Stocker les mégadonnées dans un monde fini

François KÉPÈS (Académie des Technologies)
Morgane CHAMPLEBOUX (Université d'Évry)
Carlo REITA (CEA Leti)

& Groupe de Travail "ADN : lire, écrire, stocker l'information" (2018-2020)
Académie des Technologies

Synopsis :

Les "centres de données", “cloud” inclus, stockent les mégadonnées numériques de l'humanité sur des disques durs et bandes magnétiques dont la durée de vie est très limitée ; ils représentent des "gouffres" pour les ressources en terrain, électricité et eau. En comparaison, le stockage à l’échelle infra-moléculaire, en particulier sur l’ADN, permettrait une densité supérieure d’un facteur 10 millions, une conservation prolongée d’un facteur 10 mille, pour une consommation électrique quasi-nulle. En effet, l'ADN est stable à température ordinaire durant plusieurs millénaires ; il peut être aisément dupliqué ou volontairement détruit. Les technologies requises existent ; cependant, pour devenir viable pour le stockage de l'information, ces technologies nécessitent encore des progrès qui pourraient voir le jour sous 10 ans.

Anne-Laure Boch : *"Les prétentions infinies [des technosciences] pourraient se fracasser sur les limites d'un monde fini."* [1]

1. CONTEXTE ET MOTIVATION

L'humanité accumule des données à un rythme jamais vu et qui va croissant. Les données considérées ici sont celles de vos connexions familiales, amicales et professionnelles, vos livres, vidéos et photos, vos données médicales, celles de la recherche scientifique, de l'industrie etc. On parle parfois de "big data" ou "mégadonnées". Et bien plus est à venir : par exemple voitures autonomes, capteurs et autres objets connectés, télésurveillance, réalité virtuelle, déserts médicaux compensés par du diagnostic et même de la chirurgie à distance. En 2025 il est estimé que les 3/4 d'entre nous serons connectés, et que nous interagirons chacun avec des données toutes les 18 secondes en moyenne [2]. Presque toutes ces données passent par des traitements informatiques, ce qui impose qu'elles soient représentées comme de longues suites de deux éléments, typiquement notés '0' et '1' : on parle de données numériques. Ces longues suites sont souvent subdivisées en groupes successifs de 8 éléments '0' ou '1' qui sont appelés "octets".

1.1. La Sphère Globale des Données (SGD)

L'ensemble des données numériques créées par l'humanité, la "Sphère Globale des Données" ou SGD, contient environ autant de caractères (d'octets) [3] que le nombre d'étoiles dans l'univers observable, ou que le nombre de grains de sable sur la terre. Cette SGD était estimée en 2018 à 33 Zettaoctets (Zo ; soit 33 x 10^{21} caractères [4]). Elle double tous les 2-3 ans, et atteindra environ 175 Zo en 2025 [5]. Au rythme actuel, la SGD atteindrait plus de 5.000 Zo (5 x 10^{24} caractères) en 2040. Une façon d'appréhender un nombre aussi élevé consiste à dire qu'il vous faudrait 50 millions d'années pour télécharger cette SGD avec une connexion internet de vitesse moyenne.

Une part majoritaire des données créées par l'humanité est ensuite stockée, parfois au long terme. Outre le stockage local (sur votre ordinateur ou téléphone), actuellement en décroissance, le stockage centralisé est en rapide augmentation, en particulier dans les centres de données et maintenant le "cloud" ou "nuage". Le nuage permet aux utilisateurs

---

[1] https://sciences-critiques.fr/ou-va-la-science-devoiler-le-monde-naturel-ou-creer-un-nouveau-monde/

[2] HiPEAC Vision 2015 (Commission Européenne, FP7, 2015).

[3] Un "octet" est une suite de **8** "bits". Un bit ne peut prendre que **2** valeurs désignées usuellement par les chiffres '0' et '1' (d'où le terme de codage "binaire" ou "numérique"). Donc il existe 256 ($\mathbf{2^8}$) octets possibles. Nous allons considérer ici qu'un octet représente un caractère (une lettre, un chiffre, ou un symbole) parmi 256. Par exemple l'octet '00100011' code habituellement le caractère '#'.

[4] Préfixes du Système International d'unités, et nombres correspondants :

kilo	10^3	1 000
mega	10^6	1 000 000
giga	10^9	1 000 000 000
tera	10^{12}	1 000 000 000 000
peta	10^{15}	1 000 000 000 000 000
exa	10^{18}	1 000 000 000 000 000 000
zetta	10^{21}	1 000 000 000 000 000 000 000
yotta	10^{24}	1 000 000 000 000 000 000 000 000

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

individuels de profiter de ressources informatiques à la demande sans avoir à recourir aux administrateurs informatiques, et d'augmenter le niveau d'automatisation grâce à la virtualisation des serveurs. En 2021 le nuage stockera autant d'information que les traditionnels centres de données. Dans la suite de ce texte, nous considérerons l'ensemble des centres de données, nuage inclus. Des hangars dédiés à ce stockage centralisé ne cessent de pousser dans le monde, souvent dans les pays froids car ce stockage est grand consommateur d'électricité et demande un fort refroidissement.

1.2. Centres de données

Pour concrétiser cela, prenons le cas d'un petit centre de données de 300 m^2 construit en 2008, et comprenant seulement 2.000 serveurs informatiques pour une puissance totale de 1 megawatt (MW). Durant sa durée de vie d'environ 20 ans [6], il aura englouti 66 tonnes de cuivre, 15 tonnes de plastique, 33 tonnes d'aluminium, 152 tonnes d'acier. Chaque année, il aura consommé 23 millions de litres d'eau et, en incluant le refroidissement des serveurs, 18 millions de kilowatt-heures (0,018 terawatt-heures ou 0,018 TWh) d'électricité [7]. En revanche, un gros centre de données représente plusieurs milliards d'euros en investissement, un million de m^2, un million de serveurs. Il consomme un gigawatt (GW) d'électricité (dont 40-50 % environ pour le refroidissement), soit environ 10 TWh par an, soit encore plus qu'une ville française de 100.000 habitants. Bien entendu, ces centres sont reliés au reste du monde par d'importants réseaux de connexions, également consommateurs de diverses ressources dont l'électricité. Au total, en incluant ceux des entreprises, il y aurait 8,6 millions de centres de données au monde en 2017, avec une surface totale de plus de 170 millions de m^2 – l'équivalent de 25.000 terrains de football [5]. Cette surface représente environ un millionième des terres émergées de la planète puisque ces dernières couvrent approximativement 150 millions de km^2. Si le rythme actuel d'un doublement tous les 2 ans se poursuivait, un millième des terres émergées serait occupé par ces centres avant 2040 ; cependant, ceci est probablement une surestimation, car l'efficacité des centres de données continue d'augmenter aux plans de l'énergie et de la superficie.

