ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

Archiver les mégadonnées au-delà de 2040 : la piste de l'ADN

— ADN : lire, écrire, stocker l'information —

Rapport de l'Académie des technologies

Académie des technologies
Grand Palais des Champs-Élysées - Porte C
Avenue Franklin D. Roosevelt -
75008 Paris
+33(0)1 53 85 44 44
secretariat@academie-technologies.fr
www.academie-technologies.fr

ISBN : 979-10-97579-17-3

— ADN : lire, écrire, stocker l'information —

Rapport

de l'Académie des technologies

Archiver les mégadonnées au-delà de 2040 :

la piste de l'ADN

Sommaire

Résumé

Le stockage et archivage des mégadonnées numériques (« *big data* », le carburant de l'intelligence artificielle) par l'approche actuelle des centres de données ne sera pas soutenable au-delà de 2040. Il est donc urgent de focaliser des efforts soutenus en recherche et développement (R&D) pour l'avènement d'approches alternatives, dont aucune n'est présentement assez mature.

La sphère globale des données (SGD) créées par l'humanité était estimée en 2018 à trente-trois mille milliards de milliards de caractères (« octets »), du même ordre que le nombre estimé de grains de sable sur terre. Ces données proviennent de la recherche et de l'industrie, mais aussi de nos connexions amicales et professionnelles, livres, vidéos et photos, informations médicales ; sans oublier dans le futur voitures autonomes, capteurs, télésurveillance, réalité virtuelle, diagnostic et chirurgie à distance. La SGD augmente d'un facteur environ mille tous les vingt ans.

Une part majoritaire de ces données est ensuite stockée dans plusieurs millions de centres de données (en incluant ceux des entreprises et le « cloud »), qui fonctionnent au sein de réseaux de transmission. Ensemble, ceux-ci consomment déjà environ 2 % de l'électricité dans les pays avancés. Leur coût de construction et exploitation est globalement de l'ordre de mille milliards d'euros. Ces centres couvrent un millionième de la surface émergée du globe ; au rythme actuel, ils couvriraient un millième vers 2040.

Les technologies de stockage utilisées par ces centres sont rapidement frappées d'obsolescence aux plans du format, du dispositif de lecture/écriture, et aussi du support, lequel nécessite des copies tous les cinq à sept ans pour garantir l'intégrité des données. Elles posent aussi des problèmes croissants d'approvisionnement en ressources rares comme le silicium de qualité électronique.

Une alternative prometteuse est offerte par les supports moléculaires porteurs d'information, tel que l'ADN utilisé ici comme agent chimique en dehors du vivant, ou d'autres hétéropolymères non-ADN très prometteurs. Potentiellement, l'ADN permet des densités informationnelles dix millions de fois supérieures aux mémoires traditionnelles : toute la SGD actuelle tiendrait dans une fourgonnette. L'ADN est stable à température ordinaire durant plusieurs millénaires, sans consommation énergétique. Il peut être aisément multiplié ou détruit à volonté. Certains calculs peuvent être physiquement implémentés avec des fragments d'ADN. Enfin, sa technologie ne deviendra pas obsolète car il constitue notre matériel héréditaire.

Pour archiver et retrouver des données dans l'ADN, il convient d'enchaîner cinq étapes : coder le fichier de données binaires dans l'alphabet de l'ADN qui possède quatre lettres, puis écrire, stocker, lire l'ADN et, enfin, décoder l'information lue. Un prototype réalisant ces opérations fonctionne depuis mars 2019 chez Microsoft aux États-Unis. Actuellement, plusieurs ordres de grandeur manquent pour atteindre la viabilité économique de cette approche : un facteur environ mille pour le coût et la vitesse de lecture, et cent millions pour ceux d'écriture. Ces facteurs peuvent sembler faramineux. Ce serait oublier la célérité des progrès des technologies ADN, proches d'un facteur mille tous les cinq ans, donc bien plus rapides que dans les domaines électronique et informatique.

Le rapport se termine par une réflexion prospective qui permet au lecteur de se faire une opinion en quelques pages sur les perspectives techniques, économiques et nationales sur la piste ADN pour archiver les mégadonnées.

Recommandations

Face aux limites physiques qu'atteignent les centres de données, la technologie moléculaire d'archivage des mégadonnées a le potentiel de devenir économiquement viable entre 2025 et 2040, progressant de marchés de niche vers des marchés plus globaux. Dans le futur proche, le handicap principal de l'ADN résidera en la lenteur des procédés de lecture et surtout d'écriture. Son usage se cantonnera donc initialement à l'archivage de données nécessitant d'être conservées longtemps, où ses avantages sont évidents, en compétition ou complémentarité avec l'actuelle solution, la bande magnétique.

Plusieurs laboratoires académiques, des jeunes pousses et quelques grandes entreprises se sont explicitement positionnés sur ce défi, au Royaume-Uni, en Allemagne, Irlande, Suisse, et probablement Chine. Les États-Unis y investissent environ 150 millions US$, avec des objectifs techniques précis à deux et quatre ans.

En France, au moins un laboratoire et deux petites entreprises (dont l'une vient d'être dotée par une agence des États-Unis) y ont des positions originales et fortes dans les segments-clé que sont les polymères non-ADN, la synthèse enzymatique de l'ADN et son stockage de très longue durée. Au-delà, il existe en France un gisement de compétences pertinentes en biologie, chimie, informatique et sciences de l'ingénieur, qui pourraient être mobilisées dans une nécessaire synergie entre secteurs public et privé.

L'archivage moléculaire de données constitue un enjeu majeur et stratégique à horizon proche. Il serait donc souhaitable de capitaliser sur le laboratoire et les deux sociétés identifiés, et sur le gisement plus large de compétences, afin de permettre à ces entités françaises de devenir des acteurs significatifs et d'ouvrir une perspective européenne. Dans ce but, voici deux recommandations.

A. Lancer une action concertée au plan national

Ceci passerait par une vigoureuse programmation pluriannuelle de subventions publiques explicitement dédiées, qui pourrait user des instruments suivants :

- avant tout, des appels d'offres visant spécifiquement à susciter des propositions ambitieuses de ruptures technologiques ; incitant aux synergies public/privé et trans-disciplinaires ; abaissant les risques pris en se lançant dans cette approche émergente ; s'appuyant sur un comité scientifique incluant les pionniers français et des experts internationaux ;
- une plate-forme technologique transdisciplinaire, lieu d'expérimentation et de réflexion ; fédérant les secteurs public et privé ; ayant vocation à ultérieurement s'insérer comme le nœud français dans le réseau européen pertinent ;
- une conférence annuelle et internationale, initialement à dominante française.

B. Proposer une programmation européenne

La France pourrait proposer d'identifier ce thème comme un domaine à part entière dans le futur programme de recherche de la Commission européenne. Cette dernière pourrait user des instruments suivants :

- des appels d'offres dédiés, récurrents, transdisciplinaires, et plurinationaux ;
- la mise en place d'un réseau européen de laboratoires publics et privés, facilitant la circulation des personnes, compétences et savoirs.

Chapitre I

Contexte et motivation

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

La sphère globale des données (SGD)

L'ensemble des données numériques créées par l'humanité, la « sphère globale des données » (SGD), contient environ autant de caractères (d'octets)[3] que le nombre d'étoiles dans l'univers actuellement observable, ou que le nombre estimé de grains de sable sur la terre[4]. Cette SGD était estimée en 2018 à 33 zettaoctets (Zo ; soit 33 x 10^{21} caractères)[5]. Elle double tous les 2 à 3 ans, et atteindra environ 175 Zo en 2025. Par exemple, chaque minute dans le monde, environ 400 heures de vidéo (200 Go) sont ajoutées à la

1 L'Académie des technologies s'est penchée à plusieurs reprises sur les *big data*, et deux contributions récentes en ont résulté, portant l'une sur les aspects technologico-stratégiques (*Big data : un changement de paradigme peut en cacher un autre*, EDP Sciences 2015), l'autre sur les aspects éthiques (*Big data — Questions éthiques*, Académie des technologies 2019).

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

3 Un « octet » est une suite de 8 « bits ». Un bit ne prend que 2 valeurs désignées usuellement par les chiffres '0' et '1' (d'où le terme de codage « binaire » ou « numérique »). Donc il existe 256 (28) 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 Archimède a calculé le volume de la terre (et même de l'univers) en prenant comme unité de mesure le volume du grain de sable (L'Arénaire).

5 Préfixes du Système International d'unités, et nombres correspondants :

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

SGD. Autre exemple, pris dans le monde de la recherche, l'Organisation européenne pour la recherche nucléaire (Cern) a produit plus de 100 Po de données qui doivent être conservées pour les générations suivantes de physiciens. Au rythme actuel, la SGD atteindrait plus de 5 000 Zo en 2040. Une autre façon d'appréhender un nombre aussi élevé consiste à dire qu'il 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 au long terme. Outre le stockage local (sur un ordinateur ou téléphone), dont la croissance ralentit actuellement, le stockage centralisé est en rapide augmentation, en particulier dans les centres de données y compris dans le *cloud* ou *nuage*. Celui-ci permet aux utilisateurs individuels de profiter de ressources informatiques à la demande et d'augmenter le niveau d'automatisation grâce à la virtualisation des serveurs. En 2021, le nuage stockera autant d'information que les centres de données traditionnels. Dans la suite de ce rapport, nous considérerons l'ensemble des centres de données, nuage inclus. Des locaux dédiés à ce stockage centralisé ne cessent d'être construits dans le monde, souvent dans les pays froids car ce stockage est grand consommateur d'électricité et demande un refroidissement important. Paradoxalement, les capacités disponibles de stockage en centres de données ne représentent en 2020 que 40 % des données numériques créées[6] .

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 d'un mégawatt (MW). Au cours de sa durée de vie d'environ vingt ans[7], il aura fait usage de 66 tonnes de cuivre, 15 tonnes de plastiques, 33 tonnes d'aluminium et 152 tonnes d'acier. Chaque année, il aura été traversé par 23 millions de litres d'eau ; en incluant le refroidissement des serveurs, il aura consommé 18 millions de kilowatt-heures (0,018 terawatt-heures ou 0,018 TWh)[8] d'électricité. En revanche, en 2018,un gros centre de données représente plusieurs milliards d'euros en investissement, un million de m^2 et un million de serveurs[9]. Il consomme un gigawatt (GW) d'électricité (dont 40-50 % environ pour le refroidissement), soit environ 10 TWh par an, soit 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[10]. Au total, en incluant ceux des entreprises, il y avait 8,6 millions de centres de données dans le monde en 2017, pour une surface totale de plus de 170 millions de m^2, soit l'équivalent de 25 000 terrains de football[11]. Cette surface représente environ un millionième des

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

9 La densité en serveurs apparaît plus faible dans le gros centre de 2018 que dans le petit centre de 2008. Ceci reflète l'évolution des serveurs, en termes d'architecture, mais aussi en termes d'usage : l'accent est mis sur le calcul et la gestion de bases de données en 2008, et plutôt sur les services internet en 2018.

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

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-millions-de-datacenters-dans-le-monde-en-2017

terres émergées de la planète puisque ces dernières couvrent approximativement 150 millions de km^2. Si le rythme actuel d'un doublement tous les deux ans se poursuivait, un millième des terres émergées serait occupé par ces centres avant 2040. Cependant, ceci est probablement une surestimation car l'efficacité énergétique des centres de données augmente continuellement[12] et leur empreinte au sol diminue. Les centres de données et leurs réseaux de connexions équivaudraient au cinquième 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 TWh, et engendré l'émission de 423 mégatonnes d'équivalent-CO_2. En 2012, ils étaient responsables de 2 % des émissions de gaz à effet de serre produits globalement par les technologies de l'information. L'investissement mondial annuel pour construire de nouveaux centres de données est de l'ordre de plusieurs dizaines de milliards US$[13]. Par exemple Google aurait à lui seul investi dix milliards US$ par an sur la période 2015-2017.

Conclusion provisoire

La lecture de ces quelques éléments d'information montre clairement que la croissance des données numériques ne serait pas soutenable au-delà d'environ 2040 si le rythme actuel se poursuivait et si la technologie restait constante.

Photographie aérienne d'un centre de données (Farrat Irlande).

Source : The agency creative

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

Pourrait-on faire appel à une technologie qui permettrait d'archiver l'ensemble des mégadonnées réutilisables dans une fourgonnette, quasiment sans dépense énergétique ?

La réponse est oui : c'est là le sujet de ce rapport : il synthétise les travaux du groupe de travail transdisciplinaire (2018-2020) ***ADN : lire, écrire, stocker l'information***, qui a auditionné vingt-six spécialistes mondiaux.

Mais avant de présenter cette technologie, il est important de faire le point sur l'état de l'art dans le domaine du stockage et de l'archivage des données numériques, en faisant ressortir ses limites et perspectives.

Chapitre II

État de l'art du stockage et archivage de données numériques

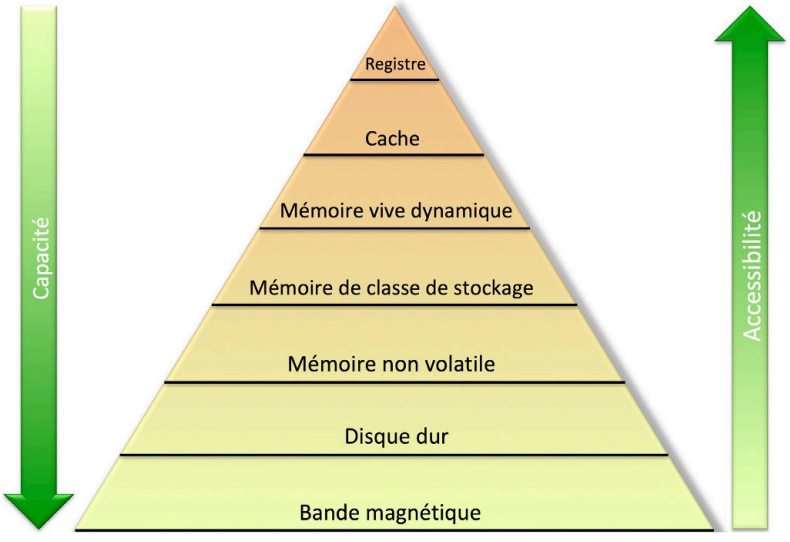
Le problème

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

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

Hiérarchie de la mémoire 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.



Pyramide des types de mémoires dans les systèmes informatiques.

Crédit : François Képès et Carlo Reita.

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 (quelques milliers d'octets).

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 kilooctets (Ko) à mégaoctets (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 (plusieurs Téraoctets — To) ;
- les bandes magnétiques, utilisées pour l'archivage à très long terme (> 10 ans) des informations. Le coût de la mémoire de masse est relativement faible, mais sa vitesse d'accès est inférieure aux autres types de mémoires.

Les technologies de stockage d'information en électronique

Pour satisfaire les besoins du stockage d'information numérique, une large gamme de technologies est employée. On distingue cinq générations de stockage : physique, magnétique, optique, à mémoire flash et virtuel. 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.

Technologies de stockage magnétique

Supports

Ces supports comprennent la bande magnétique, le disque dur et les mémoires MRAM[14]. Ils possèdent une grande capacité de stockage et une durée de conservation de l'information de l'ordre de la décennie. Ainsi, c'est une technologie de choix pour l'archivage de données. La méthode par bande magnétique présente la densité de stockage la plus élevée, mais l'accès intrinsèquement séquentiel des données limite les vitesses de lecture et d'écriture. Les disques permettent un accès aléatoire beaucoup plus rapide au prix d'une densité moindre.

Principe

L'écriture sur ces supports se fait par aimantation et la lecture est magnétique. Le courant dans une

14 MRAM : *Magnetoresistive random-access memory* (mémoire vive magnéto-résistive)

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.

Technologies de stockage par charge

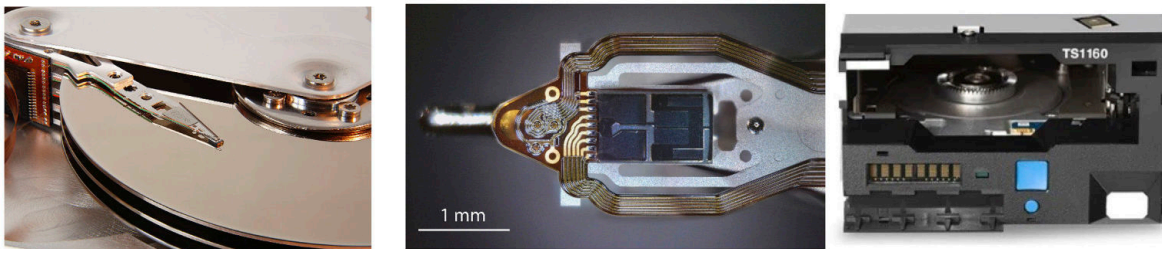
Supports

Cette technologie comprend les mémoires SRAM[15], DRAM[16] 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.

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.



Supports de stockage magnétique

Crédits de gauche à droite :
Eric Gaba - Wikimedia Commons user: Sting
Roman Starkov - Creative Commons Attribution-Share Alike 4.0
https://www.ibm.com/it-infrastructure/storage/tape/drives

Le cas des mémoires Flash

Il existe deux classes de stockage par charge : les mémoires volatiles[17] (DRAM et SRAM) et mémoires non-volatiles (Flash). La mémoire Flash est la mémoire non-volatile la plus utilisée aujourd'hui (disques statiques, cartes mémoires pour appareils portables, clés USB).

La durée de vie d'une mémoire Flash est calculée par le nombre d'écritures et d'effacements de données (typiquement dix mille à un million) que peut subir le support avant de se dégrader. Sa structure particulièrement régulière a permis une réduction de taille progressive, afin d'obtenir des densités de stockage comparables à celles des disques durs. De plus, les procédés microélectroniques de fabrication

15 SRAM : *Static random-access memory* (mémoire vive statique)

16 DRAM : *Dynamic random-access memory* (mémoire vive dynamique)

17 Perte de l'information en absence d'alimentation

en série en ont réduit drastiquement les coûts.

Actuellement, les systèmes de mémoires Flash NAND[18] 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.

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

Comparaison des technologies de stockage en électronique

Les technologies de stockage et leur évolution sont comparées dans ce tableau :

	Type de stockage	Stockage magnétique		Stockage par charge			
				mémoires volatiles		mémoires non-volatiles	
	Supports	Bandes magnétiques	Disques durs magnétiques (HDD)	SRAM	DRAM	PCM*	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	$5\text{-}8\times10^{6}$	<10-50	10-50	20-70	25 000
	Temps d'écriture (ns)	N/A	$5\text{-}8\times10^{6}$	<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	10^{15}	$>10^{17}$	10^{17}	10^{7}-10^{8}	10^{4}-10^{6}
	Coût de production ($/Go)	0,016	0,039	10^{2}-10^{3}	10	1	0,32
	Revenus générés (milliard$)	0,65	26,8	N/A	N/A	N/A	38,7
* PCM : forme de mémoire vive non-volatile.							

Deux conclusions se dégagent :

- la bande magnétique reste le meilleur compromis pour l'archivage des données à long terme ; en effet, la durée de stockage est élevée et le coût de production reste le moins cher, toutes technologies de stockage confondues ; en outre, la bande magnétique consomme moins de 1 % de l'énergie électrique totale d'un centre de données ;
- les disques durs magnétiques (HDD) et à état solide (Flash) sont les meilleurs compromis pour le stockage de masse, car ils possèdent la meilleure densité de stockage tout en offrant un accès rapide aux données stockées.

18 NAND : porte logique « NOT-AND », ou en français « NON-ET ».

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 d'accès, longévité, durabilité et coûts. Poursuivre selon la loi de Moore[19] est un défi de plus en plus difficile. Notons aussi que la production de silicium de qualité électronique (composant abondant des disques durs) est inférieure d'un facteur cent aux besoins futurs.

En outre, quoique les principes, par exemple magnétisme, sont inchangés, ces supports traditionnels sont rapidement frappés d'obsolescence, sur trois plans[20] :

- **le format de stockage** : Par exemple, l'usage des disquettes 3,5 pouces s'est progressivement éteint entre 2000 et 2010. Actuellement, les bandes magnétiques continuent comme par le passé à subir des mutations, justifiant leur remplacement par de nouvelles générations incompatibles avec les précédentes ;
- **le dispositif de lecture / écriture** : pour reprendre le même exemple, les lecteurs fonctionnels de disquettes 3,5 pouces sont devenus rares ; et les lecteurs de bande magnétique poursuivent leur évolution ;
- **le support lui-même** : du fait de leur constitution physique, tous les supports de stockage ont une durée de vie limitée, au plus une dizaine d'années, entraînant le risque de perdre de l'information. Pour s'en protéger, il faut régulièrement 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.

En conclusion, le monde est confronté à un grave problème de stockage de données qui ne peut pas être résolu par les technologies actuelles. Tous ces inconvénients ne sont pas partagés par la technologie de stockage de l'information sur l'ADN, que nous allons maintenant analyser.

---

19 La *loi de Moore* a été exprimée en 1965 dans le magazine *Electronics* par Gordon E. Moore, un des trois fondateurs d'Intel. Cette loi empirique énonce que le nombre de transistors des microprocesseurs (et non plus de simples circuits intégrés moins complexes) 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 deux ans jusqu'à présent.

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

Chapitre III

Stockage de l'information numérique sur l'ADN

De nouvelles méthodes de stockage 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 ce polymère naturel comme support d'archives numériques. L'ADN permet de stocker *1 bit pour environ 50 atomes*, alors que le stockage magnétique requiert environ *un million d'atomes,* support mécanique non compris.

