LES NOTES SCIENTIFIQUES DE L'OFFICE
OFFICE PARLEMENTAIRE D'ÉVALUATION DES CHOIX SCIENTIFIQUES ET TECHNOLOGIQUES

Note n° 29

Le stockage de données sous la forme d'ADN

Décembre 2021

N° 4793 ASSEMBLÉE NATIONALE – N° 285 SÉNAT

© iStock_ym german

Résumé

- *Le passage à l'ère numérique a été associé à une production toujours plus importante de données à conserver, posant aujourd'hui plusieurs problématiques, notamment énergétiques, environnementales ou de sécurité.*
- *L'utilisation d'ADN pour stocker des données pourrait constituer une innovation de rupture, susceptible de répondre à plusieurs de ces problématiques.*
- *Cette technologie ne pourra cependant pas constituer une solution à elle seule et devra vraisemblablement s'accompagner de changements de nos pratiques.*

M. Ludovic Haye, sénateur

L'impasse du stockage de données

- **Des données toujours plus importantes…**

L'ère informatique, qui fait suite à la révolution numérique intervenue dans les années 1970, est parfois qualifiée d'ère de l'information. En effet, les nouvelles technologies ont profondément modifié notre rapport à l'information, en permettant une croissance exponentielle des données échangées et stockées. Ainsi, le volume de données conservées dans les *data centers*[1] devrait être amené à croître toujours plus rapidement avec l'augmentation du nombre d'utilisateurs, la multiplication des terminaux et le développement de nouvelles pratiques (internet des objets, approches « *cloud* » et « *big data* »…) ; la sphère globale des données, de 33 zettaoctets (Zo, 10^{21} octets) en 2018, devrait passer à 175 Zo en 2025[2], et pourrait même atteindre 5 000 Zo en 2040[3].

- **… face à des supports de stockage dépassés**

Cette croissance n'est cependant pas sans conséquences. Sous-estimés par les utilisateurs du fait de l'invisibilité des infrastructures utilisées, les impacts environnementaux du numérique sont pourtant loin d'être anecdotiques : en 2019, ce secteur représentait 3,6 % de la consommation énergétique mondiale et devrait atteindre une part comprise entre 4,8 et 5,9 % en 2025[4]. Les *data centers*, traitant et stockant les données numériques, représentaient 19 % de la consommation d'énergie du numérique en 2017[5], et nécessitent des métaux et terres rares pour leur construction ainsi que d'importantes quantités d'eau pour leur refroidissement.

Malgré l'amélioration des performances des *data centers*, les gains d'efficacité énergétique ne devraient très probablement pas suffire à compenser l'accroissement exponentiel des usages, conduisant par conséquent à une augmentation de l'empreinte carbone liée au stockage des données[6], et ce, d'autant plus que les technologies de stockage actuelles s'approchent de leur optimum théorique[7], laissant présager un ralentissement prochain des gains d'efficacité[6].

En outre, la capacité de produire des équipements en quantité suffisante pour répondre à l'accroissement prévu de la demande de stockage est aujourd'hui remise en question[8,9]. Aussi, une innovation de rupture dans le domaine du stockage de données apparait nécessaire.

L'ADN, une solution séduisante et prometteuse

- **Le disque dur de la nature…**

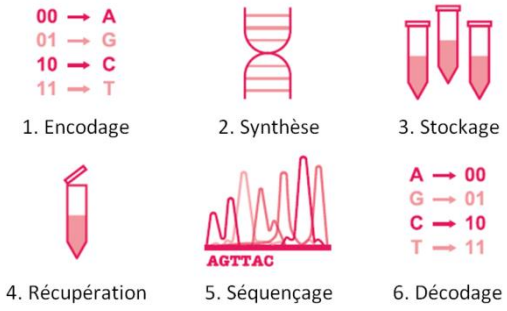
Depuis plusieurs milliards d'années, les molécules d'ADN sont utilisées par les êtres vivants pour le stockage de leur information génétique. Ces molécules sont constituées de deux brins antiparallèles, enroulés l'un autour de l'autre pour former une structure en double hélice et composés d'un assemblage de nucléotides. Chaque nucléotide inclut l'une des quatre bases azotées (adénine (A), cytosine (C), guanine (G), thymine (T)) liée à un sucre désoxyribose, lui-même lié à un groupe phosphate. Grâce à ces quatre bases azotées, l'ADN offre un système de numération quaternaire pour le codage d'information.