Les centres de données et leurs réseaux de connexions équivaudraient au 5^{eme} pays le plus consommateur d'électricité au monde, entre Inde et Japon. Il était estimé qu'en 2007, les centres de données et leurs réseaux de connexions associés avaient au niveau mondial consommé 623 terawatt-heures (623 TWh), et engendré l'émission de 423 mégatonnes d'équivalent-CO_2 ; qu'en 2012, ils étaient responsables de 2% des gaz à effet de serre produits globalement par les technologies de l'information ; et qu'en 2012 l'investissement mondial pour construire de nouveaux centres de données atteignait 450 milliards de dollars US [8]. Par exemple Google seul sur la période 2015-2017 aurait ainsi investi 10 milliards de dollars US par an.

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

[7] Guideinformatique (2008). https://www.guideinformatique.com/dossiers-actualites-informatiques/consommation-electrique-des-data-centers-29.html

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

Centre de données stockant 1 Eo (10^{18} octets, soit 1/33.000^{eme} de la SGD actuelle) d'information numérique.

1.3. Conclusion provisoire

La lecture de ces quelques éléments d'information montre clairement que la croissance des données numériques, au rythme actuel et à technologie constante, n'est pas soutenable au-delà d'environ 2040.

Existe-t-il une technologie qui permettrait d'archiver l'ensemble des mégadonnées, la SGD, dans une fourgonnette, quasiment sans dépense énergétique ?

Oui : c'est là le sujet de cet article, qui synthétise les travaux du groupe de travail trans-disciplinaire (2018-2020) "ADN : lire, écrire, stocker l'information" [9], qui en a auditionné 26 spécialistes mondiaux.

Mais avant de se lancer, il est important de brosser un succinct tableau de l'état de l'art dans le domaine du stockage et de l'archivage des données numériques, en faisant ressortir ses limites et perspectives.

2. ÉTAT DE L'ART DU STOCKAGE ET ARCHIVAGE DE DONNÉES NUMÉRIQUES

2.1. Problématique

La problématique du stockage et archivage d'informations numériques se définit par plusieurs paramètres :

- la quantité d'information et son codage ;
- la durée de stockage, c'est à dire combien de temps l'information sera conservée ;
- la fréquence d'accès à l'information ; on distinguera "stockage" ou "sauvegarde" à court terme, et "archivage" à long terme ;
- le coût de production, de conservation et de gestion de l'information.

2.2. Hiérarchie des mémoires dans les systèmes informatiques

En informatique, la mémoire est un dispositif électronique qui sert à stocker l'information. Elle est organisée de manière hiérarchique. Cette hiérarchie est représentée sous une forme pyramidale, composée de plusieurs niveaux selon les besoins d'accès aux données et la capacité de stockage. Plus la capacité de stockage augmente, plus le temps d'accès aux données est long.

[9] Animateur François Képès ; Secrétaire scientifique Morgane Champleboux.

Au sommet de la pyramide, il y a le registre qui est une mémoire interne au processeur. Il s'agit de la mémoire la plus rapide d'un ordinateur (0,1 nanoseconde pour l'accès aux données), mais dont le coût de fabrication est le plus élevé et donc réservé à une très faible quantité de données de l'ordre de quelques Kilooctets (Ko).

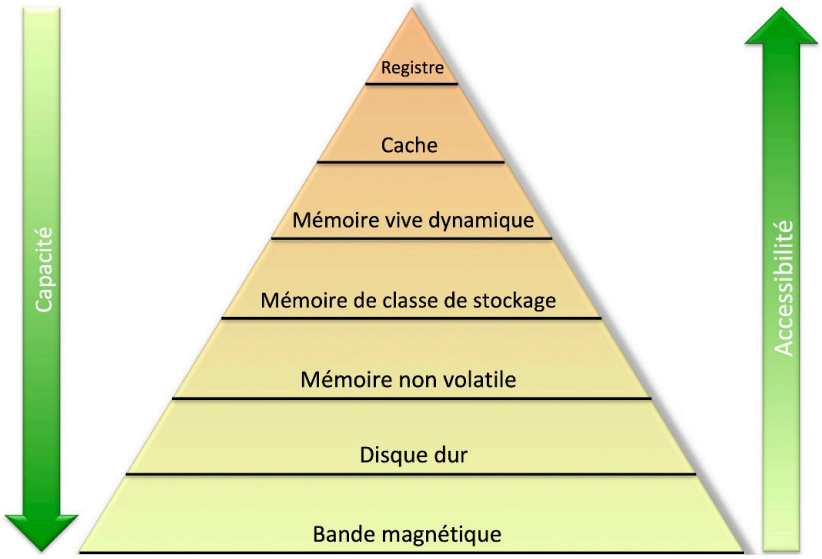
En dessous du registre, il y a la mémoire cache. Cette mémoire conserve un court instant des informations fréquemment consultées. Ces mémoires sont très rapides (1 à 10 nanosecondes pour l'accès aux données), mais également très coûteuses et réservées à une petite quantité de données (quelques Ko à Mo).

En dessous de la mémoire cache, il y a la mémoire vive, dans laquelle sont stockées puis effacées les informations traitées par l'appareil informatique. Il s'agit de l'espace principal de stockage du microprocesseur, mais dont le contenu disparait lors de la mise hors tension de l'ordinateur. C'est une mémoire relativement rapide (10 à 1.000 nanosecondes pour l'accès aux données) et réservée à quelques Gigaoctets (Go) de données.

Enfin, il y a la mémoire de masse qui comprend :

- les disques durs et les mémoires de type Flash, qui stockent à long terme une grande quantité d'information de l'ordre de plusieurs Teraoctets (To) de données ;
- les bandes magnétiques, utilisées pour l'archivage des informations à très long terme (plus de 10 ans).

Le coût de la mémoire de masse est relativement faible, mais sa vitesse d'accès est inférieure aux autres types de mémoires.



Pyramide des types de mémoires dans les systèmes informatiques.

Pour satisfaire les besoins du stockage d'information digitale, une large gamme de technologies est employée.

2.3. Technologies de stockage d'information en électronique

Après une période où le stockage sur supports optiques, tels que CD ou DVD, avait trouvé une place dans ce marché très disputé, aujourd'hui, la quasi-totalité du stockage dans le domaine des technologies de l'information se base sur le stockage magnétique et le stockage par charge.

2.3.1. Technologie de stockage magnétique

2.3.1.1. Supports

Ces supports comprennent la bande magnétique, le disque dur et les mémoires MRAM[10]. Ils offrent une grande capacité de stockage et une grande durée de conservation de l'information. Ainsi, ils représentent des technologies de choix pour l'archivage de données. La méthode sur bande magnétique présente la densité de stockage la plus élevée, mais l'accès

[10] MRAM : mémoire vive magnéto-résistive.

intrinsèquement séquentiel aux données limite les vitesses de lecture et d'écriture. Les disques permettent un accès aléatoire beaucoup plus rapide, au prix d'une densité moindre.

