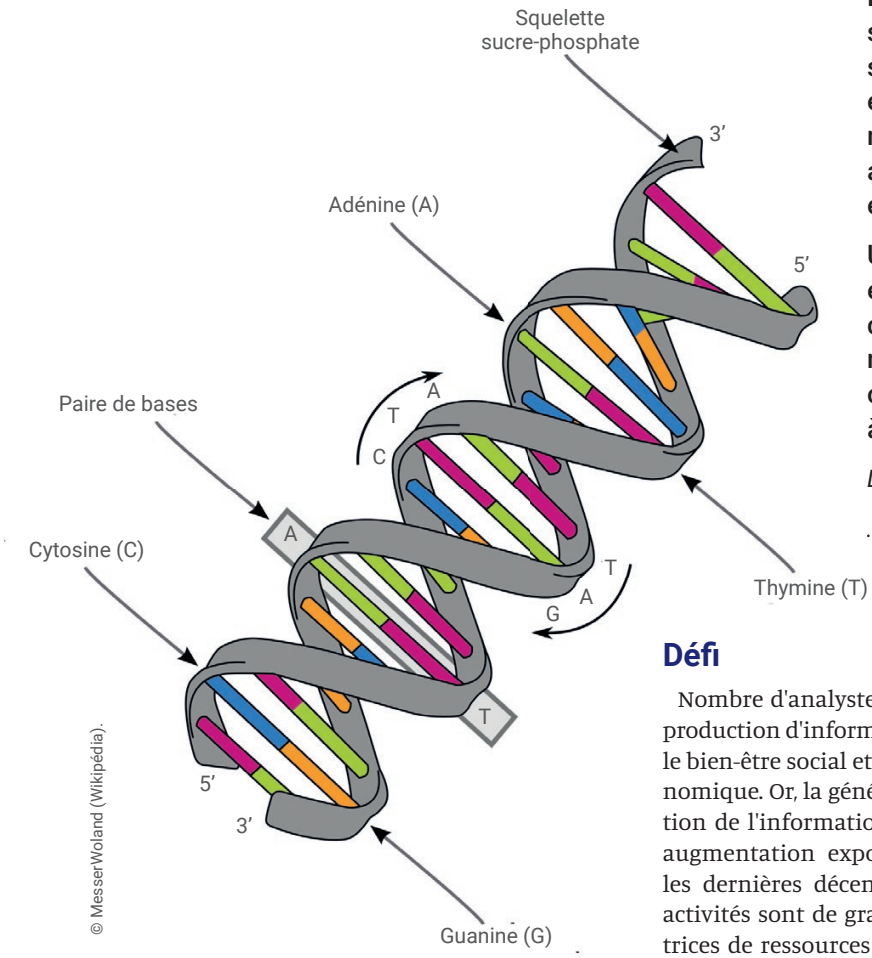
Archiver les mégadonnées numériques à l'échelle moléculaire

François Képès (Francois.Kepes@academie-technologies.fr)
Académie des technologies et Académie d'agriculture de France



Structure de la double hélice d'ADN, où sont indiquées les quatre bases azotées : adénine (en vert), cytosine (en bleu), guanine (en orange) et thymine (en violet).

L'information a été le moteur de la croissance socio-économique de la civilisation depuis ses débuts. Actuellement, son stockage, archivage et traitement par les centres dédiés n'offre plus de marges suffisantes d'optimisation pour faire face au déluge des données numériques et à son problématique impact environnemental.

Un récent rapport de l'Académie des technologies explore une alternative prometteuse au modèle conventionnel : l'archivage des mégadonnées numériques à l'échelle moléculaire dans l'ADN ou d'autres polymères, un chantier pour les vingt ans à venir.

Les acronymes utilisés sont définis dans le glossaire, p. 36.

Défi

Nombre d'analystes estiment que la production d'information corrèle avec le bien-être social et la croissance économique. Or, la génération et l'utilisation de l'information ont connu une augmentation exponentielle durant les dernières décennies. Comme ces activités sont de grandes consommatrices de ressources et d'énergie, elles sont devenues non soutenables au fil des ans, et ne seront plus supportables d'ici 2040. Il est donc essentiel de diminuer leur impact sur notre environnement. Cet article examine la piste « moléculaire », après avoir évoqué les défis que présente la situation actuelle.