- **… et du futur ?**

L'élucidation de la structure moléculaire de l'ADN, récompensée par le prix Nobel de médecine en 1962, a

joué un rôle clef dans le développement de la biologie moléculaire. La seconde moitié du XX^{e} siècle a alors vu se développer des techniques pour la synthèse comme pour le séquençage[10] de l'ADN. Ainsi, puisque des technologies permettent aujourd'hui d'« écrire » comme de « lire » une séquence d'ADN, il est possible de s'inspirer de la nature et d'utiliser l'ADN comme un support de données.

Pour ce faire, la séquence binaire correspondant à un fichier à conserver doit être encodée en une séquence ADN (une suite de nucléotides), qui peut alors être synthétisée et stockée. Pour recouvrer le fichier, la molécule d'ADN doit être séquencée et cette séquence décodée.



Crédit : DNA Data Storage Alliance[12]

L'ADN pourrait présenter de nombreux avantages en tant que support pour le stockage de données. Tout d'abord, il offre une densité informationnelle considérable : une cinquantaine d'atomes seulement sont nécessaires pour stocker 1 bit, alors que le stockage magnétique en requiert 1 million[3,11]. La longévité de l'ADN est aussi l'un de ses points forts : stockée dans des conditions appropriées, une molécule d'ADN peut être conservée des dizaines de milliers d'années, alors que les données stockées sur des supports classiques nécessitent d'être recopiées tous les 5 à 10 ans pour éviter leur dégradation. En outre, en tant que support de l'information génétique, l'ADN n'est soumis à aucun risque d'obsolescence et restera un support universel. Enfin, on peut citer la grande facilité à dupliquer les données stockées sous la forme d'ADN à l'aide de technologies maitrisées basées sur des enzymes.

Ainsi, parmi les diverses alternatives qui ont pu être proposées pour faire face à la problématique du stockage de données, l'utilisation de l'ADN semble être l'une des plus prometteuses et est la première à mobiliser internationalement de nombreux acteurs académiques et industriels[12]. Néanmoins, il existe encore divers défis technologiques à relever avant de voir une réelle utilisation de l'ADN comme support d'information.

■ Développements actuels

Le développement d'un système de stockage utilisant l'ADN comme support fait appel à un large panel de disciplines scientifiques[13] et requiert de fait une synergie entre scientifiques, à la fois académiques et industriels, travaillant sur des thématiques variées. Les recherches se structurent donc actuellement autour de projets collaboratifs et interdisciplinaires comme le consortium états-unien *Molecular Information Storage* (MIST)[14], le programme européen OligoArchive[15] ou les projets français dnarXiv[16] et MoleculArXiv[17].

- **Méthode de codage**

En définissant la séquence qui sera synthétisée, stockée puis séquencée, la méthode de codage a une influence directe sur ces différentes étapes, et à travers elles sur l'ensemble du processus. Ainsi, le développement d'une méthode de codage efficace (minimisant le nombre de nucléotides nécessaires) permet à la fois d'aboutir à une plus importante densité informationnelle mais aussi à un processus global plus efficient (les étapes de synthèse et de séquençage étant à la fois lentes et coûteuses).

Le système de codage doit cependant être aussi développé de manière à minimiser l'occurrence d'erreurs lors des étapes de synthèse et de séquençage, et dépend donc directement des technologies utilisées. À cet effet, le système de codage doit notamment respecter certaines contraintes quant à l'ordre d'enchaînement et aux proportions des différentes bases azotées[18], et inclure des nucléotides de contrôle pour la détection et la correction d'erreurs. Il est en outre préférable de ne pas utiliser des brins d'ADN de plus de 200 nucléotides (pour lesquels les erreurs de synthèse sont, avec les technologies actuelles, plus fréquentes[19]) et donc de segmenter le fichier en différents fragments. La méthode de codage doit donc inclure un système d'indexation et d'adressage pour permettre de réassembler le fichier lors de l'étape de décodage[20]. Enfin, cette étape permet également le chiffrement des données stockées pour éviter tout problème de cybersécurité[21].