2.3.1.2. Principe

L'écriture sur ces supports se fait par aimantation et la lecture est magnétique. Le courant dans une bobine génère un champ magnétique qui induit des dipôles dans un substrat (disque ou bande) recouvert d'un matériau magnétique. La lecture s'effectue en mesurant le courant induit par le mouvement des dipôles. Au fil du temps, cette technologie s'est améliorée grâce à la qualité des matériaux magnétiques et à la miniaturisation des têtes de lecture et d'écriture.

2.3.2. Technologie de stockage par charge

2.3.2.1. Supports

Cette technologie comprend les mémoires SRAM[11], DRAM[12] et Flash. Les dispositifs de stockage par charge s'intègrent facilement dans des circuits électroniques complexes. Les performances recherchées sont :

- la vitesse de lecture, d'écriture et d'accès à l'information ;
- la capacité de stockage ;
- le nombre de cycles de lecture et d'écriture de données.

2.3.2.2. Principe

L'information stockée sur ces supports est représentée par l'état de charge d'une capacité (chargé/pas chargé). Cette charge peut être lue directement, ou affecter la conduction d'un transistor. Au fil du temps, cette technologie s'est miniaturisée grâce aux méthodes de la microélectronique.

2.3.2.3. Le cas des mémoires Flash

Il existe deux classes de stockage par charge : les mémoires volatiles[13] (DRAM et SRAM) et mémoires non-volatiles (Flash). La mémoire Flash est la mémoire non-volatile la plus utilisée aujourd'hui (disques SSD, cartes mémoires pour appareils portables, clefs USB).

La durée de vie d'une mémoire Flash est calculée par le nombre de cycles de lecture et d'écriture de données que peut subir le support. Sa structure particulièrement régulière a permis la réduction de taille progressive, afin d'obtenir des densités de stockage comparables à celles des disques durs. De plus, les procédés microélectroniques de fabrication en série en ont réduit drastiquement les coûts.

Actuellement, les systèmes de mémoires Flash NAND 3D se développent. Au lieu d'être disposées sur des surfaces planes, les cellules de stockages sont disposées sur des surfaces repliées, ce qui permet de stocker beaucoup plus de cellules par unité de volume (jusqu'à 72 niveaux d'empilement) et donc d'augmenter la densité de stockage d'information.

Aujourd'hui, il n'est plus possible de réduire davantage la taille des dispositifs de mémoire Flash. Ainsi, les chercheurs travaillent sur de nouvelles mémoires basées sur le changement de résistance.

2.4. Comparaison des évolutions des différentes technologies de stockage

Les différentes technologies de stockage et archivage, et leur évolution, peuvent être comparées grâce au tableau ci-dessous.

Deux conclusions se dégagent :

[11] SRAM : mémoire vive statique.
[12] DRAM : mémoire vive dynamique.
[13] Perte de l'information en absence d'alimentation électrique.

• La bande magnétique reste le meilleur compromis pour l'archivage des données à long terme. En effet, la durée de stockage/archivage est élevée et le coût de production reste le moins cher, toutes technologies confondues.

• Les disques durs magnétiques et à état solide (Flash) sont les meilleurs compromis pour le stockage de masse, car ils possèdent la plus haute densité de stockage.

	Type de stockage	***Stockage magnétique***		***Stockage par charge***			
	Supports	*Bandes magnétiques*	*Disques durs magnétiques*	*mémoires volatiles*		*mémoires non-volatiles*	
				SRAM	*DRAM*	*PCM [14]*	*Flash (NAND)*
En 2008	*Densité de stockage (Gbits/cm2)*	0,14	59	N/A	N/A	N/A	31
	Temps de lecture (ns)	N/A	N/A	N/A	N/A	N/A	N/A
	Temps d'écriture (ns)	N/A	N/A	N/A	N/A	N/A	N/A
	Durée du stockage	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
	Cout de production ($/Go)	0,091	0,272	N/A	N/A	N/A	3,33
	Revenus générés ($)	1	34	N/A	N/A	N/A	10,1
En 2016	*Densité de stockage (Gbits/cm2)*	3,89	**170**	N/A	N/A	N/A	**310**
	Temps de lecture (ns)	N/A	5-8x10^6	<10-50	10-50	20-70	25.000
	Temps d'écriture (ns)	N/A	5-8x10^6	<10-50	10-50	50-500	200.000
	Durée du stockage	**> 10 ans**	10 ans	<seconde	<seconde	<10 ans	10 ans
	Endurance (cycles)	N/A	1015	>10^{17}	10^{17}	10^7-10^8	10^3-10^6
	Cout de production ($/Go)	**0,016**	0,039	10^2-10^3	10	1	0,32
	Revenus générés (milliards dollars US)	0,65	26,8	N/A	N/A	N/A	38,7

2.5. Conclusion

Actuellement, les technologies de stockage et archivage de données numériques citées jusqu'ici, dites traditionnelles, sont toutes proches de leur optimum théorique. En d'autres termes, les gains à venir seront faibles en termes de densité, vitesse, longévité, durabilité et coûts. Notons aussi que la production de silicium est largement inférieure aux besoins futurs.

En outre, ces supports traditionnels sont rapidement frappés d'obsolescence, sur trois plans (voir à ce propos le rapport 2010 conjoint de l'Académie des Sciences et de l'Académie des Technologies [15]).

a) Le format de stockage : que l'on songe par exemple aux disquettes 3,5 pouces dont l'usage s'est progressivement éteint entre 2000 et 2010 ; et actuellement, les bandes magnétiques continuent comme par le passé à subir des mutations justifiant leur remplacement par de nouvelles générations incompatibles avec les précédentes.

b) Le dispositif de lecture / écriture : pour reprendre le même exemple, peu de personnes possèdent encore un lecteur fonctionnel de disquettes 3,5 pouces ; et actuellement les lecteurs de bande magnétique poursuivent leur évolution.

[14] PCM : forme de mémoire vive non-volatile.

[15] Hourcade J-C, Laloë F, Spitz E (2010). Longévité de l'information numérique. Académie des Technologies & Académie des Sciences (EDP Sciences).

c) Le support lui-même : du fait de leur constitution physique, tous les supports de stockage ont une durée de vie limitée, entraînant le risque de perdre de l'information. Pour s'en affranchir, il faut constamment les vérifier et recopier les données pour les sauvegarder sur des supports fiables. Par exemple, du fait de la dégradation des signaux magnétiques au cours du temps, la pratique veut que les bandes et disques soient recopiés tous les 5 à 10 ans.

Enfin, il convient de rappeler que ces systèmes traditionnels sont gros consommateurs d'énergie, d'une part à tout instant, d'autre part via le jeu de l'obsolescence.

Tous ces inconvénients ne sont pas partagés par la technologie de stockage de l'information sur l'ADN, que nous allons maintenant analyser.

3. STOCKAGE DE L'INFORMATION SUR L'ADN

Poursuivre selon la loi de Moore [16] est un défi de plus en plus difficile à relever. De nouvelles méthodes de stockage infra-moléculaire d'informations sont envisagées afin d'augmenter la capacité de stockage, réduire la taille des supports et augmenter la durée de conservation des données. Les progrès récents des technologies de lecture et d'écriture de l'ADN ont amené les chercheurs à envisager l'ADN lui-même comme support d'archives numériques. L'ADN permet de stocker 1 bit pour environ 50 atomes.