Bref historique du stockage de données sur l'ADN

Richard Feynman en 1959, et Mikhail Neiman en 1964, ont été les premiers à envisager l'ADN comme support de stockage de l'information numérique. 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[21] a conçu et synthétisé un fragment d'ADN de dix-huit nucléotides contenant un message numérisé symbolisant l'icône *MicroVenus*, 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[22]. En 2013, l'équipe de Nick Goldman (Institut Européen de Bioinformatique, Royaume-Uni) a converti quatre fichiers en séquence d'ADN, pour un total de 0,7 Mo[23]. L'information a été retranscrite sans erreurs.

En 2018, Microsoft Corp. et l'Université de Washington, aux États-Unis d'Amérique, ont stocké sur l'ADN 1 Go d'information venant de fichiers de types variés[24]. Ils détiennent depuis le record.

En 2024, il est projeté d'archiver 1 To (équivalent à environ mille films) en vingt-quatre heures pour un coût de 1 000 US$[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. 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

L'ADN : un support de stockage performant

L'ADN

L'ADN dans le monde vivant est un des supports de l'information héréditaire, comme rappelé dans l'encadré.

ADN ET INFORMATION BIOLOGIQUE

L'acide désoxyribonucléique (ADN) est dans nos cellules le support de l'information héréditaire. La question de l'information biologique est devenue un sujet scientifique en 1972 avec le livre d'Henri Atlan[1]. L'information est largement distribuée au sein d'une cellule vivante :

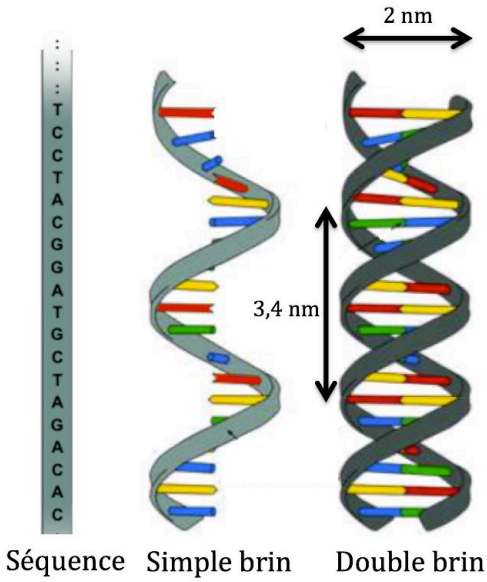
- une part de cette information réside dans l'organisation dynamique même que cette cellule entretient activement et transmet à sa descendance ;
- une seconde part, héritable aussi, réside dans l'ensemble de ses états épigénétiques (état d'un interrupteur ou d'un oscillateur etc.) ;
- une troisième part est portée par ses chromosomes, sous la forme de la séquence de l'ADN qui les constitue. Cette dernière part est de loin la plus aisément descriptible. Malgré cette facilité relative, mesurer la quantité d'informations portée par l'ADN dans un système naturel comme une cellule vivante est actuellement impossible dans sa généralité, et difficile dans des cas limités. Il n'en est pas de même dans un système artificiel, au travers du prisme de l'information ; ainsi l'ADN et son alphabet à quatre lettres (A, C, G et T) peut être directement utilisé comme un support d'information binaire codant sur deux bits, selon un principe similaire à celui d'un support magnétique à deux lettres (0 et 1) codant sur un bit. L'objet de cet encadré est de rappeler que ce principe est extrêmement réducteur par rapport au rôle complexe, malléable et multiforme que joue le matériel héréditaire dans une cellule vivante. D'autant plus que, comme support d'information numérique, l'ADN est majoritairement utilisé hors de la cellule, en éprouvette (*in vitro*). L'ADN sera donc vu par la suite comme un polymère aux propriétés intéressantes, aisément manipulable grâce à la remarquable boîte à outils que les biologistes ont accumulée depuis le milieu du xx^{e} siècle.

1 Henri Atlan, *L'organisation biologique et la théorie de l'information,* Seuil, 1972.

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 polymère[26] linéaire (non branché) composé d'un assemblage de nucléotides. Chaque nucléotide est composé d'une des quatre bases azotées, adénine (A), guanine (G), thymine (T), cytosine (C), liée à un sucre désoxyribose, lui-même lié à un groupe 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, l'adénine et la thymine s'apparient avec deux liaisons hydrogène, tandis que la guanine et la cytosine s'apparient avec trois liaisons hydrogène.

L'ADN peut, depuis 1869 (Friedrich Miescher), être manipulé en dehors des cellules, *in vitro* ; c'est principalement *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.

26 Un polymère est une grande molécule constituée de nombreuses sous-unités répétées, appelées monomères. Si toutes les sous-unités sont identiques, on parle d'homopolymère. Si les sous-unités ne sont pas toutes identiques, on parle d'hétéropolymère ou de copolymère. Un copolymère est un polymère issu de la copolymérisation d'au moins deux types de monomères chimiquement différents..



Représentations de l'ADN. À droite, hélice d'ADN double brin (montrant les appariements AT et CG) ; au milieu, ADN en simple brin ; à gauche, séquence du même simple brin, maintenant déroulé pour une représentation linéaire.

Crédit : adapté par François Képès.

Avantages du stockage d'information numérique sur l'ADN

La densité informationnelle

La densité informationnelle de l'ADN est environ dix millions de fois supérieure à celle des meilleurs systèmes traditionnels. L'ADN peut en principe stocker un demi Zo d'information par gramme (g). Ainsi, les chercheurs estiment que la SGD de l'humanité entière tiendrait actuellement dans moins de 100 g d'ADN.

- Cependant, en pratique ce n'est pas une seule molécule d'ADN qui est synthétisée pour coder 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[27].

Tenant compte de ces pertes de densité, on peut estimer que la SGD, stockée sur ADN, tiendrait plus réalistement dans une fourgonnette.

La consommation

Le stockage de l'ADN à température ordinaire n'implique aucune consommation de ressources, et les opérations sur l'ADN sont beaucoup moins énergivores qu'en électronique ; un gain d'un facteur mille a été évoqué[28].

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.

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

La longévité

La longévité de l'ADN est environ dix 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[29]. En laboratoire, une demi-vie de 52 000 ans a été démontrée en accélérant artificiellement son vieillissement[30].

L'obsolescence

L'obsolescence du support ADN ne se produira pas tant que l'homme disposera des technologies nécessaires à l'écriture et à la lecture de molécules d'ADN, qui font partie intégrante de la médecine moderne. Reste la question du codage/décodage, qui serait résolue si un standard émergeait.

La copie ou multiplication

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). La PCR est effectuée à partir de deux amorces. Ces dernières sont des fragments d'ADN simple-brin d'une vingtaine de nucléotides qui se fixent par homologie sur deux zones spécifiques délimitant la région d'intérêt à amplifier. 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 pour une fraction d'euro[31]. Ceci représente un avantage considérable par rapport à la lourdeur et au coût de duplications de données sur les supports traditionnels.

La destruction à volonté

La destruction à volonté de l'ADN est réalisable aisément et rapidement. 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. Les ADNases sont disponibles dans le commerce pour un coût modique. Des traitements physiques plus brutaux mais moins sophistiqués, par exemple à haute température, permettraient aussi en une fraction de seconde de détruire l'ADN, même protégé dans un tube ou une nanobille.

Calculs

Les propriétés physico-chimiques de l'ADN se prêtent à y implémenter directement certains calculs. Le principe d'un tel calcul (développé plus avant chapitre IV) est de coder un problème combinatoire avec des brins d'ADN que l'on fait synthétiser sur mesure, de manipuler ces brins par les outils de la biologie moléculaire pour simuler les opérations qui isolent la solution, puis de lire cette dernière par séquençage[32].

---

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

32 Adelman, LM (1994). Molecular computation of solutions to combinatorial problems, *Science* 266, 5187. pp1021-1024.

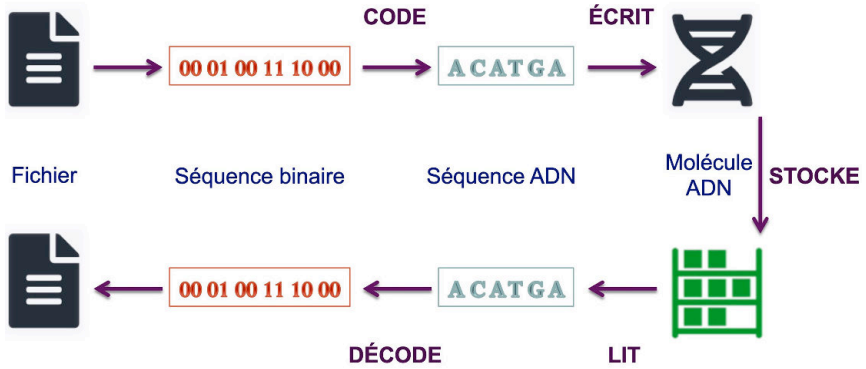
Technologies sous-jacentes au stockage d'information sur ADN

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.

Codage de l'information

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 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 ces '1'. Un octet est une série de huit bits.



Étapes du processus de stockage des mégadonnées numériques sur l'ADN. Ici sont représentés pour exemple 12 bits successifs extraits du fichier numérique. Ces 12 bits sont codés sous la forme de 6 nucléotides qui sont écrits en succession dans une molécule d'ADN. Cet ADN est ensuite stocké. Puis il est lu, et la séquence de nucléotides ainsi obtenue est décodée pour reconstituer le fichier numérique d'origine.

Crédit : François Képès.

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 de 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 G et 11 à T. 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 (cf. chapitre IV).

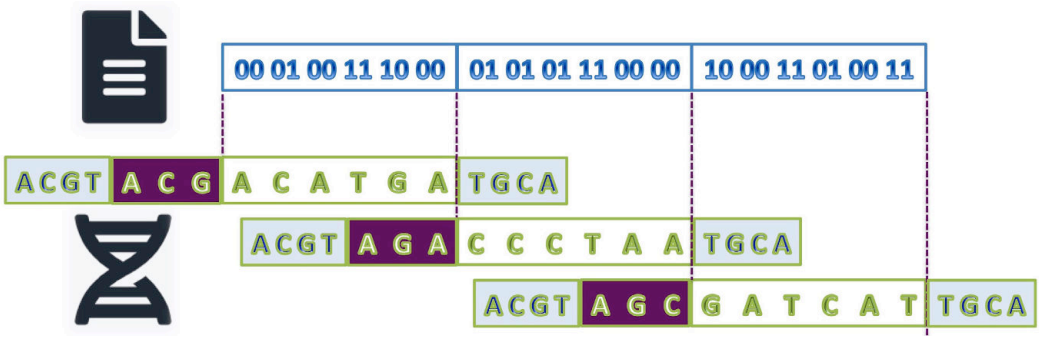
La technologie actuelle de synthèse d'ADN est limitée à des fragments de l'ordre de 200 nucléotides au maximum, donc très courts au regard des fichiers informatiques qui sont volumineux. Cela est dû au fait que plus un brin d'ADN est long, plus il est difficile de le construire chimiquement. On peut 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. 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.

Découpage du fichier informatique en fragments

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.

Conversion des fragments d'octets en nucléotides d'ADN

Les fragments d'une vingtaine d'octets sont convertis en nucléotides d'ADN. Chaque segment d'ADN possède environ 200 nucléotides et contient la charge utile *(payload)* et l'étiquette. Cette dernière 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. Une fois le processus de codage d'information numérique sur l'ADN effectué, les fragments d'ADN sont synthétisés.



Conversion des fragments d'octets en nucléotides d'ADN. Le fichier numérique est divisé en segments d'une vingtaine d'octets (symbolisés en haut par des suites de 12 bits). Chaque segment donne lieu à une synthèse d'ADN (en bas) contenant la charge utile représentative du fichier numérique (vert). Les autres éléments (bleu et pourpre) permettent l'indexation et la correction d'erreurs.

Crédit : adapté par François Képès depuis la présentation de Karin Strauss (Microsoft Corp.).

Écriture de l'ADN

La synthèse d'ADN permet d'écrire des fragments d'acides nucléiques relativement courts, possédant un enchaînement défini de nucléotides. Il existe deux voies de synthèse, chimique ou enzymatique. La séquence de l'ADN synthétisé peut ensuite être modifiée par mutagenèse.

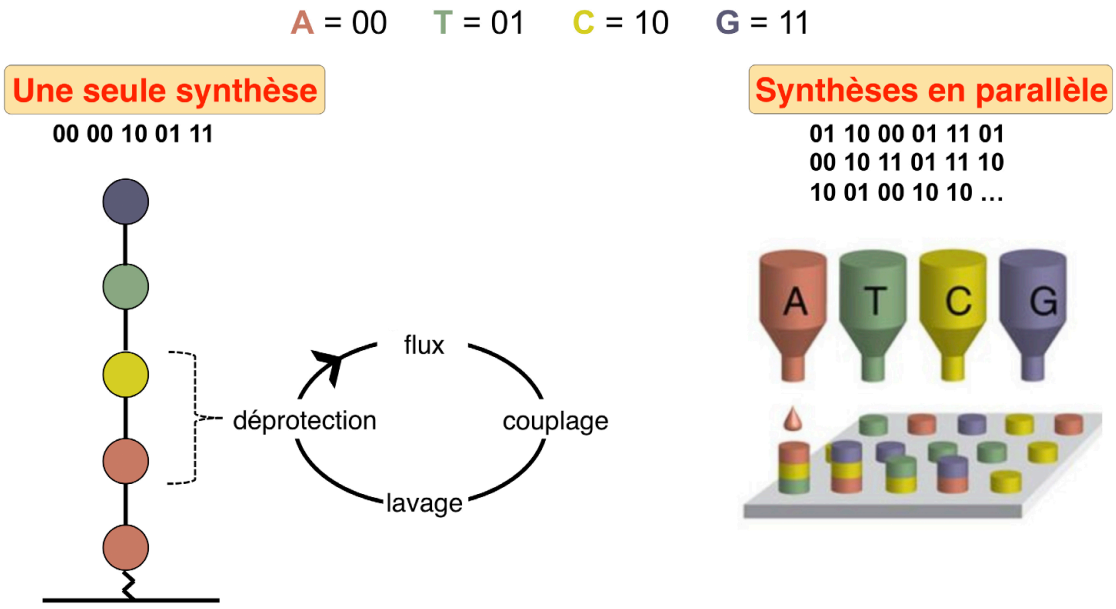
Principe de la synthèse chimique

Depuis 1983, il existe une méthode de synthèse de l'ADN par voie chimique. C'est aujourd'hui la seule méthode commerciale d'écriture de l'ADN. Son principe basé sur la chimie des phosphoramidites a très peu évolué depuis. Ce processus implique l'ajout au polymère ADN en cours de synthèse de nucléotides successifs, protégés par un groupe terminal bloquant tout ajout d'un second nucléotide. Ceci permet l'ajout à la séquence d'un seul nucléotide à la fois. L'excédent de nucléotide est enlevé par lavage, le groupe terminal protecteur est alors enlevé par réaction chimique. Les nucléotides suivants sont ajoutés l'un après l'autre selon le même cycle. Pendant toute sa synthèse, l'ADN reste accroché sur une résine. Il est décroché lors de la déprotection finale. Lors de la synthèse d' « un » fragment d'ADN, ce n'est en réalité

pas une unique molécule qui est synthétisée, mais une population de copies identiques : typiquement 10^{12} à 10^{15} copies, qu'il est possible de réduire en conditions contrôlées à une dizaine de copies afin d'augmenter la densité informationnelle[33].

Au début de la synthèse d'ADN par voie chimique, le processus d'ajout des nucléotides était manuel. Des synthétiseurs automatiques ont vu le jour dans les années 1990. Ils ont évolué et peuvent posséder jusqu'à 200 colonnes, permettant de synthétiser simultanément 200 fragments d'ADN différents. La société Twist Bioscience (États-Unis)[34], spécialisée dans la synthèse d'ADN, a miniaturisé ce processus. Elle réalise la synthèse d'ADN dans des puits gravés dans une micro-puce de silicium, synthétisant simultanément 10 000 fragments d'ADN différents comprenant jusqu'à 200 nucléotides. Elle atteindra rapidement le million de puits.

Synthèse chimique par phosphoramides :
Taux d'erreur par nucléotide ≥ 0,5 % fragments d'ADN ≤ 200 nucléotides en pratique



Synthèse de l'ADN par voie chimique sur support solide. À gauche, une synthèse séquentielle ; l'addition d'un nucléotide à la fois s'effectue selon le cycle montré au milieu. À droite, ce même processus est exécuté simultanément sur plusieurs sites.

Crédit : adapté par François Képès depuis la présentation de Nick Gold (Catalog DNA).

Le taux d'erreur pour chaque nucléotide ajouté est d'environ 0,1 %, ce qui explique qu'en pratique la longueur du fragment utilisable est limitée à 200 nucléotides. Pour écrire des morceaux d'ADN bien plus longs, la méthode usuelle est d'assembler bout à bout de nombreux fragments d'environ 200 nucléotides,

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

par exemple en jouant sur le chevauchement de leurs extrémités prévues pour être complémentaires deux à deux afin d'éviter toute erreur d'assemblage. Un autre inconvénient de la synthèse par voie chimique est que la chimie des phosphoramidites est polluante.

Principe de la synthèse enzymatique

En raison de ces limites, une méthode alternative de synthèse de l'ADN par voie enzymatique a été inventée au début des années 2010. Cette synthèse fait usage d'une ADN-polymérase spéciale présente dans les cellules immunitaires, appelée *Terminal désoxynucleotidyl Transférase* (TdT). *In vivo*, la TdT ajoute les nucléotides aléatoirement[35], contrairement à la plupart des ADN-polymérases qui dépendent d'une matrice simple-brin et allongent le brin antiparallèle en complémentarité de cette matrice. Dans le procédé décrit ici, l'utilisation de la TdT permet d'allonger l'ADN avec le seul nucléotide désiré en fournissant uniquement celui-ci à une étape donnée. Comme pour la synthèse chimique d'ADN, les chercheurs ajoutent un groupe chimique protecteur pour chaque nucléotide, empêchant ainsi 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.

En conséquence, cette approche biologique a plusieurs avantages par rapport à la traditionnelle voie chimique. La TdT possède une vitesse de synthèse élevée avec un taux d'erreur très faible. Le groupe chimique protecteur des nucléotides est différent ; il permet de conserver la solubilité du nucléotide dans l'eau, rendant l'approche enzymatique moins polluante que la voie chimique qui recourt à des solvants organiques.

Une demi-douzaine de sociétés se sont lancées dans cette nouvelle approche : Nuclera Nucleics, Ansa Biotech, Spindle Biotech, Molecular Assemblies, Merck, et en France DNA Script[36]. Il est trop tôt pour estimer à quel horizon cette approche biologique deviendra commercialement attractive.

Modification de séquence post-synthèse

Disposer d'un accès évolutif aux données offre une flexibilité appréciable dans le monde de l'informatique, car cela permet la réécriture d'une partie d'un fichier numérique sans entraîner la coûteuse nécessité de tout réécrire. De même serait-il intéressant de réécrire certaines parties de l'ADN sans toucher aux autres zones. De nombreuses méthodes de réécriture ont été développées depuis les débuts du génie génétique en 1973[37]. On parle globalement de méthodes de « mutagenèse dirigée » pour signifier qu'elles visent à obtenir une séquence prédéterminée par l'opérateur. Depuis 2009 existent des techniques permettant d'introduire simultanément de multiples mutations dans une longue double hélice d'ADN, y compris des chromosomes[38].

Comme le montre l'encadré, l'approche moléculaire présente un désavantage au regard des méthodes informatiques pour l'accès évolutif aux données.

---

35 Cette enzyme est ainsi responsable de la variabilité de portions spécifiques de gènes codant les chaînes protéiques d'immunoglobulines (anticorps) en voie de sélection pour s'adapter à un nouvel antigène.

36 http://www.dnascript.com/

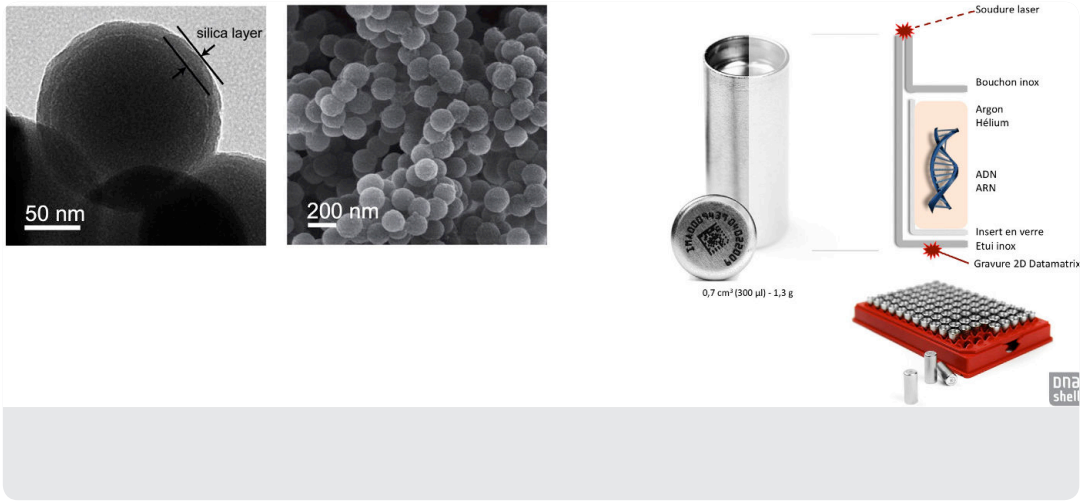
37 Esvelt KM, Wang HH. (2013) Genome-scale engineering for systems and synthetic biology. *Mol Syst Biol.* 9:641.

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.

Stockage au long terme 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 (ETH Zürich, Suisse), l'ADN synthétisé est encapsulé dans des nanobilles de silice, ensuite référencées et réparties dans des plaques à micro-puits. Avant d'être lu, l'ADN stocké dans les nanobilles doit être extrait par un réactif chimique capable de dissoudre les silices tout en préservant l'ADN.