- **Synthèse**

Pour synthétiser des fragments d'ADN, la technique actuellement la plus répandue repose sur le principe de la synthèse séquentielle, où les nucléotides sont ajoutés un à un. Pour éviter que plusieurs nucléotides ne soient adjoints concomitamment, les nucléotides utilisés portent un groupement protecteur, empêchant l'ajout d'un autre nucléotide à sa suite. L'addition d'un nucléotide à la chaîne en construction peut se faire de différentes manières : par réaction chimique (la synthèse est alors dite « chimique ») ou par l'intermédiaire d'une enzyme, une désoxynucléotidyl transférase terminale[22] (on parle dans ce cas de synthèse « enzymatique »). Après chaque ajout, le groupement protecteur est clivé afin que le nucléotide suivant puisse à son tour être additionné.

Méthode de synthèse historique, la voie chimique a aujourd'hui été automatisée et miniaturisée ; la parallélisation permet de construire jusqu'à un million de fragments d'ADN différents comprenant 200 nucléotides en 24 heures[23]. La voie enzymatique, bien que non commercialisée à l'heure actuelle, est développée par plusieurs entreprises (dont notamment la société française DNA Script, leader du domaine) et devrait permettre une

réelle avancée : moins polluante, elle permet à la fois une plus grande vitesse de synthèse, un coût plus faible et un taux d'erreur amoindri.

Il est aussi possible, plutôt que d'ajouter les nucléotides un à un, d'assembler directement de courts segments d'oligonucléotides présynthétisés. Grâce à des extrémités simple-brin cohésives, ces segments peuvent être liés les uns aux autres par l'action d'une enzyme : une ADN ligase. Cette méthode permet de synthétiser de longs brins d'ADN, tout en conservant un faible taux d'erreur et en accélérant considérablement la vitesse de synthèse. Avec cette méthode, l'entreprise états-unienne Catalog DNA a développé une machine capable de synthétiser l'équivalent de 500 Ko/s[3].

L'étape de la synthèse est celle qui doit relever le plus grand nombre de défis pour espérer parvenir à une généralisation du stockage de données sous la forme d'ADN : sa vitesse doit encore être considérablement améliorée et les coûts qui lui sont liés drastiquement baissés. Si les efforts actuels se concentrent surtout sur la parallélisation massive des techniques existantes, le développement de nouveaux processus enzymatiques pourrait apporter d'importants progrès. En outre, les technologies actuelles ont été développées en vue d'applications médicales ; or, les besoins diffèrent entre ces applications (nécessitant un faible taux d'erreur, quitte à transiger sur les questions d'échelle, de coûts et de vitesse) et le stockage de données (pouvant s'accommoder d'un plus grand nombre d'erreurs mais pour lequel les questions d'échelle, de coûts et de vitesse sont primordiales). L'émergence récente d'entreprises totalement dédiées à l'utilisation d'ADN pour le stockage[24] devrait permettre de répondre à ces besoins spécifiques et de développer des solutions adaptées.

- **Stockage**

Pour être conservées sur de longues durées, les molécules d'ADN doivent être tenues à l'écart de l'eau et de l'oxygène mais aussi de la lumière et des hautes températures[25]. Les techniques classiques permettant d'éviter l'altération de l'ADN reposent sur un stockage à basse température – méthode à la fois coûteuse (en espace, équipement, énergie et maintenance) et exposant à des risques de perte en cas de dysfonctionnement – qui ne permet pas de tirer pleinement parti des bénéfices offerts par l'ADN comme support de stockage vis-à-vis des technologies actuellement utilisées.

Pour faire face à cette situation, l'entreprise française Imagene a développé des capsules en acier inoxydable, contenant un insert en verre, pour la conservation de molécules d'ADN (celles-ci étant préalablement séchées sous vide puis placées sous atmosphère inerte). Les extrapolations de dégradation pour cette technologie prévoient des durées de conservation jusqu'à plusieurs dizaines de milliers d'années à température ambiante. Le processus d'encapsulation a de plus été entièrement automatisé et pourrait donc être facilement intégré dans un système de stockage.

La possibilité de conserver les fragments d'ADN *in vivo* dans des cellules ou des organismes vivants a aussi été étudiée. Néanmoins, bien que certaines bactéries puissent survivre plusieurs millions d'années, la quantité d'informations qui peut être conservée dans chaque hôte est limitée, empêchant d'atteindre une importante densité informationnelle *via* cette méthode. De plus, ce mode de stockage pose la question de la tolérance des fragments injectés par l'organisme hôte, des séquences d'ADN pouvant se révéler toxiques pour le porteur ou le biotope environnant[26]. Pour ces raisons de performance et bioéthiques, le stockage *in vivo* n'a été étudié qu'à l'échelle expérimentale, les alternatives *in vitro* leur étant préférées.