3.1. Bref historique du stockage de données sur l'ADN

Richard Feynman a été le premier à envisager l'ADN comme support de stockage de l'information numérique en 1959. Mais c'est en 1977 qu'a été mise au point la première méthode de lecture de l'ADN, et en 1983 une technique d'écriture de l'ADN.

En 1988 et pour la première fois, Joe Davis [17] a conçu et synthétisé un fragment d'ADN de 18 nucléotides (soit 4,5 octets) contenant un message numérisé, qu'il a ensuite transféré chez une bactérie intestinale, le colibacille.

En 2012, l'équipe de George M. Church (Université de Harvard, États-Unis d'Amérique) a stocké 0,6 Mo d'information sur l'ADN, sous forme de fragments synthétiques. En 2013, l'équipe de Nick Goldman (Institut Européen de Bioinformatique, Royaume-Uni) a converti 4 fichiers de différents formats en séquence d'ADN, pour un total de 0,7 Mo. L'information a été retranscrite sans erreurs.

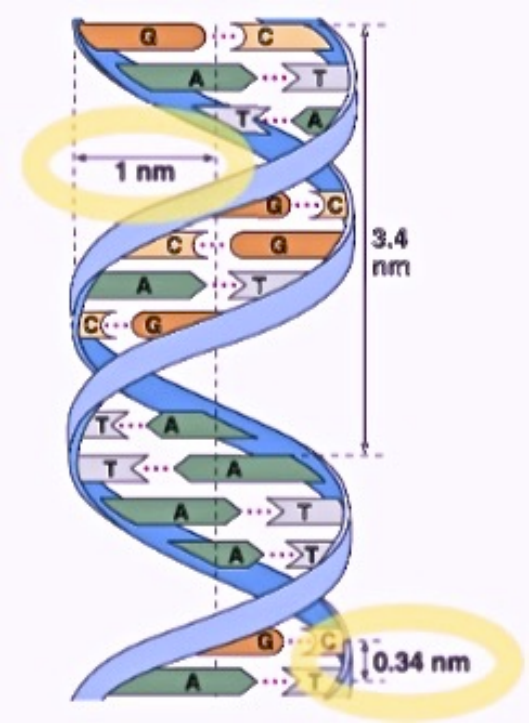
En 2018, Microsoft Corp. et l'Université de Washington, aux États-Unis d'Amérique, ont stocké 1 Go d'information sur l'ADN. Ils détiennent depuis le record.

3.2. L'ADN : un support de stockage performant

L'ADN est dans nos cellules le support de l'information héréditaire. Pour rappel, l'acide désoxyribonucléique (ADN) est formé de deux brins antiparallèles enroulés l'un autour de l'autre pour former une structure en double hélice. Chaque brin d'ADN est un hétéropolymère composé d'une suite, une séquence, de nucléotides. Il existe 4 nucléotides différents : A, C, G, T. Chaque nucléotide est composé d'une des quatre bases azotées, adénine (A), guanine (G),

[16] La *loi de Moore* a été exprimée en 1965 par Gordon E. Moore, un des trois fondateurs d'Intel. Cette loi empirique énonce que la densité de transistors sur une puce de silicium double tous les deux ans. Une autre loi empirique semble régir l'évolution des capacités de stockage. Néanmoins, là aussi le doublement intervient environ tous les 2 ans jusqu'à présent.

[17] https://en.wikipedia.org/wiki/Joe_Davis_(artist)



thymine (T), cytosine (C), liée à un sucre désoxyribose, lui-même lié à un groupe chimique phosphate. Les bases nucléiques d'un brin d'ADN peuvent interagir avec les bases nucléiques d'un autre brin d'ADN à travers des liaisons hydrogène en respectant des règles d'appariement. Ainsi, A et T s'apparient avec deux liaisons hydrogène, tandis que G et C s'apparient avec trois liaisons hydrogène.

Structure schématique de la double hélice antiparallèle d'ADN.

Cependant, l'ADN peut depuis 1869 (Friedrich Miescher) être manipulé en dehors des cellules, dans l'éprouvette ; c'est principalement sous cette forme, également appelée *in vitro*, qu'il a été envisagé de l'utiliser pour stocker des données numériques. Il présente à ce titre de nombreux avantages, comparé aux systèmes traditionnels.

a) La densité informationnelle de l'ADN est environ 10 millions de fois supérieure à celle des meilleurs systèmes traditionnels. L'ADN peut en principe stocker un demi Zo d'information par gramme. Ainsi, les chercheurs estiment que la SGD de l'humanité entière tiendrait actuellement dans moins de 100 grammes d'ADN. Cependant, en pratique ce n'est pas une seule molécule d'ADN qui est synthétisée pour capter un fichier, mais de nombreux exemplaires identiques. En outre, des zones de cet ADN devront porter des signaux de contrôle qualité et d'indexation, en sus des données. Enfin, l'ADN doit être préservé dans des conteneurs macroscopiques. Tenant compte de ces pertes de densité, on peut estimer que la SGD, stockée sur ADN, tiendrait plus réalistement dans une fourgonnette.
b) Le stockage de l'ADN à température ordinaire n'implique aucune consommation de ressources, et les opérations sur l'ADN sont 1.000 fois moins énergivores qu'en informatique.
c) La longévité de l'ADN est environ 10 mille fois supérieure à celle des supports traditionnels. Des molécules d'ADN vieilles de plus de 560.000 ans ont été analysées à partir d'échantillons historiques [18]. En laboratoire, une demi-vie de 52.000 ans a été démontrée en accélérant artificiellement son vieillissement [19].
d) L'obsolescence du support ADN ne se produira pas tant qu'il y aura une technologie, et de la vie dont les mécanismes héréditaires reposent largement sur cette macromolécule.

[18] 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.

