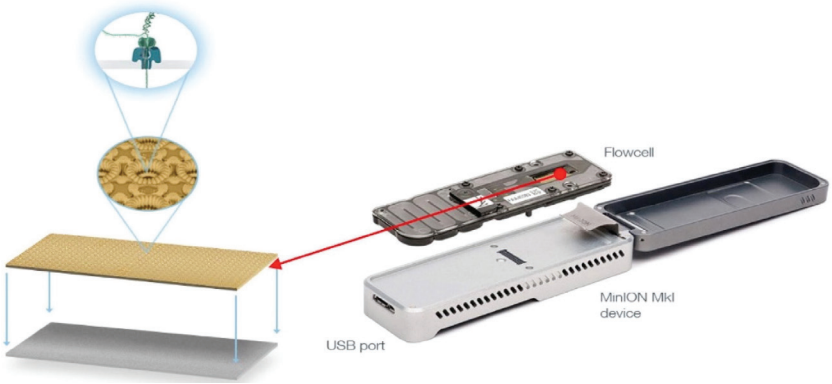
Sequencing principle with Illumina's technology.

Credit: Illumina

The final step is to computer-align the sequences of the different DNA fragments using the overlapping parts of their sequences to order them and reconstruct the overall sequence. This phase is computer-intensive and time-consuming[2].

2 See video of this process at https://www.youtube.com/watch?v=HMyCqWhwB8E

b) Principle of the Oxford Nanopore technology



The "MinIon" device from Oxford Nanopore Technologies weighs 90 g and can read more than 10 GB of data in two days for 700 US$.

Credit: Oxford Nanopore Technologies.

The third-generation sequencers developed by Oxford Nanopore Technologies[43] pass DNA or other polymers through pores and record its composition as it passes through. Long DNA sequences can be read in a single pass (the record is 2.2 million nucleotides), which avoids

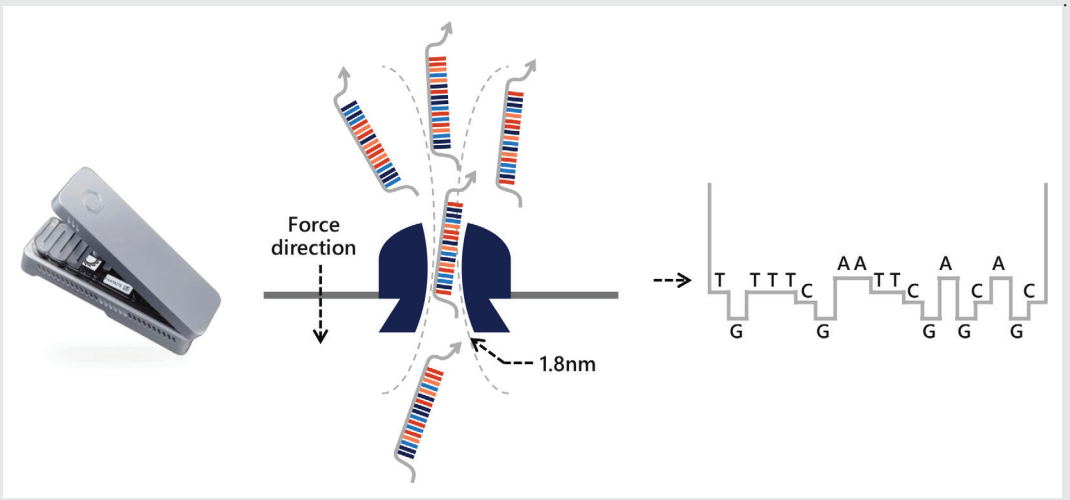
43 https://nanoporetech.com/

the need for fragmentation and alignment, and allows data analysis in real time. Despite an error rate of more than 10 % for each read, the parallel sequencing of many copies of the DNA allows correction by cross-referencing, which results in a near-perfect sequence. The method is detailed in the next box.

Nanopore sequencing technology

Each sequencing device contains several hundred membrane-proteins, called nanopores, inserted into a synthetic membrane. By applying an electric potential difference across the membrane, an ion flux is measured in real time for each nucleotide as it passes through the nanopore. This ion flux is measured in pico-amperes, which represents a sensitivity corresponding to about two hydrogen atoms. Each nucleotide thus corresponds to a reproducible and specific electrical signal, and successive signals are converted into a sequence after analysis.

Due to its steric hindrance, the nanopore only sequences one molecule at a time. Without molecular restraint, DNA would travel through the nanopore at a speed of one million nucleotides per second. In order to limit the flow to 450 nucleotides per second and to sequence in real time, the original native DNA is bound to a motor protein separating the two DNA strands and acting as a molecular brake for one of these strands at the entrance to the nanopore. Thus, each nanopore is capable of sequencing 450 nucleotides per second in a controlled manner.



In the case of the "MinIon" device from Oxford Nanopore Technologies, the membrane operates for about two days before it needs to be replaced at a cost of less than a thousand euros. The electronic part of this device, which fits in the palm of one's hand, can be connected to a computer via a USB3 interface. However, at full speed, the sequencing rate provided by hundreds of nanopores operating in parallel is such that high-performance electronic devices, such as Graphic Processing Units (GPUs), are required for real-time computational processing.

Decoding information

The sequences of the DNA fragments, taken from the same digital file, are grouped together by their labels and then by their common parts. The DNA sequences of these fragments are transcribed into bytes and those within the same digital file are ordered using their identifiers so as to reconstitute the global sequence in bits.

Chapter IV

Evolution and progress of technologies required to archive data in DNA

Challenges

The archiving of information in DNA is experimental. Before becoming viable on a large scale, it must be fully automated. In addition, the speed and cost of DNA reading and writing processes need to be improved. At present, the cost and time required to store 1 GB (10^9 bytes) of DNA data is comparable to that for 1 PB (10^{15} bytes) of data on a computer medium. According to Illumina, this cost must be divided by a factor of 10,000 before the DNA approach can be widely adopted. Moreover, a storage architecture system must be created that would allow selected information to be accessed and computational tasks to be performed directly on it.

In this section, we discuss in detail the limitations of the technologies used to store DNA information and the advances made by scientists in public and private sector organizations.

Improvement of DNA writing technologies

Limiting factors in DNA synthesis

The storage of DNA information requires accurate and large-scale DNA synthesis. Current approaches are inadequate because:

- it is difficult to synthesize a DNA fragment longer than 150 nucleotides. This limit results from the error rate during DNA synthesis, which is typically 0.5 % per nucleotide;
- certain DNA sequences are difficult to synthesize (due to repetition of the same nucleotide several times in a row or to sequences rich in C and G);
- DNA synthesis is time-consuming: current systems synthesize DNA by adding one nucleotide after another at a rate of 30 seconds per nucleotide. In addition, it can take several weeks to months to assemble a long piece of DNA of reasonable quality from constituent fragments of 100-200 nucleotides;
- DNA synthesis is expensive (about 8 cents per nucleotide in 2020). This cost needs to be reduced by a factor of 10^8 (100 million);
- DNA synthesis is not "democratic". It is generally not carried out by laboratories, but by specialized platforms. The laboratory typically orders the desired DNA sequences from

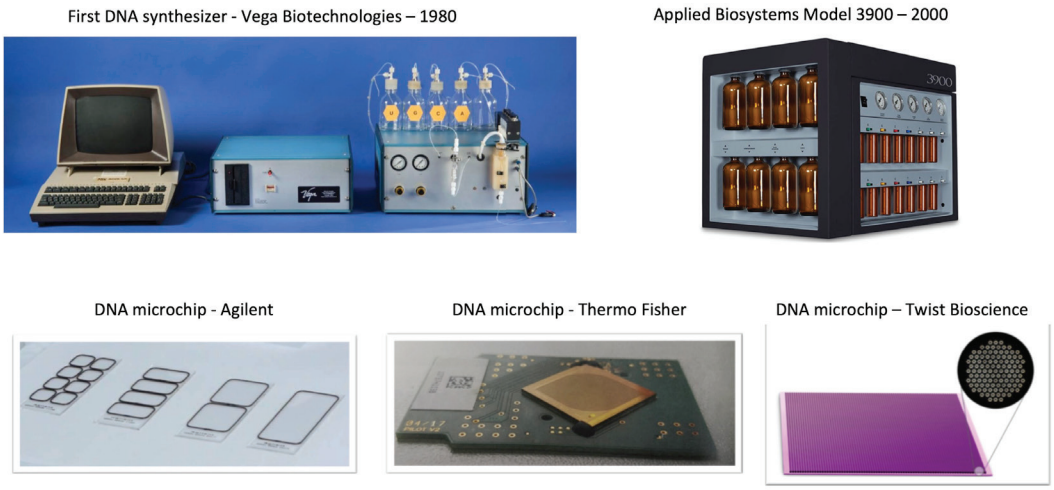
platforms which deliver the DNA fragments after a few days for sequences of 25 nucleotides to a few weeks for sequences over 2,000 nucleotides;

– chemical synthesis of DNA is polluting. The process uses acetonitrile, a harmful chemical solvent, along with other substances. It accounts for 72 % of the reaction volume for the synthesis of a DNA fragment; hence, potentially several billion liters of acetonitrile would be needed to synthesize enough DNA to contain the entire GDS.

History and progress in chemical DNA synthesis

History and evolution

Ever since the discovery of the structure of DNA in 1953 by J. Watson and F. Crick, chemists have wanted to synthesize DNA. In 1965, the first DNA fragment was chemically synthesized. In 1970, a 70-nucleotide gene was synthesized for the first time. In 1983, synthesizing longer fragments was facilitated by the use of the powerful and stable phosphoramidites.



Evolution of chemical DNA synthesis devices.

Sources (top left, then in reading direction):
1) National Museum of American History - https://americanhistory.si.edu/collections/search/object/nmah_1451158
2) McLuen Design - http://www.mcluendesign.com/portfolio-item/applied-biosystems-3900/
3) Agilent Technologies - https://www.agilent.com/cs/library/slidepresentation/public/New%20Agilent%20CGH%20microarrays%20focused%20on%20exons%20for%20clinical%20applications.pdf
4) European Biotechnology Magazine, Autumn 2018 issue, Dr Martin Laqua, Pioneers in the Synbio Revolution
5) Twist Bioscience

In the following decades, the phosphoramidite method has undergone considerable improvements. In the early days of chemical DNA synthesis, nucleotides were added manually. In the 1990s, automatic synthesizers were introduced. They have evolved to have up to 200 columns, making it possible to synthesize 200 different DNA fragments simultaneously. Agilent, Twist Bioscience and Thermo Fisher miniaturized and massively parallelized this process to reduce the cost. These companies perform DNA synthesis on microchips. Agilent is capable

of synthesizing 244,000 DNA fragments simultaneously. Thermo Fisher synthesizes 35,000 DNA fragments on a 100 mm^2 chip. Twist Bioscience has further miniaturized this synthesis.

Advances in chemical synthesis: Twist Bioscience

Twist Bioscience (USA) is one of the world's leading companies in chemical DNA synthesis. They have designed a chip that allows the simultaneous synthesis of one million — possibly different — DNA fragments of 200 nucleotides in 24 hours. This chip is composed of a silicon wafer and has 10,000 wells. In each well, 100 different fragments of 200 nucleotides are synthesized. The price of a chip is US$ 5,000. Twist Bioscience intends to further miniaturize the synthesis process. Currently, each of the 10,000 wells that make up the chip measures 50 microns. By reducing the size of the wells to 0.3 micron, Twist Bioscience could soon synthesize many more fragments simultaneously and thus further reduce the cost of synthesis.

Many suppliers of DNA have an error rate of 0.5 % in chemical synthesis. The longer the sequence, the higher the probability of it containing errors; this limits the length of useful sequences to about 150 nucleotides. Twist Bioscience has reduced the error rate to 0.1 %, effectively allowing error-free synthesis of fragments of 200 nucleotides in length.