De quelles données parlons-nous ? 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. Qu'est ce que représenterait l'ensemble des données accumulées par l'humanité, la « sphère globale des données » ? Notre unité sera l'octet, soit une suite de huit nombres '0' ou '1' dans un fichier qualifié de numérique pour cette raison.

Article disponible sur le site https://www.refletsdelaphysique.fr ou https://doi.org/10.1051/refdp/202170032

> **“ Pour stocker un *bit*, l’ADN permet de n’utiliser qu’une cinquantaine d’atomes, alors que le stockage magnétique en requiert environ un million.”**

La sphère globale des données comprend actuellement autant d'octets qu'il y a d'étoiles dans l'univers observable : quelques dizaines de milliers de milliards de milliards d'octets. Pour transférer tout cela par Internet à une vitesse moyenne, il faudrait cinquante millions d'années. Pourquoi ces données sont-elles de nature numérique ? Nos vieilles photos « argentiques » ne sont pas de cette nature. Mais tandis qu'elles virent inexorablement au sépia, nous mitraillons avec des appareils numériques. Le format numérique a l'avantage de pouvoir être traité directement par l'informatique, et aussi d'être en principe inaltérable. Pourquoi notre sphère globale des données augmente-t-elle encore ? Elle double tous les deux à trois ans environ [1], soit un facteur cent à mille en vingt ans, parce que son usage s'intensifie dans ses domaines traditionnels, et s'étend à de nouveaux secteurs : diagnostic et chirurgie à distance, voitures autonomes, capteurs, télésurveillance, réalité virtuelle. Notons incidemment que l'actuelle pandémie a encore augmenté nos interactions numériques : télétravail, téléconférences, téléconsultations. Comment conserve-t-on toutes ces données ? De moins en moins sur son propre ordinateur ou téléphone. De plus en plus dans des centres de données qui fonctionnent au sein de réseaux mondiaux de transmission. Ceci inclut le "cloud" ou « nuage », qui n'en est qu'une modalité plus virtuelle et automatisée, offrant des ressources à la demande.

Qu'est-ce qu'un centre de données ? Typiquement, un gros centre de données moderne contient un exaoctet (1 Eo : un milliard de milliards d'octets) ; un million de serveurs (des ordinateurs sans clavier ni écran) ; des disques durs ou mémoires statiques, et des bandes magnétiques. Il loge dans un immense hangar planté sur un terrain de plusieurs centaines de milliers de m^2. Il consomme plus d'électricité qu'une ville de cent mille habitants, dont environ 40% pour le refroidissement des serveurs (d'où l'intérêt de positionner ces centres dans des régions froides du globe) ; des milliers de tonnes de métaux chers, de plastiques issus du pétrole, et des terres rares, ainsi que des ressources couteuses à produire comme le silicium de grade électronique ; soit un investissement de quelques milliards d'euros [2]. Sa durée de vie est de vingt ans. 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. Le marché global des dispositifs de stockage de données (disques, bandes) est en 2020 de 57 milliards de dollars, en augmentation de 2 à 3% par an. Il faut ajouter au tableau les réseaux de connexions qui relient le centre au reste du monde, également consommateurs de ressources.