Technologies de stockage de l'ADN à température ambiante

Panneau de gauche : Nanoparticules de silice. L'ADN encapsulé dans des nanoparticules de silice est protégé des réactifs environnementaux potentiels (par exemple l'oxygène et l'eau) par une couche de silice imperméable, qui ne fait que quelques nanomètres d'épaisseur. Il en résulte une stabilité de l'ADN considérablement accrue, et les informations codées dans l'ADN ont une durée de vie prévue de plusieurs décennies à température ambiante.

Panneau de droite : Capsules de la société Imagene. L'extérieur de ces capsules est en acier inoxydable, l'intérieur en verre contient jusqu'à 0,8 g d'ADN. Une plaque (en rouge) peut recevoir 96 de ces capsules. Les informations codées dans l'ADN ont une durée de vie prévue de plusieurs dizaines de milliers d'années à température ambiante.

Crédits

- image de gauche : Robert Grass (ETH Zürich)

- image de droite : Sophie Tuffet (Imagene)

MUTAGENÈSE DIRIGÉE

Schématiquement, la mutagenèse dirigée *in vitro* fait souvent appel à la PCR, les mutations étant portées sur les brins d'ADN synthétique qui servent d'amorce à l'ADN-polymérase. La dernière méthode en date de mutagenèse dirigée *in vivo* est basée sur l'usage de CRISPR-Cas9[1] et ses dérivés, qui permet des modifications précises et plus rapides qu'auparavant, à moindre coût, même sur des génomes complexes comme ceux des mammifères. Il est constitué d'un «ARN guide», qui cible une séquence d'ADN particulière, associé à la protéine-enzyme «Cas9» qui, comme des «ciseaux moléculaires», coupe l'ADN au pointDans le

1 CRIPSR-Cas9 est un système naturel de défense des bactéries, qui garde mémoire de l'agression d'un virus. Les ARN codés par CRISPR se lient à la protéine-enzyme Cas9 qui peut alors couper l'ADN du virus afin de l'inactiver. Les travaux d'Emmanuelle Charpentier et Jennifer Doudna ont permis en 2012 de dériver de ce système naturel des outils très efficaces et précis de génie génétique : 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.

cadre de l'archivage de mégadonnées sur l'ADN, l'accès évolutif impliquerait plusieurs étapes, différentes selon que l'ADN est stocké *in vivo* ou *in vitro*. *In vitro,* il serait nécessaire de :

- récupérer l'ADN depuis son tube ou sa nanobille ;
- le modifier localement par mutagenèse dirigée *in vitro*, notamment la PCR mutagénique [2] ;
- vérifier par séquençage local que les mutations désirées ont bien

2 Comme décrit précédemment, la PCR permet d'amplifier un segment d'ADN délimité par deux courtes "amorces" synthétiques d'ADN simple-brin choisies par l'opérateur. À fin de mutagenèse dirigée, il suffit que l'une des amorces porte les mutations désirées dans la zone ciblée. À supposer que la PCR amplifie l'ADN d'un facteur 1 million, toutes ces copies sauf l'original (1 pour 1 million) porteront les mutations désirées.

Twist Bioscience[39](États-Unis) utilise des capsules de stockage physique conçues par la société Imagene[40] (France). L'extérieur de ces capsules est en acier inoxydable, l'intérieur en verre, et elles ont la taille d'une pile bouton. Chaque capsule contient jusqu'à 0,8 g d'ADN (potentiellement 1,4 Eo de données en tenant compte de la redondance). Elle est à usage unique. Son ouverture permet de récupérer l'ADN. La durée de conservation de l'ADN est estimée à plus de 50 000 ans dans ces systèmes de stockage qui protègent l'ADN de l'eau, 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 ou dans une nanobille. Les fragments d'ADN contenant des fichiers d'information moins volumineux seront regroupés dans une même capsule, où ils seront différenciés par leurs séquences-étiquettes.

Lecture de l'ADN

Réaction en chaîne de la polymérase (PCR)

Pour retranscrire l'information stockée sur l'ADN, il faut la séquencer. Une étape de multiplication de l'ADN par PCR (définie plus haut) est nécessaire pour que les technologies de séquençage puissent lire sa séquence de nucléotides.

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 disponibles dans l'échantillon, la lecture de l'ADN n'est pas destructrice.

Séquençage

Le séquençage de l'ADN consiste à déterminer l'enchaînement des nucléotides A, T, G et C au sein de ce polymère linéaire.

Principe de la technologie Illumina

La société Illumina, leader du marché du séquençage de l'ADN, commercialise des appareils pouvant séquencer jusqu'à 4 milliards de nucléotides par expérience (pour comparaison, le génome humain en contient de l'ordre de 3 milliards), soit l'équivalent de 1 Go ou d'un film. Cependant, cette longue séquence

39 https://www.twistbioscience.com/

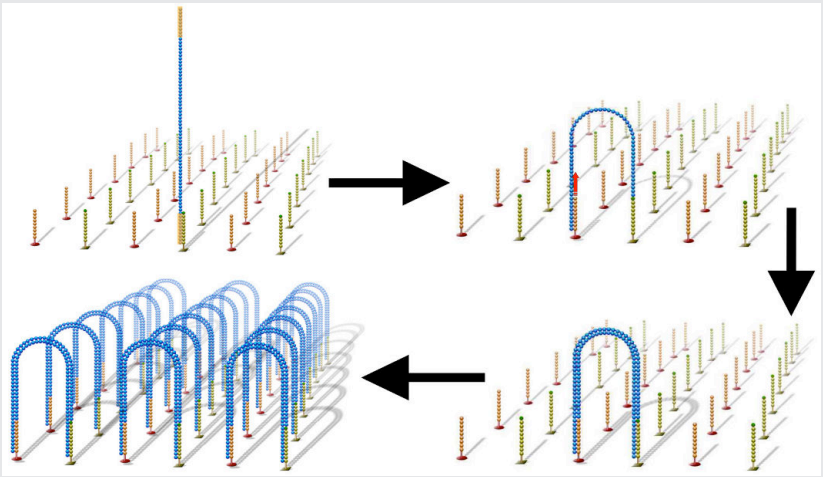
40 http://www.imagene.fr/

est en fait obtenue sous la forme d'un grand nombre de fragments dont la longueur est typiquement 300 nucléotides. Le taux d'erreur de ce séquençage est de 1 %. C'est la technologie qu'a utilisé Microsoft Corp. pour sa preuve de concept décrite plus haut portant sur 1 Go. Elle est décrite en détail dans l'encadré.

TECHNOLOGIE DE SÉQUENÇAGE PAR SYNTHÈSE

La première étape du séquençage consiste à préparer une bibliothèque de fragments d'ADN. Pour cela, l'ADN est extrait de l'échantillon puis fragmenté. Il est chauffé pour dissocier ses deux brins l'un de l'autre. Aux extrémités de ces fragments d'ADN sont ajoutés des «adaptateurs», courtes séquences connues d'ADN.

Le mélange en phase liquide des ADN simple-brin ainsi marqués par une combinatoire d'adaptateurs est alors mis en contact avec une lame de verre standardisée. Cette lame contient un milliard de dépôts d'ADN simple-brin longs de quelques dizaines de nucléotides, qui sont complémentaires des adaptateurs ajoutés lors de la préparation de la bibliothèque d'ADN. Les deux extrémités de l'ADN simple-brin se fixent à la lame de verre pour former un pont. Dans ce mélange est ajoutée une enzyme appelée ADN-polymérase qui va alors synthétiser le brin d'ADN complémentaire à ce pont, à raison d'un nucléotide par cycle. Les quatre nucléotides A, T, G et C, marqués par quatre fluorophores différents, sont ajoutés au mélange de manière séquentielle. Après l'incorporation d'un nucléotide par l'ADN-polymérase, la lame de verre est lavée de l'excédent de nucléotides. Ensuite une photographie numérique de la lame de verre est effectuée. Puis la fluorescence du nucléotide est enlevée par réaction chimique. Cela permet à un nouveau nucléotide marqué d'être incorporé au cycle suivant. Ce processus est effectué 300 fois de suite. Pour chaque photographie, la fluorescence est quantifiée en chaque point. L'ensemble des 300 photographies permet de reconstituer une séquence de 300 nucléotides successifs pour chaque fragment d'ADN ayant été ponté sur la lame de verre.

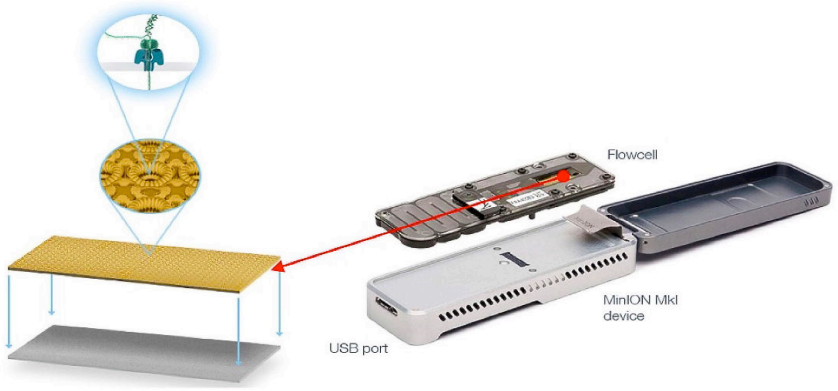


Principe de séquençage avec la technologie de la société Illumina

Crédit: Illumina

La dernière étape consiste à aligner informatiquement les séquences des différents fragments d'ADN en usant de leurs zones chevauchantes afin de les ordonner et de reconstituer la séquence globale. Cette phase demande une grande puissance de calcul et est particulièrement chronophage.

b) Principe de séquençage avec la technologie Oxford Nanopore.



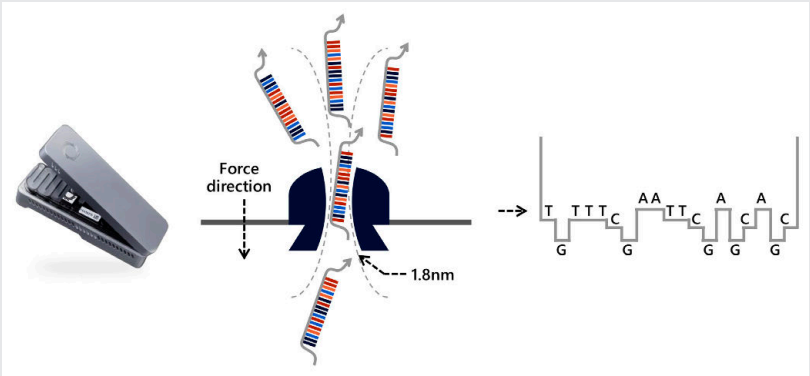
Le dispositif « MinIon » de Oxford Nanopore Technologies pèse 90 g et permet de lire plus de 10 Go de données en deux jours pour 700 US$.

Crédit : Oxford Nanopore Technologies

Les séquenceurs de troisième génération développés par Oxford Nanopore Technologies[41], font passer l'ADN ou autre polymère par des pores en enregistrant sa composition au passage. Ils permettent de lire de longues séquences d'ADN d'une traite (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. La méthode est détaillée dans l'encadré.

TECHNOLOGIE DE SÉQUENÇAGE PAR NANOPORES

Chaque dispositif de séquençage contient plusieurs centaines de protéines membranaires d'interface, appelées nanopores, insérées au travers d'une membrane synthétique grâce à un champ électrique. En appliquant une différence de potentiel de part et d'autre de la membrane, un flux ionique est mesuré en temps réel pour chaque acide nucléique qui traverse le nanopore. Le flux ionique est mesuré en pico-ampère ce qui représente une sensibilité de l'ordre de deux atomes d'hydrogène. Chaque nucléotide correspond ainsi à un signal électrique reproductible et spécifique, qui sera converti en séquence après analyse.



Principe de séquençage avec la technologie Oxford Nanopore.

Crédit : Oxford Nanopore Technologies.

41 https://nanoporetech.com/

De par son encombrement stérique, le nanopore ne séquence qu'une molécule à la fois. Sans frein moléculaire, l'ADN défilerait dans le nanopore à une vitesse d'un million de nucléotides par seconde. Afin de contrôler le flux à 450 nucléotides par seconde et de séquencer en temps réel, l'ADN natif de départ est lié à une protéine motrice séparant les deux brins d'ADN, et jouant le rôle de frein moléculaire pour l'un de ces brins à l'entrée du nanopore. Ainsi, chaque nanopore est capable de séquencer de façon contrôlée 450 nucléotides par seconde.

Dans le cas du dispositif «MinIon» de Oxford Nanopore Technologies, la membrane fonctionne environ 2 jours avant de devoir être remplacée pour un coût inférieur au millier d'euros. La partie électronique de cet instrument qui tient dans le creux de la main peut être reliée à un ordinateur par une interface USB3. Cependant, à plein régime, le débit de données venant en parallèle des centaines de nanopores est tel que des appareils électroniques de hautes performances sont nécessaires, par exemple des unités de traitement graphiques («Graphic Processing Units» ou «GPU»).

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.

Chapitre IV

Évolution et progrès des technologies requises pour archiver des données sur l'ADN

Défis

L'archivage d'information sur l'ADN est expérimental. Avant de devenir viable à grande échelle, il doit être complètement automatisé. De plus, les processus de lecture et d'écriture de l'ADN doivent être améliorés en termes de vitesse et de coût. Actuellement le coût et le temps requis pour stocker 1 Go (10^9 octets) de données sur l'ADN, est comparable à ceux pour 1 Po (10^{15} octets) de données sur un support informatique. Selon l'entreprise Illumina, ce coût doit être divisé par un facteur 10 000 avant que l'approche ADN puisse être largement adoptée. En outre, il est nécessaire de mettre en œuvre un système d'architecture du stockage. L'objectif est d'accéder à une portion d'information choisie et d'effectuer des tâches informatiques directement sur les données moléculaires de cette portion.

Dans ce chapitre, nous aborderons en détail les limites des technologies utilisées dans le stockage d'information sur l'ADN, et les progrès réalisés par les scientifiques, incluant les organismes des secteurs public et privé.

Amélioration des technologies d'écriture de l'ADN

Les facteurs limitants dans la synthèse d'ADN

Le stockage d'information sur l'ADN requiert une synthèse d'ADN précise et à grande échelle. Les approches actuelles ne sont pas assez performantes car :

- il est difficile de synthétiser un fragment d'ADN long, supérieur à 150 nucléotides. Cette limite résulte du taux d'erreur lors de la synthèse d'ADN, qui est typiquement 0,5 % par nucléotide ;
- certaines séquences d'ADN sont difficiles à synthétiser (répétition d'un même *nucléotide* plusieurs fois, séquence riche en C et G) ;
- la synthèse d'ADN est longue : les systèmes actuels synthétisent l'ADN nucléotide par nucléotide au rythme de 30 secondes/nucléotide. En outre, il faut parfois plusieurs semaines à mois pour assembler un long morceau d'ADN avec une qualité raisonnable, à partir des fragments bruts de 100-200 nucléotides.
- la synthèse d'ADN est coûteuse (environ 8 centimes d'euros le nucléotide en 2020). Il serait nécessaire de réduire le coût de la synthèse d'ADN d'un facteur 10^8 (100 millions) ;
- la synthèse d'ADN n'est pas « démocratisée ». Elle n'est généralement pas effectuée par les laboratoires, mais par des plates-formes spécialisées. Le scientifique commande les séquences d'ADN. Les plates-formes fournissent les fragments d'ADN après un délai de quelques jours pour des séquences de 25 nucléotides à quelques semaines pour des séquences supérieures à 2 000 nucléotides ;
- la synthèse d'ADN par voie chimique est polluante. Le processus utilise, entre autres, l'acétonitrile,

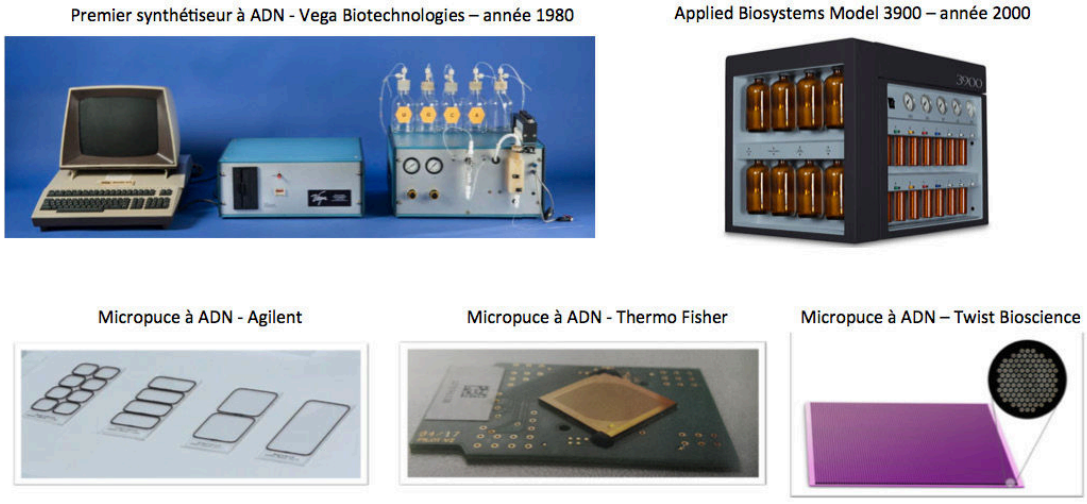
un solvant chimique nocif. Il représente 72 % du volume réactionnel de synthèse d'un fragment d'ADN. Dans le cas du stockage d'information numérique sur l'ADN, il faudrait potentiellement plusieurs milliards de litres d'acétonitrile pour synthétiser l'ADN contenant la SGD.

Histoire et progrès dans la synthèse d'ADN par voie chimique

Histoire et évolution

Depuis la découverte de la structure de l'ADN en 1953 par J. Watson et F. Crick, les chimistes ont voulu synthétiser de l'ADN. En 1965, le premier fragment d'ADN est synthétisé chimiquement. En 1970, les scientifiques synthétisent pour la première fois un gène de 70 nucléotides. En 1983, ils inventent les phosphoramidites, réactifs puissants et stables permettant de synthétiser plus facilement des fragments d'ADN plus longs.

Depuis, 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. Ils ont évolué afin de posséder jusqu'à 200 colonnes, permettant de synthétiser simultanément 200 fragments d'ADN différents. Agilent, Twist Bioscience et Thermo Fisher ont miniaturisé et massivement parallélisé ce processus afin d'en réduire le coût. Ces compagnies effectuent la synthèse d'ADN sur des micropuces. Agilent est capable de synthétiser simultanément 244.000 fragments d'ADN. Thermo Fisher synthétise 35.000 fragments d'ADN pour une puce de 100 mm^2. Twist Bioscience a miniaturisé davantage ce processus de synthèse.



Évolution des dispositifs de synthèse d'ADN par voie chimique

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

Progrès dans la synthèse chimique : Twist Bioscience

La société Twist Bioscience (États-Unis) est une des entreprises leader en synthèse d'ADN par voie chimique dans le monde. Elle a conçu une puce permettant de synthétiser simultanément un million de fragments d'ADN, éventuellement différents, de 200 nucléotides en vingt-quatre heures. Cette puce est composée d'une plaque de silicium et possède 10 000 puits. Dans chaque puits, 100 fragments différents de 200 nucléotides sont synthétisés. Le prix d'une puce est de 5 000 US$. Twist Bioscience souhaite miniaturiser davantage le processus de synthèse. Actuellement chacun des 10 000 puits qui composent la puce mesure 50 microns. En réduisant la taille des puits à 0,3 micron, Twist Bioscience pourrait dans un proche avenir synthétiser beaucoup plus de fragments simultanément et donc réduire le coût de synthèse.

Le taux d'erreur lors de la synthèse chimique d'ADN est 0,5 % chez de nombreux fournisseurs. Plus la séquence est longue, plus les probabilités de contenir des erreurs sont élevées, limitant la longueur de séquences viables à 150 nucléotides environ. Twist Bioscience a réduit le taux d'erreur à 0,1 %, permettant ainsi la synthèse de fragments de 200 nucléotides sans erreur.

Progrès dans la synthèse d'ADN par voie enzymatique

Depuis 2010, des techniques de synthèse d'ADN par voie enzymatique ont émergé (voir le principe chaoitre III). Le processus de la synthèse enzymatique est plus simple, plus rapide et plus efficace que la synthèse chimique. Les extrapolations réalisées par les sociétés impliquées suggèrent qu'il sera possible, d'ici quelques années, de synthétiser des fragments jusqu'à 2 000 nucléotides d'un seul tenant grâce à la synthèse enzymatique. Dans une cellule vivante, la lecture/écriture de l'ADN est plus rapide qu'une mémoire Flash (moins de 100 microseconde par bit). Ceci nous donne une idée du potentiel de l'approche ADN, quoiqu'en règle générale les processus du vivant sont moins efficaces lorsqu'ils sont portés hors de la cellule comme ici. De plus, la synthèse enzymatique ne nécessite pas l'utilisation de produits chimiques dangereux, ce qui en diminue l'impact environnemental. Actuellement six entreprises travaillent sur la synthèse d'ADN enzymatique : DNA Script, Nuclera Nucleics, Molecular Assemblies, Merck, Ansa Biotechnologies et Spindle Biotech.