Advances in enzymatic DNA synthesis

Since 2010, techniques for DNA synthesis by enzymatic means have been developed (see principle in chapter III). The process of enzymatic synthesis is simpler, faster and more efficient than chemical synthesis. Extrapolations made by the companies involved suggest that it will be possible, within a few years, to synthesize fragments of up to 2,000 nucleotides in one piece using enzymatic synthesis. In a living cell, the reading/writing of DNA is faster than Flash memory (less than 100 microseconds per bit). This gives an idea of the potential of the DNA approach. It is worth noting that cellular processes are generally more efficient when they take place inside rather than outside cells. Moreover, enzymatic synthesis does not require the use of dangerous chemicals, which reduces its environmental impact. At present, six companies are working on enzymatic DNA synthesis: DNA Script, Nuclera Nucleics, Molecular Assemblies, Merck, Ansa Biotechnologies and Spindle Biotech Inc. The first two of these companies were consulted for this report.

DNA Script

DNA Script[44] (France) is currently the world leader in enzymatic DNA synthesis. The company's goal is to revolutionize DNA writing with a faster, more efficient and ultimately less expensive technology based on enzymatic synthesis. DNA Script is developing this technology by using reversible terminator nucleotides and DNA polymerases. The company is currently able to synthesize DNA fragments of more than 250 nucleotides under laboratory conditions, with an error rate of about 0.7 %. Optimizations are underway to reduce this error rate still further and to increase the length of the synthesized DNA fragments. By 2024, the company plans to be

44 https://www.dnascript.com

able to synthesize and sequence 1 TB (4 trillion nucleotides) of data in 24 hours.

One of the company's objectives is to "democratize" DNA synthesis in laboratories, so that it is no longer carried out exclusively by specialized platforms. To do this, DNA Script is designing a benchtop machine capable of synthesizing DNA. The aim is to allow researchers to synthesize DNA sequences in their own laboratories in a few hours, thus avoiding delays in ordering and delivery. DNA Script plans to market 10,000 machines within the next ten years. They are intended for use in research, specialized medicine, and personalized gene therapy. According to the company, it is still too early to set a definitive sales price for the machine, but it is anticipated that it could be marketed for around 50,000 euros.

Nuclera Nucleics

Nuclera Nucleics[45] (United Kingdom) is developing three technologies: DNA polymerase enzymes, reversible terminator nucleotides[46] and an automation system.

Four generations of enzymes have been designed to optimize their ability to incorporate nucleotides. In addition, Nuclera Nucleics designs terminating nucleotides that are compatible with the enzyme that synthesizes the DNA strand. These nucleotides possess protective chemical groups that block any interaction, in order to extend the sequence by only one nucleotide at a time. These chemical groups also protect the nucleotide from chemical changes and prevent the unwanted interaction of one nucleotide with another, thus preventing the formation of secondary structures[47] in the DNA. The company is developing a system to automate the enzymatic process: a microfluidic electronic system that controls the movement of liquid drops. The machine incorporating this technology should be able to synthesize a single sequence of 10,000 nucleotides, quickly, without error and at a reasonable cost.

DNA Script

Kilobaser

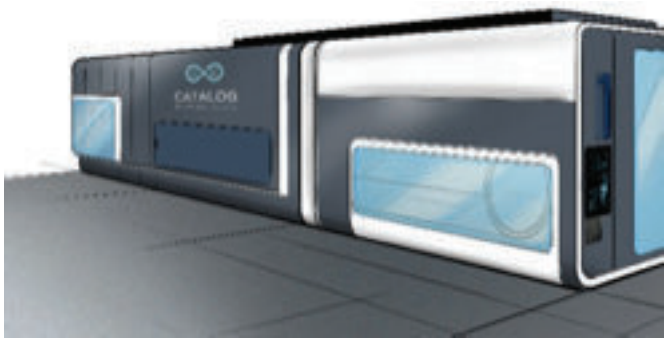
Evonetix

45 https://www.nuclera.com/

46 As discussed in chapter III, the nucleotides used are protected by a terminal chemical group blocking any addition of a second nucleotide, in order to extend the sequence by only one nucleotide at a time. Here, the terminal group is different from that used in chemical synthesis.

47 DNA conformation obtained by interactions between regions of the double helix.

Catalog DNA



DNA synthesis devices under development

Credits: Thomas Ybert, Alexander Murer, Tim Brears, Hyunjun Park.

New DNA synthesis technologies

Kilobaser

Kilobaser[48] (Austria) is designing a benchtop machine capable of synthesizing a DNA fragment of about 100 nucleotides per experiment. The use of this machine is reserved for the synthesis of PCR primers. The aim is to enable a researcher to obtain the desired primer, synthesized chemically in his laboratory, in a few hours, and thus avoid delays in ordering and delivery.

For each nucleotide added, the probability of error is approximately 2 %, which limits the sequence length of the fragment to about 100 nucleotides. The synthesis time of the DNA fragment is two minutes per added nucleotide plus 20 minutes for the final processing, or about one hour for the full synthesis of a 20-nucleotide fragment.

Kilobaser started commercializing this machine in 2020. Its launch price is around 15,000 euros. The price of consumables should be approximately 450 euros per month. The objective is to obtain a price of 75 cents per nucleotide synthesized. This price is significantly higher than that of the synthesis platforms (about 8 cents in 2019). These characteristics of the Kilobaser system imply that it will not directly contribute to synthesizing the large quantities of DNA needed to store digital information, but rather be part of the arsenal of a storage unit, where it would be used in particular to supply the numerous and varied primers required for a wide variety of DNA amplifications.

Helixworks

Helixworks[49] (Ireland) designs, manufactures and sells DNA-based data storage technologies. It is developing the "Molecular Storage System" (MoSS) technology, whose early stage aims at converting digital files into DNA sequences.

Helixworks manufactures its own DNA using proprietary enzyme technology. The company

48 https://www.kilobaser.com/

49 https://helix.works/

reports that it can now synthesize DNA sequences longer than 2,000 nucleotides in one run. This is an order of magnitude better than the 200 nucleotide average synthesized using other techniques. The company also synthesizes DNA with nucleotides modified by the addition of chemical groups, such as acetylated nucleotides.

Helixworks uses the Oxford Nanopore Technologies sequencing technology described above.

Catalog DNA

Catalog DNA[50] (USA) aims to use DNA as a storage medium for digital information. To do so, researchers have designed a fast and relatively cheap technology to generate DNA fragments for data storage purposes.

Current systems synthesize DNA one nucleotide at a time at a rate of 30 seconds per nucleotide. Catalog DNA, however, synthesizes and checks small fragments of DNA (called "components") that then form an ordered library; these components can be selected and combined to form a specific, much longer sequence. This combination determines their meaning, as in many languages where there is a limited set of letters but a very large number of words. Such assembly requires less DNA synthesis, which is the most expensive and slowest part of the process.

Each component assembled by Catalog DNA is double-stranded but has two single-stranded ends, allowing them to be joined together like the bricks of a construction game. With each cycle, a new component is added to the previous one through an enzymatic process.

Catalog DNA has designed a machine capable of assembling DNA fragments in a programmable and automated way in order to generate unique DNA sequences. This machine generates 500,000 DNA assembly reactions per second. It contains an incubator that maintains ideal conditions for the assembly of DNA fragments by enzymes. The reactions are carried out on a long moving belt, which moves between the compartments at a speed of 16 meters per minute. Using this machine, Catalog DNA can synthesize DNA coding for 500 KB of information per second. In 2019, they coded the English-language version of Wikipedia in DNA, *i.e.* 16 GB.

Evonetix Ltd.

Evonetix Ltd.[51] (United Kingdom) is developing an innovative DNA synthesis system combining silicon chips and finely adjusted thermal regulation. It is designing a benchtop machine capable of synthesizing several thousand DNA nucleotides in a single experiment. A promising improvement brought by Evonetix reduces the error rate by a factor of 100 to 1,000, by thermally controlling each of the 10,000 reaction micro-sites. The underlying technology is summarized in the box. The Evonetix device is composed of the machine, a silicon MEMS (micro-electromechanical system) chip and thermolabile[52] reagents. The whole device is coordinated by computer software developed by the company. The machine and technologies are currently under development.

50 https://www.catalogdna.com/about

51 https://www.evonetix.com

52 Sensitive to temperature.

According to Evonetix, it is still too early to set a selling price, but the machine could be marketed for around 10,000 euros.

DNA synthesis by thermal control

The principle adopted by Evonetix Ltd. is to synthesize short fragments of single-stranded DNA in each well of the silicon chip with thermolabile reagents, using a very precise control of the temperature of each of these wells. The idea is then to assemble these short single-stranded DNA fragments into longer DNA fragments. The device also includes an error correction system during DNA synthesis.

As in conventional chemical synthesis, the different DNA fragments are constructed nucleotide by nucleotide; each well on the chip contains a particular fragment. Four solutions, each containing one of the nucleotides A, T, G, C with a heat-labile protective group (different for each type of nucleotide), are circulated successively on the chip. The controlled temperature change in a given well of the chip induces deprotection of the DNA fragment being synthesized in this well. The DNA fragment, now deprotected, incorporates the thermolabile nucleotide from the circulating solution. Conversely, if the change in temperature is not sufficient to induce deprotection of the DNA fragment, the latter will not incorporate the nucleotide from the circulating solution. This cycle is repeated until the desired sequence is obtained. This very precise temperature of each individual well in the chip is therefore an essential element of this approach.

The synthesized DNA fragments are then assembled in pairs to obtain a longer DNA fragment. This entails releasing the DNA fragments from the chip by cleaving the thermolabile chemical bond between the first nucleotide in the chain and the chip. These bonds are of differing thermolabilities, which allows a temperature control of the release of DNA fragments. The movement of the released DNA fragment on the chip is controlled electrophoretically until it meets the right second DNA fragment. The homologous parts of the two single-stranded DNA fragments then lead to their pairing. These homologous sequences are different for each pair of DNA fragments and these differences determine the fusion temperature at which the DNA fragments pair. Once these two, single-stranded DNA fragments have paired, the process is repeated so as to assemble an even longer DNA fragment.

The assembly process includes an error correction mechanism. The accuracy and precision of the temperature control system is such that if a single-stranded DNA fragment has a mutation in its sequence, it can no longer fuse with the second single-stranded DNA fragment at the appropriate temperature and is therefore eliminated. As a result, the error rate anticipated by Evonetix is two to three orders of magnitude below the 0.5 % rate obtained in the conventional chemical synthesis process.

Conclusion

Several companies are working in DNA synthesis technologies to improve performance (automation, parallelization, cost, error rate, velocity, DNA fragment length). Substantial progress has been made rapidly.

Twist Bioscience, one of the leading companies in chemical synthesis, is massively parallelizing this method of synthesis to reduce its cost. DNA Script and Nuclera Nucleics, leaders in enzymatic synthesis, are optimizing this process in order to rapidly generate long DNA fragments in a single step. Within four years, DNA Script plans to be able to synthesize 1 TB of digital information in 24 hours, *i.e.* the equivalent of 1,000 films. It should be noted that Twist Bioscience and Evonetix technologies are compatible with the enzymatic route of synthesis, even if they currently use the standard chemical route.

Very promising alternative DNA synthesis systems are also under development. Evonetix's device, based on temperature-controlled assembly, is one of the most advanced and inexpensive systems on the market. Helixworks and Catalog DNA are the only companies to have optimized their DNA synthesis technology for DNA information storage purposes only. Their technique consists of combinatorial assembly of pre-synthesized — and thus verified — DNA fragments,

thus partially overcoming the limitations set out above in this chapter. It should be noted that the Catalog DNA machine, which is capable of synthesizing 0.5 MB of information per second, constitutes a major asset for the storage of digital information in DNA.