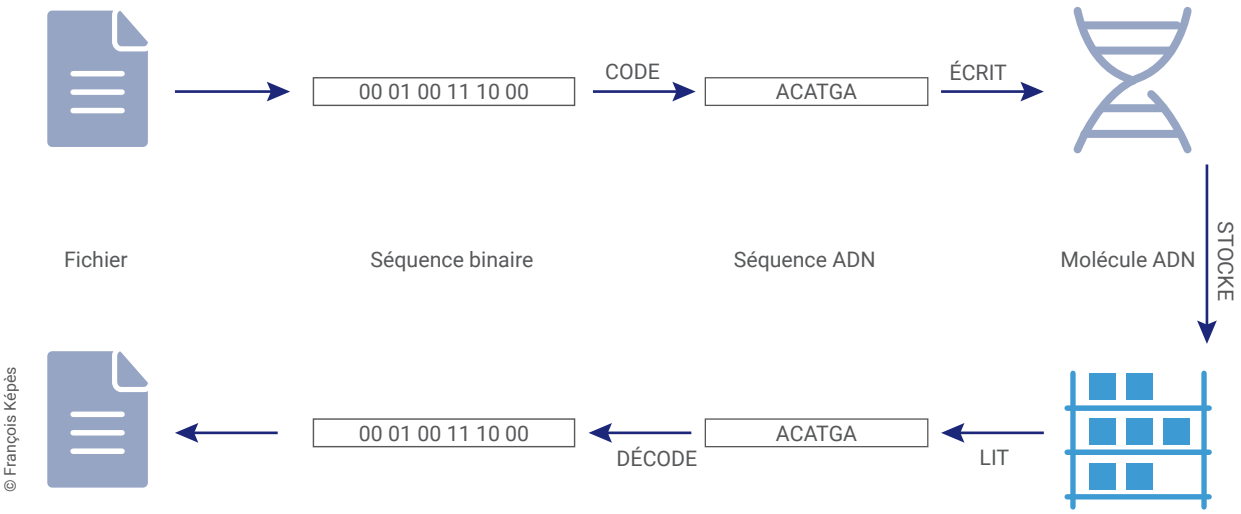
Combien existe-t-il de centres de données ? Plusieurs millions, en incluant ceux des compagnies. Mais s'ils contenaient tous un exaoctet de données, il en suffirait aujourd'hui d'environ 50 000 sur la planète. Ensemble, ces millions de centres couvrent actuellement un millionième des terres émergées de la planète. Avec les réseaux, ils consomment 2 à 4 % de l'électricité dans les pays avancés ; si tous ces centres formaient un seul pays par hypothèse, alors ce pays serait le cinquième plus gros consommateur d'électricité au monde, entre l'Inde et le Japon.

Sachant qu'en 2040 nous aurions cent à mille fois plus de données à conserver, ces chiffres démontrent que le modèle actuel de conservation sera d'ici là devenu insupportable au plan environnemental. Si encore il rendait les services souhaités ! Mais ce n'est plus le cas. Il a été calculé que, déjà en 2020, l'offre en capacité de stockage ne représente que 60% de la demande. Ce pourcentage chute rapidement avec le temps. Même en ignorant l'impact environnemental, il reste donc que le système actuel ne remplit même plus son office.

Pistes

Les réponses face à ce redoutable défi multiforme sont de plusieurs ordres. Un chantier de vingt ans se dresse devant nous pour révolutionner ces technologies. L'Académie des technologies s'est penchée sur cette question durant deux années et a publié un rapport fin 2020 [3]. Son groupe de travail s'est focalisé sur les supports moléculaires porteurs d'information, tel que l'acide désoxyribonucléique (ADN) utilisé ici comme agent chimique en dehors du vivant, ou d'autres polymères non-ADN très prometteurs. En effet, de nouvelles méthodes de stockage d'information à l'échelle moléculaire sont envisagées afin d'augmenter la capacité de stockage et la durée de conservation des données, ainsi que de réduire la taille des supports. Les progrès récents des technologies d'écriture et de lecture de l'ADN ont amené les chercheurs à envisager ce polymère naturel comme support d'archives numériques. L'ADN permet optimalement de stocker un *bit* pour environ cinquante atomes, alors que le stockage magnétique requiert environ un million d'atomes, support mécanique non compris.

>>>



1. Étapes du processus de stockage des mégadonnées numériques sur l'ADN.
Ici sont représentés pour exemple douze *bits* successifs extraits du fichier numérique. Ces douze *bits* sont codés sous la forme de six 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.