[19] 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/

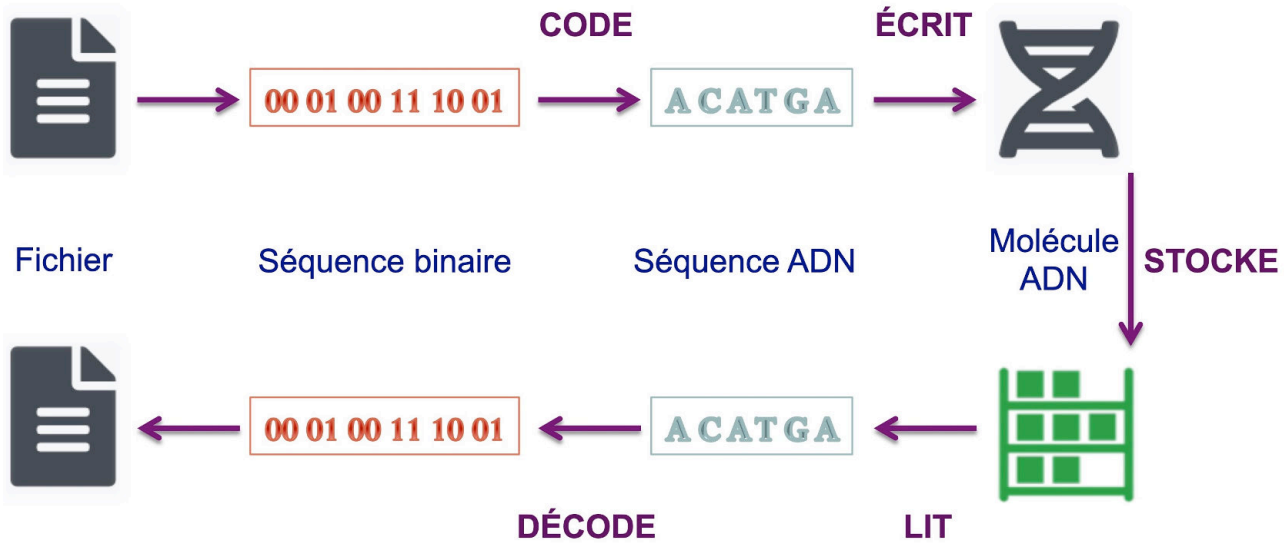
e) La copie ou multiplication de l'ADN, et donc de l'information qu'il contient, est rapide et à faible coût. En effet, l'ADN est naturellement répliqué dans les cellules avant leur division. Ce phénomène de réplication est reproduit *in vitro* par la "réaction en chaîne de la polymérase" (PCR). Ainsi, un seul fragment d'ADN peut être dupliqué en chaîne par des thermocycleurs de paillasse, engendrant par ce processus exponentiel plusieurs milliards de copies en quelques heures [20]. Ceci représente un avantage considérable par rapport à la lourdeur et au coût de duplications de données sur les supports traditionnels.

f) La destruction à volonté de l'ADN est réalisable en un tournemain. En effet, quoique cette macromolécule soit peu réactive chimiquement, le monde vivant s'est doté de catalyseurs protéiques (enzymes appelées ADNases) extrêmement efficaces pour détruire l'ADN en ses composants ou nucléotides. Des approches plus brutales mais moins sophistiquées, faisant appel à des conditions extrêmes, par exemple de pH ou de température, permettraient aussi de détruire l'ADN.

g) Certains calculs peuvent être physiquement implémentés directement avec l'ADN. L'idée est de coder une instance d'un problème combinatoire avec des brins d'ADN que l'on fait synthétiser sur mesure, et de manipuler ces brins par les outils classiques de la biologie moléculaire pour simuler les opérations qui isolent la solution du problème, puis de lire cette dernière [21].

3.3. Principe du stockage des données numériques sur l'ADN

Le processus enchaîne plusieurs étapes : codage des données, écriture sur l'ADN, stockage de l'ADN, lecture de l'ADN, et décodage. Peuvent s'y greffer : modification, amplification ou destruction de l'ADN.



Étapes du processus de stockage des mégadonnées numériques sur l'ADN.

3.3.1 Codage des données

3.3.1.1. Principe

Pour rappel, le système binaire est le système de numération utilisant la base 2. On nomme "bit", de l'anglais "binary digit" (chiffre binaire), les chiffres de la numération binaire. Un bit

[20] Chaque cycle duplique l'existant. Donc 30 cycles produisent 2^{30} copies, soit plus d'un milliard.

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

peut prendre deux valeurs, notées par convention '0' et '1'. Les supports traditionnels tels que les disques durs, les clés USB ou les DVD stockent des données numériques en modifiant les propriétés magnétiques, électriques ou optiques d'un matériau afin de stocker ces '0' et '1'. Un octet est une série de 8 bits.

Pour stocker des données dans l'ADN, le concept est le même, mais le processus est différent. Plutôt que de créer des séquences de '0' et '1', comme pour les données numériques, le stockage de données sur l'ADN utilise des séquences de nucléotides. Il existe plusieurs méthodes, mais l'idée générale est d'attribuer des valeurs numériques aux nucléotides d'ADN. Par exemple, le couple de bits '00' pourrait être équivalent au nucléotide 'A', '01' à 'C', '10' à 'T' et '11' à 'G'. Ainsi, un nouveau code est inventé, où les "bits" sont convertis en nucléotides pour former un fragment d'ADN ensuite synthétisé *in vitro*. Cependant, des méthodes de codage plus élaborées commencent à apparaître.

La technologie actuelle de synthèse d'ADN est limitée à des fragments de l'ordre de 200 nucléotides au maximum, donc très courts au regard des fichiers informatiques qui sont volumineux. Cela est dû au fait que plus un brin d'ADN est long, plus il est difficile de le construire chimiquement. On peut donc dresser une analogie entre les "paquets de nucléotides" (les courts brins d'ADN) et les paquets d'octets qui sont expédiés lors d'une transaction internet, par exemple l'envoi d'un courriel; pour filer la métaphore, dans les deux cas, le message complet sera correctement reconstitué à l'arrivée à partir des paquets grâce à leurs signaux d'appartenance, indexation et adressage, et de contrôle qualité.

3.3.1.2. Découpage du fichier informatique en fragments d'octets

Le fichier informatique est découpé en fragments d'une vingtaine d'octets. Chaque fragment possède un identifiant de quelques bits à son extrémité. L'identifiant permet d'ordonner les fragments au sein d'un même fichier numérique.

3.3.1.3. Conversion des fragments d'octets en nucléotides d'ADN

Les fragments d'octet sont convertis en nucléotides d'ADN. Chaque fragment possède 200 nucléotides et contient la charge utile ("payload", soit typiquement 150 nucléotides) et l'étiquette (typiquement 50 nucléotides). L'étiquette permet de regrouper les fragments d'ADN contenant l'information issue des mêmes fichiers numériques. Elle est utilisée pour l'accès sélectif à l'information, selon un principe d'indexation. En outre, les fragments sont chevauchants, avec un recouvrement suffisant pour limiter les erreurs lors de la reconstitution de l'information.

3.3.2. Écriture de l'ADN

Depuis 1983 existe une méthode de synthèse de l'ADN par voie chimique. Son principe basé sur la chimie des phosphoramidites, a très peu évolué depuis. Il implique l'ajout de nucléotides successifs, protégés par un groupe terminal bloquant toute interaction chimique, afin d'étendre la séquence d'un seul nucléotide à la fois. Le groupe terminal est enlevé, puis le nucléotide suivant est ajouté. La synthèse chimique s'effectue dans le sens inverse de la biosynthèse naturelle. Pendant toute la synthèse, l'ADN en extension reste accroché sur une résine. Il est décroché lors de la déprotection finale.