Optimization and densification of digital information encoding into DNA

Challenges

Let us recall that the general principle of digital information storage in DNA is to assign numerical values to the four nucleotides A, T, G and C and then incorporate these nucleotides into a DNA fragment with a specific sequence. Each subsequence of four nucleotides represents a byte or, as a first approximation, a letter, number or other symbol. DNA synthesis is the slowest and most expensive part of the information storage process. In order to reduce the cost and increase the speed, scientists are working on new methods to densify information. The goal is to reduce the amount of DNA needed to store the same information and thus synthesize fewer nucleotides per bit of data.

MOSLA project

Researchers from the MOSLA (Molecular Storage for Long-term Archiving)[53] project at the Universities of Marburg, Darmstadt and Giessen in Germany are establishing a method for "compressing" digital information converted into nucleotides. In addition, they are developing a new alphabet of nucleotides. Their aim is to be able to distinguish nucleotides modified by the addition of chemical groups from unmodified nucleotides (A, T, G, C). Different types of chemical modifications to nucleotides, such as methylation, can be detected by third generation sequencers using Oxford Nanopore or Pacific Biosciences technologies.

Project of the Technion and IDC Herzliya

Zohar Yakhini's group[54] (Technion, IDC Herzliya - Israel) is designing a new nucleotide alphabet containing both the four standard nucleotides A, T, G, C and composite nucleotides. The principle of this new nucleotide alphabet is based on the following characteristics of current techniques:

- in writing a DNA fragment, it is not a single DNA molecule that is synthesized, but a population of identical DNA molecules;
- in reading a DNA fragment, it is not a single DNA molecule that is sequenced, but a population of identical DNA molecules.

This necessary multiplicity can be exploited when designing new letters. In addition to the four standard nucleotides A, T, G, C, the alphabet could contain a mixture of nucleotides composed for example of 50 % A and 50 % C. If all combinations of nucleotides could be in a 50:50

53 https://mosla. mathematik.uni-marburg.de/gb/

54 Anavy L, Vaknin I, Atar O, Amit R, Yakhini Z (2019). Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nature Biotechnology* 37,1229-1236 https://www.technion.ac.il/en/home-2/

ratio, the alphabet of available 'nucleotide-equivalents' would comprise ten letters[55]. This process can be extended to create an even richer mixed nucleotide alphabet. For example, a letter could be defined as 33 % A, 33 % T and 33 % C; hence, at a particular position in the sequence of the synthesized fragments, one third of the population would have nucleotide A, one third T and one third C. However, the imperfection of the current methods of both DNA sequencing and synthesis limits the number of possible combinations that can be distinguished in practice.

By increasing the size of the nucleotide alphabet, fewer nucleotides per byte of digital information need synthesizing. However, more copies of the same DNA fragment must be sequenced to decode the information with 100 % success. The cost of synthesizing a DNA nucleotide is estimated to be 5,000 times higher than the cost of sequencing a DNA nucleotide. Thus, by using a more diversified alphabet, the overall cost of the operation is significantly reduced since it involves less writing for more reading.

Unconventional nucleotides

New nucleotides that are chemically different from the natural nucleotides A, T, G, C, have been made by modifying the sugar or nitrogenous bases of the nucleotides. As a reminder, a nucleotide is always composed of the same five-carbon sugar (deoxyribose), a phosphate group, and a nitrogenous base responsible for the identity of the nucleotide A, T, G, or C.

Modifications of nucleotide sugars

Piet Herdewijn's group (Université Catholique de Louvain, Belgium) has synthesized new nucleotides by artificially modifying their sugars. These unconventional nucleic acids are called XNA (xeno-nucleic acids)[56]. XNAs were designed to avoid any hybridization with natural DNA, while being incorporated and tolerated by a living organism. The objective is to construct cells that would store all or part of their genetic information in an alternative informational polymer, without the risk of recombination with the original genomes.

These XNAs could also be used to store digital information in DNA:

- *in vitro*: in order to increase the information density with new nucleotides;
- *in vivo*: to store digital information in the form of an artificial polymer, in living cells or organisms, without the risk of this information interacting with the natural DNA of the organism.

Changes in the chemical group of nucleotides

Steven Benner's group (University of Florida, USA) has synthesized eight new nucleotides, named S, B, J, V, K, X, Z and P, by modifying the chemical groups of natural nucleotides. The aim is to avoid problems related to "DNA imperfections"[57]. Indeed, natural DNA has some

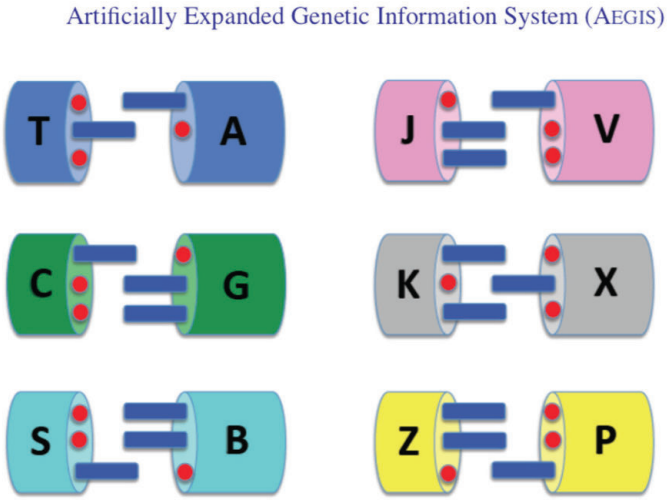
55 In addition to A, T, G, C, the following mixtures in equal parts are also letters: A/T, A/G, A/C, T/G, T/C, G/C.

56 Chaput JC, Herdewijn P, Hollenstein M. Orthogonal Genetic Systems (2020). *Chembiochem* 21(10):1408-1411.

57 Hoshika H, Leal N, Kim MJ, Kim MS, Karalkar NB, Kim HJ, Bates AM, Watkins Jr. NE, SantaLucia HA, Meyer AJ, DasGupta S, Piccirilli JA, Ellington AD, SantaLucia Jr. J, Georgiadis MM, Benner SA (2019). Hachimoji DNA and RNA: A genetic system with eight building blocks. Science 363:884-887.

chemical characteristics that can compromise its ability to function as a digital information storage molecule:

- chemical reactions can make unwanted modifications to its sugars or to its nitrogenous bases and therefore affect its information content;
- DNA has only four different nucleotides, limiting information density;
- The bond between A and T is based on two hydrogen bonds and therefore weaker than the bond between G and C based on three hydrogen bonds. This can create hybridization problems and lead to artefacts and errors during DNA amplification (a mandatory step for the storage in DNA of digital information).
- G-rich DNA fragments are very difficult to synthesize. Indeed, Gs can pair with one another to form a three-dimensional structure that impedes the synthesis of DNA fragments.



Conventional (TA, CG) and non-conventional nucleotide pairings designed by Steven Benner's group (SB, JV, KX, ZP).
Credit: Steven Benner

These new nucleotides have been designed in such a way that S pairs with B, J with V, K with X, Z with P. These pairs are all linked by three hydrogen bonds. In addition, Benner and co-workers have chemically modified A to pair with T with three hydrogen bonds instead of two, and have modified G to no longer form a three-dimensional structure with other Gs. This facilitates the synthesis of G-rich sequences.

This "rectified DNA" has many advantages for the storage of digital information:

- the modified nucleotides facilitate the synthesis of G-rich DNA fragments. Hence, the difficulty of synthesizing this DNA fragment is no longer a factor in coding;
- the eight new nucleotides allow digital data to be converted into a denser polymer sequence;
- the PCR step that allows the DNA containing the information to be multiplied and copied is more specific. This reduces the risk of errors and hence preserves the information and facilitates accessing it.

Conclusion

Some researchers propose to densify information, using extended alphabets, so as to reduce the amount of DNA needed to store the same quantity of information.

The project led by Zohar Yakhini consists in using the four standard nucleotides, as well as mixtures of these nucleotides to code information in a denser way. A major advantage of this method lies in the fact that it uses only standard nucleotides, and therefore widely proven reading and writing methods. However, many copies of this DNA must be sequenced to decode the information it contains without error.

The MOSLA project uses nucleotides modified by the addition of chemical groups. Currently, only third-generation sequencing technologies (*e.g.*, the Oxford Nanopore Technologies device) are capable of reading information using these nucleotides.

Finally, some researchers such as Steven Benner or Piet Herdewijn are designing unconventional nucleotides. These nucleotides have been deliberately modified for various purposes and can in particular facilitate the processes of writing, storing and accessing information. However, current sequencing processes limit the reading of these non-conventional nucleotides. Further progress needs to be made in order for them to be widely used.

Improvement of DNA reading technologies

Limiting factors in DNA sequencing

The storage of digital information in DNA requires accurate and large-scale DNA reading. Current approaches are successful but further progress is still needed because:

- DNA reading is expensive. According to the Illumina company, the cost of DNA reading needs to be reduced by a factor of 1,000;
- the error rate when reading a DNA sequence is too high. According to Microsoft Corp., most information encoding/decoding errors in the process of storing digital information come from DNA sequencing;
- DNA reading is not fast enough. In order to use DNA as a medium for information storage effectively, a near-instantaneous accessing and reading of the information is needed;
- not all sequencing technologies allow the reading of modified conventional and of non-conventional nucleotides;
- DNA reading is not yet "democratized" enough. It is generally not carried out by laboratories, but by specialized platforms.

History and progress of sequencing technologies

History

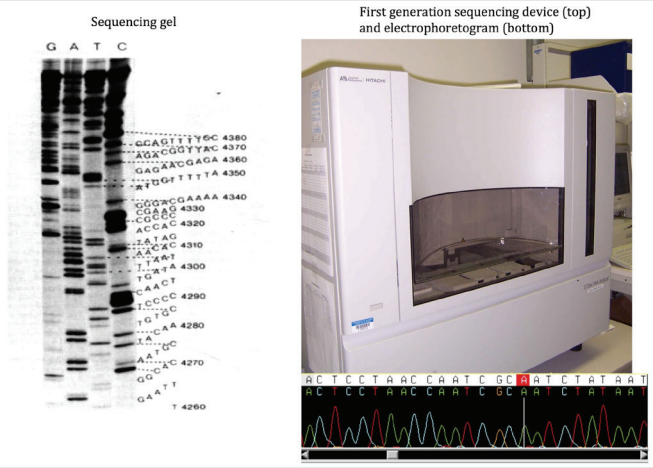
Advances in genetic engineering in the 1970s enabled researchers two to three decades later to sequence the entire genome of an organism. There are three generations of sequencing methods, which are summarized in the box.

A BRIEF HISTORY OF DNA READING TECHNOLOGY

a) 1st generation sequencing

Two DNA sequencing methods were developed in 1977, one by Frederick Sanger's group and the other by Walter Gilbert's group. Both methods use the DNA replication function. Sanger's method was preferred to Gilbert's method in the years that followed.

Sanger's principle consists of replicating the complementary strand of the single-stranded DNA to be sequenced with a DNA polymerase. All four deoxyribonucleotides are added to the reaction (dATP, dCTP, dGTP, dTTP), as well as a low concentration of one of the four di-deoxyribonucleotides (ddATP, ddCTP, ddGTP, ddTTP). These di-deoxyribonucleotides are analogous to deoxyribonucleotides, but a small chemical difference makes them chain terminators; if they are incorporated, with a low probability dictated by their low concentration, they prevent further elongation of the DNA complementary strand. Thus, elongation stops, for example where a ddATP has been incorporated instead of a dATP. For the complete sequencing of the same fragment, the reaction is performed separately for each of the four di-deoxyribonucleotides. The synthesized DNA fragments are analyzed by electrophoresis of each of the four reactions on an acrylamide gel, which separates the fragments by size, so allowing the sequence to be read. The actual detection of the synthesized fragments is done by incorporating a radioactive tracer, which is detected by a sensitive film placed on the gel.



First generation sequencing

Sources (from left to right): Nucleotide sequence of Phi174 DNA. F. Sanger et al Nature **265,** *687-695 (1977) -* https://doi.org/10.1038/265687a0
Applied Biosystem (Thermo Fisher Scientific), and Philippe Glaser for the electrophoretogram.