>>>

Bref historique du stockage de données dans 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 a conçu et synthétisé un fragment d'ADN de 18 nucléotides (les monomères de l'ADN symbolisés par 'A', 'C', 'T' et 'G') contenant un message numérisé, qu'il a ensuite transféré chez une bactérie intestinale, le colibacille. En 2012, l'équipe de George M. Church (Université de Harvard, États-Unis d'Amérique) a stocké 0,6 Mo d'information sur l'ADN, sous forme de fragments synthétiques. En 2013, l'équipe de Nick Goldman (Institut Européen de Bioinformatique, Royaume-Uni) a retranscrit sans erreur quatre fichiers en séquence d'ADN, pour un total de 0,7 Mo. En 2018, Microsoft Corp. et l'Université de Washington, aux États-Unis d'Amérique, ont stocké sur l'ADN un Go d'informations venant de fichiers de types variés. Depuis, ils détiennent le record qui a fait l'objet d'une publication validée. En 2024, il est projeté d'archiver un To (équivalent à environ 1000 films) en vingt-quatre heures pour un cout de 1000 US$ [4].

Avantages du stockage moléculaire

L'ADN dans le monde vivant est un des supports de l'information héréditaire. Pour rappel, l'ADN est formé de deux brins antiparallèles enroulés l'un autour de l'autre pour former une structure en double hélice (voir la figure, p. 32). Chaque brin d'ADN est un polymère 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 interagissent avec les bases nucléiques de l'autre brin à travers des liaisons hydrogène : l'adénine et la thymine s'apparient avec deux liaisons hydrogène, tandis que la guanine et la cytosine s'apparient avec trois liaisons hydrogène.

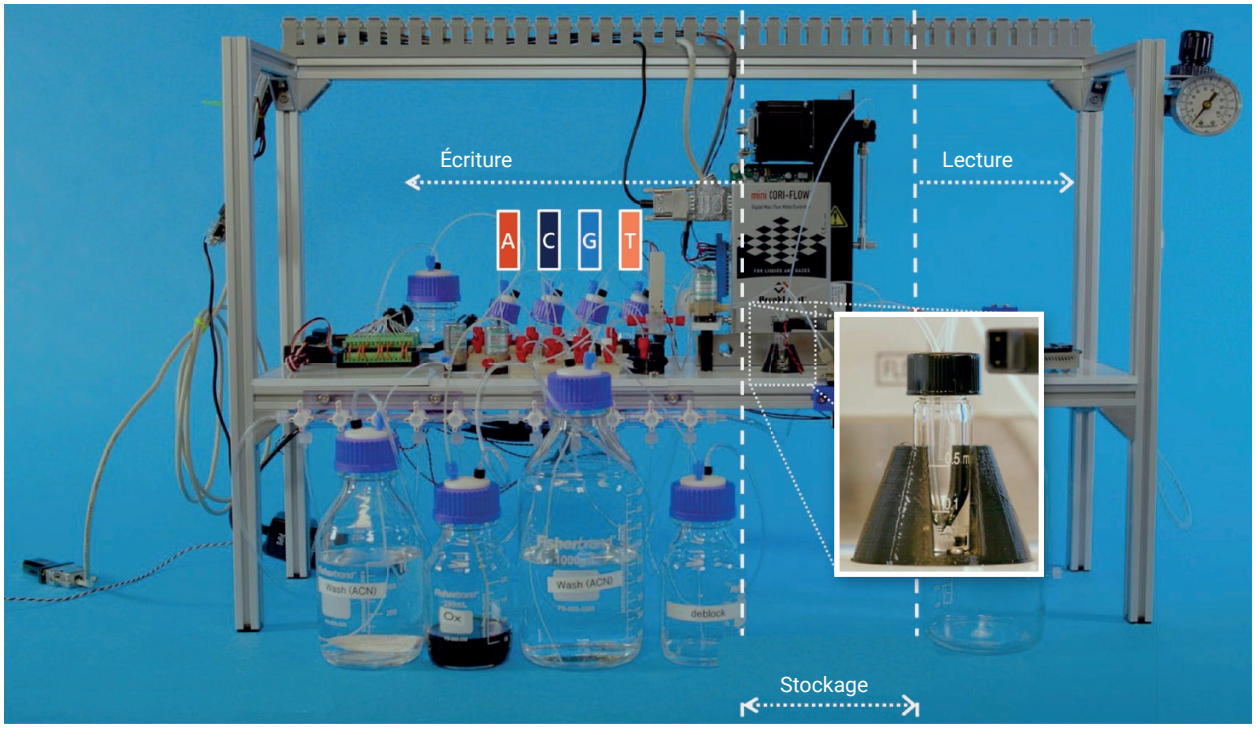
L'ADN peut, depuis sa découverte en 1869 (Friedrich Miescher), être manipulé en dehors des cellules, *in vitro* ; c'est principalement *in vitro* qu'il a été envisagé de l'utiliser pour stocker des données numériques. Il présente à ce titre de nombreux avantages, comparé aux systèmes traditionnels.