- **Séquençage**

Au cours des dernières années, les technologies de séquençage ont réalisé de prodigieux progrès, bien plus rapides que ceux décrits par la loi de Moore[27], mise en évidence en informatique. Alors qu'il avait fallu entre 500 millions et 1 milliard de dollars américains pour séquencer le premier génome humain (opération achevée en 2001)[28], le coût actuel est proche de 700 dollars. Néanmoins, des améliorations – moins importantes que pour la synthèse – restent encore nécessaires pour envisager un réel développement du stockage de données sous la forme d'ADN.

La technologie la plus utilisée à l'heure actuelle, développée par la société Illumina, dérive directement de la méthode historique[29], améliorée et massivement parallélisée. La molécule d'ADN d'intérêt est d'abord fragmentée et convertie en ADN simple brin. Après greffage et amplification de ces fragments sur une surface solide, une ADN polymérase est utilisée pour construire les brins complémentaires en présence de nucléotides portant un fluorochrome clivable (d'une couleur différente pour chaque base azotée), servant également de groupement protecteur. Après chaque ajout, une photographie numérique est réalisée pour identifier la nature du nucléotide ainsi additionné. Le fluorochrome peut alors être retiré, pour permettre l'ajout d'un nouveau nucléotide. La séquence est *in fine* obtenue par traitement informatique à partir des différentes images, en utilisant les chevauchements pour réordonner les fragments.

Une innovation susceptible de devenir un réel atout pour le stockage de données sous la forme d'ADN, et utilisant un tout autre principe, a récemment été développée par l'entreprise Oxford Nanopore Technologies. Grâce à l'application d'un champ électrique, les fragments d'ADN d'intérêt – préalablement convertis en ADN simple brin – sont entraînés à travers une membrane contenant des nanopores. La mesure du flux ionique traversant chaque nanopore permet de déterminer en temps réel la nature du nucléotide le traversant. Bien qu'associée à un taux

d'erreurs plus élevé[30], cette nouvelle méthode s'avère bien plus rapide (il est actuellement possible de séquencer jusqu'à 450 nucléotides par seconde) et permet de s'affranchir des étapes d'amplification et de calcul informatique pour reconstituer la séquence.

- **Alternatives à l'utilisation d'ADN**

L'utilisation de polymères non-ADN (c'est-à-dire des polymères utilisant comme monomères des molécules autres que des nucléotides) est également étudiée pour le stockage d'informations numériques[31]. Le principal avantage de cette alternative repose sur la liberté fournie par le choix des monomères. En effet, ceux-ci peuvent être conçus de manière à atteindre une densité informationnelle plus importante que dans le cas de l'ADN[32], à conférer une plus grande stabilité ou des propriétés particulières au polymère utilisé, ou à faciliter les étapes de synthèse ou de séquençage qui sont aujourd'hui limitantes. Si cette approche apparait prometteuse et pourrait éventuellement surpasser les opportunités offertes par l'ADN, elle requiert encore d'importants efforts de recherche, les techniques de synthèse et de séquençage pour ces polymères n'étant à l'heure actuelle pas encore performantes et compétitives au regard de celles de l'ADN[33].

■ Perspectives

- **Des impacts modérés**

Pour devenir viable et être couramment utilisé, le stockage d'information sous forme d'ADN doit être complétement automatisé et inclure des étapes de synthèse et de séquençage aux coûts et durées mesurés. Si une première preuve de concept a pu être réalisée en 2019 par une équipe de l'Université de Washington et de Microsoft, avec un prototype capable de réaliser l'ensemble du processus de manière autonome[34], d'importantes avancées restent encore à réaliser. On estime actuellement qu'il faudrait réduire les coûts d'un facteur mille pour le séquençage et de cent millions pour la synthèse[3]. Bien que considérables, ces objectifs doivent être mis en regard des récentes avancées obtenues dans ces secteurs, permettant d'espérer des progrès significatifs dans les prochaines années. Le marché associé devrait d'ailleurs connaître une importante croissance, comme estimé par BCC Research[35].