Numerous genes have been sequenced in this way. For over twenty years, Sanger sequencing technology has been optimized: automation and robotization, introduction of fluorescent tracers replacing radioactive markers, use of electrophoresis in capillaries instead of between plates. However, this method is still not very efficient for sequencing whole genomes because the DNA fragments are sequenced one at a time rather than in parallel. In 2003, it took one year with this technology to sequence the human genome at a cost of around US$3 billion.

b) 2nd generation sequencing

A call for proposals from the National Human Genome Research Institute (NHGRI) within the National Institutes of Health (NIH) in the USA was broadcast to promote technological innovations in sequencing. The objective was to sequence the human genome at a cost of US$ 1,000. Many ideas for new DNA sequencing technologies emerged and three sequencing technologies, 454 Titanium ROCHE, SOLID v3 ABI, GAIIX Illumina, appeared in the years 2000 to 2010. They marked the beginning of second-generation sequencing. Since then, the Illumina company has perfected its devices.

►►

Second generation sequencing devices

Credit: Illumina.

c) 3rd generation sequencing

Third generation sequencing devices.

Credits: Pacific Biosciences and Oxford Nanopore Technologies.

The third-generation sequencers are developed by Pacific Biosciences and Oxford Nanopore Technologies Ltd. The first products were commercialized in 2011 by Pacific Biosciences and in 2015 by Oxford Nanopore Technologies. With these technologies it is possible to:

- sequence single molecules in order to avoid the amplification step, and thus to detect any chemical modifications to the nucleotides that alter the electrical signal generated by the passage of the polymer through the nanopore;
- reduce the cost of sequencing;
- obtain long read sequences and analyze data in real time because sequence fragmentation and alignment steps are unnecessary.

Illumina technology

The Illumina company[58] (USA) is a leader in the sequencing market. Currently, 96 % of the world's sequencing data is produced by Illumina devices. A total of 100 PB (10^{17} bytes) of data has been generated by their sequencers. Recent advances in their technologies are described in the box.

58 https://www.illumina.com/

"ILLUMINA" DEVICES

Since the appearance of the first second-generation sequencers in 1996, Illumina's devices have evolved to provide faster sequencing, higher capacity and lower sequencing error rates. Several sequencing devices are commercially available.

The most recent device is the NovaSeq6000, which has been on the market for three years. There are currently 600 NovaSeq6000 devices worldwide. This device has been designed to be scalable and to adapt itself to the methods and scale of different sequencing projects. It has two independent measurement cells for sequencing. There are four types of measuring cells: SP, S1, S2 and S4. The smallest measuring cell reads 80 to 400 billion nucleotides by lecture cycle whilst the largest reads 2,000 to 4,000 billion nucleotides. This device reads up to 6,000 billion nucleotides in 48 hours, *i.e.* the sequencing of 50 human genomes with high overlap between sequenced fragments. Illumina's smallest device is a compact benchtop device. It reads up to 1.2 billion nucleotides per cycle in 17 hours. Its cost is about US$20,000.

Illumina sequencing devices

Credit: Illumina.

Each new device requires innovations in fluid mechanics, organic chemistry, surface chemistry, mathematics, physics, molecular biology, optics, computer science, etc. A total of seventy innovations and patents were required to design Illumina's latest device, the NovaSeq6000. However, the sequencing market is evolving with the commercialization of the Oxford Nanopore Technologies and Pacific Biosciences sequencing devices. The next step for Illumina is to reduce costs. The current devices can sequence a human genome for US$700. Over the next few years, Illumina hopes to bring this down to less than US$100.

Oxford Nanopore technology

Oxford Nanopore Technologies[59] developed the world's first real-time, long-reading, nanopore DNA and RNA sequencer in 2015. Recent advances in their technologies are detailed in the box.

"OXFORD NANOPORE TECHNOLOGIES" DEVICES

This technology is the only one capable of sequencing native molecules without prior PCR amplification. Thus, the chemical modifications present on the nucleic acids, which constitute additional information, are preserved. A modification is detected by a change in the ion flow as the modified nucleotide passes through the nanopore, since this differs from that of the same nucleotide without the modification. Finally, unlike other sequencing methods currently on the market, the Oxford Nanopore technology can sequence very long nucleic acids. Indeed, molecules of the order of a million nucleotides have been sequenced, the record being 2.2 million.

There are several devices of different sizes and capacities. The MinION device is the only portable, real-time DNA/RNA sequencing device. It weighs less than 100 g and connects to a computer using a USB cable. It is composed of a sequencing box, which costs US$1,000, and a consumable measuring cell, which costs US$500-900. Each measuring cell interrogates up to 512 nanopores simultaneously and generates 10-30 GB of sequencing data per experiment.

The GridION X5 is a compact benchtop device. It has five measuring cells of the MinION type. Up to five experiments can be carried out individually and simultaneously. It generates up to 150 GB of data and analyzes it in real time. The PromethION was marketed in 2018. It offers the same real-time, long DNA/RNA sequences reading technology as the MinION and GridION, but on a much larger scale. The current system uses up to 24 measurement cells at a time. Each measurement cell uses up

▶ ▶

59 https://nanoporetech.com/

to 3,000 nanopores, for a total theoretical yield of 15 TB of data in 48 hours. This device is designed to sequence complete human genomes for less than US$1,000. In 2019, a PromethION system with 48 measurement cells was commercialized.

Oxford Nanopore Technologies devices

Credit: Oxford Nanopore Technologies - https://nanoporetech.com/products

Recently, devices smaller than the MinION have been marketed. The Flongle is an adaptor for the MinION or GridION and sequences DNA/RNA in real time on much smaller disposable measuring cells. It is designed to be the fastest and cheapest sequencing system on the market. The SmidgION, still under development, will be the smallest commercialized device and it will possible to plug it in a mobile phone.

An ionic flux is measured for each nucleotide that passes through the nanopore. It is converted into a DNA sequence in real time thanks to the operating software of the company: MinKNOW. A sequenced DNA molecule is represented by a file whose format follows the international databases standard. These files are stored locally and can only be accessed by the user. In order to be able to sequence DNA or other polymers without a computer and without internet access, Oxford Nanopore Technology has designed MinIT, a stand-alone unit that accompanies the MinION sequencer. It is pre-configured with MinKNOW software and performs high-volume data acquisition, analysis and storage. Due to the fast data flow generated by nanopore technology, this MinIT unit is equipped with a 500 GB hard disk drive (SSD) and powerful graphics processor units (GPUs).

Improvements are underway to increase the longevity of nanopores and reduce the error rate. The enzymatic reaction between the DNA-binding motor protein and the reading protein requires a source of energy, which is depleted after 48 hours. By optimizing the energy source, the same nanopore will be able to sequence longer, up to 160 hours. In addition, thanks to a new electronic circuit integrated in the membrane, there is no longer any loss of current, which preserves the nanopores. With these improvements, the sequencing capacity of a nanopore has doubled in just a few years and is expected to continue to increase.

The reading head of the current nanopore reads one nucleotide at a time, but the immediate environment (the neighboring five nucleotides) influences the nature of the signal. Thus, errors are frequent when sequencing a polymer with sequences of more than five consecutive identical nucleotides. A new type of nanopore was designed with two reading heads, which improves the quality of sequencing by a factor of ten. This new nanopore was commercialized in 2019.

The company is currently working on nucleic acid analogues. These analogues are added to the DNA sequence to calibrate the electrical signal measured by the nanopore. As a result, the sequencing error rate will be further reduced.

Conclusion

Illumina, the world market leader in sequencing, has a reliable and robust technology. Their high-capacity devices sequence a human genome quickly and for less than US$ 1,000. The Oxford Nanopore technology is very promising. It is miniaturized and therefore accessible to any laboratory at low cost (less than US$ 1,000 for a MinION). It has the important advantage of being able to read native DNA molecules containing modified nucleotides, or even to read non-nucleotide polymers if the components (monomers) of the molecule (polymer) can be differentiated by their electrical signal. According to its users, the Oxford Nanopore technology still has too high a sequencing error rate (10 % sequencing errors compared to 1 % for the Illumina technology), but it should be noted that these errors are largely corrected during the

data analysis stage by comparing the readings of the multiple copies passing through the numerous nanopores. Large corporations, such as Microsoft Corp., seem to estimate that this technology represents a major asset for reading information stored in DNA. Some companies, such as Helixworks, have even optimized their DNA synthesis technology so as to use Oxford Nanopore Technologies devices.

Improvements in systems for archiving digital information in DNA

Limiting factors

Light, water and oxygen have a deleterious effect on nucleic acids. They cause chemical reactions that are responsible for breaks or mutations in DNA, making its information content impossible to decipher. However, DNA kept safe from its "enemies" can be preserved for several thousand years. It should be noted that DNA molecules have been recovered from fossil bones more than 560,000 years old buried in dry, cool soil; their analysis has made it possible to compare the DNA sequence of extinct animals with that of their modern descendants. Conventional methods of DNA storage mainly use low-temperature storage (-20°C to -196°C). Unfortunately, these methods are difficult to automate and are expensive in terms of space, equipment, energy and maintenance. In addition, they expose samples to risks of degradation, contamination or loss in the event of hardware or electrical failures. Freeing DNA preservation from a reliance on low temperature would therefore constitute a considerable technological, economic and ecological advance in information storage, provided that the stability of the samples was ensured.

Improving DNA long-term storage technologies

Chemical storage: Robert Grass' technology

Based on the observation that DNA can be preserved in fossil bones, the group of Robert Grass (ETH Zurich, Switzerland)[60] came up with the idea of encapsulating DNA by a chemical process to improve its preservation. They made nanobeads out of glass (*i.e.,* silicon dioxide) that store DNA at room temperature and preserve it from oxygen through a chemical encapsulation process. Each nanobead is about 100 nm in diameter. Recent advances and planned improvements in this technology are discussed in the box.

60 https://fml.ethz.ch/the-lab/people/lecturer.html

DNA storage in glass nanobeads

Robert Grass'group modifies the surface of the glass beads so that it becomes positively charged. Negatively charged DNA is attracted to the surface of the beads. A chemical molecule, possessing a positive charge on one side and a silica precursor compound on the other, is added. The positively charged side attaches to the DNA to coat it and the side containing the silica precursor compound forms a solid glass layer enclosing the DNA. This process takes place in the aqueous phase, which means that the encapsulated DNA remains in contact with a small number of water molecules. The de-encapsulation process, *i.e.*, the release of the DNA, consists in treating the glass beads with a fluoride solution. At the low concentration used, the fluoride degrades the glass without damaging the DNA. The released DNA is amplified by PCR and then sequenced to decode the information it contains.

The DNA that has been encapsulated in the nanobeads for several years is completely preserved, with no change whatsoever in its nucleotide sequence. As the technology is recent, it is impossible to directly evaluate the conservation of DNA in the nanobeads over more than a few years. Degradation kinetics were therefore carried out at high temperatures to mimic the effect of ageing on the samples. These experiments showed that the DNA stored in the nanobeads would be preserved and analyzable for several decades at room temperature and for 1 million years at -18°C[1]. In short, this technology makes it possible to significantly increase the stability of the DNA and therefore the shelf life of the digital information it contains. However, the density of stored information decreases as the encapsulated DNA constitutes only 0.1 % of the mass of the nanobead.

The research group is collaborating with Microsoft Corp. to increase the density or amount of DNA stored per nanobead. They are making magnetic nanoparticles[2] capable of storing several successive layers of DNA. A layer of DNA is deposited on a positively charged magnetic nanoparticle. A layer of positively charged polymer is then deposited on the DNA. Finally, a new layer of DNA is deposited on the positively charged polymer. The cycle is repeated to obtain several successive layers of DNA on the nanobead. Thanks to this multi-layer storage technology, the group has increased the amount of DNA stored per bead. The DNA then represents 3 % of the mass of the nanobead. The group currently works on the reading of DNA information from these multi-layered beads.