Potentiellement, l'ADN permet des densités informationnelles dix millions de fois supérieures aux mémoires traditionnelles : la sphère globale actuelle des données tiendrait dans une fourgonnette. L'ADN est stable à température ordinaire durant plusieurs millénaires. Il peut être aisément multiplié ou détruit à volonté. En outre, 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 l'ADN, qui font partie intégrante de la médecine moderne.

Pratique du stockage moléculaire

Pour archiver et retrouver des données dans l'ADN, il convient d'enchainer cinq étapes (fig. 1) : 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 (fig. 2) réalisant ces opérations fonctionne depuis mars 2019 chez Microsoft aux États-Unis [5]. L'Académie des technologies a évalué les techniques actuelles pour chacune de ces cinq étapes.

Codage et décodage sont des opérations informatiques simples faisant correspondre un nucléotide à deux *bits* (arbitrairement, 'A' peut être attribué à '00' ; 'C' à '01' ; 'G' à '10' ; 'T' à '11'). Quoique des codages plus sophistiqués existent, celui-ci fonctionne parfaitement (fig. 3).

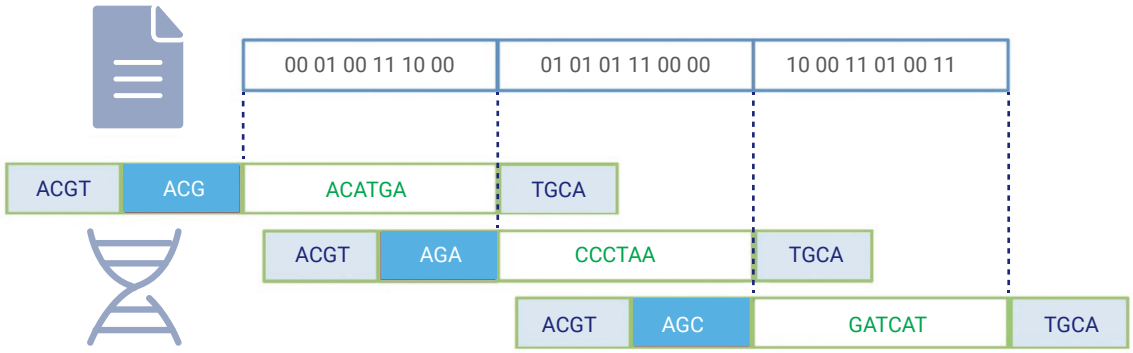


© Microsoft Corp. / Université de Washington, États-Unis. Adapté par François Képès.

2. Premier prototype entièrement automatisé de stockage de données sur l'ADN. Parmi les cinq étapes décrites en figure 1, le codage et le décodage sont opérés par un ordinateur non présenté dans cette photographie. Les trois autres étapes sont effectuées sur le prototype montré ici : écriture par synthèse chimique traditionnelle, stockage sur nanobilles de verre, et lecture par séquençage à travers des nanopores. Les flacons de verre contiennent les réactifs nécessaires à réaliser l'opération de synthèse d'ADN décrite dans le texte, qui est un cycle à trois stades : incorporation/lavage/déprotection. Le flacon agrandi contient les nanobilles de stockage.