DNA Script

DNA Script[42] (France) est actuellement le leader de la synthèse d'ADN par voie enzymatique à l'échelle mondiale. L'entreprise a pour objectif de révolutionner l'écriture de l'ADN grâce à une technologie de synthèse enzymatique plus rapide, de meilleure qualité, plus efficace et à terme moins coûteuse que les technologies actuelles. DNA Script développe deux technologies essentielles de la synthèse d'ADN par voie enzymatique : les nucléotides terminateurs réversibles et les enzymes ADN-polymérases. L'entreprise parvient actuellement en conditions de laboratoire à synthétiser des fragments d'ADN de plus de 250 nucléotides, et le taux d'erreur avoisine 0,7 %. De nouvelles optimisations sont en cours afin de diminuer le taux d'erreur et d'augmenter la longueur des fragments d'ADN synthétisés. D'ici 2024, la société envisage de pouvoir synthétiser et séquencer 1 To (quatre mille milliards de nucléotides) de données en vingt-quatre heures.

42 https://www.dnascript.com

Un des objectifs de la société est de « démocratiser » la synthèse d'ADN dans les laboratoires, afin qu'elle ne soit plus exclusivement effectuée par des plates-formes spécialisées. Pour cela, DNA Script conçoit une machine de paillasse capable de synthétiser de l'ADN. Le but est de permettre à un chercheur d'obtenir en quelques heures le fragment d'ADN désiré, synthétisé dans son laboratoire, et donc d'éviter les délais de commande et de livraison. D'ici dix ans, DNA Script envisage de commercialiser 10 000 machines. Elles seront utilisées dans la recherche, la médecine spécialisée, et les méthodes de traitements par thérapie génique personnalisée. Selon l'entreprise, il est trop tôt encore pour fixer un prix de vente de la technologie, mais la machine pourrait être commercialisée pour environ 50 000 euros.

Nuclera Nucleics

Nuclera Nucleics[43] (Royaume-Uni) développe trois technologies : les enzymes ADN-polymérases, les nucléotides terminateurs réversibles[44] et un système d'automatisation.

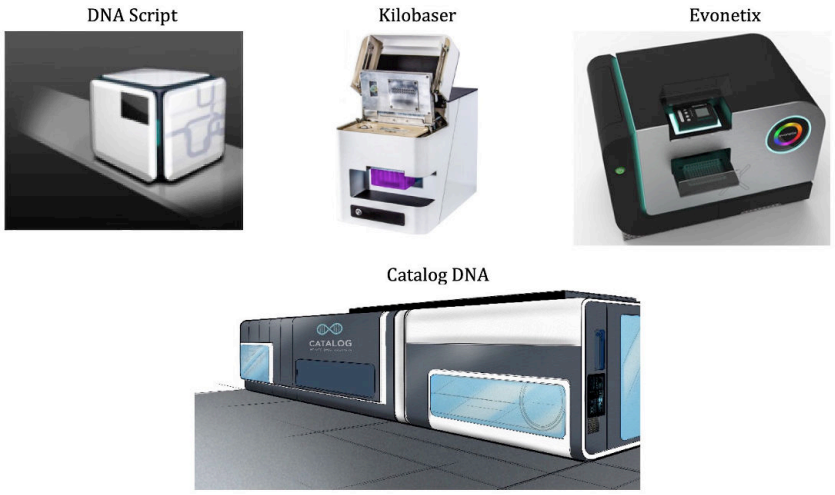
Quatre générations d'enzymes ont été conçues afin d'optimiser leur capacité à incorporer les nucléotides. En outre, Nuclera Nucleics conçoit des nucléotides terminateurs compatibles avec l'enzyme qui synthétise le brin d'ADN. Ces nucléotides possèdent des groupements chimiques protecteurs bloquant toute interaction, afin d'étendre la séquence par un seul nucléotide à la fois. Ces groupements chimiques protègent également le nucléotide de modifications chimiques et évite l'interaction non désirée d'un nucléotide avec un autre, et donc la formation de structures secondaires[45] dans l'ADN. L'entreprise développe un système d'automatisation du processus enzymatique. Il s'agit d'un système électronique microfluidique qui contrôle le mouvement de gouttes de liquide.

La machine, en cours de développement, a pour objectif de synthétiser des morceaux d'ADN longs de 10 000 nucléotides d'un seul tenant, rapidement, sans erreur et à un coût raisonnable.

43 https://www.nuclera.com/

44 Comme discuté au chapitre III, les nucléotides utilisés sont protégés par un groupe chimique terminal bloquant tout ajout d'un second nucléotide, afin d'étendre la séquence d'un seul nucléotide à la fois. Ici, le groupe terminal est différent de celui utilisé en synthèse par voie chimique.

45 Conformation de l'ADN obtenue par les interactions entre les nucléotides.



Dispositifs de synthèse d'ADN en cours de conception

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

Nouvelles technologies de synthèse d'ADN

Kilobaser

Kilobaser[46] (Autriche) conçoit une machine de paillasse capable de synthétiser un fragment d'ADN d'une centaine de nucléotides par expérience. L'usage de cette machine est réservé à la synthèse d'amorces de PCR. Le but est de permettre à un chercheur d'obtenir en quelques heures l'amorce désirée, synthétisée par voie chimique dans son laboratoire, et donc d'éviter les délais de commande et de livraison.

Pour chaque nucléotide ajouté, la probabilité d'erreur est d'environ 2 %, limitant la longueur de séquence du fragment à 100 nucléotides. Le temps de synthèse du fragment d'ADN est de deux minutes par nucléotide ajouté et vingt minutes de traitement final, soit 1 heure environ pour la synthèse d'un fragment de vingt nucléotides.

Kilobaser commercialise la machine depuis cette année 2020, pour environ 15 000 euros. Le prix des consommables, incluant la puce microfluidique et le réservoir, constituerait une dépense d'environ 450 euros par mois. L'objectif est d'obtenir un prix de 75 centimes d'euros par nucléotide synthétisé. Ce prix est nettement supérieur à celui des plates-formes de synthèse (environ 8 centimes en 2019), mais, avec le système Kilobaser, le client évite le délai de commande et de livraison. Ces faits conduisent à estimer que la machine Kilobaser ne contribuera pas directement à synthétiser les grandes quantités d'ADN nécessaires au stockage d'informations numériques. Mais elle pourrait faire partie de l'arsenal d'une unité de stockage, en particulier faisant face à une grande diversité de travaux d'amplification de l'ADN nécessitant des amorces nombreuses et variées.

46 https://www.kilobaser.com/

Helixworks

Helixworks[47] (Irlande) conçoit, fabrique et vend des technologies de stockage de données basées sur l'ADN. Ils développent la technologie MoSS (Molecular Storage System) dont une étape précoce permet de convertir les fichiers numériques en séquences d'ADN.

Helixworks fabrique son propre ADN en utilisant une technologie exclusive par voie enzymatique. Actuellement, la société annonce être capable de synthétiser des fragments d'ADN d'une longueur supérieure à 2 000 nucléotides d'un seul trait. Cela dépasserait largement les autres techniques de synthèse qui ne peuvent synthétiser que 200 nucléotides en moyenne. L'entreprise synthétise également de l'ADN possédant des nucléotides modifiés par l'addition de groupements chimiques, par exemple des nucléotides acétylés.

Helixworks utilise la technologie de séquençage d'Oxford Nanopore Technologies, décrite plus haut. L'ADN synthétisé par l'entreprise est lu de manière optimale par les séquenceurs à nanopore : sa vitesse de lecture est augmentée, le signal pour chaque nucléotide séquencé est plus fort et plus précis. Cela réduit les erreurs lors de son séquençage, en comparaison avec un fragment d'ADN synthétisé par les plate-formes concurrentes.

Catalog DNA

Catalog DNA[48] (États-Unis) a pour objectif de faire de l'ADN un support de stockage d'information numérique. Pour cela, les chercheurs ont conçu une technologie plus rapide et moins coûteuse que les technologies actuelles, pour générer des fragments d'ADN à des fins de stockage de données.

Les systèmes actuels synthétisent l'ADN nucléotide par nucléotide au rythme de 30 secondes/nucléotide. Au lieu de synthétiser les nucléotides un par un, l'entreprise prépare par avance une bibliothèque ordonnée de petits fragments d'ADN déjà synthétisés et vérifiés (appelés ***composants***). Ces composants sont ensuite combinés pour former des molécules d'ADN plus longues. Ainsi, le processus nécessite moins de synthèse d'ADN, ce qui représente la partie la plus coûteuse et la plus lente du travail. C'est la disposition de ces composants qui détermine leur signification. Cela ressemble à une langue. Par exemple, en français, il n'y a que 26 lettres, mais grâce à divers arrangements, nous pouvons créer un très grand nombre de mots.

Chaque composant assemblé par Catalog possède deux extrémités simple-brin, leur permettant de s'assembler les uns aux autres, comme les briques d'un jeu de construction. À chaque cycle, un nouveau composant est ajouté au composant précédent grâce à un processus enzymatique.

Catalog a conçu une machine capable d'assembler les fragments d'ADN de manière programmable et automatisée, afin de générer des molécules d'ADN uniques. Cette machine génère 500 000 réactions d'assemblage d'ADN par seconde. Elle contient un incubateur qui maintient les conditions idéales pour l'assemblage des fragments d'ADN par les enzymes. Les réactions sont effectuées sur une longue bande mobile, qui se déplace entre les compartiments à la vitesse de 16 mètres par minute. Grâce à cette machine, Catalog synthétise de l'ADN codant pour 500 Ko d'information par seconde. En 2019, ils ont codé la version anglophone de Wikipedia sur l'ADN, soit 16 Go.

---

47 https://helix.works/

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

Evonetix Ltd.

Evonetix Ltd[49] (Royaume-Uni) développe un système de synthèse d'ADN innovant, combinant des puces en silicium et une régulation thermique finement localisée. Elle conçoit une machine de paillasse capable de synthétiser en une seule expérience plusieurs milliers de nucléotides d'ADN. Une amélioration prometteuse portée par Evonetix permet de réduire le taux d'erreur d'un facteur 100 à 1 000, en contrôlant thermiquement chacun des 10 000 micro-sites de réaction. La technologie sous-jacente est résumée dans l'encadré.

Le dispositif d'Evonetix est composé de la machine, d'une puce MEMS (microsystème électromécanique) en silicium et de réactifs thermolabiles. L'ensemble est coordonné par des logiciels informatiques développés par l'entreprise. La machine et les technologies sont en cours de développement. Selon Evonetix, il est encore trop tôt pour fixer un prix de vente de la technologie, mais la machine pourrait être commercialisée pour environ 10 000 euros.

SYNTHÈSE D'ADN PAR CONTRÔLE THERMIQUE

Le principe suivi par Evonetix Ltd. consiste à synthétiser des courts fragments d'ADN simple-brin dans chaque puits de la puce de silicium avec les réactifs thermolabiles, grâce à un contrôle précis de la température. En outre, le système prévoit d'assembler les courts fragments d'ADN simple-brin pour construire des fragments d'ADN plus long. Le dispositif inclut également un système de correction d'erreur lors de la synthèse d'ADN.

Comme pour la synthèse chimique classique, les fragments d'ADN différents sont construits nucléotide après nucléotide, chacun dans un puits de la puce. Quatre solutions contenant les nucléotides A, T, G, C possédant un groupement protecteur thermolabile (différent pour chaque type de nucléotide) circulent successivement sur la puce. Le changement de température dans un puits donné de la puce induit la déprotection du fragment d'ADN en cours de synthèse dans ce puits. Le fragment d'ADN, maintenant déprotégé, incorpore le nucléotide thermolabile de la solution circulante. À l'inverse, si le changement de température n'est pas suffisant pour induire la déprotection du fragment d'ADN, alors il ne pourra pas incorporer le nucléotide de la solution circulante. Ce cycle est répété jusqu'à obtenir la séquence désirée. Ce contrôle très précis de la température indépendamment pour chaque puits dans la puce est donc un élément essentiel de cette approche.

Les fragments d'ADN synthétisés sont ensuite assemblés deux à deux pour obtenir un fragment d'ADN plus long. Une fois les fragments d'ADN synthétisés, ils sont décrochés de la puce en clivant la liaison chimique thermolabile du premier nucléotide de la chaîne. Il existe plusieurs liaisons de diverses thermolabilités, permettant de contrôler par la température quel fragment d'ADN libérer. Le fragment d'ADN clivé est transporté dans la puce vers un deuxième fragment d'ADN via des pièges diélectrophorétiques. Les deux fragments d'ADN simple-brin s'assemblent par homologie d'une partie de leur séquence. Cette séquence homologue est différente pour chaque paire de fragments d'ADN, et détermine la température de fusion à laquelle les fragments d'ADN seront assemblés deux à deux. Une fois les deux fragments d'ADN assemblés, le processus se répète pour former un fragment d'ADN encore plus long.

Le processus d'assemblage inclut un système de correction d'erreur. En effet, si un fragment d'ADN simple-brin possède une mutation dans sa séquence, la température de fusion au deuxième fragment d'ADN simple-brin change. La précision du système de contrôle de température est telle qu'il sera impossible pour le fragment d'ADN muté de l'assembler à cet autre fragment d'ADN. Il sera donc éliminé du processus. De ce fait, le taux d'erreur anticipé par Evonetix se situe 2 à 3 ordres de grandeur en-dessous de celui (0,5 %) obtenu dans le procédé de synthèse chimique classique.

49 https://www.evonetix.com

Conclusion

Plusieurs entreprises travaillent sur les technologies de synthèse d'ADN afin d'en améliorer les performances (automatisation, parallélisation, coût, taux d'erreur, vélocité, longueur des fragments d'ADN). Des progrès conséquents ont été réalisés rapidement.

Twist Bioscience, une des entreprises leader en synthèse chimique, parallélise massivement cette méthode de synthèse pour en réduire le coût. DNA Script et Nuclera Nucleics, leaders de la synthèse enzymatique, optimisent ce processus afin de rapidement générer des longs fragments d'ADN en une seule étape. D'ici quatre ans, DNA Script envisage de pouvoir synthétiser en vingt-quatre heures 1 To d'information numérique, soit l'équivalent de 1 000 films. À noter, les technologies de Twist Bioscience et de Evonetix sont compatibles avec la voie enzymatique de synthèse, même si elles font appel aujourd'hui à la voie chimique.

Des systèmes de synthèse d'ADN alternatifs très prometteurs sont également en cours de développement. Le dispositif d'Evonetix, basé sur l'assemblage par contrôle de température est un des systèmes les plus aboutis et les moins chers du marché. Il permet de réduire le taux d'erreur d'un facteur 100 à 1 000, en contrôlant thermiquement chacun des 10 000 micro-sites de réaction. Les entreprises Helixworks et Catalog DNA sont les seules à avoir optimisé leur technologie de synthèse d'ADN à des fins de stockage d'information sur l'ADN uniquement. Leur technique consiste à assembler combinatoirement des fragments d'ADN pré-synthétisés et donc vérifiés, s'affranchissant ainsi partiellement des limites exposées ci-dessus. Notons que la machine de Catalog, capable de synthétiser 0,5 Mo d'information par seconde, représente un atout majeur pour le stockage d'information numérique sur l'ADN.

Optimisation du codage de l'information numérique en ADN et densification de l'information

Défis

Rappelons que le principe général du stockage d'information numérique sur l'ADN consiste à attribuer des valeurs numériques aux quatre nucléotides A, T, G et C, puis à synthétiser ces nucléotides sous forme de fragment. Chaque suite de quatre nucléotides représente un octet, soit en première approximation une lettre, un chiffre ou autre symbole. La synthèse d'ADN représente la partie la plus lente et la plus coûteuse du stockage d'information sur l'ADN. Afin d'en diminuer le coût et d'en augmenter la vitesse, les scientifiques travaillent sur des nouvelles méthodes de densification de l'information. Le but consiste à réduire la quantité d'ADN nécessaire pour stocker une même information et ainsi, synthétiser moins de nucléotides par bits de données.

Projet MOSLA

Les chercheurs du projet MOSLA (*Molecular Storage for Long-term Archiving*)[50] sont affiliés aux universités de Marburg, Darmstadt et Giessen, en Allemagne. L'un de leurs sous-projets vise à densifier l'information portée par l'ADN, en utilisant deux approches. Premièrement, ils tentent d'établir une méthode de « compression » de l'information numérique convertie en nucléotides. Deuxièmement, ils conçoivent un nouvel

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

alphabet de nucléotides, dont le principe consiste à différencier les nucléotides modifiés par l'addition de groupements chimiques, des nucléotides non modifiés (A, T, G, C). Différents types de modifications chimiques peuvent être présents sur les nucléotides, comme la méthylation. Ces modifications peuvent être détectées par les séquenceurs de troisième génération utilisant les technologies Oxford Nanopore ou Pacific Biosciences.

Projet du Technion et d'IDC Herzliya

L'équipe de Zohar Yakhini[51] (Technion, IDC Herzliya – Israël) tente aussi de densifier l'information portée par l'ADN. Elle propose d'appliquer un principe différent, basé sur un alphabet élargi, tout en ne recourant qu'aux quatre nucléotides standard A, T, G, C. Pour élargir l'alphabet, elle fait appel à la notion de nucléotide composite : par exemple, une position donnée sur l'ADN sera occupée dans la population de molécules d'ADN par un mélange d'un tiers de A et deux tiers de C. Le principe de ce nouvel alphabet de nucléotides repose sur les constats suivants :

- lors de l'écriture d'un fragment d'ADN, ce n'est pas un fragment d'ADN unique qui est synthétisé, mais une population de fragments d'ADN identiques ;
- de même, lors de la lecture d'un fragment d'ADN, ce n'est pas un fragment d'ADN unique qui est séquencé, mais une population de fragments d'ADN identiques.

Cette équipe tire avantage de cette nécessaire multiplicité pour concevoir de nouvelles lettres. En effet, en plus de disposer des quatre nucléotides standards A, T, G, C, les scientifiques disposent également de nucléotides mixtes composés par exemple à 50 % de A et 50 % de C. Dans cet exemple, la moitié de la population du fragment d'ADN synthétisé possédera le nucléotide A, à un endroit défini de sa séquence, et l'autre moitié possèdera le nucléotide C au même endroit. L'alphabet de nucléotides disponibles est alors de dix lettres[52]. Ce processus peut être étendu pour créer un alphabet de nucléotides mixtes encore plus riche. Par exemple, la lettre composée à 33 % de A, 33 % de T et 33 % de C peut être créée. Dans cet exemple, un tiers de la population du fragment d'ADN synthétisé possédera, à un endroit défini de sa séquence, le nucléotide A, un autre tiers le nucléotide T et le dernier tiers le nucléotide C. Cependant, l'imperfection des méthodes de séquençage comme de synthèse de l'ADN limite en pratique le nombre de combinaisons possibles qui pourront être distinguées.

En augmentant la taille de l'alphabet de nucléotides, les scientifiques doivent synthétiser moins de nucléotides par octet d'information numérique. Cependant, ils doivent séquencer un plus grand nombre de copies du même fragment d'ADN afin d'augmenter la redondance du séquençage et de décoder l'information avec 100 % de réussite. Le coût de la synthèse d'un nucléotide d'ADN est estimé 5 000 fois plus élevé que le coût du séquençage d'un nucléotide d'ADN. Ainsi, en utilisant un alphabet plus diversifié, le coût global de l'opération est nettement réduit puisqu'il implique moins d'écriture pour plus de lecture.

Nucléotides non conventionnels

Certains scientifiques conçoivent de nouveaux nucléotides, chimiquement différents des nucléotides naturels A, T, G, C. Pour cela, ils modifient le sucre ou les bases azotées des nucléotides. Pour rappel, un

51 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/

52 Outre A, T, G, C, sont aussi des lettres les mélanges à parts égales suivants : AT, AG, AC, TG, TC, GC.

nucléotide est composé toujours du même sucre à cinq carbones, le désoxyribose, d'un groupe phosphate, et d'une base azotée responsable de l'identité du nucléotide A, T, G, ou C.

Modifications du sucre des nucléotides

L'équipe de Piet Herdewijn (université catholique de Louvain - Belgique) a synthétisé de nouveaux nucléotides en modifiant artificiellement leur sucre. Ces acides nucléiques non conventionnels sont appelés AXN (acides xéno-nucléiques)[53]. Les AXN ont été conçus pour éviter toute hybridation avec l'ADN naturel, tout en étant incorporés et tolérés par un organisme vivant. L'objectif est de construire des cellules qui stockeraient tout ou partie de leur information génétique dans un polymère informationnel alternatif, sans risque de recombinaison avec les génomes originels.

Ces AXN pourraient être utilisés pour le stockage d'information numérique sur l'ADN :

- *in vitro* : afin d'augmenter la densité informationnelle avec de nouveaux nucléotides ;
- *in vivo* : afin de stocker l'information numérique sous forme de polymère artificiel, dans des cellules ou des organismes vivants, sans risque d'interaction de cette information avec l'ADN naturel de l'organisme.