One of the applications of this technology concerns the storage of computer data in DNA, at a cost of one euro per KB. Grass' group has synthesized DNA containing the digital information from a 1.4 MB film, encapsulated this DNA in the nanobeads and printed glasses containing these nanobeads. Similarly, with the participation of the music group Massive Attack, they encapsulated the DNA containing the soundtrack of their latest album, which they then added to a spray paint can. One of the band's singers was thus able to "paint his music".

Since nanobeads can be safely swallowed they can also be used to identify food fraud. In addition, they are used for tracing gemstones; the Haelixa company[3] has developed a solution of tracer nanobeads that are applied to the gemstone during the polishing process. Nanobeads can contain a DNA message longer than a barcode, such as a product data sheet or user manual. They are then printed into the product using a 3D printer, and may be read by sequencing. The idea of storing the information associated to an object in the object itself, and not in its packaging, has certain advantages: the information (user manual, technical notice) is not likely to be lost and the quantity of the product packaging is reduced.

1 Temperature of the Global Seed Vault in Spitsbergen, Norway.

2 Magnetic particles are easier to handle in solution than glass nanoparticles.

3 http://www.haelixa.com/

Physical storage: the Imagene technology

The Imagene company[61] (France) makes conservation capsules, called DNAshell®, which store DNA and preserve it from water, oxygen and light. Each capsule can contain up to 0.8 g of DNA, *i.e.* 1.4 EB of data including redundancies[62]. The capsules consist of a stainless-steel case of a few millimeters in size, enclosing a glass insert in which the DNA is deposited. Each capsule is marked for traceability and is compatible with the standard 96-well plate format used in lab-

61 http://www.imagene.fr

62 1 Exabyte (EB) represents 1 million 1 Terabyte (TB) hard disks.

oratories. Details of the ongoing improvements and performance of the Imagene technology are summarized in the box below.

DNA storage in steel capsule

The DNA encapsulation process, developed by Imagene, is fully automated. The DNA in solution is deposited in the glass insert of the stainless-steel capsule. In a first drying step, it is dried under vacuum. A second drying step is carried out under an atmosphere of neutral, anoxic and anhydrous gases (a mixture of argon and helium). This DNA is then encapsulated by sealing the metal plug onto the capsule by laser welding. Capsule tightness checks are carried out: the gas mixture contained inside the capsule is easily detectable by mass spectrometry in case of leakage.

This technology thus shields DNA from its "enemies": it is not in contact with water, oxygen or light. In addition, this system is autonomous and does not consume energy during storage. The sequence of the DNA encapsulated by Imagene can be completely preserved for several years without any modification. Since this technology is relatively recent, it is impossible to directly determine the degree of conservation of the DNA in the capsules over more than a few years. That said, the degradation kinetics carried out by Imagene at different temperatures and extrapolated according to the Arrhenius law[1], give an estimated half-life of 52,000 years for DNA, stored in these capsules at room temperature.

Imagene capsules can also be used to store other biological samples, such as RNA and blood, and the company is continuing to develop this technology so as to conserve molecular biology reagents (enzymes, reaction media) and microorganisms (viruses, yeasts, bacteria).

1 In chemical kinetics, the Arrhenius law describes the variation of the speed of a chemical reaction as a function of temperature.

Storage of digital information in DNA *in vivo*

Project

The majority of the research into the storage of digital information in DNA focuses on *in vitro* storage: digital data is encoded as a nucleotide sequence that is then synthesized and stored in capsules or beads. Some researchers, however, focus on storing digital DNA information *in vivo*, so that the DNA would be protected in cells or other organisms. This approach has several advantages including the fact that some types of cells can grow and divide in cheap growth media, which results in the amplification not only of their own DNA but also of any foreign DNA.

Several scientific groups have therefore been studying the storage of digital information *in vivo*. By 2007, the researchers who subsequently launched the MOSLA project had already integrated synthetic oligonucleotides, encoding logos and company names, into the genome of bacteria. In 2016, George Church's group integrated dozens of bytes of information into *E. coli* using CRISPR technology[63]. In 2020, an artist, Joe Davis, searched for the sturdiest possible naturally occurring container to preserve the DNA archives of humanity after our species has gone extinct. He chose *Halobacterium salinarum*, an archaea that carries more than 20 copies of its chromosome per cell and that can survive in saline deposits for several hundred million years[64].

One of the objectives of the MOSLA project already mentioned above in this chapter is to construct artificial chromosomes and plasmids[65] containing digital information. The aim of this

63 Shipman SL, Nivala J, Macklis JD, Church GM (2016). Molecular recordings by directed CRISPR spacer acquisition. Science 353,6298

64 https://www.sciencemag.org/news/2020/02/hardy-microbe-s-dna-could-be-time-capsule-ages

65 Circular DNA fragment, independent of chromosomal DNA, natural or artificial.

project is to generate a "cloud" of bacterial cells each containing a fragment of this information in the form of a mega-chromosome. This project therefore requires the ability to design and build large bacterial chromosomes.

Synthetic chromosome design

Designing synthetic chromosomes requires high-performance molecular biology tools. Moreover, to avoid the loss of artificial chromosomes from cells, the natural mechanisms of maintenance and replication of DNA in cells must be well understood and taken into account. This is because the synthetic chromosome could be degraded by the cell, or undergo mutations, or partially recombine with and perhaps completely integrate into the cell's natural chromosome, or not be maintained during cell division.

To this end, the group of Torsten Waldminghaus (SYNMIKRO, Philipps Universität Marburg, Germany)[66] is developing techniques for assembling small fragments of DNA into synthetic chromosomes, which are stored in cells. They have introduced two 100,000 nucleotide synthetic chromosomes into *E. coli* to study how they are affected by the normal repair and maintenance processes.

Synovance (Genopole Évry, France)[67] designs and builds large synthetic DNA chromosomes and bacterial strains optimized for bioproduction. To do this, the company has developed appropriate computational biology and DNA assembly methods.

Storage of information in spores

Another of the objectives of the MOSLA project (see above) is to exploit the properties of bacterial spores so as to store the digital information contained in DNA over a long period. This is because spores have remarkable resistance characteristics and can survive for several thousand years, even under unfavourable conditions. Sporulation occurs in some species of bacteria when conditions become unfavourable for growth, such as a lack of nutrients, water, etc. Sporulation is characterized by a thickening of the cell wall which is accompanied by dehydration since the presence of water is an important factor in the degradation of DNA by hydrolysis.

The objective is therefore to introduce the synthetic chromosome into a cell and then to induce sporulation.

Limitations of **in vivo** *information storage*

There are disadvantages to storing information *in vivo*. On the one hand, the volume of a cell is relatively small, so the quantity of DNA it can contain is limited; this problem may partly be solved by the distributive approach described above (fragmenting the information and distributing the different fragments to different cells). On the other hand, the DNA carried by cells naturally undergoes mutations, which would alter any stored digital information; this problem

66 https://synmikro.com

67 https://synovance.com

may partly be solved by incorporating error correction codes into the stored DNA.

Finally, the DNA containing the digital information must be tolerated by the organism. Even if the nucleotides, A, T, C, G, used to encode this information are the ones used by the cell (and not analogues), this information must not correspond to nucleotide sequences that are toxic in some way. This is because a particular sequence of encoded information may have a biological meaning for the host cell, unlike the vast majority of such digital information which would be meaningless for the cell. Such rare, unfortunate coincidences may occur if extensive use is made of the *in vivo* approach for storing big data. These coincidences might lead to cells interpreting a DNA sequence containing digital information as a chromosome maintenance signal, a binding site for regulatory elements, a code for an RNA, or for a protein which could act as a catalyst. Among these rare cases, some may be toxic to the carrier cell or dangerous to the surrounding biotope. If they are toxic for the carrier itself, random mutations making them harmless will be preferentially selected and rapidly fixed in the carrier population; the consequence will be a rapid drift of this sequence away from the original and thus the loss of the stored information. If they are interpreted but are not toxic to the carrier, there is still a low probability that the carrier will produce a compound that is toxic to humans or to other parts of the surrounding biotope.

This potential safety issue with *in vivo* storage, however tenuous, is not encountered in the predominantly *in vitro* approach described in the previous sections. The drawbacks of the *in vivo* approach could slow down its development[68].

Non-DNA molecular information storage systems

Polymers are macromolecules with numerous repeating subunits (monomers) in their structures. In principle, any polymer with at least two different monomers could be used to store digital information. In practice, it must additionally be possible to write this polymer as an arbitrary sequence (as determined by the digital file to be archived), *i.e.*, by iterative solid-phase chemistry. There must also be methods to keep it for a long time and to read it easily. Ideally, this polymer should have an even higher information density than DNA.

Data archiving systems using such non-DNA polymers are thus being investigated.

Artificial polymers

The academic project led by Jean-François Lutz (Charles Sadron Institute, University of Strasbourg, France) uses non-DNA copolymers to store digital information[69]. Such polymers may be:

– natural, such as DNA or polysaccharides; or
– artificial, man-made materials, such as nitrocellulose;

Synthetic polymers, such as plastics, have considerable potential for storing digital information. They allow for greater information density and better data retention than current electronic storage media. To store digital information in synthetic polymers, the general idea is to translate

68 It is noteworthy that the *in vivo* approach was the only mention in response to the "Questionnaire on Ethics and Technology" approved by the NATF. This mention was made by a very small number of the interviewees, who were all the members of the working group and the invited speakers.

69 Colquhoun H & Lutz JF (2014). Information-containing macromolecules. *Nature* Chemistry 6:455-456.

the sequence of "bits" of the digital file (sequence of '0' and '1') into a sequence of monomers, which is then synthesized as a polymer by chemistry. Thus, an alphabet based on two different monomers allows the writing of binary information in a linear polymer chain. Polymers used for archiving digital information are called "digital" polymers in what follows.

To read the information, polymers are "sequenced" by mass spectrometry. This provides an extremely precise measurement of the mass of a molecule (the mass spectrum), which can be used to detect and identify it. In the case of a digital polymer, the mass spectrum is computationally decoded to reconstruct the message in bits, and thus the initial information. The following box describes the progress being made in the field of digital polymers.

Digital polymers

The storage of digital information in synthetic polymers requires the following features:

- Monodispersity: the polymer must be uniform, that is, composed of monomers with molecular weights in the same range and with similar structures (albeit different in order to be distinguished from one another). This facilitates the reading of the polymer by mass spectrometry.
- Selectivity: the polymer must offer the possibility of orthogonal reactions, *i.e.*, of performing protection and deprotection steps of one group of atoms without influencing the protection and deprotection of another group of atoms. This facilitates the writing of the polymer.
- Possibility of binary encoding: the polymer must have at least two different units, such as a methyl group for a '1' and a hydrogen atom for a '0'.
- Ease: the polymer must be easily and quickly assembled from its monomers.

There are many synthetic polymers that could be used to store digital information more densely than in DNA. However, the conditions for the incorporation and detection of each monomer vary. Lutz's group routinely uses up to eight monomers.

Phosphodiester polymers have a major advantage over other synthetic, digital polymers as they are used in phosphoramidite chemistry for the chemical synthesis of DNA[1]. Thus, they are compatible with DNA molecule synthesizers and can be assembled in an automatic and programmable way.

Every eight monomers, a molecular separator, which may be cleavable, is introduced. For each monomer added, the yield of the chemical bonding reaction is about 99 %, which limits the length of the polymer to 100-150 monomers.

The information stored in the polymer chains can be edited using physical triggers such as temperature or light. This information can also be erased after polymerization[2]. Such erasure is straightforward because some monomers are stable at room temperature but unstable at higher temperatures. There are also light-sensitive monomers that lose some of their information after exposure. In this case, the polymer is not destroyed, but the information it contains is erased. Conversely, exposure to light can reveal the information contained in a polymer — there is a family of monomers that are unreadable before exposure to light. In this case, light releases one of the chemical groups from the monomer, allowing it to be read[3]. This property can be used, for example, in measures to prevent counterfeiting. Finally, the information can be modified after polymerization, as some monomers can be transformed into other monomers when exposed to light. This property is useful for modifying the information that is already in a digital polymer.