Écriture et lecture sont des opérations qui diffèrent fortement de celles prenant place dans la cellule vivante. Depuis 1983 jusqu'à présent, l'écriture fait appel à la synthèse d'ADN par voie chimique. Le prochain nucléotide à incorporer dans l'ADN synthétique fixé sur support solide est fourni sous forme de phosphoramidite. Après incorporation, l'excédent de réactif est lavé. Puis le nucléotide nouvellement incorporé est « déprotégé », autrement dit le groupe chimique qui empêchait qu'il soit incorporé en de multiples exemplaires est retiré. Ce cycle à trois stades, incorporation/lavage/déprotection, est répété pour le nucléotide suivant, etc. Le second brin est synthétisé complémentaire et antiparallèle au premier, puis les deux brins sont mélangés et s'hybrident en reconstituant la double hélice d'ADN. Les trois sérieux inconvénients de cette méthode de synthèse chimique de l'ADN sont les suivants : 1) elle se fait en solvant organique ; 2) elle fait appel à la chimie des phosphoramidites dont la production use des ressources limitantes et polluantes ; 3) son taux d'erreur ou de non-incorporation est au mieux de 0,1%, ce qui limite en pratique la longueur utile des segments d'ADN synthétisés à 200 nucléotides. En soi, la limite de longueur n'est pas rédhibitoire, car un ensemble de segments courts permet par indexation de reconstituer l'information complète. En revanche, le taux d'erreur reste trop élevé. Aussi l'Académie des technologies a estimé que la synthèse chimique d'ADN ne pourra pas passer à l'échelle requise par les mégadonnées.

>>>



© François Képès

3. 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 sur fond blanc ; à l'heure actuelle typiquement 75% du total). Les autres éléments (bleu foncé sur fond gris et blanc sur fond bleu) permettent l'indexation et la correction d'erreurs selon des méthodes directement issues de l'informatique.

>>>

L'alternative récente, proche de la commercialisation, est une approche par synthèse dite « enzymatique » d'ADN, qui suit le même cycle à trois stades. Le catalyseur de cette synthèse est une enzyme d'origine naturelle qui a été largement modifiée pour cet usage *ex vivo*. La synthèse enzymatique présente un meilleur potentiel pour l'avenir de ce domaine : 1) elle se pratique en phase aqueuse ; 2) les nucléotides sont aussi protégés par un groupe chimique pour éviter d'être multiplement incorporés, mais la synthèse chimique de ces monomères n'a pas les inconvénients de celle des dérivés phosphoramidites ; 3) le taux d'erreur ou de non-incorporation est quasi égal aujourd'hui à celui de la synthèse chimique, mais baisse plus rapidement.

Depuis 1977, la lecture opère par séquençage de l'ADN. La méthode la plus récente (2015), par nanopores [6], présente le meilleur potentiel car elle permet de longues lectures d'ADN d'un seul tenant (des millions de nucléotides), contrairement aux méthodes classiques limitées à quelques centaines de nucléotides. Chaque dispositif contient plusieurs centaines de protéines d'interface, appelées nanopores, préalablement insérées au travers d'une membrane synthétique. En appliquant une différence de potentiel de part et d'autre de la membrane synthétique, un flux ionique est mesuré en temps réel pour chaque acide nucléique qui traverse un nanopore. Le flux ionique est mesuré en picoampères, ce qui représente une sensibilité de l'ordre de deux atomes d'hydrogène. Chaque nucléotide possède ainsi un signal électrique reproductible et spécifique, qui est enregistré lors de son passage, ce qui permet de reconstituer la séquence du polymère. 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 très excessive d'un million de monomères par seconde. Afin de contrôler le flux à 450 monomères par seconde et de séquencer en temps réel, l'ADN natif de départ est lié à l'entrée du nanopore 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 pendant deux jours environ. Les nanopores sont utilisés en parallèle et, au total, avec un dispositif tenant dans la paume de la main et doté d'une interface USB3, jusqu'à 20 Go de données peuvent être lus en deux jours.