Mais l'application de cet unique principe s'est beaucoup améliorée. Au début de la synthèse d'ADN par voie chimique, le processus d'ajout des nucléotides était manuel. Puis des synthétiseurs automatiques ont vu le jour dans les années 1990. Les synthétiseurs ont évolué afin de posséder jusqu'à 200 colonnes, permettant de synthétiser simultanément 200 fragments d'ADN différents. La société Twist Bioscience (États-Unis) [22], spécialisée dans la synthèse d'ADN, a miniaturisé ce processus. Elle réalise la synthèse d'ADN sur une micro-puce de silicium, synthétisant simultanément 10.000 fragments d'ADN différents. Elle atteindra

[22] https://www.twistbioscience.com/technology

rapidement le million de puits.

Cependant, le taux d'erreur pour chaque nucléotide ajouté est d'environ 0,1%, ce qui limite en pratique la longueur du fragment utilisable à 200 nucléotides. En outre, la chimie des phosphoramidites a un impact négatif sur l'environnement.

En raison de ces limites entre autres, depuis les années 2010 existe une méthode alternative de synthèse de l'ADN, non par voie chimique mais par voie biologique. Cette synthèse fait usage d'une ADN polymérase spéciale présente dans les cellules immunitaires, appelée "terminal désoxynucleotidyl transférase" (TdT). La TdT, contrairement à la plupart des ADN polymérases qui dépendent d'une matrice d'ADN simple-brin, ajoute les nucléotides aléatoirement. Comme pour la synthèse chimique d'ADN, les chercheurs ont ajouté un groupe chimique protecteur pour chaque nucléotide, empêchant la TdT d'en ajouter plus d'un à la fois. Une fois le nucléotide désiré ajouté, sa protection est enlevée et le cycle se répète. L'ADN polymérase possède une vitesse de synthèse élevée avec un taux d'erreur très faible. Une demi-douzaine de sociétés se sont lancées dans cette nouvelle approche, dont DNA Script [23] en France.

Une amélioration prometteuse portée par Evonetix Ltd. [24] (Royaume-Uni) permet de réduire le taux d'erreur de 100-1000 fois, en contrôlant thermiquement chacun des 10.000 micro-sites de réaction. Une autre approche, dont Catalog DNA [25] (États-Unis) et Helixworks [26] (Irlande) sont précurseurs, consiste à assembler combinatoirement des fragments d'ADN pré-synthétisés et donc vérifiés, s'affranchissant ainsi partiellement des limites ci-dessus. Enfin, d'autres chercheurs ont proposé des alphabets étendus.

3.3.3. Stockage de l'ADN

L'ADN synthétisé est ensuite stocké par deux méthodes : physique ou chimique.

Dans le système de stockage chimique développé par Robert Grass [27] (ETH Zürich, Suisse), l'ADN synthétisé est encapsulé dans des nanobilles, ensuite référencées et réparties dans des plaques à micro-puits. L'inconvénient de cette approche est qu'elle dilue l'ADN dans de grands volumes de silices, donc la densité informationnelle en est abaissée. En outre la construction des nanobilles ne permet pas l'élimination totale de l'eau et l'oxygène, d'où une durée de vie abaissée de l'ADN. Pour usage, l'ADN stocké dans les nanobilles doit ensuite être extrait par un réactif chimique capable de dissoudre les silices.

Twist Bioscience [22] utilise des capsules de stockage physique conçues par la société Imagène [19]. L'extérieur de ces capsules est en acier inoxydable et elles ont la taille d'une pile bouton. Chaque capsule peut contenir 0,8 gramme d'ADN (potentiellement 1,4 Eo en tenant compte de la redondance), et est à usage unique. Son ouverture permet de récupérer l'ADN non dilué. La conservation de l'ADN est estimée à 100.000 ans dans ces systèmes de stockage qui protègent l'ADN de l'eau, des sels, de l'oxygène et de la lumière.

Pour ces deux approches, les fichiers d'information sont stockés sur l'ADN de manière structurée. Par exemple, l'ADN contenant une information volumineuse sera archivé indépendamment dans une capsule Imagène ou nanobille. Les fragments d'ADN contenant des fichiers d'information moins volumineux seront regroupés dans une même capsule, où ils seront différenciés via leurs séquences-étiquettes.

---

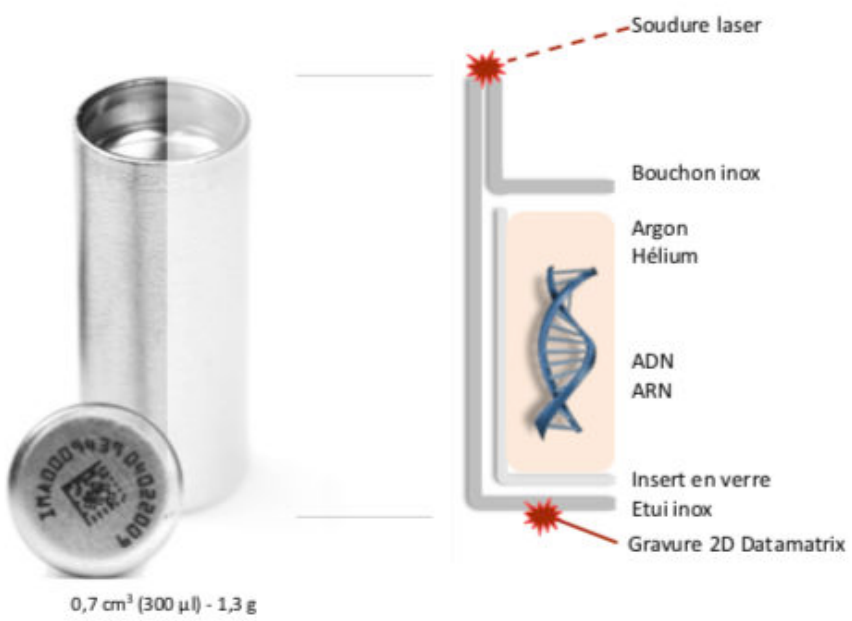
[23] http://www.dnascript.co/#1

[24] https://www.evonetix.com

[25] https://www.catalogdna.com/about

[26] https://helix.works

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



Capsule Imagène de conservation de l'ADN à très long terme.

3.3.4. Lecture de l'ADN

3.3.4.1. Réaction en chaîne de la polymérase (PCR)

L'ADN stocké dans les nanobilles est extrait par un procédé chimique. L'ADN est ensuite multiplié par PCR (défini plus haut) pour être séquencé. La PCR est effectuée à partir d'amorces. Ce sont des fragments d'ADN simple-brin d'une vingtaine de nucléotides qui se fixent par homologie sur une région spécifique d'ADN pour en amplifier une région d'intérêt. La technique de PCR est également utilisée pour accéder sélectivement à l'information. En effet, parmi un ensemble de fragments d'ADN, seuls ceux possédant une certaine étiquette complémentaire à l'amorce choisie, seront amplifiés pour être lus. Du fait du grand nombre de copies, la lecture de l'ADN n'est pas destructrice.

3.3.4.2. Technologies de séquençage

La technologie Illumina [28], leader du marché du séquençage de l'ADN, commercialise des appareils pouvant séquencer jusqu'à un total de 4 milliards de nucléotides par expérience (le génome humain en contient de l'ordre de 3 milliards), avec un taux d'erreur de 1%, et par fragments inférieurs à 1.000 nucléotides. C'est la technologie qu'a utilisé Microsoft Corp. pour leur preuve de concept décrite plus haut portant sur 1 Go.