To read the information stored in the polymers, there are three sequencing techniques:

- tandem mass spectrometry;
- nanopore sequencing;
- ion beam deposition sequencing and scanning tunnelling microscopy (STM).

The most widely used of these techniques is tandem mass spectrometry. This technique makes it possible to decipher short and long digital sequences. It detects and identifies molecules by measuring their mass to charge ratio. The result is called a mass spectrum. Some spectra are complex and time-consuming to interpret. To optimize the reading of digital polymers, they are assembled from monomers that have skeletons of the same molecular weight, but different chemical ►►

1 Al Ouahabi A, Charles L, Lutz JF (2015). Synthesis of Non-Natural Sequence-Encoded Polymers using Phosphoramidite Chemistry. *JACS* 137:5629-5635.

2 Roy RK, Meszynska A, Laure C, Charles L, Verchin C, Lutz JF (2015). Design and synthesis of digitally encoded polymers that can be decoded and erased. *Nat Comm* 6 :7237.

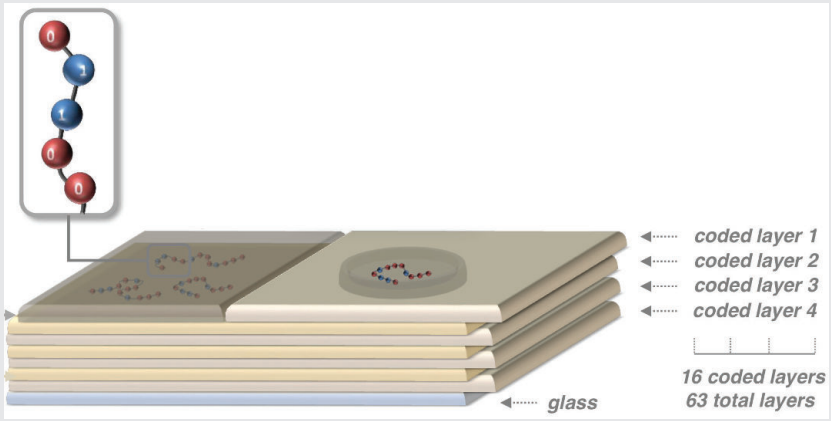
3 König NK, Al Ouahabi A, Oswald L, Szweda R, Charles L, Lutz JF (2019). Photo-editable macromolecular information. *Nat Comm* 10 :3774.

groups with readily identifiable mass to charge ratios. For long chains, separators positioned every eight monomers can be cleaved to give octamers that can be accurately identified by spectrometry. The Lutz group has developed decoding software, called MS-DECODER, which can transcribe the mass spectrum of each of these octamers into bytes[4]. In this way, the data can be analyzed by any user.

Within a few years, it should be possible to read digital polymers by nanopore sequencing such as that developed by Oxford Nanopore Technologies. The performance in terms of price and speed could then be comparable to that of DNA.

Moreover, a method is being developed for storing information in layers of polymers in order to densify the information. Each layer, separated from the others by glass slides, would contain polymer chains. Unlike DNA, synthetic polymers do not need to be encapsulated to protect their information. Polymers such as plastic are very robust and take thousands of years to be degraded.

Digital polymers are used for the traceability of certain products as they can contain information such as a barcode. They can be added to the surface of an object or incorporated into the object itself. Jean-François Lutz's group added a digital polymer containing a barcode to a plastic biomedical implant. After several years *in vivo* in rats, the implant was removed and the barcode was shown to be intact. Surface mass spectrometry can also be used to decode polymers on the surface of objects. One can imagine a near future where digital polymers, containing barcodes, would be printed on the surface of banknotes for traceability and counterfeit resistance.



Layers of synthetic polymers for digital information storage

Credit: Jean-François Lutz.

4 Burel A, Carapito C, Lutz JF, Charles L (2017). MS-DECODER: Milliseconds Sequencing of Coded polymers. *Macromolecules* 50:8290-8296.

Organometallic compounds

In the MOSLA project, a new storage medium based on organometallic compounds is being developed to increase information storage capacity.

These compounds consist of carbon and metal atoms and are completely artificial. They are capable of reflecting light by emitting different wavelengths. In contrast to the linear polymers discussed so far, these compounds will be printed on a surface, such as that of a compact disc. With this technology, more information will be stored on the same surface.

In order to read the information printed on these media, special scanners and readers are being developed.

Conclusion

The storage of digital information in DNA obviates the need for low temperatures, which constitutes a considerable technological, economic and ecological advantage. To this end, two storage technologies are being developed:

– chemical — with Robert Grass' storage system, the synthesized DNA is encapsulated in silica nanobeads. The drawbacks of this approach are low information density, and incomplete removal of water and oxygen, which can degrade DNA.
– physical — Imagene's system stores huge amounts of information in a small capsule. It protects DNA from water, salts, oxygen and light to conserve it for thousand of years at room temperature.

The *in vivo* digital DNA information storage proposed by the MOSLA project has various disadvantages not shared by *in vitro* systems.

Very promising systems for storing digital information in non-DNA 'digital' polymers are under development in Jean-François Lutz's group.

Finally, unorthodox approaches have also been proposed, for example, based on the formation and recognition of secondary stem-loop DNA structures (of two distinctly different lengths representing '0' and '1'), with the advantage of a lower error rate, and the disadvantage of a loss of information density[70].

Indexing and computing with synthetic DNA

Indexing

As a general rule, it is not possible to access a small portion of the information stored in the DNA without reading all the information present. New architectures for information storage in DNA are therefore being developed.

The MOSLA project (Germany) uses specific nucleotides and oligonucleotides to create a storage system that can be used to selectively access information.

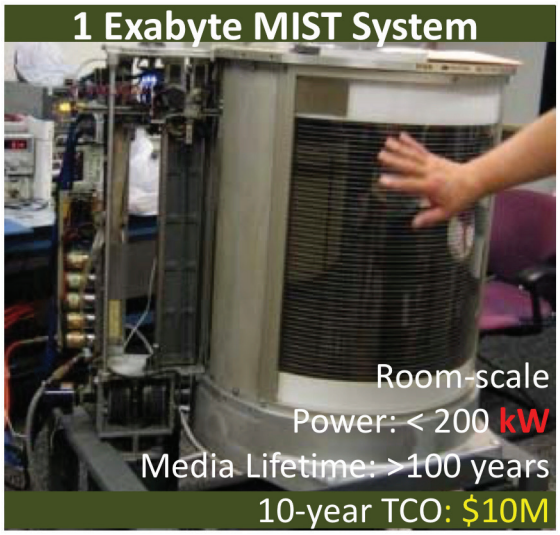
Catalog DNA (USA) is working on nucleotide labels in which each assembled DNA fragment can be used as a label. These labels make it possible to group together DNA molecules containing information from the same digital files.

Researchers at the University of Illinois are working on a new architecture[71] that will allow to access blocks of data and to rewrite information already stored in these locations. It is based on DNA sequences with specialized addresses that can be used for selective access to information.

Finally, the MIST program (IARPA, USA) is considering building a benchtop device capable of random access to the information stored on the molecular support. An operating system would also be designed for the storage device to coordinate indexing, addressing, data compression, error correction and translation of binary information into DNA sequences and *vice versa*.

70 Chen K, Kong J, Zhu J, Ermann N, Predki P, Keyser UF (2019). Digital Data Storage Using DNA Nanostructures and Solid-State Nanopores. *Nano Lett.* 19(2):1210-1215.

71 Hossein Tabatabaei Yazdi SM, Gabrys R & Milenkovic O (2017). Portable and Error-Free DNA Based Data Storage. Scientific Reports 7(1):5011.



Objectives for the MIST information storage device.
Credit: IARPA and https://commons.wikimedia.org/wiki/File:IBM_350_RAMAC.jpg (license: CC BY-SA 2.5)

Computing with synthetic DNA

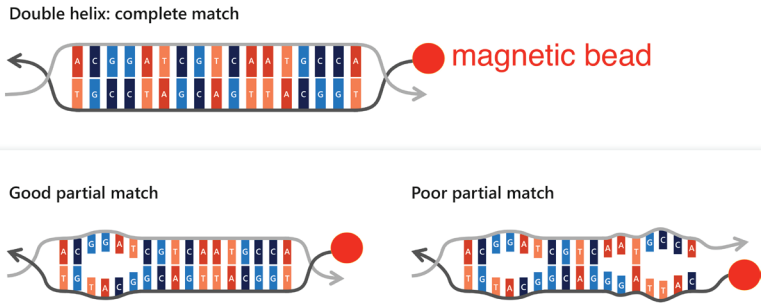
Microsoft Corp. has demonstrated that it is possible to exploit data stored in DNA. It has designed a computer system that combines the storage and processing of molecular data; it takes the form of a hybrid electronic-molecular architecture that exploits the strengths of both the electronic and chemical domains[72].

The basic principle of DNA computing is based on the physico-chemical properties of DNA and, in particular, nucleotide pairing. When a single-stranded DNA fragment meets its complementary strand, these two strands hybridize to form a double-stranded DNA. However, sometimes two strands of DNA can hybridize without having perfectly complementary nucleotide sequences. This is called partial hybridization. Using the principle of hybridization, it has proved possible to find images similar to a ***target*** image among a wide range of images in a library. Further details are in the box.

As a solution to combinatorial problems, DNA computing has the disadvantage of slowness but the advantages of high parallelization and high energy efficiency[73]. Finally, it should be noted that the same DNA library can be reused to solve other problems that have different target images.

72 Organick L, Ang SD, Chen YJ, Lopez R, Yekhanin S, Makarychev K, Racz MZ, Kamath G, Gopalan P, Nguyen B, Takahashi CN, Newman S, Parker HY, Rashtchian C, Stewart K, Gupta G, Carlson R, Mulligan J, Carmean D, Seelig G, Ceze L, Strauss K (2018). Random access in large-scale DNA data storage. *Nature Biotechnology* 36(3) :242-248.

73 Adelman LM (1994). Molecular computation of solutions to combinatorial problems, Science 266, 5187:1021-1024.



Pairing two DNA strands according to their respective sequences.
Credit: adapted from the presentation by Karin Strauss (Microsoft Corp.) by François Képès.

DNA computing

Suppose that the objective of the calculation is to find in a database containing hundreds of pictures, the one that most closely resembles a picture of interest (the target picture). The idea is to convert the characteristics of each picture in the database into vectors, and then code these vectors into single-stranded DNA fragments. Thus, the picture database is represented by a library of different single-stranded DNA fragments mixed in a liquid solution. Pictures from the database with similar characteristics are represented by similar DNA sequences in this library. The target picture is likewise represented by a single-stranded DNA fragment. However, to constitute a probe, its complementary strand is synthesized, to whose end a magnetic bead is chemically attached.

The objective of finding the pictures in the database that most closely resemble this target picture is implemented as a calculus with DNA: which single-stranded DNA fragments in the library best match the magnetic complementary single-stranded DNA probe? The computing is performed by adding the probe into the library. If a picture perfectly resembles the target picture, the corresponding DNA fragment will fully hybridize (reconstitute a double-stranded DNA) with that of the probe DNA. If a picture slightly resembles the target picture, the corresponding DNA fragment will partially hybridize to that of the probe DNA. Finally, if there is no resemblance, there will be hardly any hybridization. To test for these different possibilities, the probe attached to the magnetic bead is extracted from the solution with a magnet. In the case where the probe has hybridized to a DNA fragment in the library, this fragment will be extracted along with the probe; sequencing this fragment will then reveal the picture in the initial database that resembles the target picture. By modulating the temperature or salinity , the operator can vary the stringency of the hybridization conditions of the two DNA strands, which would correspond to altering the threshold of detectable similarity between the target picture and any picture in the database.