Concernant le stockage proprement dit, la méthode la plus performante est celle des capsules de la société Imagene [7]. L'extérieur de ces capsules est en acier inoxydable, l'intérieur en borosilicate ; elles ont la taille d'une pile bouton et sont à usage unique. Chaque capsule contient jusqu'à 0,8 g d'ADN (potentiellement 1,4 Eo de données en tenant compte de la redondance) en atmosphère neutre. À température ambiante — donc sans consommation d'énergie ou autre ressource — la demi-vie de l'ADN est estimée à 52 000 ans dans ces capsules qui protègent l'ADN de ses trois ennemis que sont l'eau, l'oxygène et la lumière. D'autres méthodes moins protectrices mais plus compactes existent, comme celle utilisant des nanobilles de silice qui préservent l'ADN seulement quelques années [8].

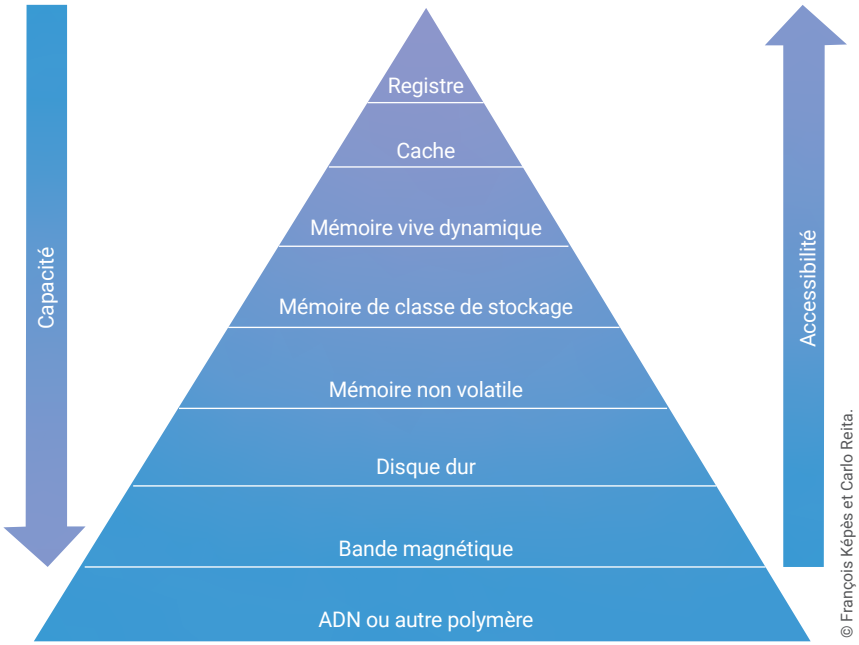
« Polymères numériques » non-ADN

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, donc par chimie itérative en phase solide. Il faut aussi qu'il existe des méthodes pour le conserver longtemps et le lire aisément. Le projet académique porté par Jean-François Lutz (Institut Charles Sadron, Université de Strasbourg) utilise des hétéropolymères non-ADN pour stocker de l'information numérique [9]. Ces polymères de synthèse ont un potentiel considérable pour le stockage moléculaire. Ils permettent une plus grande densité d'information par l'élargissement de l'alphabet, et une meilleure conservation des données que les supports de stockage électroniques actuels. Pour lire l'information, les polymères sont « séquencés » par spectrométrie de masse, technique qui permet de détecter, identifier et caractériser les molécules d'intérêt. Le résultat, appelé spectre de masse, est décodé informatiquement pour reconstituer le message en *bits*, et donc l'information initiale. La spectrométrie de masse est rapide et très précise. Son inconvénient réside dans la taille importante de la machine utilisée. Il serait intéressant de pouvoir séquencer certains polymères numériques par la technique des nanopores décrite plus haut, qui met en œuvre des machines portables et versatiles. Ceci suppose que les monomères se distinguent par leur signal électrique.

Pour finir, mentionnons plusieurs approches moins radicales, conservant la structure de l'ADN en double brin. Il est possible de les classer en deux

ACRONYMES UTILISÉS

A	adénine
ADN	acide désoxyribonucléique
C	cytidine
CERN	Organisation Européenne pour la Recherche Nucléaire
Eo	Exa-octet (10^{18} octets)
g	gramme
G	guanine
Go	Giga-octet (10^{9} octets)
m	mètre
Mo	Mega-octet (10^{6} octets)
o	octet
Po	Peta-octet (10^{15} octets)
T	thymine
To	Tera-octet (10^{12} octets)
US$	Dollars des États-Unis d'Amérique



4. 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.

catégories : celles qui substituent un autre sucre au désoxyribose, et celles qui usent d'autres bases azotées que A, C, G et T [10].