Modifications du groupement chimique des nucléotides

L'équipe de Steven Benner (université de Floride – États-Unis) a synthétisé huit nouveaux nucléotides, nommés S, B, J, V, K, X, Z et P, en modifiant les groupements chimiques des nucléotides naturels. L'objectif est d'éviter les problèmes liés aux « imperfections de l'ADN »[54]. En effet, l'ADN naturel possède quelques caractéristiques chimiques qui peuvent compromettre sa capacité à fonctionner comme une molécule de stockage d'information numérique :

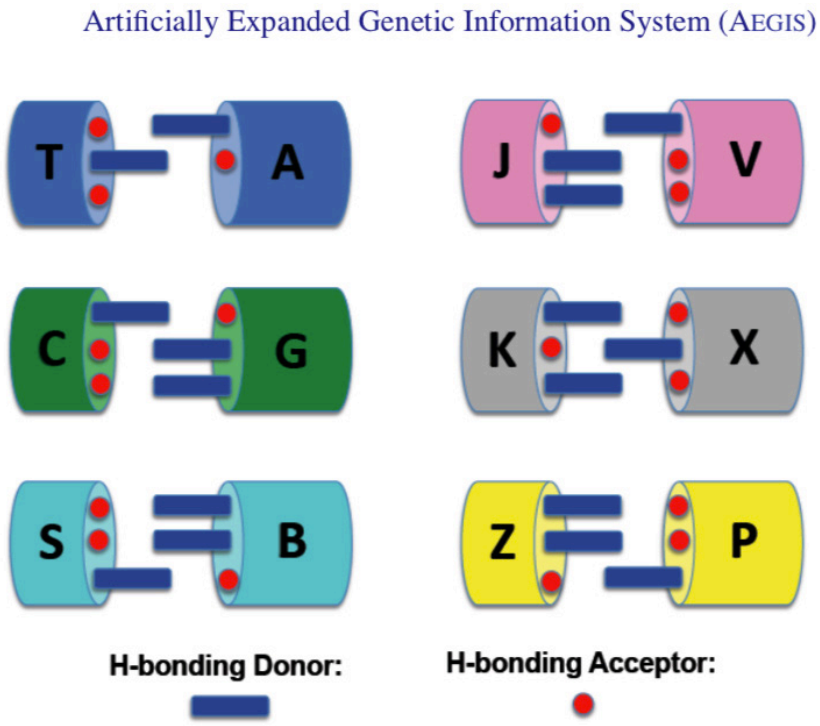
- les réactions chimiques peuvent modifier ses bases azotées donc son contenu informationnel, ou modifier son sucre ;
- l'ADN ne possède que quatre nucléotides différents, limitant la densité informationnelle comme discuté ci-dessus ;
- la liaison entre A et T est fondée sur deux liaisons hydrogène, donc plus faible que celle entre G et C fondée sur trois liaisons hydrogène. Cela peut créer des problèmes d'hybridation, responsables d'artefacts et d'erreurs lors de l'amplification de l'ADN (étape obligatoire pour le stockage d'information numérique sur l'ADN) ;
- les fragments d'ADN riches en G sont très difficiles à synthétiser. En effet, les G peuvent s'apparier entre eux, formant une structure tri-dimensionnelle qui obère la synthèse des fragments d'ADN.

Ces nouveaux nucléotides ont été conçus de telle manière que S s'apparie avec B, J avec V, K avec X, Z avec P. Ces paires sont toutes liées par trois liaisons hydrogènes. En outre, les scientifiques ont chimiquement modifié A pour s'apparier à T avec trois liaisons hydrogènes au lieu de deux. Ils ont également modifié G pour ne plus former de structure tri-dimensionnelle avec d'autres G. Cela facilite la synthèse des séquences riches en G.

---

53 Chaput JC, Herdewijn P, Hollenstein M. Orthogonal Genetic Systems (2019). *Chembiochem* 2019 Dec 30, in press.

54 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.



Appariements de nucléotides conventionnels (TA, CG), et non-conventionnels conçus par l'équipe de Steven Benner (SB, JV, KX, ZP).

Crédit : Steven Benner

Cet « ADN rectifié » présente de nombreux avantages pour le stockage d'information numérique :

- les nucléotides modifiés facilitent la synthèse de fragments d'ADN riches en G. Les chercheurs encodent donc les bits en nucléotides sans se soucier de la difficulté à synthétiser ce fragment d'ADN ;
- les huit nouveaux nucléotides permettent de convertir les données numériques de manière plus dense ;
- l'étape de PCR qui permet de multiplier et copier l'ADN contenant l'information est plus robuste. Cela diminue le risque d'erreurs et préserve l'information.

Conclusion

Certains chercheurs proposent de densifier l'information, grâce à des alphabets étendus, afin de réduire la quantité d'ADN nécessaire pour stocker une même information.

Le projet porté par Zohar Yakhini consiste à utiliser les quatre nucléotides standards, ainsi qu'un mélange de ces nucléotides pour coder l'information de façon plus dense. Un avantage majeur de cette méthode réside dans le fait qu'elle fait appel exclusivement à des nucléotides standard, donc à des méthodes de lecture et écriture largement éprouvées. Cependant, de nombreuses copies de cet ADN doivent être séquencées pour décoder l'information qu'il contient sans erreur.

Le projet MOSLA utilise des nucléotides modifiés par l'addition de groupements chimiques. Actuellement, seules les technologies de séquençage de troisième génération (exemple : le dispositif de Oxford Nanopore Technologies) sont capables de lire l'information utilisant ces nucléotides.

Enfin, certains chercheurs comme Steven Benner ou Piet Herdewijn conçoivent des nucléotides non conventionnels. Ces nucléotides ont été volontairement modifiés pour divers objectifs, et peuvent

en particulier faciliter les processus d'écriture et de stockage d'information. Cependant, les processus de séquençage actuels limitent la lecture de ces nucléotides non-conventionnels. Des progrès doivent encore être effectués afin de démocratiser leur utilisation.

Amélioration des technologies de lecture de l'ADN

Facteurs limitants dans le séquençage de l'ADN

Le stockage d'information numérique sur l'ADN requiert une lecture de l'ADN précise et à grande échelle. Les approches actuelles sont performantes, mais des progrès doivent encore être réalisés car :

- la lecture d'ADN est coûteuse. Selon l'entreprise Illumina, il serait nécessaire de réduire le coût de lecture de l'ADN d'un facteur 1 000 ;
- le taux d'erreur lors de la lecture d'une séquence d'ADN est trop élevé. Selon Microsoft Corp., la majorité des erreurs de codage/décodage de l'information lors du processus de stockage d'information numérique sur l'ADN proviennent du séquençage ;
- la lecture de l'ADN n'est pas assez rapide. Dans le cadre du stockage d'information sur l'ADN, il serait nécessaire d'avoir une lecture de l'information quasi-instantanée ;
- toutes les technologies de séquençage ne permettent pas la lecture de nucléotides conventionnels modifiés et de nucléotides non-conventionnels ;
- la lecture d'ADN n'est pas encore assez « démocratisée ». Elle n'est généralement pas effectuée par les laboratoires, mais par des plates-formes spécialisées.

Histoire et progrès des technologies de séquençage

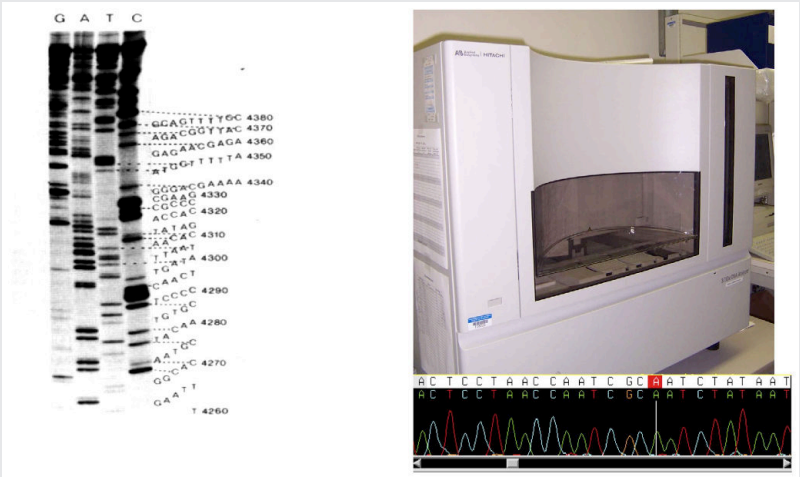
Histoire

Les progrès réalisés en génie génétique dans les années 1970 ont conduit deux à trois décennies plus tard les chercheurs à séquencer tout le génome d'un organisme. On distingue trois générations de séquençage, qui sont résumées dans l'encadré.

BREF HISTORIQUE DE LA TECHNOLOGIE DE LECTURE DE L'ADN

a) Séquençage de 1ère génération

Deux méthodes de séquençage d'ADN ont été développées en 1977, l'une par l'équipe de Frederick Sanger et l'autre par l'équipe de Walter Gilbert. Ces deux méthodes utilisent la fonction de réplication de l'ADN. La méthode de Sanger a pris le pas sur l'autre dans les années qui ont suivi.



Séquençage de 1ère génération

Sources (de gauche à droite) : Nucleotide sequence of Phi174 DNA. F. Sanger et al. Nature ***265,*** *687–695 (1977) - https://doi.org/10.1038/265687a0*

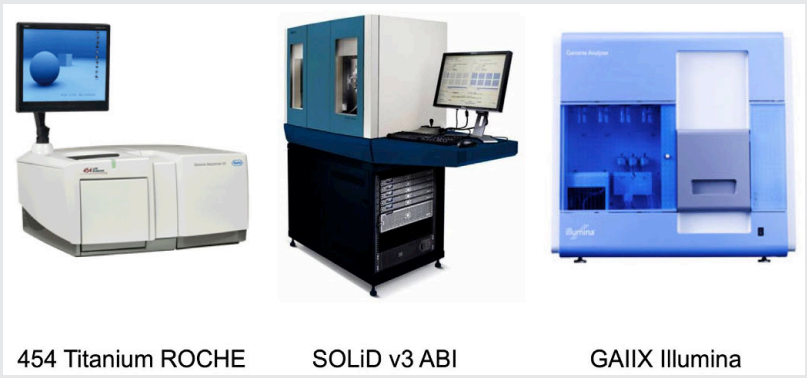
Applied Biosystem (Thermo Fisher Scientific) et Philippe Glaser pour l'électrophorégramme.

Le principe de Sanger consiste à répliquer le brin complémentaire de l'ADN avec une ADN-polymérase. Les quatre désoxyribonucléotides sont ajoutés à la réaction (dATP, dCTP, dGTP, dTTP), ainsi qu'une faible concentration de l'un des quatre di-désoxyribonucléotides (ddATP, ddCTP, ddGTP, ddTTP). Ces di-désoxyribonucléotides sont analogues aux désoxyribonucléotides, mais une petite différence chimique en fait des terminateurs de chaîne ; s'ils se trouvent incorporés, avec une faible probabilité dictée par leur basse concentration, ils empêchent la poursuite de l'élongation du brin d'ADN. Ainsi, des arrêts d'élongation s'étagent là où par exemple un ddATP a été incorporé à la place d'un dATP. Pour le séquençage complet d'un même fragment, la réaction est répétée quatre fois, avec les quatre di-désoxyribonucléotides. Les fragments d'ADN synthétisés sont analysés par électrophorèse de chacune des quatre réactions sur gel d'acrylamide, qui sépare les fragments par leur taille, permettant ainsi de lire la séquence. La détection des fragments synthétisés se fait en incorporant un traceur radioactif, qui sur le gel est détecté par un film sensible.

De nombreux gènes ont ainsi été séquencés. Pendant vingt ans, la technologie de séquençage Sanger a été optimisée : automatisation et robotisation, introduction de traceurs fluorescents remplaçant les marqueurs radioactifs, utilisation de l'électrophorèse en capillaire plutôt qu'entre deux plaques. Cependant, cette méthode reste peu performante pour le séquençage des génomes entiers car les fragments sont séquencés un par un. En 2003, il a fallu avec cette technologie une année pour séquencer le génome humain pour un coût avoisinant 3 milliards de US$.

b) Séquençage de 2nde génération

Un appel d'offre du *National Human Genome Research Institute* (NHGRI) au sein de l'Institut national de la santé (NIH) aux États-Unis a été effectué pour promouvoir des innovations technologiques de séquençage. L'objectif était de séquencer le génome humain pour un coût de 1 000 US$. De nombreuses idées de nouvelles technologies de séquençage de l'ADN ont émergé. Trois technologies de séquençage, 454 Titanium ROCHE, SOLID v3 ABI, GAIIX Illumina, sont apparues dans les années 2000 à 2010. Elles marquent le début du séquençage de seconde génération. Depuis, la société Illumina a perfectionné ses dispositifs de séquençage.



Dispositifs de séquençage de seconde génération

Crédit : Illumina.

c) Séquençage de 3ème génération

Les séquenceurs de troisième génération sont développés par Pacific Biosciences et Oxford Nanopore Technologies Ltd. Les premiers produits ont été commercialisés en 2011 par Pacific Biosciences, et en 2015 par Oxford Nanopore Technologies. Ces technologies permettent notamment de :

séquencer des molécules uniques afin d'éviter l'étape d'amplification et ainsi de pouvoir enregistrer les modifications chimiques éventuellement présentes sur les nucléotides dans les cas où elles altèrent le signal électrique engendré par le passage du polymère dans le nanopore ;

diminuer le coût de séquençage ;

obtenir des séquences à lecture longue afin d'éviter les étapes de fragmentation et d'alignement des séquences, et d'analyser les données en temps réel.

Dispositifs de séquençage de 3ème génération.

Crédits : Pacific Biosciences et Oxford Nanopore Technologies.

La technologie Illumina

L'entreprise Illumina (États-Unis) est leader sur le marché du séquençage. Actuellement, 96 % des données de séquençage mondiales sont produites par les dispositifs Illumina. Au total, 100 Po (10^{17} octets) de données ont été générées par leurs séquenceurs. Les progrès récents de leurs technologies sont décrits dans l'encadré.

APPAREILS «ILLUMINA»

Depuis l'apparition des premiers séquenceurs de deuxième génération en 1996, les dispositifs d'Illumina ne cessent d'évoluer pour permettre un séquençage plus rapide, de plus grande capacité et avec un taux d'erreur de séquençage plus faible. Plusieurs appareils de séquençage sont commercialisés.

Le dispositif le plus récent est le NovaSeq6000, commercialisé depuis trois ans. Il y a actuellement 600 appareils NovaSeq6000 dans le monde. Ce dispositif a été conçu pour être évolutif et pour s'adapter à la méthode et à l'échelle des différents projets de séquençage. Il possède deux cellules de mesure indépendantes permettant le séquençage. Une cellule de mesure lit 80 à 400 milliards de nucléotides pour la plus petite, et 2 000 à 4 000 milliards de nucléotides pour la plus grande. Ce dispositif lit jusqu'à 6 000 milliards de nucléotides en quarante-huit heures, soit le séquençage de 50 génomes humains avec un fort taux de recouvrement entre fragments séquencés. Le plus petit dispositif d'Illumina est un appareil compact de paillasse. Il lit jusqu'à 1,2 milliards de nucléotides par cycle en dix-sept heures. Son coût est d'environ 20 000 US$.

Chaque nouveau dispositif demande des innovations en mécanique des fluides, chimie organique, chimie de surface, mathématique, physique, biologie moléculaire, optique, informatique etc. Au total, 70 innovations et brevets ont été nécessaires pour concevoir le dernier dispositif d'Illumina, le NovaSeq 6000. Cependant, le marché du séquençage évolue avec la commercialisation des dispositifs de séquençage d'Oxford Nanopore et de Pacific Biosciences. Le prochain progrès à réaliser pour Illumina concerne le coût. Les dispositifs actuels permettent de séquencer un génome humain pour 700 US$. Dans les prochaines années, Illumina souhaite réduire ce coût à moins de 100 US$.

La technologie Oxford Nanopore

Oxford Nanopore Technologies a mis au point le premier séquenceur d'ADN et d'ARN à nanopores en temps réel et à lecture longue en 2015. Les récents progrès de leurs technologies sont détaillés dans l'encadré

APPAREILS «OXFORD NANOPORE TECHNOLOGIES»

Cette technologie est la seule capable de séquencer des molécules natives, sans amplification préalable par PCR. Ainsi, les modifications chimiques présentes sur les acides nucléiques sont conservées. Cette information supplémentaire est mesurée par une modification du flux ionique au passage du nucléotide modifié, en comparaison du même nucléotide non modifié. Cela nécessite un calibrage préalable, effectué lors du développement du logiciel. Enfin, contrairement aux autres méthodes de séquençage actuellement sur le marché, la technologie Oxford Nanopore permet de séquencer des acides nucléiques très longs. Des molécules longues de l'ordre du million de nucléotides ont été séquencées, avec un record à 2,2 millions.

Il existe plusieurs dispositifs de tailles et de capacités différentes. Le dispositif MinION est le seul appareil de séquençage d'ADN/ARN en temps réel et portatif. Il pèse moins de 100 g et se branche sur un ordinateur à l'aide d'un câble USB3. Il est composé d'un boitier de séquençage accessible pour un coût de 1 000 US$ et d'une cellule de mesure consommable, accessible pour un coût de 500-900 US$. Chaque cellule de mesure interroge jusqu'à 512 nanopores en simultané et génère 10 à 30 Go de données de séquençage par expérience.

Le dispositif GridION X5 est un appareil compact de paillasse. Il possède cinq cellules de mesure de type MinION. Il permet d'effectuer individuellement et simultanément jusqu'à cinq expériences. Il génère jusqu'à 150 Go de données et les analyse en temps réel. Le PromethION a été commercialisé en 2018, il en existe quatre en France. Il offre la même technologie de séquençage d'ADN/ARN en temps réel et à lecture longue que le MinION et le GridION, à une échelle beaucoup plus grande. Le

système actuel utilise jusqu'à 24 cellules de mesure à la fois. Chaque cellule de mesure utilise jusqu'à 3 000 nanopores, pour un rendement total théorique de 15 To de données en quarante-huit heures. Cet appareil est conçu pour séquencer des génomes humains complets pour moins de 1 000 US$. En 2019, un système PromethION à 48 cellules de mesure a été commercialisé.

Dispositifs de séquençage Oxford Nanopore Technologies

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

Très récemment, des dispositifs plus petits que le MinION ont été commercialisés. Le Flongle est un adaptateur pour MinION ou GridION et séquence en temps réel l'ADN/ARN sur des cellules de mesure beaucoup plus petites et à usage unique. Il est conçu pour être le système de séquençage le plus rapide et le moins cher. Le SmidgION, encore en développement, sera le dispositif le plus petit, pouvant être branché sur son téléphone portable.

Un flux ionique est mesuré pour chaque acide nucléique qui traverse le nanopore. Il est converti en séquence d'ADN en temps réel grâce au logiciel d'exploitation de la technologie nanopore : MinKNOW. Une molécule d'ADN séquencée est représentée par un fichier dont le format est au standard des bases de données internationales. Ces fichiers sont stockés localement et accessibles uniquement par l'utilisateur. Oxford Nanopore Technologie propose aussi son unité informatique autonome, MinIT, pour accompagner le séquenceur MinION. Il est préconfiguré avec le logiciel MinKNOW et effectue l'acquisition, l'analyse et le stockage des données en grande quantité. Étant donnée la rapidité du flux de données engendré par la technologie nanopore, cette unité MinIT est dotée d'un disque dur (SSD) de 500 Go et de puissants processeurs graphiques (GPU).

Des améliorations sont en cours pour augmenter la longévité des nanopores et réduire le taux d'erreur. La réaction enzymatique entre la protéine motrice qui se lie à l'ADN et la protéine de lecture, a besoin d'une source d'énergie, épuisée au bout de 48h. En optimisant la source d'énergie, le même nanopore pourra séquencer plus longtemps, jusqu'à 160 h. De plus, grâce à un nouveau circuit électronique intégré dans la membrane, il n'y a plus de perte de courant, ce qui préserve les nanopores. Avec ces améliorations, la capacité de séquençage d'un nanopore a doublé en quelques années, et devrait continuer d'augmenter.

La tête de lecture du nanopore actuel lit un nucléotide à la fois mais l'environnement immédiat (cinq nucléotides) influe sur la nature du signal. Ainsi, les erreurs sont fréquentes lors du séquençage d'un polymère comportant des suites de plus de cinq nucléotides identiques. Les chercheurs ont conçu un nouveau type de nanopore possédant deux têtes de lecture. Son utilisation améliore la qualité du séquençage d'un facteur dix. Ce nouveau nanopore a été commercialisé en 2019.

La société travaille actuellement sur des analogues d'acides nucléiques. Ces analogues, ajoutés à la séquence d'ADN, permettent de calibrer le signal électrique mesuré par le nanopore. Avec cet enrichissement, le taux d'erreur de séquençage sera encore diminué.

Conclusion

Illumina, leader du marché du séquençage dans le monde, possède une technologie fiable et robuste. Notamment, leurs dispositifs grande capacité séquencent un génome humain rapidement et pour moins de 1 000 US$. La technologie Oxford Nanopore est très prometteuse. Elle est miniaturisée et donc accessible pour tout laboratoire, peu onéreuse (moins de 1 000 US$ pour un MinION). Elle a l'important avantage de lire des molécules d'ADN natives et, donc, des nucléotides modifiés, voire des polymères non nucléotidiques ; en tout cas, il faut que les composants (monomères) de la molécule (polymère) puissent être différenciés par leur signal électrique. Selon les utilisateurs, la technologie Oxford Nanopore possède

encore un taux d'erreur de séquençage trop important (10 % d'erreurs de séquençage contre 1 % pour la technologie Illumina), mais, rappelons-le, ces erreurs sont largement corrigées lors de l'étape d'analyse des données par comparaison des lectures des multiples copies passant par les nombreux nanopores. Les grands groupes, comme Microsoft Corp., sont convaincus que cette technologie représente un atout majeur pour la lecture d'information stockée sur l'ADN. Certaines entreprises comme Helixworks ont même optimisé leur technologie de synthèse d'ADN en fonction des dispositifs d'Oxford Nanopore.

Améliorations des systèmes d'archivage d'information numérique sur l'ADN

Facteurs limitants

La lumière, l'eau et l'oxygène ont un effet délétère sur les acides nucléiques. Ils engendrent des réactions chimiques qui sont responsables de cassures ou mutations dans l'ADN, rendant son contenu informationnel impossible à déchiffrer. Cependant, l'ADN maintenu à l'abri de ses trois « ennemis » peut être conservé plusieurs milliers d'années. Rappelons que des molécules d'ADN ont pu être récupérées dans des os fossiles de plus de 560 000 ans enterrés dans un sol sec et frais[55] ; leur analyse a permis de comparer la séquence d'ADN d'animaux disparus avec celle de leurs descendants modernes. Les méthodes conventionnelles de stockage de l'ADN recourent essentiellement au stockage à basse température (de -20 °C à -196 °C). Hélas, ces méthodes sont difficilement automatisables et sont coûteuses en espace, équipement, énergie et maintenance. En outre, elles exposent les échantillons à des risques de dégradation, de contamination ou de perte en cas de pannes matérielles ou électriques. Affranchir du froid la conservation de l'ADN représenterait donc un avantage technologique, économique et écologique considérable pour le stockage d'information, si la stabilité des échantillons est assurée.