Par ailleurs, les étapes d'écriture et de lecture de l'ADN devraient rester relativement lentes par rapport aux technologies actuellement utilisées pour le stockage de données numériques. De fait, l'ADN se positionnera – au moins initialement – comme support stockage de données « froides », c'est-à-dire les données nécessitant d'être conservées sur des temps longs mais n'ayant pas besoin d'être consultées ou modifiées régulièrement[36].

La plupart des données conservées n'étant que très rarement consultées après une centaine de jours, les données « froides » représentent 60 % des données numériques[37]. Cependant, les dispositifs à bandes magnétiques, qui sont le principal support pour la conservation de ces données à l'heure actuelle, ne représentent qu'une faible part de l'électricité consommée par les centres de données[3,37]. L'utilisation d'ADN pour le stockage de données « froides » apportera donc principalement des bénéfices en termes de longévité et de densité, plutôt que sur le plan énergétique.

- **Des évolutions de pratiques nécessaires**

Si l'utilisation d'ADN ne devrait donc pas apporter – à moyen terme – de solution sur le plan énergétique pour le stockage de données, les opportunités fournies par cette technologie pourraient *a contrario* conduire à une augmentation de consommation énergétique à travers un mécanisme d'« effet rebond »[38].

Par conséquent, le développement du stockage de données sous la forme d'ADN ne dispensera pas d'une réflexion sur les modes de consommation de données[39] et devra notamment s'accompagner de règles de bon usage afin de déterminer les données méritant d'être conservées sur le long terme (en particulier les données personnelles, pour lesquelles des problématiques éthiques[40] s'ajoutent aux problématiques environnementales).

■ Conclusion

Le stockage d'informations sous forme d'ADN constitue une technologie attrayante, susceptible d'offrir de nombreux bénéfices pour l'archivage de données. Le développement d'une filière française dédiée doit être encouragé, en favorisant le dialogue entre les différents intervenants qui disposent des expertises nécessaires[41]. Les résultats obtenus dans le cadre d'une telle démarche pourraient très vraisemblablement être valorisés pour diverses autres applications (en santé, informatique…). Il convient néanmoins de rester réaliste sur les possibilités qui seront réellement offertes par cette technologie et sur l'enjeu environnemental. Son développement ne peut s'affranchir d'une réflexion sur les modes de consommation numérique qui vont nécessiter un effort de sobriété.

Sites Internet de l'Office :
http://www.assemblee-nationale.fr/commissions/opecst-index.asp
http://www.senat.fr/opecst

Personnes consultées

M. Marc Antonini, directeur de recherche CNRS au laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis

M. Raja Appuswamy, maître de conférences en Data Science à Eurecom

M. Pascal Barbry, professeur en physiologie génomique des eucaryotes à l'Institut de Pharmacologie Moléculaire et Cellulaire

M. Dominique Lavenier, directeur de recherche CNRS à l'Institut de recherche en informatique et systèmes aléatoire

M. Jean-François Lutz, directeur de recherche CNRS à l'Institut Charles Sadron de Strasbourg

M. Thomas Ybert, président-directeur général et co-fondateur de DNA Script

Mme Sophie Tuffet, directrice générale, présidente du directoire et co-fondatrice d'Imagene

M. Erfane Arwani, président-directeur général et co-fondateur de Biomemory

M. François Képès, membre correspondant de l'Académie d'Agriculture de France, membre de l'Académie des Technologies, animateur du groupe de travail « ADN : lire, écrire, stocker l'information » et co-auteur du rapport « Archiver les mégadonnées au-delà de 2040 : la piste de l'ADN »

M. Alain Bravo, membre et président du comité « éthique, société et technologies » de l'Académie des technologies, co-auteur du rapport « Big Data – Questions éthiques »

M. Louis Dubertret, membre de l'Académie des technologies, co-auteur du rapport « Big Data – Questions éthiques »

Mme Anne Siegel, directrice de recherche au CNRS, directrice adjointe scientifique à l'Institut des sciences de l'information et de leurs interactions (INS2I) du CNRS.