Les séquenceurs de troisième génération développés par Oxford Nanopore Technologies [29], permettent d'obtenir de longues séquences d'un trait (record à 2,2 millions de nucléotides) afin d'éviter les étapes de fragmentation et d'alignement des séquences, et d'analyser les données en temps réel. Malgré un taux d'erreur sur chaque brin d'ADN supérieur à 10%, le séquençage en parallèle de nombreuses copies de l'ADN permet une correction quasi-parfaite de la séquence par recoupement.

3.3.5. Décodage de l'information

Les séquences des fragments d'ADN, issues du même fichier numérique, sont regroupées informatiquement par leurs étiquettes puis par leurs parties communes. Les séquences d'ADN sont retranscrites en fragments d'octets. Au sein d'un même fichier numérique, les fragments d'octets sont ordonnés, grâce à leurs identifiants, afin de reconstituer la séquence globale en bits.

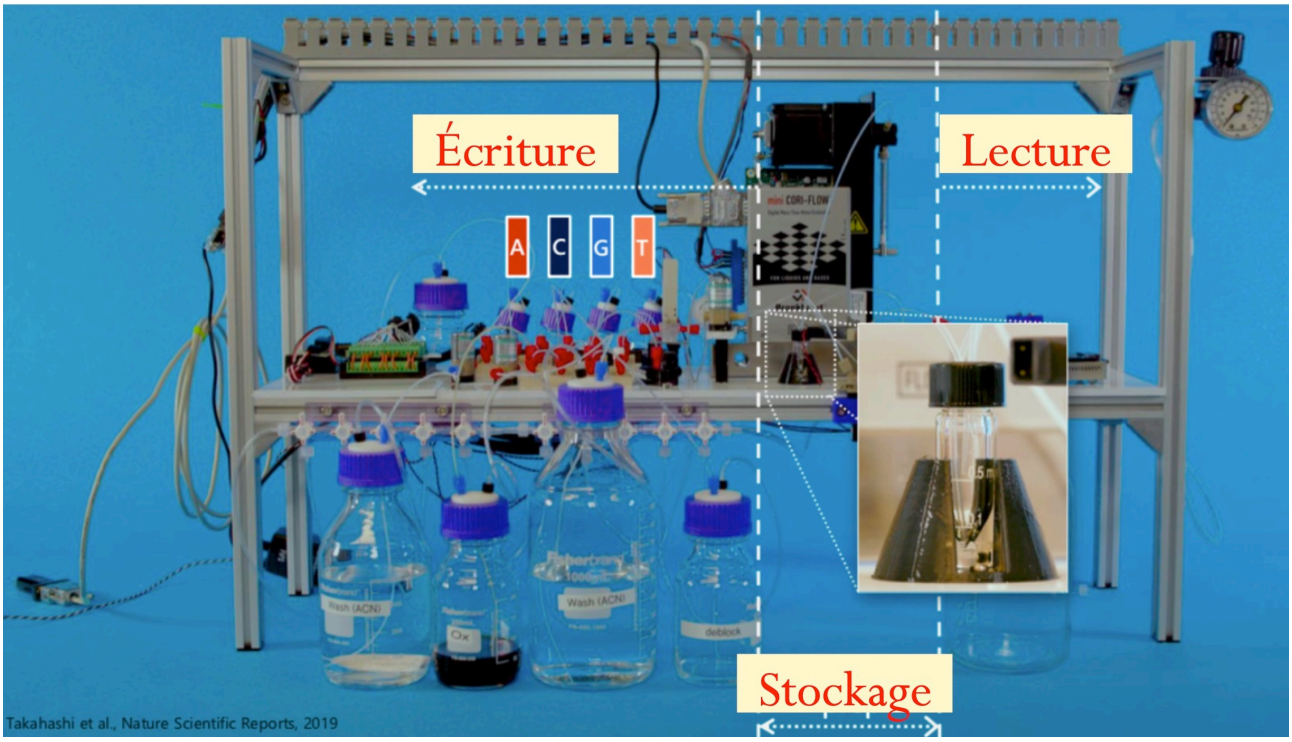
[28] https://www.illumina.com

[29] https://nanoporetech.com/

4. PROSPECTIVE

4.1. Projet avancé

Le projet le plus abouti à ce jour dans le domaine du stockage ou archivage de données numériques sur l'ADN est coordonné par Karin Strauss (Microsoft Corp. et Université de Washington, États-Unis) [30]. Ces chercheurs ont conçu un prototype de paillasse utilisant l'ADN comme support pour stocker l'information. L'appareil est entièrement automatisé et autonome. Il est composé de 3 parties : le synthétiseur, le système de stockage et le séquenceur. Le synthétiseur est capable de convertir l'information numérique en séquence d'ADN et d'écrire cette dernière. Puis l'ADN est stocké et protégé. Il est enfin extrait, et séquencé via le dispositif d'Oxford Nanopore Technologies.



Premier prototype entièrement automatisé de stockage des données sur l'ADN (Microsoft Corp. / Université de Washington).

Ce prototype est fonctionnel et a déjà permis de stocker et retrouver 1 Go de données. Les chercheurs l'optimisent pour qu'il devienne plus compact et plus rapide. Ils développent de nouveaux systèmes basés sur la microfluidique pour transporter des gouttes de réactifs sur un support électronique. Par ailleurs, ces chercheurs ont lancé un projet visant à implémenter des capacités de calcul en utilisant directement les propriétés physico-chimiques de l'ADN. À ce stade, leur système permet de regrouper les images, prises parmi un large ensemble, ressemblant le plus à une photo cible.

4.2. Initiatives mondiales

Dans ce domaine émergent, l'investissement public aux États-Unis se monterait à environ 150 millions de dollars, répartis entre les trois agences IARPA (projet MIST [31]), avec NSF (projet

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

[31] https://www.iarpa.gov/index.php/research-programs/mist

SemiSynBio [32]), et DARPA. Harvard Medical School est un des pionniers du domaine [33]. En outre, plusieurs compagnies sont actives dans ce domaine, comme Microsoft Corp., Twist Bioscience, Catalog DNA.

En Chine, il est difficile d'obtenir une image claire de la situation, mais il semble que Huawei et BGI Genomics seraient impliqués dans ce domaine.

Au Royaume-Uni, le European Bioinformatics Institute est un des pionniers du domaine [34]. En outre, plusieurs compagnies sont actives dans ce domaine, comme Oxford Nanopore Technologies, Nuclera Nucleics, Evonetix Ltd.

En Allemagne, un financement de 4,2 millions d'euros a été accordé par le ministère hessois pour le projet MOSLA [35] (Universités de Marburg, Darmstadt, Giessen, 2019-22).

L'Union Européenne (initiatives "FET") finance un projet "OligoArchive" [36] qui implique entre autres deux laboratoires français de Nice.

