médecine/sciences 2018 ; 34 : 622-5

Chroniques génomiques

L'ADN comme mémoire informatique ?

Bertrand Jordan

UMR 7268 ADÉS, Aix-Marseille, Université/EFS/ CNRS ; CoReBio PACA, case 901, Parc scientifique de Luminy, 13288 Marseille Cedex 09, France.
brjordan@orange.fr

Une idée séduisante

L'ADN que renferme chaque noyau de nos cellules porte une quantité importante d'information dans un volume microscopique : la nature s'avère, de ce point de vue, bien plus économe et efficace que les mémoires flash de nos clefs USB ou les disques durs intégrés dans nos ordinateurs. L'idée de mettre à notre service cette molécule pour y stocker des informations est aussi séduisante que médiatique, et elle a déjà été évoquée depuis plusieurs décennies. On pourrait ainsi emmagasiner de très grandes quantités de données dans un petit volume, et en assurer la conservation à long terme : alors que les supports actuels s'effacent en une ou deux décennies, l'ADN peut être stable à l'échelle du millénaire. C'est, de plus, un support qui ne risque pas d'être frappé d'obsolescence... Mais les auteurs qui ont évoqué cette possibilité par le passé [1] se sont immédiatement heurtés aux limites de la technique, tant du côté de la synthèse d'ADN qu'au niveau de sa lecture. Il y a une dizaine d'années, la synthèse de millions de nucléotides (nécessaires pour coder des mégaoctets de données) tout comme leur lecture rapide étaient hors de portée du point de vue des délais et des coûts : l'utilisation de l'ADN comme mémoire informatique restait une idée séduisante mais irréalisable. Plusieurs équipes ont néanmoins poursuivi des tentatives en ce sens et, grâce à des avancées techniques significatives, obtiennent aujourd'hui des résultats encourageants [2].

Des avancées sur plusieurs fronts

Chacun sait que les systèmes de lecture d'ADN ont fait des progrès énormes au cours de la dernière décennie, au point qu'aujourd'hui, un génome humain peut être intégralement séquencé pour environ 1 000 dollars. Les coûts de synthèse d'oligonucléotides n'ont pas baissé dans les mêmes proportions, et ils sont aujourd'hui de l'ordre de 0,1 dollar par base (pour des oligonucléotides d'une centaine de bases). Mais il s'agit là d'oligonucléotides produits individuellement (sur des microcolonnes), comme ceux que l'on utilise dans les efforts en cours pour synthétiser des chromosomes de levure [3] (➜).

(➜) Voir la Chronique génomique de B. Jordan, *m/s* n° 10, octobre 2016, page 898

Il existe un autre format de fabrication qui consiste à synthétiser un grand nombre d'oligonucléotides différents (jusqu'à un million) sur un substrat solide (comme pour un réseau d'ADN ou *microarray*) puis à les détacher du support : on obtient alors un mélange de milliers d'oligonucléotides différents (chaque séquence ayant été définie *a priori*), et le coût par base est bien inférieur, estimé à environ 0,0001 dollars par base [4-6][1]. Comme nous le verrons, c'est ce format qui est employé dans le travail que je vais décrire, et c'est en grande partie pour cela que les délais et coûts deviennent raisonnablement envisageables. Dernier développement important, la conception d'algorithmes de codage et de systèmes de correction d'erreurs a fait de grand progrès et peut être mise à profit pour assurer un codage économique et fidèle de l'information binaire d'un fichier informatique en une suite de bases pouvant prendre quatre valeurs (T, A, C ou G).

Une démonstration de faisabilité : 200 Mo inscrits dans l'ADN

Plusieurs articles récents font état du codage dans l'ADN d'un ou plusieurs fichiers représentant en géné-

[1] Le prix annoncé par les fournisseurs (voir par exemple http://www.customarrayinc.com/oligos) est plus élevé, mais ce chiffre semble être le plus réaliste pour une fabrication en masse d'après les publications citées.

m/s n° 6-7, vol. 34, juin-juillet 2018
https://doi.org/10.1051/medsci/20183406025

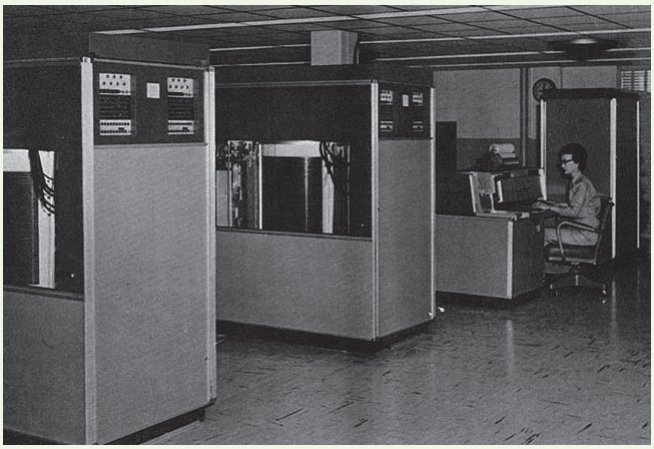


Figure 1. ***Le premier disque dur commercialisé en 1956 (IBM 350).*** Au premier plan, deux unités IBM 350, poids 1 tonne, capacité 3,75 Mo. Au fond, l'ordinateur proprement dit (IBM 305 RAMAC) (Wikipedia open commons).