M. Jean-Marc Rietsch, expert international en dématérialisation, signature et archivage électronique, expert judiciaire près la Cour d'Appel de Paris

Références

[1] Les *data centers* (en français : centre de données) correspondent à des infrastructures stockant et distribuant des données.
[2] Commission européenne, « Stratégie européenne pour les données », 2020 (https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_fr#documents).
[3] Académie des technologies, « Archiver les mégadonnées au-delà de 2040 : la piste de l'ADN », 2020 (http://academie-technologies-prod.s3.amazonaws.com/2020/10/15/10/17/37/9fbaaf5e-d5a6-4baf-921d-815da5d7983f/ADN_web.pdf).
[4] En ce qui concerne l'empreinte carbone, cela correspond à 3,5 % des émissions mondiales en 2019 et une part entre 5,5 et 6,9 % de celles-ci en 2025. Voiadr : The Shift Project, « Impact environnemental du numérique : tendances à 5 ans et gouvernance de la 5G », 2021 (https://theshiftproject.org/article/impact-environnemental-du-numerique-5g-nouvelle-etude-du-shift/).
[5] The Shift Project, « Lean ICT : Pour une sobriété numérique », 2018 (https://theshiftproject.org/article/pour-une-sobriete-numerique-rapport-shift/). Une étude réalisée en 2020 par les cabinets Citizing et KPMG pour la Commission de l'aménagement du territoire et du développement durable du Sénat a pour sa part estimé que les *data centers* représentaient 14 % de l'empreinte carbone du numérique en France en 2019 (voir la note de fin n°6).
[6] « Pour une transition numérique écologique », Rapport d'information n°*555* (2019-2020) de MM. Guillaume Chevrollier et Jean-Michel Houllegatte, sénateurs, fait au nom de la Commission de l'aménagement du territoire et du développement durable par la Mission d'information sur l'empreinte environnementale du numérique (http://www.senat.fr/commission/dvpt_durable/mission_dinformation_sur_lempreinte_environnementale_du_numerique.html).
[7] La taille des transistors est aujourd'hui proche de l'échelle atomique, barrière physique infranchissable. Les performances des semi-conducteurs (dont l'évolution est décrite par la loi Moore, voir la note de fin n°27) devraient donc plafonner dans les prochaines années. Voir : a) M. M. Waldrop, Nature 2016, 530, 144 (https://www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338); b) A. Shehabi, S. J. Smith, N. Horner, I. Azevedo, R. Brown, J. Koomey, E. Masanet, D. Sartor, M. Herrlin, W. Lintner, « United States Data Center Energy Usage Report », Lawrence Berkeley National Laboratory, Berkeley, California, 2016 (https://www.osti.gov/servlets/purl/1372902/).
[8] DNA Data Storage Alliance, « Preserving our digital legacy: an introduction to DNA data storage », 2021 (https://dnastoragealliance.org/dev/wp-content/uploads/2021/06/DNA-Data-Storage-Alliance-An-Introduction-to-DNA-Data-Storage.pdf).
[9] a) V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, W. L. Hughes, Nat. Mater. 2016, 15, 366; b) Mark Whitby (Senior Vice President, Branded Products Group at Seagate Technology, EMEA) to TechRadar, 2015 (https://www.techradar.com/news/computing-components/storage/the-data-capacity-gap-why-the-world-is-running-out-of-data-storage-1284024).
[10] Le séquençage d'une molécule d'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides la constituant.

[11] Cet avantage semble particulièrement remarquable, à l'heure où la quantité de données stockées augmente exponentiellement et où la question de place occupée par les data centers se pose. Voir : C. Diguet, F. Lopez, « L'impact spatial et énergétique des data centers sur les territoires », Rapport ADEME, 2019 (https://librairie.ademe.fr/urbanisme-et-batiment/908-impact-spatial-et-energetique-des-data-centers-sur-les-territoires-l-.html). D'après François Képès : « Aujourd'hui, les centres de données occupent un millionième des terres immergées. Au rythme actuel de développement, un millième des terres immergées sera occupé par des entrepôts de données en 2040. » alors que l'ADN « pourrait permettre de conserver l'ensemble de la sphère des données actuelle de l'humanité dans un volume correspondant à une petite camionnette. En 2040, ce sera le volume d'un camion ». Voir : a) A. Schwyter, Challenges 2021 (https://www.challenges.fr/high-tech/fini-les-data-centers-place-aux-donnees-stockees-sur-l-adn_744391); b) A. Couto, Industrie & Technologies 2020 (https://www.industrie-techno.com/article/l-adn-pourrait-nous-permettre-de-conserver-l-ensemble-des-donnees-mondiale-dans-le-volume-d-une-camionnette-clame-francois-kepes-membre-de-l-academie-des-technologies.62299). En outre, la densité offerte par l'ADN peut s'avérer particulièrement utile en termes de cryptographie puisqu'il est de fait facilement possible de dissimuler une information dans une longue séquence de nucléotides.