Perspectives

Au-delà des approches théoriques, empiriques ou éducatives visant à limiter la quantité d'information engendrée par l'humanité, son archivage moléculaire constitue un enjeu majeur et stratégique à horizon proche.

À l'heure actuelle, le stockage d'information dans l'ADN reste expérimental, mais sa preuve de principe est acquise. 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 du patrimoine culturel (par exemple films, livres, monuments), scientifique (expériences de physique nucléaire — le CERN conserve environ 100 Po de données expérimentales pour les physiciens de la prochaine génération) et d'informations sensibles (renseignement) ou imposées par la règlementation (banques). Ces applications valorisent certains atouts maitres du stockage moléculaire : densité informationnelle, longévité, durabilité, économie en énergie et autres ressources, facilité à multiplier l'ADN et instantanéité de sa destruction volontaire. Elles ne souffrent que peu des actuels handicaps de cette solution puisque l'écriture se produit une fois, les modifications jamais, et la lecture — bien moins couteuse que la synthèse — occasionnellement.

Pour entrer en compétition avec les marchés plus globaux de l'archivage des mégadonnées numériques, il faudra peut-être dix ou vingt ans. En ce cas, le stockage sur l'ADN entrera en complémentarité puis en compétition avec la bande magnétique, actuellement la solution de choix pour l'archivage à long terme (fig. 4). La viabilité économique de cette approche ne sera atteinte qu'en améliorant les couts et vitesses des technologies de l'ADN : d'un facteur mille pour la lecture, et cent millions pour l'écriture. Ces facteurs peuvent sembler rédhibitoires. Ce serait oublier la célérité des progrès des technologies de l'ADN, proches d'un doublement tous les six mois, soit d'un facteur 1000 tous les cinq ans, donc bien plus rapides que l'évolution des capacités de stockage magnétique ou électronique qui jusqu'à présent doublent environ tous les deux ans. ▮

1• D. Reinsel *et al.*, *The Digitization of the World – From Edge to Core*, International Data Corporation & SeaGate (2018). https://cutt.ly/seagate-whitepaper-pdf

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

3• *Archiver les mégadonnées au-delà de 2040 : la piste de l'ADN*, rapport de l'Académie des technologies (2020). www.academie-technologies.fr/blog/categories/publications-de-l-academie/posts/archiver-les-megadonnees-au-dela-de-2040-la-piste-de-l-adn

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

5• C.N. Takahashi *et al.*, "Demonstration of End-to-End Automation of DNA Data Storage", *Scientific Reports* **9** (2019) 4998.

6• https://nanoporetech.com

7• J. Bonnet *et al.*, "Chain and conformation stability of solid-state DNA: implications for room temperature storage", *Nucleic Acids Res.* **38** (2010) 1531-1546. www.imagene.fr/dnashell-rnashell/dnashell/

8• W.D. Chen *et al.*, "Combining Data Longevity with High Storage Capacity –Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles", *Advanced Functional Materials* **29** (2019) 1901672.

9• H. Colquhoun et J.F. Lutz, "Information-containing macromolecules", *Nature Chemistry* **6** (2014) 455-456.
A. Al Ouahabi *et al.*, "Synthesis of non-natural sequence-encoded polymers using phosphoramidite chemistry". *J. Am. Chem. Soc.* **137** (2015) 5629-5635.
R.K. Roy *et al.*, "Design and synthesis of digitally encoded polymers that can be decoded and erased", *Nature Communications* **6** (2015) 7237.

10• J.C. Chaput *et al.*, "Orthogonal Genetic Systems", *Chembiochem* **21** (2020) 1408-1411. H. Hoshika *et al.*, "Hachimoji DNA and RNA: A genetic system with eight building blocks", *Science* **363** (2019) 884-887.