Amélioration des technologies de stockage au long terme de l'ADN

Stockage chimique : la technologie de Robert Grass

Partant de l'observation que l'ADN peut être préservé dans des os fossiles, l'équipe de Robert Grass (ETH Zürich, Suisse)[56] a eu l'idée d'encapsuler l'ADN par un processus chimique afin d'améliorer sa conservation. Ils ont conçu des nanobilles de verre (dioxyde de silicium) qui stockent l'ADN à température ambiante et le préservent de l'oxygène, grâce à un processus d'encapsulation chimique. Chaque nanobille mesure environ 100 nm de diamètre. Les progrès récents et améliorations envisagées pour cette technologie sont discutés dans l'encadré.

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

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

STOCKAGE DE L'ADN EN NANOBILLES DE VERRE

L'équipe de Robert Grass a modifié la surface des billes de verre pour la charger positivement. L'ADN chargé négativement est attiré sur la surface de la bille. Une molécule chimique, possédant d'un côté une charge positive et de l'autre côté un composé précurseur de la silice, est ajoutée. Le côté chargé positivement se fixe à l'ADN pour le recouvrir et le côté contenant le composé précurseur de la silice constitue une couche de verre solide enfermant l'ADN. Ce processus se déroule en phase aqueuse, ce qui implique que l'ADN encapsulé reste au contact d'un petit nombre de molécules d'eau. Le processus de désencapsulation, c'est à dire de libération de l'ADN, consiste à traiter les billes de verre avec une solution de fluorure. À la faible concentration choisie, le fluorure dégrade le verre sans endommager l'ADN. L'ADN libéré est amplifié par PCR puis séquencé pour décoder l'information qu'il contient.

L'ADN encapsulé dans les nanobilles depuis plusieurs années est conservé à 100 %, sans modification de sa séquence en nucléotides. La technologie étant récente, il est impossible d'évaluer directement la conservation de l'ADN dans les nanobilles au-delà de quelques années. Des cinétiques de dégradation ont donc été effectuées à des températures élevées pour mimer l'effet du vieillissement sur les échantillons. Les expériences menées montrent que l'ADN stocké dans les nanobilles sera conservé et analysable pendant plusieurs décennies à température ambiante, et 1 million d'années à -18°C[1]. Dans le cadre du stockage d'information numérique sur l'ADN, la technologie de Robert Grass permet d'augmenter significativement la stabilité de l'ADN et donc la durée de conservation de l'information. Cependant, la densité volumique d'information stockée diminue. En effet, l'ADN encapsulé ne représente que 0,1 % de la masse de la nanobille.

L'équipe de recherche collabore avec Microsoft Corp. afin d'augmenter la densité, c'est à dire la quantité de l'ADN stocké par nanobille. Ils ont conçu des nanoparticules magnétiques[2] capables de stocker plusieurs couches d'ADN successives. Une couche d'ADN est déposée sur une nanoparticule magnétique chargée positivement. Une couche de polymère cationique est ensuite déposée sur l'ADN. Enfin, une nouvelle couche d'ADN est déposée sur le polymère cationique. Le cycle se répète afin d'obtenir plusieurs couches successives d'ADN sur la nanobille. Grâce à cette technologie de stockage multi-couche, l'équipe a augmenté la quantité d'ADN stocké par bille. L'ADN représente alors 3 % de la masse de la nanobille. Actuellement, les chercheurs travaillent sur la lecture de l'information d'ADN à partir de la nanobille multi-couche.

Une des applications de cette technologie concerne le stockage de données informatiques sur l'ADN, pour un coût de 1 € par Ko. L'équipe de Robert Grass a synthétisé l'information numérique d'un film de 1,4 Mo sur l'ADN, encapsulé l'ADN dans les nanobilles et imprimé des verres de lunettes contenant ces nanobilles. De même avec la participation du groupe de musique Massive Attack, ils ont encapsulé l'ADN contenant la bande son de leur dernier album, qu'ils ont ensuite ajouté à une bombe de peinture. L'un des chanteurs du groupe a ainsi pu «peindre sa musique».

Comme les nanobilles sont ingérables sans risque, elles peuvent aussi servir à identifier des fraudes alimentaires. En outre, elles sont utilisées pour la traçabilité de pierres précieuses ; la société Healixa[3] a développé une solution de nanobilles traceurs s'appliquant sur la pierre durant le processus de polissage. Les nanobilles peuvent contenir un message d'ADN plus long qu'un code-barres, comme la fiche technique d'un produit ou le manuel d'utilisation. Elles sont ensuite imprimées dans le produit grâce à une imprimante 3D. L'idée de stocker l'information attachée à un objet dans l'objet lui-même, et non dans son emballage, présente certains avantages : l'information (manuel d'utilisation, notice technique) n'a pas de risque d'être perdu et les emballages du produit sont réduits.

1 Température de la réserve mondiale de graine (Global Seed Vault) à Spitsberg en Norvège.

2 Les particules magnétiques sont plus faciles à manipuler en solution que les nanoparticules de verre.

3 http://www.haelixa.com/

Stockage physique : la technologie d'Imagene

La société Imagene[57] (France) conçoit des capsules de conservation, appelées DNAshell®, qui stockent l'ADN et le préservent de l'eau, de l'oxygène et de la lumière. Chaque capsule peut contenir jusqu'à 0,8 g d'ADN, soit 1,4 Eo de données en tenant compte des redondances[58]. Les capsules sont composées d'un étui en acier inoxydable de quelques millimètres, enfermant un insert en verre dans lequel l'ADN est déposé. Chaque capsule est marquée pour la traçabilité et compatible avec le format standard de plaques

57 http://www.imagene.fr

58 1 Exaoctet (Eo) représente 1 million de disques durs de 1 Teraoctet (To).

à 96 puits, utilisé dans les laboratoires. Le détail des améliorations en cours et des performances de la technologie Imagene est résumé dans l'encadré.

STOCKAGE DE L'ADN EN CAPSULE D'ACIER

Le processus d'encapsulation de l'ADN, développé par Imagene, est entièrement automatisé. L'ADN en solution est déposé dans l'insert en verre de la capsule en acier inoxydable. Il est ensuite séché sous vide par une première étape de dessiccation. Une seconde étape de dessiccation est effectuée sous atmosphère de gaz neutres anoxiques et anhydres (mélange d'argon et d'hélium). Le bouchon métallique de la capsule est ensuite scellé par soudure laser afin d'encapsuler l'ADN. Des contrôles d'étanchéité des capsules sont effectués : le mélange de gaz contenu à l'intérieur de la capsule est facilement détectable par spectrométrie de masse en cas de fuite.

Ainsi, au total, cette technologie évite à l'ADN ses trois « ennemis » : il n'est pas au contact de l'eau, de l'oxygène ou de la lumière. En outre ce système de stockage est autonome et ne consomme pas d'énergie. L'ADN encapsulé par Imagene depuis plusieurs années est conservé à 100 %, sans modification de sa séquence en nucléotides. La technologie étant récente, il est impossible d'analyser directement la conservation de l'ADN dans les capsules au-delà de quelques années. Cependant, les cinétiques de dégradation réalisées par Imagene à différentes températures, extrapolées grâce à la loi d'Arrhenius , permettent d'estimer que l'ADN, stocké dans les capsules à température ambiante, aura une demi-vie de 52 000 ans.

Notons au passage que les capsules Imagene conservent également de l'ARN et des échantillons sources, comme du sang. Les chercheurs optimisent la technologie afin de conserver également des réactifs de biologie moléculaire (enzymes, milieux réactionnels) et des microorganismes (virus, levures, bactéries).

Stockage d'information numérique sur l'ADN *in vivo*

Projet

Le stockage d'information numérique sur l'ADN est expérimenté *in vitro* : les données numériques sont converties en séquences de nucléotides et les fragments d'ADN sont synthétisés et conservés dans des capsules. Certains scientifiques envisagent également de stocker l'information numérique sur l'ADN *in vivo*, protégé dans des cellules ou des organismes vivants. Ce système présenterait des avantages. En particulier, ces cellules se divisent — dans des milieux de culture bon marché — ce qui entraîne l'amplification de leur ADN, y compris celui qui a été introduit par le biologiste.

Plusieurs groupes scientifiques étudient le stockage de l'information numérique ***in vivo***. En 2007, des chercheurs ayant ensuite lancé le projet MOSLA avaient déjà intégré des oligonucléotides synthétiques, codant des logos et des noms d'entreprises, dans le génome des bactéries. En 2016, l'équipe de George Church a intégré des dizaines d'octets d'information chez le colibacille, grâce à la technologie CRISPR[59]. En 2020, l'artiste Joe Davis a recherché la capsule naturelle la plus résistante possible pour conserver sous forme d'ADN des archives de l'humanité après sa disparition. Il a choisi pour capsule ***Halobacterium salinarum***, une archébactérie qui porte plus de vingt copies de son chromosome et survit dans des dépôts salins durant des centaines de millions d'années[60].

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

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

L'un des objectifs du projet MOSLA déjà mentionné au chapitre IV est de concevoir des chromosomes artificiels et des plasmides[61] indépendants, en sus du chromosome naturel de la bactérie, contenant les informations numériques. La finalité de ce projet est de générer un « *cloud* » bactérien où l'information numérique sera fragmentée, chaque fragment étant stocké sous forme de méga-chromosome dans une multitude de cellules bactériennes. Ce projet nécessite donc une capacité à concevoir et réaliser de grands chromosomes bactériens.

Conception de chromosomes synthétiques

Concevoir des chromosomes synthétiques nécessite de disposer d'outils de biologie moléculaire performants. De plus, si la finalité du projet consiste à stocker de l'information numérique sous forme de chromosome artificiel dans les cellules, il est nécessaire de comprendre les mécanismes de maintenance et de réplication du chromosome, afin de ne pas perdre l'information qu'il contient. En effet, le chromosome synthétique pourrait être dégradé par la cellule, subir des mutations, s'intégrer dans le génome naturel des cellules ou ne pas être maintenu au cours de la division cellulaire.

À cette fin, l'équipe de Torsten Waldminghaus (SYNMIKRO, Philipps Universität Marburg, Allemagne)[62] développe des techniques d'assemblage de petits fragments d'ADN en chromosomes synthétiques, stockés dans les cellules. Les scientifiques ont construit deux chromosomes synthétiques de 100 000 nucléotides chez le colibacille, afin d'étudier les mécanismes de réparation et de maintenance des chromosomes.

La société Synovance (Genopole Évry, France)[63] conçoit et construit de grands chromosomes synthétiques d'ADN, et des souches bactériennes optimisées pour la bioproduction. Pour cela, l'entreprise développe des technologies de biologie computationnelle et des méthodes d'assemblage d'ADN.

Stockage d'information dans des spores

Les chercheurs du projet MOSLA souhaitent exploiter les propriétés des spores bactériennes pour stocker l'information numérique contenue dans l'ADN à long terme. En effet, les spores présentent des caractéristiques de résistance remarquables : elles survivent pendant plusieurs milliers d'années, même dans des conditions défavorables. La sporulation intervient chez certaines bactéries lorsque les conditions deviennent défavorables à la croissance, comme par exemple une carence en nutriments, en eau etc. Le processus de sporulation se caractérise par un épaississement de la paroi et une déshydratation complète de la cellule. Or, la présence d'eau est un facteur important de dégradation de l'ADN par hydrolyse.

Le principe consisterait donc à induire la sporulation dans une souche bactérienne après y avoir introduit l'ADN désiré dans son patrimoine génétique.

Limites du stockage d'information in vivo

Le stockage d'information *in vivo* présente des inconvénients. D'une part, la capacité d'un organisme à contenir de grands volumes d'information reste limitée ; ce fait peut être partiellement contré par l'approche distribuée décrite plus haut, selon laquelle l'ensemble de l'information est fragmentée et répartie en un grand nombre de cellules différentes, et peut être reconstituée après séquençage global grâce à une préalable indexation des divers fragments. D'autre part, les phénomènes évolutifs de l'organisme

61 Fragment d'ADN circulaire, indépendant de l'ADN chromosomique, naturel ou artificiel.

62 https://synmikro.com

63 https://synovance.com

conduisent à des mutations de l'information ; ce problème peut être pallié par l'usage de codes de correction d'erreur incorporés dans l'ADN sauvegardé.

Enfin, l'ADN contenant l'information numérique doit être toléré par l'organisme, et donc contenir des séquences nucléotidiques non toxiques pour celui-ci. Plus précisément, si la toxicité d'une séquence d'ADN ne peut résulter de ses nucléotides A, T, G et C qui sont naturels, elle peut l'être à cause de l'interprétation biologique qui en est faite par l'organisme porteur. Les séquences nucléotidiques destinées à porter une information numérique peuvent être vues comme aléatoires pour l'organisme porteur. Aussi resteront-elles non interprétées dans l'immense majorité des cas. Cependant, si l'idée est de recourir massivement à l'approche *in vivo* pour les mégadonnées, le hasard peut faire apparaître un petit nombre d'occurrences de séquences signifiantes pour le porteur : signaux de maintenance, séquences exprimées en ARN ou même protéine, ou repliées pour former des catalyseurs. Parmi ces cas rares, certains pourraient se révéler toxiques pour le porteur ou dangereux pour le biotope environnant. S'ils sont toxiques pour le porteur, les mutations aléatoires les rendant inoffensives seront préférentiellement sélectionnées et rapidement fixées dans la population porteuse ; la conséquence en sera une dérive rapide de cette séquence, donc la perte de l'information que l'on souhaitait archiver. Si sans être toxiques, ils sont interprétés par le porteur, il existe une faible probabilité qu'ils instruisent le porteur à produire un composé toxique pour l'homme ou pour le biotope environnant.

Ce problème potentiel de sécurité, aussi ténu soit-il, dans le stockage *in vivo* ne se rencontre pas dans l'approche majoritaire *in vitro* décrite dans les chapitres précédents. Les inconvénients de l'approche *in vivo* pourraient ralentir son développement futur[64] .

Systèmes de stockage d'information numérique non-ADN

En principe, tout polymère comportant au moins deux monomères différents pourrait être utilisé pour stocker l'information numérique. En pratique, il faut que ce polymère puisse être écrit selon une séquence arbitraire (déterminée par le fichier numérique à archiver), donc par chimie itérative en phase solide. Il faut aussi qu'il existe des méthodes pour le conserver longtemps et le lire aisément. Idéalement, ce polymère pourrait présenter une densité informationnelle encore supérieure à l'ADN.

Des systèmes d'archivage de données sur des polymères non-ADN sont étudiés.

Polymères artificiels

Le projet académique porté par Jean-François Lutz (Institut Charles Sadron, université de Strasbourg, France) utilise des copolymères non-ADN pour stocker de l'information numérique[65]. Les polymères sont des macromolécules possédant, dans leurs structures, de nombreuses sous-unités (monomères) se répétant. Il existe des polymères :

– **naturels** tels l'ADN ou les polysaccharides ;
– **artificiels**, fabriqués par l'homme, telle la nitrocellulose ;

Les polymères de synthèse, comme le plastique, ont un potentiel considérable pour le stockage d'information numérique. Ils permettent une plus grande densité d'information et une meilleure conservation

64 Il est notable que l'approche *in vivo* a fait l'objet de la seule mention en réponse au « Questionnaire éthique et technologie » approuvé par l'Académie des technologies. Encore cette mention a-t-elle été portée par un très petit nombre des personnes interrogées qui consistaient en l'ensemble des membres du groupe de travail et des personnalités auditionnées.

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

des données que les supports de stockage électroniques actuels. Pour stocker l'information numérique sur des polymères synthétiques, l'idée générale est de convertir les « bits » (séquence de '0' et de '1') en monomères pour former un polymère, ensuite synthétisé par chimie. Ainsi, un alphabet basé sur deux monomères différents permet l'écriture d'informations binaires dans une chaîne polymère linéaire. Les polymères utilisés pour l'archivage d'information numérique sont ici appelés : polymères « numériques ».

Pour lire l'information, les polymères sont « séquencés » par spectrométrie de masse. Cette technique permet de détecter et d'identifier les molécules d'intérêt par mesure extrêmement précise de leur masse et de caractériser leur structure chimique. Le résultat, appelé spectre de masse, est décodé informatiquement pour reconstituer le message en bits, et donc l'information initiale. L'encadré décrit les progrès en cours dans le domaine des polymères « numériques ».

POLYMÈRES « NUMÉRIQUES »

Le stockage d'information numérique sur des polymères synthétiques nécessite de disposer de polymères possédant des caractéristiques bien particulières :

la monodispersité : le polymère doit être uniforme, composé de monomères ayant la même constitution et des masses moléculaires du même ordre de grandeur (mais différentes afin d'être distinguées). Cela facilite la lecture du polymère par spectrométrie de masse ;

- pouvoir être utilisés pour des réactions orthogonales : possibilité d'effectuer des étapes de protection et déprotection d'un groupe d'atomes sans influencer les étapes de protection et déprotection d'un autre groupe d'atomes. Cela facilite la synthèse du polymère ;
- pouvoir être utilisés pour l'encodage binaire, c'est-à-dire pouvant posséder au moins deux motifs différents, comme par exemple un groupe méthyle pour un '1' ou un atome d'hydrogène pour un '0' ;
- posséder des monomères qui s'assemblent facilement et rapidement.

Il existe une multitude de polymères «numériques» pouvant être utilisés pour le stockage d'information numérique. Cette diversité permet de convertir les données numériques de manière dense. Cependant, pour chaque monomère, il est nécessaire d'adapter la synthèse et la lecture. Actuellement les chercheurs de l'équipe de Jean-François Lutz utilisent jusqu'à huit monomères de manière routinière.

Les polymères phosphodiesters possèdent un avantage majeur que n'ont pas les autres polymères « numériques » synthétiques : ils sont utilisés dans le cadre de la chimie des phosphoramidites (synthèse chimique d'ADN)[1]. Ainsi, ils sont compatibles avec les synthétiseurs de molécules d'ADN et sont assemblés de manière automatique et programmable.

Tous les huit monomères, un séparateur moléculaire, éventuellement clivable, est introduit. Le polymère ainsi créé contient autant d'octets d'information qu'il y a de groupes de huit monomères. Pour chaque monomère ajouté, le rendement de la réaction chimique de liaison est d'environ 99 %, limitant la longueur de synthèse du polymère à 100-150 monomères.

Les informations stockées dans les chaînes de polymères sont éditées à l'aide de déclencheurs physiques tels que la température ou la lumière. L'information contenue dans un polymère peut être effacée après polymérisation[2]. Certains monomères sont stables à température ambiante mais se dégradent à haute température. Il existe également des monomères sensibles à la lumière qui perdent une partie de leur information après exposition. Dans ce cas, le polymère n'est pas détruit, mais l'information qu'il contient est effacée. À l'inverse, l'information contenue dans un polymère peut être révélée. Il existe une famille de monomères illisibles avant exposition à la lumière. La lumière libère un des groupements chimiques du monomère, lui permettant d'être lu[3]. Ce processus est utile pour des applications liées à la contrefaçon par exemple. Enfin l'information peut être modifiée après polymérisation. Certains monomères se transforment en d'autres monomères grâce à la lumière. Ce processus est utile pour modifier une information déjà synthétisée.

Pour lire les informations stockées dans les polymères, il existe trois techniques de séquençage :

- la spectrométrie de masse en tandem ;

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.

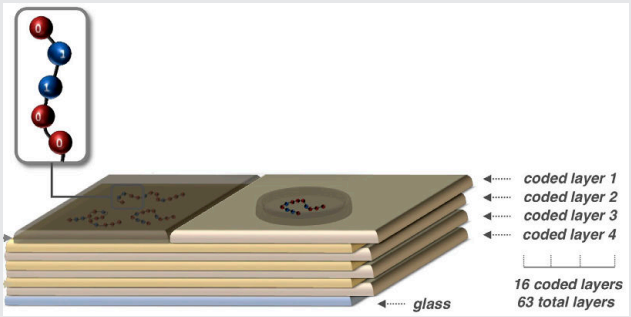
– le séquenceur nanopore ;
– le séquençage par dépôt de faisceau d'ion et microscopie à effet tunnel (STM).

La technique la plus utilisée est la spectrométrie de masse en tandem. Cette technique permet de déchiffrer des séquences numériques courtes et longues. Elle détecte et identifie des molécules d'intérêt par mesure de leur masse. Le résultat est appelé spectre de masse. Certains spectres sont complexes et longs à interpréter. C'est pour optimiser leur lecture que les scientifiques ont conçu les polymères « numériques ». Ils sont composés de monomères dont le squelette a la même masse moléculaire. Afin de différencier le '0' du '1', ces monomères possèdent deux motifs chimiques différents qui induisent un changement de masse détectable par le spectromètre. Pour les longues chaînes, les séparateurs positionnés tous les 8 monomères peuvent être clivés, et ce sont donc les sous-groupes de 8 monomères qui sont lus avec précision par spectrométrie. Les scientifiques ont mis au point un logiciel de décodage, appelé MS-DECODER, capable de retranscrire le spectre de masse de chaque groupe de polymères en bits[4]. Ainsi, les données sont analysées par n'importe quel utilisateur.