[12] En octobre 2020, les sociétés Illumina, Microsoft, Twist Bioscience et Western Digital ont notamment créé une « DNA Data Storage Alliance » ayant pour objectif de fédérer les entreprises et les institutions travaillant dans des domaines liés à l'utilisation d'ADN pour stocker des données (https://dnastoragealliance.org/).

[13] Ce domaine fait notamment appel aux mathématiques et à la théorie du signal (pour le développement d'un système de codage efficace), à la chimie et à la biologie moléculaire (pour la synthèse d'ADN), à la génomique et à l'informatique (pour l'étape de séquençage), à la microfluidique et à la robotique (pour la construction de dispositifs automatisés).

[14] Le programme états-unien *Molecular Information Storage* (MIST), financé par l'*Intelligence Advanced Research Projects Activity* (IARPA) à hauteur de 48 millions de dollars américains (soit environ 40 millions d'euros), a pour objectif de pouvoir synthétiser dès 2025 l'équivalent d'un téraoctet et de séquencer l'équivalent de 10 téraoctets en 24 heures, pour un coût de 1 000 US$ (https://www.iarpa.gov/index.php/research-programs/mist).

[15] Le projet européen OligoArchive est financé pour une durée de 3,5 ans par le programme H2020 de l'Union européenne à hauteur de 3 millions d'euros (https://oligoarchive.github.io).

[16] Le projet français dnarXiv, financé par l'INRIA et le LabEx CominLabs, regroupe divers laboratoires de recherche bretons (https://project.inria.fr/dnarxiv/).

[17] Le projet MoleculArXiv, porté par le CNRS et impliquant plus de 20 laboratoires, a obtenu en septembre 2021 un financement de 20 millions d'euros sur 84 mois dans le cadre des programmes et équipements prioritaires de recherche (PEPR) exploratoires. L'objectif à 5 ans est de pouvoir réaliser le cycle de lecture/écriture à un rythme de 1 bit/s (soit 10 Go de données en 24h) afin de pouvoir ensuite déployer des démonstrateurs.

[18] La proportion de bases C et G ne doit pas excéder celle de A et T et la répétition successive d'un même nucléotide plus de 3 fois ou d'un motif doit être évitée. De fait, l'encodage le plus simple (00 → A ; 01 → C ; 10 → G ; 11 → T) n'est pas forcément le plus adapté. Voir : G. M. Church, Y. Gao, S. Kosuri, Science 2012, 337, 1628 (https://science.sciencemag.org/content/337/6102/1628).

[19] Plus un fragment est long, plus le risque de trouver une erreur sur l'un de ses nucléotides est élevé. Actuellement, le taux d'erreur pour chaque nucléotide étant au mieux d'environ 0,1 %, il est préférable de ne pas utiliser des fragments de plus de 200 nucléotides pour garder un taux d'erreur acceptable.

[20] Un tel système est également primordial pour permettre un accès aléatoire aux données stockées.

[21] En addition des opportunités fournies par la partie numérique du processus, les étapes biologiques pourraient également permettre de renforcer la sécurité du système de stockage. Le projet dnarXiv (voir note de fin n°16) travaille notamment sur cet aspect. Une des pistes envisagées consiste à utiliser les amorces comme « clefs » : sans connaissance de la séquence des amorces, l'étape d'amplification est impossible, empêchant le séquençage et donc la « lecture ».

[22] Contrairement à la plupart des ADN polymérases, les désoxynucléotidyl transférases terminales n'ajoutent pas de nucléotides à partir d'une matrice simple brin mais de manière aléatoire. Dans le cas de la synthèse enzymatique d'ADN, les nucléotides sont ajoutés un à un (avec un groupement chimique protecteur) pour contrôler la séquence.

[23] Performance réalisée par Twist Bioscience, voir la note de fin n°3.

[24] Il existe actuellement quelques entreprises ayant fait le choix de placer leurs activités exclusivement sur ce domaine naissant : Catalog DNA (Etats-Unis), Helixworks Technologies (Irlande) et Biomemory (France).