Chapter V

Global Initiatives

The European Commission published a *European Data Strategy* in February 2020[74]: it is notable that the terms "DNA" and "polymer" were absent from this document.

United States of America

In this emerging field, the public investment in the USA is about US$150 million. This is divided between the three agencies: DARPA, IARPA (MIST project[75]) and NSF. NSF and IARPA contribute to the SemiSynBio project[76], which published in 2018 a roadmap with two- and four- year quantified objectives[77]. George M. Church (Harvard Medical School and MIT) has been one of the pioneers in the field since 2012[78]. In addition, several private companies are active in this field, such as Microsoft Corp., Twist Bioscience, Catalog DNA, and other companies involved with enzymatic DNA synthesis.

MIST (IARPA)

MIST (Molecular Information STorage) is a four-year US program, started in January 2019 by IARPA[79], and operating largely through competitive bidding. It has invested US$48 million to develop new information storage technologies. An initial envelope of US$23 million is funding a consortium that includes Illumina (USA), DNA Script (France), and researchers from MIT and Harvard University (USA). The remaining US$25 million has been allocated to a second consortium that brings together Microsoft Corp. and Twist Bioscience[80].

The objective of the MIST project is to store a large amount of information (of the order of 1 EB), in a minimum amount of space (of the order of mm^3), with reduced financial and energy costs compared to current storage systems. The ultimate goal is to store 1 EB of data from a

74 "*A European Data Strategy*". Communication from the European Commission to the Parliament, the Council, the Economic and Social Committee and the Committee of the Regions (2020).

75 https://www.iarpa.gov/index.php/research-programs/mist

76 https://www.src.org/program/grc/semisynbio/

77 https://www.src.org/library/publication/p095387/p095387.pdf - Chapter 1.

78 Church GM, Gao Y, Kosuri S (2012). Next-Generation Digital Information Storage in DNA. *Science* 337,1628. http://nook.cs.ucdavis.edu/~koehl/Teaching/ECS129/Reprints/Church_DNAStorage_12.pdf

79 IARPA (Intelligence Advanced Research Projects Activity) is an organization of the Office of the Director of National Intelligence of the United States of America. Its research is applied by the CIA, FBI and NSA.

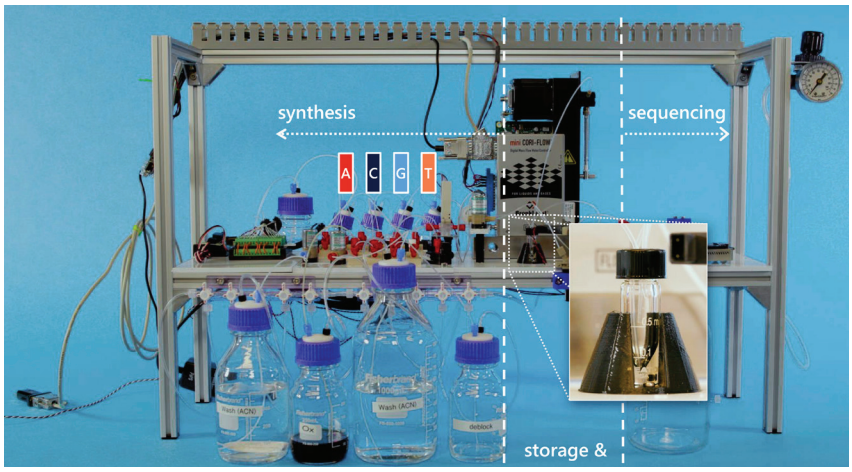
80 https://www.lemonde.fr/economie/article/2020/01/22/le-gouvernement-americain-investit-dans-le-stockage-de-donnees-dans-l-adn_6026763_3234.html

datacenter in the MIST device, so that, compared with the information stored in current systems, the volume is 20,000 times smaller, the maintenance requires 1,000 times less energy, the lifetime is at least 20 times longer, and the cost is 100 times less.

To achieve this, MIST will use polymers as data storage media. It will also design the devices and operating systems necessary to interface with this medium. Technologies will be developed to optimize the writing and reading of information in these polymers, and to allow random access to the information stored in them.

University of Washington / Microsoft Corp.

The most advanced project to date in the field of storage or archiving of digital DNA data is coordinated by Karin Strauss (Microsoft Corp. and University of Washington)[81]. They have designed a benchtop prototype using DNA as a medium for storing information[82]. The device is fully automated and self-contained. It is composed of three parts: the synthesizer, the storage system and the sequencer. The synthesizer encodes digital information that it then synthesizes as DNA sequences. The storage system encapsulates this DNA in nanobeads. The DNA is then extracted from the nanobeads and sequenced using the Oxford Nanopore Technologies device.



First fully automated prototype for DNA data storage
Credit: Microsoft Corp./University of Washington, USA. Adapted by François Képès.

This prototype is functional and has already made it possible to store and retrieve 1 GB of data[83]. It is now being optimized so as to be more compact and faster. New systems based on microfluidics are also being developed to transport drops of reagents on an electronic support[84]. The same group has also launched a project to perform computation directly using

81 https://www.microsoft.com/en-us/research/blog/storing-digital-data-in-synthetic-dna-with-dr-karin-strauss/

82 Takahashi CN, Nguyen BH, Strauss K, Ceze L (2019). Demonstration of End-to-End Automation of DNA Data Storage. *Scientific Reports* 9(1):4998.

83 Ceze L, Nivala J, Strauss K (2019). Molecular digital data using DNA. Nat Rev Genet 456:466.

84 Newman S, Stephenson AP, Willsey M, Nguyen BH, Takahashi CN, Strauss K, Ceze L (2019). High density data storage library via dehybridization with digital microfluidic retrieval. *Nat Commun* 10(1):1706

the physico-chemical properties of DNA (see chapter III). At this stage, their system makes it possible to find the group of images that are similar to a target image in a large set of images (as described in the above box).

Twist Bioscience

Twist Bioscience is an American start-up listed on the NASDAQ. In five years, the company has raised US$190 million. It specializes in the synthesis of DNA on microchips by chemical means, even though its technologies are compatible with future developments in enzymatic synthesis. It is participating in several projects on the storage of digital information in DNA:

- The Iconem[85] project, which consists of digitizing models of heritage sites in three dimensions and archiving this information in DNA.
- The ARCH mission[86], in collaboration with Microsoft Corp. and the University of Washington. The mission of this project is to archive a collection of photographs from around the world in DNA for future generations.
- DNA storage of several audio recordings of the Montreux Jazz Festival in 2017.

Catalog DNA

Catalog is an American start-up whose goal is to turn DNA into a storage medium for digital information. In 2018, the company received US$9 million from various private companies.

Catalog has built a machine capable of synthesizing DNA encoding 0.5 MB of information every second. It converted the entire Wikipedia library (14 GB of information) into a virtual DNA sequence, which was then synthesized by this machine. This is the absolute record at the moment. As it is based on the use of pre-synthesized DNA fragments, it is not directly comparable to the 1 GB record held by Microsoft Corp./Univ. Washington which assembles DNA one nucleotide at a time.

This machine is being modified in order to increase its capacities and performance. The goal is to build a new version of the machine that will be able to synthesize DNA several orders of magnitude faster than the present model and to process 0.12 GB of information per second.

China

In China, it is difficult to get a clear picture of the situation, but it seems that Huawei and BGI Genomics are involved in this area.

Israel

In Israel, Technion's project, led by Zohar Yakhini, involves storing digital information in DNA more densely. This group is working on a new alphabet of composite letters. They have converted a

85 http://iconem.com/

86 https://www.archmission.org/

6.4 MB file into DNA nucleotides, using an alphabet of five or six composite nucleotides[87]. They are also working on an alphabet of 20 nucleotides.

United Kingdom

Located in the United Kingdom, the European Bioinformatics Institute (EBI) is one of the pioneers in this field[88]. In addition, several private companies are also active in this field, such as Oxford Nanopore Technologies, Nuclera Nucleics and Evonetix Ltd.

Researchers from Nick Goldman's group (EBI) demonstrated as early as 2013 that it is possible to use DNA to store and retrieve digital information. They converted four digital files corresponding to documents of varied nature (Text, JPEG, PDF, mp3) into DNA sequences. The 0.7 MB contained in the total of the four files was reconstructed without error. Their results were published in 2013[89].

Ireland

Helixworks is an Irish start-up that manufactures and sells DNA-based data storage technologies. They are developing the Molecular Storage System (MoSS) technology to convert and synthesize digital files into DNA. The company carried out the Fusion project[90] in partnership with Ubisoft, Ambey and AKQA. The objective of this project is to store the digital information of a video game in a 250 mL energy drink can.

In addition, Helixwork participates in the OligoArchive project[91], which was financed in 2019 by the European Innovation Council (EIC) for a period of three years and for a total of 3 million euros. Its objective is to design a benchtop device using DNA as an information storage medium. This device will be capable of converting digital information into a virtual DNA sequence, synthesizing the corresponding sequence, storing it for a long time and sequencing it to extract the stored information. The other participants are Imperial College London (UK), the Institute of Molecular and Cellular Pharmacology (IPMC, Sophia-Antipolis, France), the Computer, Signals and Systems Laboratory (I3S, Sophia-Antipolis) and Eurecom's Data Science Department (Sophia-Antipolis).

87 Anavy L, Vaknin I, Atar O, Amit R, Yakhini Z (2019). Data storage in DNA with fewer synthesis cycles using composite DNA. *Nat Biotechnol* 1229-1236

88 https://www.ebi.ac.uk/research/goldman/dna-storage

89 Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, Birney E (2013). Toward Practical, high-capacity, low maintenance storage of digital information in synthesized DNA. *Nature* 494(7435):77-80.

90 https://www.fusiondna.com.br/

91 https://oligoarchive.github.io

Germany

In Germany, 4.2 million euros of funding has been granted by the Hessian Ministry for the MOSLA[92] project (Universities of Marburg, Darmstadt, Giessen, 2019-2022).

The objective of the MOSLA project is to develop new approaches and solutions for long-term information archiving based on molecular and chemical storage systems. The vision of the project is to increase the current storage capacity. The project has four parts:

- optimize digital data storage;
- develop a system for storing information in DNA, *in vitro* and *in vivo*;
- develop a storage system based on organometallic compounds;
- organize the storage of man-made information.

France

An academic project on the use of non-DNA copolymers is led by Jean-François Lutz (Charles Sadron Institute, CNRS and University of Strasbourg)[93]. The company DNA Script (Paris) is well positioned in the field of enzymatic DNA synthesis; in addition to private investment, it received significant funding from the USA in 2020 (see above). The Imagene company (Bordeaux and Évry) has a strong position in the field of very long-term DNA storage. Finally, three laboratories (Sophia-Antipolis) are involved in an international project financed by the EIC (see above).

It should also be noted that in 2014, the best equipped countries in the world in terms of large datacenters were, in decreasing order, the United States, the United Kingdom, Germany and France[94]. Up-to-date quantitative information on this issue may be found on the datacenter map webpage[95].

92 https://mosla.mathematik.uni-marburg.de/gb/

93 http://recherche.unistra.fr/index.php?id=30740

94 https://www.journaldunet.com/solutions/cloud-computing/1141294-data-center-la-france-quatrieme-pays-le-mieux-equipe-au-monde/

95 https://www.datacentermap.com

Chapter VI

Perspectives

Techno-scientific perspectives

Proof of concept for DNA data archiving *in vitro* (*i.e.*, not in living cells) has been established. Several studies have shown that such archiving can support selective and scalable access to data, as well as error-free storage and retrieval of information. However, technical challenges remain to make this process economically viable for a broad spectrum of data types. These relate to improving the cost, speed and efficiency of technologies for reading, and especially writing and editing, DNA or other polymers.