En France n'existe aucun investissement public visant directement ce domaine. Cependant, il faut noter un projet universitaire portant sur l'usage de copolymères non-ADN, porté par Jean-François Lutz (It Charles Sadron, Université de Strasbourg) [37]. En outre, la compagnie DNA Script est bien positionnée dans le domaine de la synthèse enzymatique d'ADN. Enfin, la compagnie Imagène a une position forte dans le domaine du stockage d'ADN à très long terme.

4.3. Limites et perspectives

Plusieurs études ont montré que l'archivage de données numériques sur l'ADN peut prendre en charge l'accès aléatoire et évolutif (réécriture de certaines parties) aux données, et le stockage d'information sans erreur. Cependant, des défis techniques subsistent pour que ce procédé devienne viable économiquement. Ils concernent l'amélioration des coûts, de la vitesse et de l'efficacité des technologies de lecture et surtout d'écriture, y compris l'accès sélectif aux données.

Concernant l'écriture, plusieurs acteurs du domaine placent beaucoup d'espoir dans la synthèse enzymatique d'ADN qui leur semble être la voie du futur. Quant à la lecture, l'usage de nanopores offre un bon potentiel. Notons aussi que, quoique l'écriture et la lecture de l'ADN soient d'allures limitantes, cet inconvénient est pallié dans certaines d'applications par la possibilité de parallélisation massive. Concrètement, d'ici 2024 une seule machine pourrait probablement écrire et lire 1 To par jour.

Pourtant, plusieurs ordres de grandeur manquent actuellement pour atteindre la viabilité économique de la solution ADN pour l'archivage de mégadonnées : environ mille pour le coût de lecture, et 100 millions pour celui d'écriture. Ces facteurs peuvent sembler faramineux. Ce serait oublier la célérité des progrès des technologies ADN. Ainsi, George M. Church, dans son exposé de mars 2019, a estimé que les coûts de la lecture et de l'écriture de l'ADN avaient chuté en 10 ans d'un facteur supérieur au million, soit un facteur 2 tous les 6 mois. Ceci est à comparer aux progrès dans les domaines électronique et informatique. D'une part la "loi" de Moore, déjà mentionnée, constate le doublement des densités des semi-conducteurs tous les 2 ans entre 1971 et 2016. Et d'autre part, la société Seagate a signalé qu'elle avait fait descendre

---

[32] https://www.src.org/program/grc/semisynbio/

[33] 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

[34] https://www.ebi.ac.uk/research/goldman/dna-storage

[35] https://mosla.mathematik.uni-marburg.de/gb/

[36] https://oligoarchive.github.io

[37] http://recherche.unistra.fr/index.php?id=30740

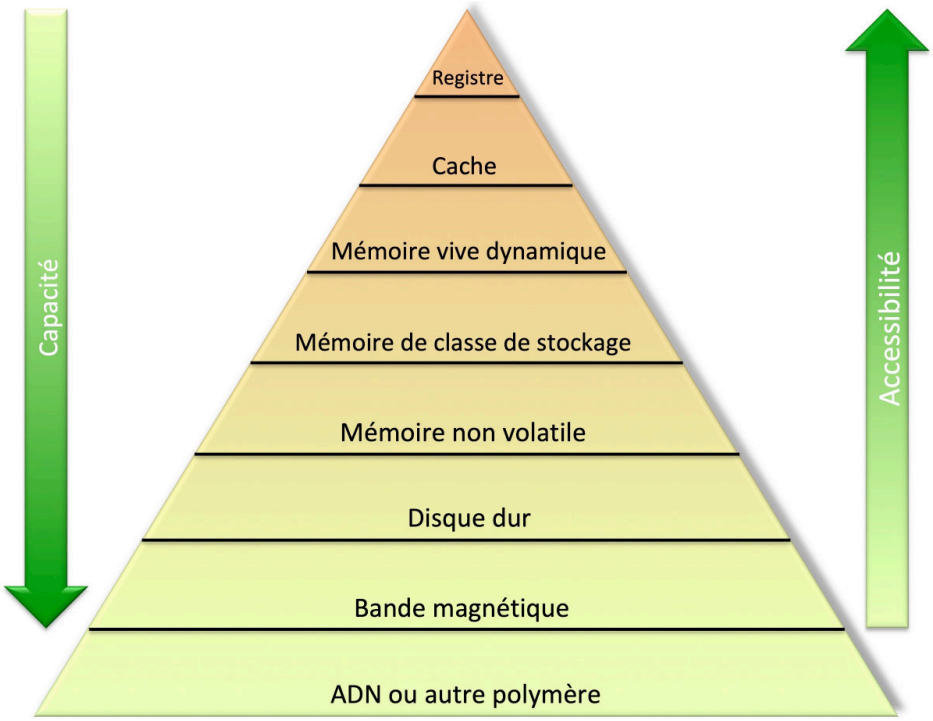
en 29 ans le coût du Mo sur disque d'un facteur 1.300.000. On voit donc que, selon ces critères, les technologies de l'ADN évoluent bien plus vite que celles de l'information.

Des approches hétérodoxes ont aussi été proposées, par exemple basées sur la formation et reconnaissance de structures secondaires de l'ADN de type tige-boucle (de deux longueurs nettement différentes représentant '0' et '1'), avec l'avantage dun taux d'erreur abaissé, et l'inconvénient d'une perte de densité informationnelle. D'autres approches encore se libèrent de l'usage de l'ADN pour envisager d'autres hétéropolymères ou copolymères présentant des avantages théoriques. Lorsqu'ils rejoindront les performances de l'ADN en termes de lecture et écriture, ce qui pourrait prendre une décennie, ils feront probablement irruption sur le marché.

4.4. Marché potentiel

Un certain consensus s'est dessiné parmi quelques acteurs du domaine pour considérer que la viabilité économique du stockage inframoléculaire d'information pourrait être atteinte sous 5-10 ans pour des marchés de niche. Citons à titre d'exemple l'archivage à long terme d'informations sensibles : les atouts maîtres seraient la facilité à multiplier l'ADN pour répartir géographiquement des copies de l'information, ainsi que l'instantanéité de sa destruction volontaire.

Pour entrer en compétition avec les marchés plus globaux de l'archivage de mégadonnées, il faudra peut-être 10-20 ans. Le handicap principal de l'ADN résidant dans le futur proche en la lenteur des procédés d'écriture et lecture, il est raisonnable de supposer que son usage se cantonne encore longtemps dans l'archivage à long terme, où ses avantages sont évidents. En ce cas, la principale compétition résidera dans l'usage de la bande magnétique, actuellement la solution de choix pour l'archivage à long terme. Il est possible aussi que l'ADN se positionne en complémentarité de la bande magnétique, pour l'archivage à très long terme d'énormes mégadonnées numériques.



Pyramide des types de mémoires dans les systèmes informatiques. En bas de la pyramide a été ajouté à titre hypothétique l'usage de l'ADN ou d'un autre hétéropolymère.