[25] On peut de plus ajouter les irradiations, les enzymes, les microorganismes, l'ozone ainsi que divers autres polluants atmosphériques et xénobiotiques. Ces sensibilités peuvent cependant être également vues comme des atouts dans le cadre de stockage d'informations sensibles, en permettant la destruction instantanée des informations si nécessaire (notamment grâce à la température).

[26] En dehors du stockage *in vivo*, le stockage de données sous forme d'ADN ne semble pas poser de problématiques éthiques puisque les molécules synthétisées ne sont pas introduites dans des cellules et donc susceptibles d'être interprétées.

[27] a) G. E. Moore, « Cramming More Components Onto Integrated Circuits », Electronics 1965 (https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf); b) G. E. Moore, « Progress in Digital Integrated Electronics », IEEE Text Speech, 1975 (https://www.eng.auburn.edu/~agrawvd/COURSE/E7770_Spr07/READ/Gordon_Moore_1975_Speech.pdf).

[28] National Human Genome Research Institute, « The Cost of Sequencing a Human Genome » 2020 (https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost).

[29] F. Sanger, S. Nicklen, A.R. Coulson, Proc. Natl. Acad. Sci. USA 1977, 74, 5463 (https://pubmed.ncbi.nlm.nih.gov/271968/).

[30] Bien qu'il soit souhaitable de minimiser le taux d'erreurs, il est possible de corriger en partie celles-ci grâce à des codes correcteurs d'erreurs tels qu'utilisés pour les disques compacts par exemple (pour pallier les défauts microscopiques du support, les traces de doigts, les rayures…).

[31] Il est également possible d'utiliser des nucléotides non conventionnels, permettant notamment d'aboutir à une plus grande densité informationnelle.

[32] Les polymères non-ADN ne nécessitent pas d'être constituées de deux brins et les monomères choisis peuvent être d'une taille inférieure aux nucléotides, permettant donc d'atteindre une plus grande densité d'information. De plus, le nombre de monomères distincts pouvant être utilisés n'est pas limité, il est donc possible d'utiliser un système de numération permettant une écriture plus compacte que le système quaternaire fourni par l'ADN (un système octal ou hexadécimal, par exemple).

[33] De plus, contrairement à l'ADN, l'utilisation de tels polymères ne permet pas de s'affranchir du phénomène d'obsolescence, problématique pour la conservation de données sur des temps longs.

[34] C. N. Takahashi, B. H. Nguyen, K. Strauss, L. Ceze, Sci. Rep. 2019, 9, 4998 (https://www.nature.com/articles/s41598-019-41228-8).

[35] BCC Research, « DNA Data Storage: Global Markets and Technologies », 2020 (https://www.bccresearch.com/market-research/biotechnology/dna-data-storage-market.html).

[36] Des calculs réalisés par des chercheurs du CNRS estiment qu'un coût de 1 € pour l'écriture ou la lecture de 1 Mo de données permettrait au stockage sous la forme d'ADN de pénétrer le marché des données froides pour des documents de valeur nécessitant un accès rare (par exemple, les contrats, titres de propriété, lois). À 1 € pour l'écriture ou la lecture de 1 Go, le stockage moléculaire deviendrait une solution intéressante pour l'archivage de données, grâce la réduction de volume et l'augmentation de durabilité offertes. Enfin, à 1 € pour l'écriture ou la lecture de 1 To, l'ADN deviendrait intéressante pour transférer (physiquement) des données très volumineuses, transferts aujourd'hui limités par les capacités des serveurs.

[37] Horison Information Strategies, « Tiered Storage 2020: Building the Optimal Storage Infrastructure », 2020 (https://horison.com/publications/tiered-storage-2020).

[38] Des données qui n'auraient pas été conservées pourraient l'être du fait de l'utilisation de l'ADN comme support de stockage.

[39] F. Képès, « Information et environnement », Revue politique et parlementaire 2021 (https://www.revuepolitique.fr/%ef%bb%bf%ef%bb%bfinformation-et-environnement/).

[40] Académie des technologies, « Big Data – Questions éthiques », 2019 (https://www.academie-technologies.fr/blog/categories/publications-de-l-academie/posts/big-data-questions-ethiques).

[41] Le récent financement du PEPR exploratoire MoleculArXiv (voir la note de fin n°17) démontre un certain engagement en ce sens et mérite d'être salué.