D'ici quelques années, il sera possible de lire les polymères non naturels par des technologies de type Oxford Nanopore Technologies. Les performances en termes de prix et rapidité pourraient alors être comparables à celles de l'ADN.

Les polymères synthétiques sont utilisés pour la traçabilité de certains produits. Ils peuvent contenir une information comme un code-barres. Ils sont ensuite ajoutés sur la surface d'un objet, ou composent l'objet lui-même. L'équipe de Jean-François Lutz a ajouté un polymère numérique contenant un code-barres dans un implant biomédical en plastique. Après plusieurs années *in vivo* chez le rat, l'implant a été retiré et le code-barres qu'il contenait était intact. Il est également possible d'effectuer de la spectrométrie de masse de surface, pour décoder des polymères à la surface des objets. On imagine un futur proche où des polymères numériques, contenant des codes-barres, seraient imprimés à la surface des billets de banque pour la traçabilité et non-falsifiabilité.

Les polymères synthétiques pourront être utilisés pour le stockage d'information numérique, y compris à long terme. Les chercheurs travaillent sur une méthode de stockage d'information par couche de polymères afin de densifier l'information. Chaque couche, séparée des autres par des lames de verre, contiendrait des chaînes de polymères. Contrairement à l'ADN, les polymères synthétiques n'ont pas besoin d'être encapsulés pour protéger leur information. Les polymères comme le plastique sont très robustes et des milliers d'années sont nécessaires à leur dégradation.



Couches de polymères synthétiques pour le stockage d'information numérique.

Crédit : Jean-François Lutz

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

Composés organométalliques

Les scientifiques du projet MOSLA conçoivent un nouveau support de stockage basé sur les composés organométalliques, afin d'augmenter la capacité de stockage d'information.

Ces composés sont exclusivement synthétiques. Ils sont constitués d'atomes de carbone et de métal. Ils sont capables de refléter la lumière en émettant différentes longueurs d'onde. Par contraste avec la linéarité des polymères discutés jusqu'à présent, ces composés seront imprimés sur une surface, comme par exemple celle d'un disque compact. Grâce à cette technologie, une plus grande quantité d'informations sera stockée sur une même surface.

Afin de lire les informations imprimées sur ces supports, les scientifiques développent des numériseurs et des lecteurs spécifiques.

Conclusion

Affranchir du froid la conservation de l'ADN représente un avantage technologique, économique et écologique considérable pour le stockage d'information numérique sur l'ADN. À cette fin, deux technologies de stockage sont développées :

- **chimique**. Avec le système de stockage de Robert Grass, l'ADN synthétisé est encapsulé dans des nanobilles de silice. Les inconvénients de cette approche résident dans une faible densité informationnelle, et l'élimination incomplète de l'eau et de l'oxygène susceptibles de dégrader l'ADN.
- **physique**. Le système de stockage de Imagene stocke d'énormes quantités d'information dans une petite capsule. Il protège l'ADN de l'eau, des sels, de l'oxygène et de la lumière. La conservation de l'ADN est ainsi estimée à 52 000 ans à température ambiante.

Le stockage d'information numérique sur l'ADN *in vivo* proposé par le projet MOSLA présente divers inconvénients que ne partagent pas les systèmes *in vitro*.

Des systèmes très prometteurs de stockage d'information numérique sur polymères non-ADN sont en développement dans l'équipe de Jean-François Lutz.

Enfin, 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 d'un taux d'erreur abaissé, et l'inconvénient d'une perte de densité informationnelle .

Indexage et calcul informatique avec de l'ADN synthétique

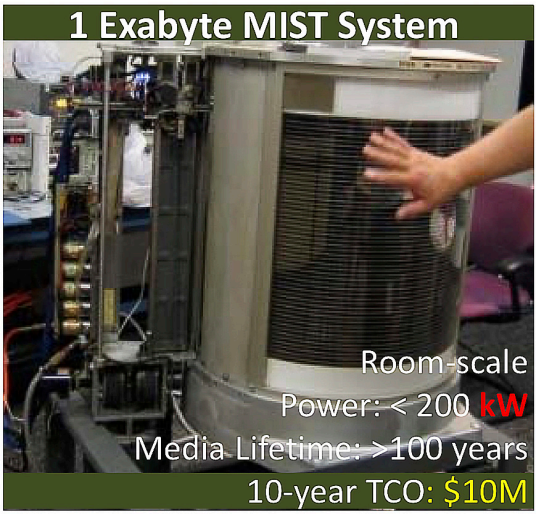
Indexage

En règle générale, il n'est pas possible d'accéder à une petite portion d'information stockée sur l'ADN sans lire toute l'information présente. Aussi, les scientifiques travaillent sur une nouvelle architecture de stockage d'information sur l'ADN.

Le projet MOSLA utilise des nucléotides et oligonucléotides spécifiques pour créer un système de stockage pouvant être utilisé pour accéder sélectivement à l'information.

Catalog DNA (État-Unis) travaille sur des étiquettes de nucléotides. Ainsi, chaque fragment d'ADN assemblé peut être utilisé comme étiquette. Ces étiquettes permettent de regrouper les molécules d'ADN contenant l'information issue des mêmes fichiers numériques.

Des chercheurs de l'université de l'Illinois travaillent sur une nouvelle architecture qui permettra d'accéder à des blocs de données et de réécrire des informations déjà stockées dans ces emplacements. Elle est basée sur des séquences d'ADN dotées d'adresses spécialisées pouvant être utilisées pour l'accès sélectif à l'information.



Objectifs pour le dispositif de stockage d'information de MIST.

Sources : IARPA et https://commons.wikimedia.org/wiki/File:IBM_350_RAMAC.jpg (licence : CC BY-SA 2.5)

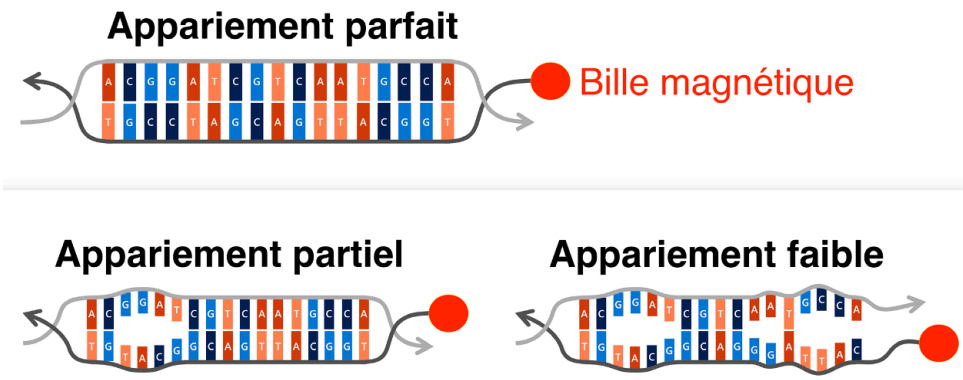
Enfin, le programme MIST (IARPA – États-Unis) envisage un appareil de paillasse capable d'accès aléatoire aux informations stockées sur le support moléculaire. Un système d'exploitation sera également conçu pour le dispositif de stockage. Il coordonnera l'indexation, l'adressage, la compression des données, les corrections d'erreurs et les traductions d'informations binaires en séquences d'ADN et inversement.

Calcul informatique avec de l'ADN synthétique

Microsoft Corp. a démontré qu'il est possible d'exploiter des données stockées sur l'ADN. Les scientifiques ont conçu un système informatique qui combine le stockage et le traitement de données moléculaires. Ils proposent une architecture hybride, électronique – moléculaire, qui exploite les atouts des deux domaines[66].

Le principe de base du calcul sur ADN *(DNA computing)* repose sur les propriétés physico-chimiques de l'ADN. En effet lorsque qu'un fragment d'ADN simple-brin rencontre son brin complémentaire, ces deux brins s'hybrident pour former un ADN double-brin. Il arrive cependant que deux brins d'ADN s'hybrident, sans avoir des séquences en nucléotides parfaitement complémentaires. On appelle ce phénomène l'hybridation partielle. En tenant compte du principe d'hybridation, les chercheurs ont trouvé et groupé des images semblables, codées en nucléotides d'ADN, parmi un large panel d'images. Des détails supplémentaires se trouvent dans l'encadré.

66 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.



Principe d'appariement des brins d'ADN selon leur propriété physico-chimique.

Crédit : François Képès.

Le calcul sur ADN a l'inconvénient de la lenteur, mais pour résoudre les problèmes combinatoires, il a les avantages d'une haute parallélisation et d'une grande efficacité énergétique[67]. Notons pour finir que la même bibliothèque peut être ré-utilisée dans de nombreux problèmes avec des cibles variées.

CALCUL SUR L'ADN

Supposons que l'objectif du calcul est de trouver dans une base de données contenant des centaines d'images, celle qui ressemble le plus à une illustration d'intérêt (l'image cible). L'idée est de convertir les caractéristiques propres à chaque image en vecteurs, puis de coder ces vecteurs en séquences d'ADN synthétique. Ainsi la base de données est représentée par une bibliothèque de fragments d'ADN simple-brin plus ou moins similaires. Les images possédant des caractéristiques proches auront une séquence d'ADN similaire. La cible est représentée aussi par un fragment d'ADN simple-brin, mais une bille magnétique est fixée chimiquement à son extrémité.

L'objectif de retrouver les images de la base de données qui ressemblent le plus à cette cible devient un calcul sur ADN : quel(s) ADN de la bibliothèque est le plus similaire à l'ADN représentant la cible ? Le calcul est réalisé en introduisant le fragment d'ADN simple-brin magnétique dans le mélange de fragments d'ADN de la bibliothèque. Si une image ressemble parfaitement à la cible, sa séquence d'ADN s'hybridera totalement avec celle de l'ADN cible.

Si une image ressemble légèrement à la cible, sa séquence d'ADN s'hybridera partiellement à celle de l'ADN cible. Enfin s'il n'y a aucune ressemblance entre les deux images, il n'y aura pas d'hybridation. Ce procédé permet d'extraire du mélange, avec un aimant, les copies de l'ADN cible magnétique, hybridé à d'éventuels ADN simple-brin complémentaires. Il s'agit ensuite de séquencer ces ADN complémentaires, ce qui identifie les images de la base de données ressemblant à l'illustration cible. En jouant sur la température ou sur la salinité, l'opérateur peut faire varier la stringence des conditions d'hybridation de deux brins d'ADN en fonction du taux de ressemblance désiré entre une image et sa cible.

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

Chapitre V

Initiatives mondiales

La Commission européenne a communiqué en février 2020 une *Stratégie européenne pour les données*[68] : il est notable que les termes *ADN* ou *polymère* en sont absents.

États-Unis

Dans ce domaine émergent, l'investissement public aux États-Unis serait d'environ 150 millions US$, répartis entre les trois agences DARPA, IARPA (projet MIST[69]), et NSF qui avec IARPA contribue au projet SemiSynBio[70] lequel a publié en 2018 une feuille de route avec des objectifs chiffrés à deux et quatre ans[71]. George M. Church (Harvard Medical School et MIT) est un des pionniers du domaine depuis 2012[72]. En outre, plusieurs compagnies privées sont actives dans ce domaine, telles que Microsoft Corp., Twist Bioscience et Catalog DNA.

MIST (IARPA)

MIST (Molecular Information STorage) est un programme américain de quatre ans, débuté en janvier 2019 par l'IARPA[73], et opérant largement par appel d'offres. Il a investi 48 millions US$ afin de développer de nouvelles technologies de stockage d'information. Une première enveloppe de 23 millions US$ finance le consortium dont fait partie DNA Script (France) avec la société Illumina et des chercheurs du MIT et de Harvard University (États-Unis). Les 25 millions US$ restants ont été alloués à un second consortium qui réunit Microsoft Corp. et Twist Bioscience[74] .

L'objectif du projet MIST est de stocker une grande quantité d'information (de l'ordre de 1 Eo), en un minimum d'espace (de l'ordre du mm^3), avec des coûts financier et énergétique réduits par rapport aux systèmes de stockage actuels. L'objectif ultime sera de stocker 1 Eo de données d'un centre de données dans le dispositif MIST, 20 000 fois plus petit, 1 000 fois moins coûteux en énergie, pour une durée de vie au moins vingt fois plus importante et pour un coût 100 fois moins important.

À cette fin, MIST utilisera des polymères comme support de stockage de données. Il concevra également les dispositifs et les systèmes d'exploitation nécessaires pour assurer une interface avec ce

68 *Une stratégie européenne pour les données*. Communication de la Commission européenne au Parlement, au Conseil, au Comité économique et social et au Comité des Religions (2020).

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

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

71 https://www.src.org/library/publication/p095387/p095387.pdf — Chapitre 1.

72 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.

73 L'IARPA *(Intelligence Advanced Research Projects Activity)* est une organisation du Bureau du directeur du renseignement national des États-Unis d'Amérique. Ses recherches sont appliquées par CIA, FBI et NSA.

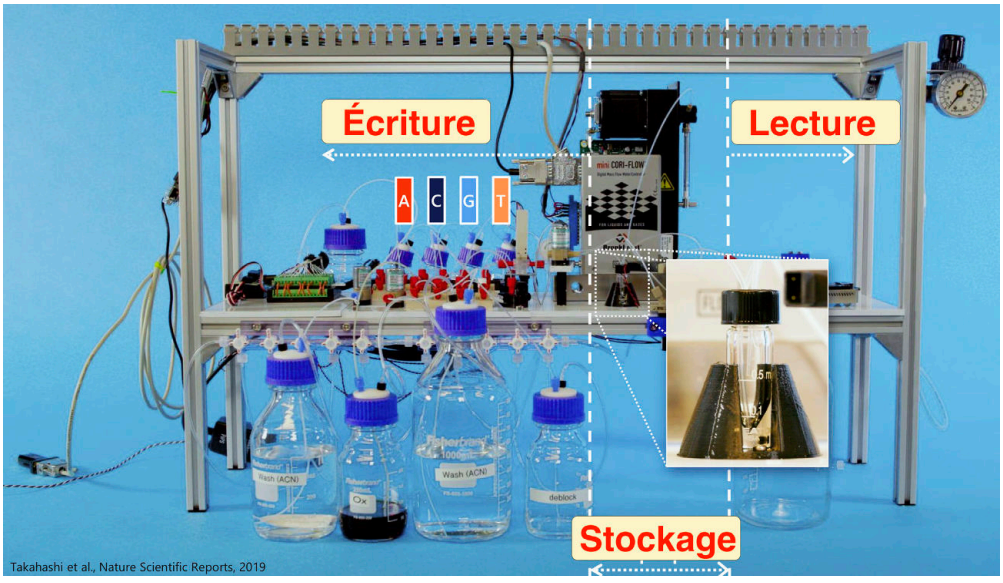
74 https://www.lemonde.fr/economie/article/2020/01/22/le-gouvernement-americain-investit-dans-le-stockage-de-donnees-dans-l-adn_6026763_3234.html

support. Des technologies seront développées pour optimiser l'écriture et la lecture d'informations sur ces polymères, et pour permettre un accès aléatoire à l'information stockée sur ces polymères.

Université de Washington / Microsoft Corp.

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)[75]. Les chercheurs ont conçu un prototype de paillasse utilisant l'ADN comme support pour stocker l'information[76]. L'appareil est entièrement automatisé et autonome. Il est composé de trois 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. Le système de stockage encapsule l'ADN dans des nanobilles. L'ADN est ensuite extrait des nanobilles et séquencé via le dispositif d'Oxford Nanopore Technologies.

Ce prototype est fonctionnel et a déjà permis de stocker et retrouver 1 Go de données[77]. Les chercheurs l'optimisent pour qu'il devienne plus compact et plus rapide. Ils développent notamment de nouveaux systèmes basés sur la microfluidique pour transporter des gouttes de réactifs sur un support électronique[78]. 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 (cf. chapitre III). À ce stade, leur système permet de regrouper les images, prises parmi un large ensemble, selon leur ressemblance à une illustration cible.



Premier prototype entièrement automatisé de stockage de données sur l'ADN

Crédit : Microsoft Corp. / Université de Washington, États-Unis.

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

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

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

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

Twist Bioscience

Twist Bioscience est une jeune pousse américaine cotée au NASDAQ. En cinq ans, la société a levé 190 millions US$. Elle est spécialisée dans la synthèse d'ADN par voie chimique sur micropuce. La jeune pousse participe à plusieurs projets sur le stockage d'information numérique sur l'ADN :

- le projet Iconem[79], qui consiste à numériser des modèles de sites patrimoniaux en trois dimensions et à archiver ces informations sur l'ADN ;
- la mission ARCH[80], en collaboration avec Microsoft Corp. et l'université de Washington. Cette mission consiste à archiver une collection de photographies du monde entier sur l'ADN, qui sera envoyée sur la lune en 2020 ;
- le stockage sur l'ADN de plusieurs enregistrements audio du Festival de jazz de Montreux en 2017.

Catalog DNA

Catalog est une jeune pousse américaine ayant pour objectif de faire de l'ADN un support de stockage d'information numérique. En 2018, l'entreprise a reçu neuf millions US$ de diverses sociétés privées.

Catalog a conçu une machine capable de synthétiser de l'ADN codant 0,5 Mo d'information par seconde. La jeune pousse a converti l'ensemble de la bibliothèque de Wikipédia (soit 14 Go d'information) en séquence d'ADN, ensuite synthétisé par cette machine. Il s'agit d'un record absolu à l'heure actuelle. Comme il repose sur l'usage de fragments d'ADN pré-synthétisés, il n'est pas directement comparable au record de 1 Go détenu par Microsoft Corp./Univ. Washington qui assemblent l'ADN nucléotide à nucléotide.

Les chercheurs perfectionnent cette machine afin d'en augmenter les capacités et les performances. Leur objectif à terme est de concevoir une nouvelle version de la machine qui sera capable de synthétiser l'ADN 1 000 fois plus rapidement et transformera 0,12 Go d'information par seconde.

Chine

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

Israël

En Israël, le projet du Technion, porté par Zohar Yakhini, consiste à stocker les informations numériques sur l'ADN de façon plus dense. Les chercheurs travaillent sur un nouvel alphabet de lettres composites. Ils ont converti un fichier de 6,4 Mo en nucléotides d'ADN, en utilisant un alphabet de cinq puis six nucléotides composites[81]. Ils travaillent également sur un alphabet de 20 nucléotides.

79 http://iconem.com/fr/

80 https://www.archmission.org/

81 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

Royaume-Uni

Au Royaume-Uni, *l'European Bioinformatics Institute* (EBI) est l'un des pionniers du domaine[82]. En outre, plusieurs compagnies privées y sont actives, comme Oxford Nanopore Technologies, Nuclera Nucleics, Evonetix Ltd.

Les chercheurs de l'équipe de Nick Goldman (EBI) ont démontré dès 2013 qu'il était possible d'utiliser l'ADN pour stocker et récupérer des informations numériques. Ils ont converti quatre fichiers numériques correspondant à des documents de nature diverse (Texte, JPEG, PDF, mp3) en séquence d'ADN. Les 0,7 Mo contenus dans le total des quatre fichiers ont été reconstitués au bit près, sans erreur. Leurs résultats ont été publiés en 2013[83].

Irlande

Helixworks est une jeune pousse irlandaise qui fabrique et vend des technologies de stockage de données basées sur l'ADN. Ils conçoivent une technologie MoSS (Molecular Storage System) permettant de convertir puis synthétiser les fichiers numériques en ADN. L'entreprise travaille au projet Fusion[84] en partenariat avec Ubisoft, Ambey et AKQA. Ce projet a pour objectif de stocker l'information numérique d'un jeu vidéo dans une cannette de boisson énergétique de 250 ml.

En outre Helixwork contribue au projet OligoArchive[85], qui a été financé en 2019 par l'EIC (European Innovation Council), pour une durée de trois ans, et pour un montant de trois millions d'euros. Son objectif est de concevoir un appareil de paillasse utilisant l'ADN comme support de stockage d'information. Cet appareil sera capable de convertir l'information numérique en séquence d'ADN, de synthétiser la séquence correspondante, de la stocker à long terme et de la séquencer pour en extraire l'information stockée. Les autres contributeurs sont l'Imperial college de Londres, l'institut de pharmacologie moléculaire et cellulaire (IPMC, Sophia-Antipolis, France), le laboratoire d'informatique, signaux et systèmes (I3S, Sophia-Antipolis) et le département *Data science* d'Eurecom (Sophia-Antipolis).

Allemagne

En Allemagne, un financement de 4,2 millions d'euros a été accordé par le ministère hessois pour le projet MOSLA[86] (universités de Marburg, Darmstadt, Giessen, 2019-2022).

L'objectif du projet MOSLA est de développer de nouvelles approches et solutions pour l'archivage pour de longues durées d'information, basées sur des systèmes de stockage moléculaires et chimiques. Ce projet vise à augmenter la capacité de stockage actuelle. Il comporte quatre parties :

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

83 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 synthetisized DNA. *Nature* 494(7435):77-80.

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

85 https://oligoarchive.github.io

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

- optimiser le stockage des données numériques ;
- développer un système de stockage d'information sur l'ADN, *in vitro* et *in vivo* ;
- développer un système de stockage sur des composés organométalliques ;
- organiser le stockage de l'information produite par l'Homme.

France

En France, il n'existe pas de politique nationale ou d'investissement public visant directement ce domaine, malgré quelques activités de premier plan. Un remarquable projet académique sur l'usage de copolymères non-ADN est porté par Jean-François Lutz (Institut Charles Sadron, CNRS et université de Strasbourg)[87]. La société DNA Script (Paris) est bien positionnée dans le domaine de la synthèse enzymatique d'ADN; outre des investissements privés, elle a reçu en 2020 un financement significatif des États-Unis (mentionné plus haut dans la description des actions de l'IARPA). La compagnie Imagene (Bordeaux et Évry) a une position forte dans le domaine du stockage de très longue durée d'ADN. Enfin, trois laboratoires (Sophia-Antipolis) sont engagés dans un projet international financé par l'EIC (cf. ci-dessus).