In the case of writing, several players in the field are pinning their hopes on enzymatic DNA synthesis, whose development potential seems to be greater than that of the traditional chemical synthesis. When changes to stored DNA are needed, two approaches are, *a priori*, possible: either to rewrite all information or to edit it. The choice between these two approaches must be based on a cost-benefit assessment, which depends on the extent and number of the changes as well as the rapidly evolving state of the art.

In the case of reading, the nanopore approach has a strong potential because it can read long sequences without the need for fragmentation, is intrinsically parallelizable, and can be readily adapted to the growing chemical diversity of polymers with "digital" applications.

It should also be noted that, although the speeds of writing and reading DNA are, as yet, limited, this disadvantage is offset in some applications by the possibility of massive parallelization. Concretely, by 2024, a single machine could write and read 1 TB per day.

However, as far as cost and speed are concerned, several orders of magnitude are currently missing for the full adoption of the DNA solution for big data archiving: reading should improve a thousand times and writing a 100 million times. These factors may seem staggering. This would be forgetting the speed of progress of DNA-related technologies. For example, George M. Church, in his March 2019 presentation, estimated that the cost of reading and writing DNA had dropped by a factor of more than a million in 10 years or, put differently, their performances have doubled every six months. This can be compared with advances in electronics and computing: Moore's "law", mentioned above, states that semiconductor densities double every two years, as they did between 1971 and 2016; this is correlated with costs and, for example, the Seagate company reported that it had reduced the cost of unit data storage on disk by a factor of 1.3 million in 29 years. By these criteria, DNA technologies are evolving much faster

than information technologies.

The performance of the overall archiving process could also be improved by optimizing the information encoding and decoding steps. This optimization could involve data compression and/or extended or combinatorial alphabets.

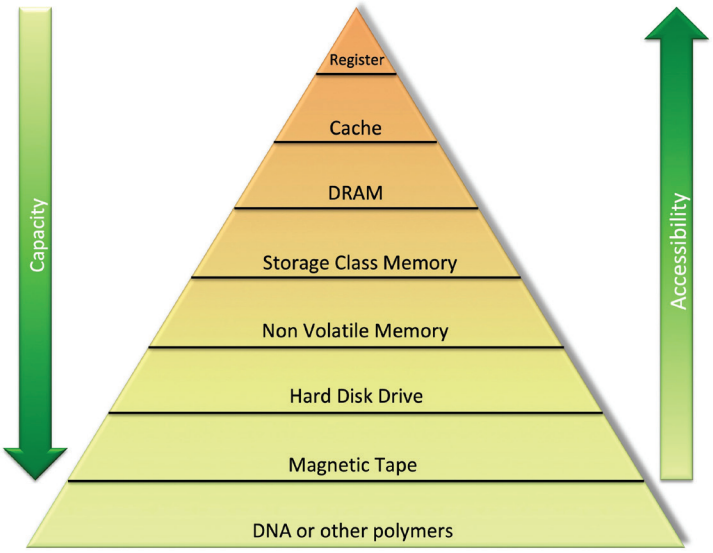
Finally, several lines of research have started from the observation that DNA is not necessarily the most efficient "digital" polymer outside the cell: either its alphabet is too limited (to four letters) or its physical chemistry is not optimal. This observation has given rise to alternative approaches that, to different degrees, move away from DNA, to other heteropolymers or linear copolymers with theoretical advantages. When their performance rivals that of DNA in terms of reading, writing and editing performance, which could take a decade, these very promising polymers will probably burst onto the market for archiving digital information.

Economic outlook

The idea of archiving digital data in DNA dates back several decades. Proof of principle has been obtained but DNA storage of information is still largely at the R&D stage.

A certain consensus has emerged among some players in the field that the economic viability of molecular information storage could be achieved within 5 to 10 years for niche markets. One example is the long-term archiving of sensitive information: two key assets would be the ease of amplifying DNA so as to distribute copies of the information geographically, and the rapidity with which it can be voluntarily destroyed.

Competing successfully with electronic storage technologies in the more global markets for big data archiving may take 10 to 20 years. Since the main handicap of the DNA-based approach is the slow writing and editing process, it is reasonable to assume that its use during this period will remain confined to long-term archiving, given its obvious advantages in terms of information density, longevity, and durability. The need for massive long-term archiving is recognized in fields as diverse as particle physics and film preservation. In these cases, storage in DNA would compete with or complement magnetic tape, currently the solution of choice for long-term archiving. As magnetic tape consumes less than 1 % of the electricity in a datacenter, it appears however that reduction in energy consumption does not constitute the best asset of molecular archiving.



Pyramid of memory types in computer systems. At the bottom of the pyramid has been added hypothetically the use of DNA or another heteropolymer.

Credit: François Képès and Carlo Reita.

Annexes

Works and contributions

This document has been prepared by the National Academy of Technologies of France working group "*DNA: reading, writing, storing information*" chaired by François Képès, with the support of Morgane Champleboux.

French-to-English translation by
Wolf Gehrisch, Victor Norris and François Képès.

Speakers

12/04/2018

Olivier Lucas, Associate Director (Europe South) - Oxford Nanopore Technologies Ltd., UK

01/08/2019

Philippe Glaser, Director of Pasteur Genopole® Île-de-France - Institut Pasteur, France
Piet Herdewijn, Professor - REGA Institute, Catholic University of Leuven, Belgium

03/12/2019

Dominik Heider, Professor - Philipps University of Marburg, Germany
Mostafa Ronaghi, CTO and Senior Vice President - Illumina Inc., USA
Nick Goldman, Head of Research and Senior Scientist - EMBL-EBI, University of Cambridge, UK

03/14/2019

George M. **Church**, Professor - Harvard Medical School and MIT, USA

[Seminar held at École Polytechnique]

03/29/2019

Emily Leproust, CEO - Twist Bioscience, USA

04/09/2019

Alexander Murer, CEO - Kilobaser, Austria
John Hoffman, Technical SETA, and **David A. Markowitz**, Manager - MIST Program, IARPA, USA *[via videoconference].*
Tim Brears, CEO, and **Matt Hayes,** CTO - Evonetix Ltd., United Kingdom

05/14/2019

Brian Jester, CEO - Synovance, France
Torsten Waldminghaus, Professor - *at the time of the interview:* SYNMIKRO, LOEWE Center for Synthetic Microbiology, University of Marburg, Germany

06/11/2019

Zohar Yakhini, Professor - Technion & InterDisciplinary Center (IDC) Herzliya, Israel [*via video-conference]*

07/07/2019

Carlo Reita, Director of Strategic Partnerships and Planning - CEA-Leti, France

09/10/2019

Sachin Chalapati, CTO - Helixworks, Ireland

10/08/2019

Nick Gold, Vice President of Marketing - Catalog DNA, USA *[via videoconference]*
Steven Benner - Ffame and University of Florida, USA

10/11/2019

Karin Strauss, Principal Research Manager - Microsoft Corp., USA

Luis Ceze, Professor - University of Washington, USA

11/12/2019

Tomas Ybert, CEO - DNA Script, France
Sophie Tuffet, Chairman of the Management Board - Imagene, France

12/03/2019

Jiahao Huang, CCO - Nuclera Nucleics, UK
Robert Grass, Professor - ETH Zürich, Switzerland
Jean-François Lutz, Research Director - Charles Sadron Institute, University of Strasbourg, France.

Members of the French workgroup "DNA: reading, writing, storing information"

Members of the National Academy of Technologies of France

René Amalberti
Pierre-Étienne Bost
Alain Boudet
Pierre Bourlioux
Leonardo Chiariglione
Patrice Courvalin
Bernard Daugeras
Pierre Feillet
Gérard Grunblatt
Bruno Jarry
François Képès *(Chair)*
Bernard Le Buanec
Patrick Ledermann
Denis Lucquin
Jean Lunel
Thierry Magnin
Pierre Monsan
Gérard Roucairol
Christian Saguez
Erich Spitz
Pierre Tambourin

Members not belonging to the NATF

Morgane Champleboux - University of Évry *(Scientific Secretary)*
Wolf Gehrisch - International Relations, Académie des technologies
Hannu Myllykallio - École Polytechnique
Victor Norris - University of Rouen
Carlo Reita - CEA-Leti, Grenoble

Abbreviations used

3D	tridimensional
A	Adenine
B	byte
bit	binary digit
C	Cytidine
CCO	Chief Commercial Officer
CD	Compact Disk
CEA	French Alternative Energies and Atomic Energy Commission
CEO	Chief Executive Officer
CERN	European Organisation for Nuclear Research
CNRS	French National Center for Scientific Research
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTO	Chief Technical Officer
dATP	deoxyAdenosine TriPhosphate
dCTP	deoxyCytosine TriPhosphate
ddATP	di-deoxyAdenosine TriPhosphate
ddCTP	di-deoxyCytosine TriPhosphate
ddGTP	di-deoxyGuanosine TriPhosphate
ddTTP	di-deoxyThymidine TriPhosphate
dGTP	deoxyGuanosine TriPhosphate
DNA	DeoxyriboNucleic Acid
DRAM	Dynamic Random Access Memory
dTTP	deoxyThymidine TriPhosphate
DVD	Digital Versatile Disk
EBI	European Bioinformatics Institute
EIC	European Innovation Council
ETH	Swiss Federal Institute of Technology
FfAME	the Foundation for Applied Molecular Evolution
g	gram

G	Guanine
GDS	Global DataSphere
GPU	Graphic Processor Unit
HDD	Hard Disk Drive
IDC	InterDisciplinary Center
IDC	International Data Corporation
IS	International System
JPEG	Joint Photographic Expert Group
km	kilometer (10^3 meters)
m	meter
MEMS	MicroElectroMechanical System
mL	milliliter (10^{-3} liter)
mm	millimeter (10^{-3} meter)
MOSLA	Molecular Storage for Long-term Archiving
MoSS	Molecular Storage System
mp3	MPEG Audio Layer 3
MPEG	Moving Picture Experts Group
MRAM	Magnetic Random Access Memory
NAND	NOT-AND (logical gate)
NHGRI	National Human Genome Research Institute
NIH	National Institutes of Health
nm	nanometer (10^{-9} meter)
PCM	Phase-Change Memory
PCR	Polymerase Chain Reaction
PDF	Portable Document Format
R&D	Research and Development
RNA	RiboNucleic Acid
SRAM	Static Random Access Memory
T	Thymine
TdT	Terminal Deoxynucleotidyl Transferase

USB	Universal Serial Bus
US$	United States dollars
W	watts
XNA	Xeno-Nucleic Acid

International System (IS) prefixes of units, and corresponding numbers

prefix	abbreviation	engineering format	direct format
K	kilo	10^3	1 000
M	mega	10^6	1 000 000
G	giga	10^9	1 000 000 000
T	tera	10^{12}	1 000 000 000 000
P	peta	10^{15}	1 000 000 000 000 000
E	exa	10^{18}	1 000 000 000 000 000 000
Z	zetta	10^{21}	1 000 000 000 000 000 000 000
Y	yotta	10^{24}	1 000 000 000 000 000 000 000 000

Datacenters, including the “cloud”, store humanity’s digital big data on hard disks and magnetic tapes whose limited lifespan requires expensive copies to be made every five to seven years; they devour resources such as land, electricity, water and scarce materials. In comparison, storage at a molecular scale in a polymer such as DNA could have a density ten million times higher, last ten thousand times longer without the need for periodic copying, and consume very little energy. Indeed, DNA is stable at ordinary temperatures for several millennia and can be easily duplicated or deliberately destroyed. The required technologies already exist. However, in order to become viable for information archiving, they must be developed further. This could be achieved within five to twenty years, and would be facilitated by synergies between the public and private sectors.

Académie des technologies
Le Ponant
19. rue Leblanc
Bâtiment A
75015 Paris
+33(0)1 53 85 44 44
secretariat@academie-technologies.fr
www.academie-technologies.fr

ISBN : 979-10-97579-19-7