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

Chapitre VI

Perspectives

Perspectives techno-scientifiques

La preuve de concept de l'archivage de données numériques sur l'ADN hors du vivant (*in vitro*) est établie. Plusieurs études ont montré que cet archivage peut prendre en charge l'accès sélectif et évolutif aux données, ainsi que le stockage et la restitution d'information sans erreur. Cependant, des défis techniques subsistent pour que ce procédé devienne viable économiquement pour un large spectre de données. Ils concernent l'amélioration des coûts, de la vitesse et de l'efficacité des technologies de lecture, et surtout d'écriture et édition, de l'ADN ou autres polymères.

Concernant l'écriture, plusieurs acteurs du domaine placent leurs espoirs dans la synthèse d'ADN par voie enzymatique, dont le potentiel de développement semble supérieur à celui de la voie chimique traditionnelle. Lorsque des modifications sur de l'ADN stocké deviennent nécessaires (accès évolutif), deux approches sont *a priori* envisageables, soit la ré-écriture, soit l'édition du stock. Le choix entre ces deux approches doit s'appuyer sur une évaluation du rapport coût/bénéfice, qui dépend de l'extension des modifications, de leur multiplicité, et de l'état de l'art qui évolue rapidement.

Quant à la lecture, l'usage de nanopores offre un bon potentiel car cette approche restitue de longues séquences d'un seul tenant, est intrinsèquement parallélisable, et montre une grande versatilité en s'adaptant à la croissante diversité chimique des polymères à applications « numériques ».

Notons aussi que, quoique les vitesses d'écriture et de lecture de l'ADN soient limitantes, cet inconvénient est pallié dans certaines applications par la possibilité de parallélisation massive. Concrètement, d'ici 2024, une seule machine pourrait écrire et lire 1 To par jour.

Pourtant, plusieurs ordres de grandeur manquent actuellement pour la pleine adoption de la solution « ADN pour l'archivage de mégadonnées » : un facteur d'environ mille pour le coût de lecture, et cent millions pour celui d'écriture. Ces facteurs peuvent sembler faramineux. Ce serait oublier la célérité des progrès des technologies associées à l'ADN. Ainsi, George M. Church, dans son exposé de mars 2019, avait estimé que les coûts de la lecture et de l'écriture de l'ADN avaient chuté en dix ans d'un facteur supérieur au million, soit de moitié tous les six 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 deux ans entre 1971 et 2016 ; d'autre part, la société Seagate a signalé qu'elle avait fait diminuer en 29 ans le coût du stockage unitaire de données sur disque d'un facteur 1,3 millions. On voit donc que, selon ce critère, les technologies de l'ADN évoluent bien plus vite que celles de l'information.

Les performances du procédé global d'archivage pourraient aussi être améliorées par des alphabets étendus ou composites.

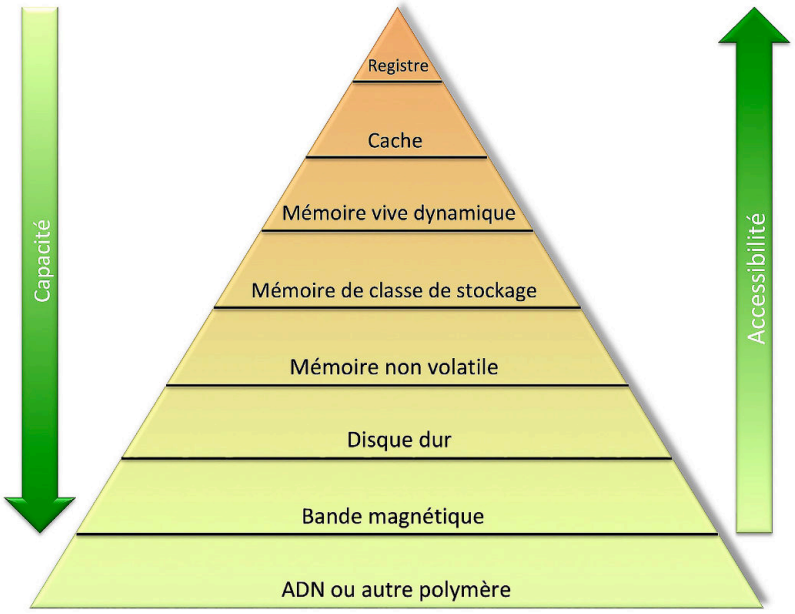
Enfin, plusieurs lignes de recherche sont parties du constat que l'ADN n'est pas nécessairement le polymère « numérique » le plus performant hors de la cellule : soit que son alphabet soit trop limité (à quatre lettres), soit que sa physico-chimie ne soit pas optimale. Ce constat a donné lieu à des approches alternatives s'éloignant plus ou moins de l'ADN, pour envisager d'autres hétéropolymères ou copolymères linéaires présentant des avantages théoriques. Lorsque leurs performances de lecture, d'écriture et d'édition rejoindront celles de l'ADN, ce qui pourrait prendre une décennie, ces polymères très prometteurs feront probablement irruption sur le marché de l'archivage de l'information numérique.

Perspectives économiques

L'idée d'archiver des données numériques dans l'ADN date de plusieurs décennies. À l'heure actuelle, les scientifiques en ont apporté la preuve de principe, mais ce stockage d'information reste expérimental.

Un certain consensus s'est dessiné parmi quelques acteurs du domaine pour considérer que la viabilité économique du stockage moléculaire d'information pourrait être atteinte sous cinq à dix ans pour des marchés de niche. Citons à titre d'exemple l'archivage à long terme d'informations sensibles : certains atouts maîtres seraient la facilité à multiplier l'ADN pour distribuer 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 à 20 ans. Le handicap principal de l'ADN résidant dans la lenteur des procédés d'écriture et d'édition, il est raisonnable de supposer que son usage se cantonnera encore longtemps à l'archivage à long terme, où sont évidents ses atouts : densité informationnelle, longévité, non-obsolescence. Cet archivage massif au long cours est par exemple un besoin reconnu des sociétés et institutions qui conservent des films, ou des laboratoires de physique des particules, pour seulement citer deux domaines disjoints. En ce cas, le stockage sur l'ADN entrera en compétition ou en complémentarité avec la bande magnétique, actuellement la solution de choix pour l'archivage à long terme. Comme les bandes magnétiques consomment moins de 1 % de l'électricité d'un centre de données, il apparaît que ce n'est pas au plan énergétique que le stockage moléculaire dispose de son meilleur atout.



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.

Crédit : François Képès et Carlo Reita.

Perspectives nationales

Le contexte français est d'une grande richesse en gros centres de données[88]. Aujourd'hui, hors entreprises, existent en France 155 gros centres plus 22 dédiés au « nuage »[89].

En France (voir chapitre 5), au moins un laboratoire universitaire et deux petites entreprises ont des positions originales et fortes dans les segments-clés très prometteurs que sont les polymères non-ADN, la synthèse enzymatique de l'ADN et son stockage de très longue durée. Au-delà, existe en France un gisement de compétences pertinentes en biologie, chimie, informatique et sciences de l'ingénieur, qui pourraient être mobilisées dans une nécessaire synergie entre secteurs public et privé. Un exemple de gisement de compétences est le trio de laboratoires de Sophia-Antipolis contribuant à un consortium européen sur l'intégration de l'ADN dans la hiérarchie de stockage des bases de données.

Il est notable que les deux petites entreprises mentionnées ci-dessus sont recrutées pour des collaborations de fond par des entités états-uniennes ; l'une d'entre elles, DNA Script, a même été richement dotée en 2020 par une agence des États-Unis. *A minima,* il semble donc indispensable de faire de la veille active dans le domaine, et d'encourager les initiatives ponctuelles prises dans ce domaine par diverses entités pionnières sur sol français. Cependant, comme ce domaine émergent constitue probablement la clé de l'archivage soutenable des mégadonnées, comme elle présente un intérêt particulier dans le cas de données sensibles, il convient de réfléchir à un positionnement plus actif .

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

89 https://www.datacentermap.com

Face aux limites physiques qu'atteignent les centres de données, la technologie moléculaire d'archivage des mégadonnées a, en effet, le potentiel de devenir économiquement viable entre 2025 et 2040, progressant de marchés de niche vers des marchés plus globaux. Il s'agit donc bien d'un enjeu majeur et stratégique à horizon proche. Aussi, il serait souhaitable de capitaliser sur le laboratoire et les deux petites sociétés qui ont été identifiés (voir chapitre V), et sur le gisement plus large de compétences, afin de de devenir des acteurs significatifs et d'ouvrir une perspective européenne. Dans ce but, voici deux recommandations détaillées.

Lancer une action concertée au plan national

Ceci passerait par une vigoureuse programmation pluriannuelle de subventions publiques explicitement dédiées, qui pourrait user des instruments suivants :

- avant tout, des appels d'offres visant spécifiquement à susciter des propositions ambitieuses de ruptures technologiques ; incitant aux synergies entre disciplines, et entre secteurs public et privé ; abaissant les risques pris en se lançant dans cette approche émergente ; reposant sur un comité scientifique s'appuyant sur l'expertise et la motivation des pionniers français, et sur des avis internationaux ;
- une plate-forme technologique transdisciplinaire, lieu d'expérimentation et de réflexion (notamment pour la rédaction de la feuille de route nationale) ; fédérant les secteurs public et privé ; ayant vocation à ultérieurement s'insérer comme le nœud français dans le consortium européen pertinent ;
- une conférence annuelle et internationale, initialement à dominante française. Les objectifs de cette conférence seraient de :
  - augmenter la visibilité des objectifs scientifiques et techniques du thème de l'archivage moléculaire,
  - montrer aux acteurs potentiels l'intérêt et la faisabilité d'une action commune,
  - relever les progrès réalisés en France et au-delà, et mettre à jour la feuille de route nationale du domaine.

Proposer une programmation européenne

La France pourrait proposer d'identifier ce thème comme un domaine à part entière dans le futur programme de recherche de la Commission européenne. Cette dernière pourrait user des instruments suivants :

- des appels d'offres dédiés, récurrents, transdisciplinaires, et plurinationaux;
- la mise en place d'un réseau européen de laboratoires publics et privés, facilitant la circulation des personnes, compétences et savoirs. Idéalement, ce réseau aurait pour bras armé un consortium de plates-formes technologiques nationales.

Travaux et Contributions

Document élaboré par le groupe de travail « ADN: lire, écrire, stocker l'information » animé par François KÉPÈS, avec l'appui de Morgane CHAMPLÉBOUX

Personnalités auditionnées

04/12/2018

Olivier Lucas, directeur associé Europe sud — Oxford Nanopore Technologies Ltd., Royaume-Uni

08/01/2019

Philippe Glaser, directeur Pasteur Genopole Île-de-France — Institut Pasteur, France

Piet Herdewijn, professeur — Institut REGA, université catholique de Louvain, Belgique

12/03/2019

Dominik Heider, professeur — université Philipps de Marburg, Allemagne

Mostafa Ronaghi, CTO et vice-président senior — Illumina Inc., États-Unis

Nick Goldman, chef de la recherche et scientifique senior — EMBL-EBI, université de Cambridge, Royaume-Uni

14/03/2019

George M. Church, professeur — Faculté de médecine de Harvard et MIT, États-Unis *[Séminaire tenu à l'École Polytechnique]*

29/03/2019

Emily Leproust, CEO — Twist Bioscience, États-Unis

09/04/2019

Alexander Murer, CEO — Kilobaser, Autriche

John Hoffman, Technical SETA, et **David A. Markowitz**, Manager — programme MIST, IARPA, États-Unis *[en téléconférence]*

Tim Brears, CEO, et **Matt Hayes**, CTO — Evonetix Ltd., Royaume-Uni

14/05/2019

Brian Jester, CEO — Synovance, France

Torsten Waldminghaus, professeur — *au moment de l'entretien :* SYNMIKRO, LOEWE center for synthetic microbiology, université Philipps de Marburg, Allemagne

11/06/2019
Zohar Yakhini, professeur — Technion & IDC Herzliya, Israël *[en téléconférence]*

07/07/2019
Carlo Reita, directeur des partenariats stratégiques et du planning — CEA-Leti, France

10/09/2019
Sachin Chalapati, CTO — Helixworks, Irlande

08/10/2019
Nick Gold, vice-président du marketing — Catalog DNA, États-Unis [en téléconférence]
Steven Benner — Ffame et université de Floride, États-Unis

11/10/2019
Karin Strauss, principale directrice de recherche — Microsoft Corp., États-Unis
Luis Ceze, professeur — université de Washington, États-Unis

12/11/2019
Tomas Ybert, CEO — DNA Script, France
Sophie Tuffet, présidente du directoire — Imagene, France

03/12/2019
Jiahao Huang, CCO — Nuclera Nucleics, Royaume-Uni
Robert Grass, professeur — ETH Zürich, Suisse
Jean-François Lutz, directeur de recherche — Institut Charles Sadron, université de Strasbourg, France.

Membres du groupe de travail « ADN: lire, écrire, stocker l'information »

Membres de l'Académie des technologies

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

Membres non académiciens

Morgane Champleboux — université d'Évry (secrétaire scientifique)
Wolf Gehrisch — relations internationales, Académie des technologies
Hannu Myllykallio — école polytechnique
Victor Norris — université de Rouen
Carlo Reita — CEA-Leti, Grenoble

Le Groupe de Travail « ADN: lire, écrire, stocker l'information »

Méthodologie suivie

Ce groupe de travail a été pensé lors de discussions informelles au sein de l'Académie des technologies. L'opportunité tenait aussi à une double actualité parlementaire et gouvernementale : une stratégie bioéconomie française a été élaborée par quatre ministères [90] ; l'office parlementaire d'étude des choix scientifiques et technologiques (OPECST) a travaillé sur la modification ciblée des génomes [91]. Le projet de ce groupe a ensuite fait l'objet de présentations et validations successives en Commission « Biotechnologies » le 10/04/2018, en Comité des travaux le 17/05/2018, et en séance plénière le 13/06/2018. L'animation du groupe a alors été confiée à François Képès. Ce dernier a proposé Morgane Champleboux comme secrétaire scientifique. Comme prévu, la commission « Biotechnologies », présidée par Bernard Le Buanec, a ensuite cessé ses activités, cependant que le nouveau pôle « Alimentation et santé », sous la houlette de René Amalberti et Alain Boudet, reprenait en janvier 2019 le parrainage de ce groupe.

Le groupe de travail, marqué par une forte multi-disciplinarité, a déterminé sa méthode et les contours de son étude lors d'une première séance le 19/09/2018. La nouveauté du sujet a permis de l'appréhender assez globalement. Le groupe a choisi les personnalités à auditionner sur des bases multidisciplinaire et internationale. Il a ensuite procédé aux auditions de vingt-six personnalités venant de dix nations au cours de quatorze séances détaillées dans une annexe précédente. Ces auditions étaient généralement suivies d'une discussion entre membres du groupe, menant à des décisions collectives. Le groupe a tenu une séance conclusive le 7/01/2020. Les questions éthiques ont été traitées en sollicitant plusieurs fois les intervenants et les membres du groupe pour réfléchir avec l'aide du « Questionnaire éthique et technologie » ***ad hoc*** validé par l'Académie des technologies, entre septembre 2019 et janvier 2020. Les rares retours ont été exploités par F. Képès ; ils ne portaient que sur les questions de sécurité liées à l'approche d'archivage *in vivo* (cf. chapitre IV).

Après amendements et validation, pour chaque audition ont été mis à disposition des membres du groupe depuis la page web de l'Académie des technologies dédiée :

- le verbatim ;
- le diaporama ;
- la synthèse écrite de l'exposé ;
- le relevé des décisions prises par les membres présents.

Restitutions

Les restitutions du travail de groupe ont été :

90 Voir le rapport gouvernemental *Une stratégie Bioéconomie pour la France*, 2017. https://agriculture.gouv.fr/une-strategie-bioeconomie-pour-la-france-plan-daction-2018-2020

91 Voir le rapport OPECST *La révolution de la modification ciblée du génome (genome editing)*, 2017. Synthèse : http://www.assemblee-nationale.fr/14/cr-oecst/4618_synthese.pdf rapport : http://www.senat.fr/rap/r16-507-1/r16-507-11.pdf

- lettre trimestrielle de la Fondation de l'Académie des technologies (22/01/2020) rédigée par François Képès, Morgane Champleboux, Carlo Reita et le groupe (16 pages) ;
- exposé de trois heures en séance plénière thématique (22/01/2020) par Carlo Reita et François Képès ;
- résumé et recommandations (02/2020) par le groupe (2 pages) ;
- rapport synthétique (le présent document) lisible par un large public, incluant le résumé et les recommandations ci-dessus (07/2020) par le groupe (85 pages), après relecture critique par le comité des travaux et le comité de la qualité ;
- traduction du rapport synthétique en langue anglaise (08/2020) par Wolf Gehrisch, Victor Norris et François Képès (85 pages) ;
- colloque dans le grand amphithéâtre du CNRS (26/10/2020) organisé par le CNRS et l'Académie des technologies, avec Bruno Jarry, Patrick Ledermann, Patrick Maestro.

Ce colloque clôturera deux années d'activités du groupe de travail, sans préjuger de leurs répercussions plus lointaines, favorisées par la large diffusion de chaque restitution. En particulier, la version anglophone du rapport synthétique étendra la visibilité des travaux du groupe et de l'Académie des technologies, d'autant que ce sera probablement le premier rapport d'envergure sur ce thème novateur.

Acronymes utilisés

3D	tridimensionnel
A	adénine
ADN	acide désoxyribonucléique
ARN	acide ribonucléique
AXN	acide xéno-nucléique
bit	binary digit (chiffre binaire)
C	cytidine
CCO	Chief commercial officer (directeur commercial)
CD	compact disk (disque compact)
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CEO	Chief executive officer (directeur exécutif)
CERN	Organisation européenne pour la recherche nucléaire
CNRS	Centre national de la recherche scientifique
CRISPR	Clustered regularly interspaced short palindromic repeats (courtes répétitions en palindrome regroupées et régulièrement espacées)
CTO	Chief technical officer (directeur technique)
dATP	désoxyadénosine triphosphate
dCTP	désoxycytosine triphosphate
ddATP	di-désoxyadénosine triphosphate
ddCTP	di-désoxycytosine triphosphate
ddGTP	di-désoxyguanosine triphosphate
ddTTP	di-désoxythymidine triphosphate
dGTP	désoxyguanosine triphosphate
DRAM	dynamic random access memory (mémoire vive dynamique)
dTTP	désoxythymidine triphosphate
DVD	digital versatile disk (disque versatile digital)
EBI	European bioinformatics institute (Institut européen de bioinformatique)
EIC	European innovation council (Conseil européen de l'innovation)
ETH	Eidgenössische technische hochschule (Institut de technologie fédéral)
FfAME	the Foundation for applied molecular evolution (la Fondation pour l'évolution moléculaire appliquée)
g	gramme

G	guanine
GPU	Graphic processor unit (unité de traitement graphique)
HDD	Hard disk drive (disque dur)
IDC	Interdisciplinary center (centre interdisciplinaire)
IDC	International Data Corporation
JPEG	Joint photographic expert group (Groupe mixte d'experts de la photographie)
m	mètre
mL	millilitre
mm	millimètre (10-3 mètre)
MEMS	Microelectromechanical system (Microsystème électromécanique)
MOSLA	Molecular storage for long-term archiving (Stockage moléculaire pour l'archivage long terme)
MoSS	Molecular storage system (Système de stockage moléculaire)
mp3	MPEG audio layer 3 (Couche 3 audio du MPEG)
MPEG	Moving picture experts group (Groupe d'experts de l'image animée)
MRAM	Magnetic random access memory (Mémoire vive magnéto-résistive)
NAND (gate)	(porte logique) NON-ET
NHGRI	National human genome research institute (institut national de recherche sur le génome humain)
NIH	National institutes of health (Instituts nationaux de la santé)
nm	nanomètre (10-9 mètre)
o	octet
OPECST	Office parlementaire d'évaluation des choix scientifiques et technologiques
PCM	Phase-change memory (Mémoire vive non-volatile)
PCR	Polymerase chain reaction (Réaction en chaîne de la polymérase)
PDF	Portable document format (Format portable de document)
R&D	Recherche et développement
SGD	Sphère globale des données
SI	Système international
SRAM	Static random access memory (Mémoire vive statique)
T	thymine
TdT	Terminal désoxynucleotidyl transférase
USB	Universal serial bus (Bus série universel)
US$	Dollars des États-Unis d'Amérique
W	watts

Préfixes du Système international (SI) d'unités, et nombres correspondants

préfixe	abréviation	notation ingénieu-rale	notation directe
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

Les centres de données, « cloud » inclus, stockent les mégadonnées *(big data)* numériques de l'humanité sur disques durs et bandes magnétiques dont la durée de vie limitée oblige à de dispendieuses recopies tous les cinq à sept ans ; ils représentent des « gouffres » pour les ressources en terrain, électricité, eau et matériaux rares. En comparaison, le stockage à l'échelle moléculaire sur un polymère tel que l'ADN, permettrait une densité supérieure d'un facteur dix millions, une conservation prolongée d'un facteur dix mille sans recopie périodique, pour une consommation électrique quasi-nulle. En effet, l'ADN est stable à température ordinaire durant plusieurs millénaires et il peut être aisément dupliqué ou volontairement détruit. Les technologies requises existent. Cependant, pour devenir viables pour l'archivage de l'information, ces technologies nécessitent encore des progrès qui pourraient voir le jour sous cinq à vingt ans et seraient favorisés par des synergies entre secteurs public et privé.