ral une fraction de mégaoctet (Mo)[2] [2]. Mais un article récemment paru dans *Nature Biotechnology* [7] constitue une avancée significative puisqu'il montre le codage de 35 fichiers représentant au total 200 Mo, l'accès direct à chaque fichier et la récupération de son contenu sans aucune erreur après lecture par deux techniques de séquençage. Notons que ce travail a été réalisé dans le cadre d'une collaboration entre plusieurs laboratoires universitaires (Washington, Stanford, Princeton) et l'entreprise *Microsoft* à laquelle appartient la majorité des vingt-deux signataires ainsi que l'un des deux auteurs correspondants.

Dans les grandes lignes, il s'est agi d'inscrire l'information dans un grand nombre d'oligonucléotides dont chacun contient, en plus d'une fraction du message codé, une adresse indiquant à quel fichier il appartient et une autre adresse donnant sa position dans ce fichier. Ce mélange d'oligonucléotides contient donc toute l'information, qui peut être récupérée en amplifiant par PCR les oligonucléotides correspondant au fichier choisi (grâce à la première adresse) puis en les séquençant et en les assemblant (grâce à la deuxième adresse). Il ne s'agit donc pas de synthétiser de très longues molécules d'ADN qui porteraient le contenu d'un fichier entier, comme lorsqu'on cherche à produire un génome synthétique [3], mais d'employer un mélange de segments relativement courts, ce qui est bien plus rapide et plus économique mais permet néanmoins de récupérer l'information. Le codage proprement dit (passage du code {0,1} informatique au code {T,A,G,C} de l'ADN a été réalisé par les auteurs en utilisant un système sophistiqué avec correction d'erreurs ce qui, avec la redondance choisie (voir plus bas), assure la récupération fidèle de l'information. Les séquences ainsi définies, découpées en segments et munies de leurs adresses, ont été transmises à l'entreprise *Twist Bioscience* pour exécution de la synthèse. Cet ensemble, qui comprend 13 millions d'oligonucléotides longs d'environ 150 bases, représentant au total 2 milliards de bases, a été synthétisé en 9 *pools* ou lots (chacun comprend donc environ 1,4 millions d'oligonucléotides différents). L'ensemble de ces *pools* contient ainsi l'information correspondant aux 35 fichiers, avec une redondance de 2,5x du point de vue informatique [3] ; notons que ceci correspond à dix bases par octet. Cet ADN peut être stocké quasi-indéfiniment sous forme lyophilisée et occupe un tout petit volume, largement inférieur au millimètre cube. Pour la lecture d'un des 35 fichiers, l'ADN est réhydraté, amplifié par PCR en utilisant les amorces correspondant au fichier puis séquencé à l'aide d'une machine *Illumina*, le standard actuel, avec une redondance moyenne de 6x (six lectures en moyenne pour chaque base). Suivent alors plusieurs étapes de décodage, afin d'obtenir la séquence de chaque oligonucléotide (en utilisant la redondance de séquençage pour réduire le taux d'erreurs) et d'en déduire le contenu du fichier, en assemblant les séquences grâce à l'adresse de position portée par chacune d'elles puis en corrigeant les erreurs grâce à la redondance de codage et au système informatique de correction d'erreurs. Les auteurs montrent qu'ils parviennent en pratique à récupérer ainsi l'ensemble des 200 Mo de données initialement codées sans aucune erreur : la différence entre les fichiers d'origine et les fichiers récupérés est nulle. Ils ont donc bel et bien fait la démonstration qu'ils peuvent emmagasiner 200 Mo de données dans de l'ADN et récupérer cette information fichier par fichier et sans perte d'information. C'est beaucoup plus que tout ce qui avait été réalisé jusqu'ici et cela inclut l'accès direct à chaque fichier, ce qui n'était généralement pas le cas [7]. Dans une dernière partie, les auteurs font aussi la démonstration qu'ils peuvent utiliser une autre approche de séquençage, celle des nanopores [8, 9] (➜) et, avec quelques modifications techniques, obtenir des résultats prometteurs – c'est important dans la perspective d'un futur système intégré car l'appareil de séquençage par nanopores est très compact et peu onéreux [8].

(➜) Voir la Chronique génomique de B. Jordan, *m/s* n° 8-9, août-septembre 2017, page 801, et la Synthèse de F. Montel, *m/s* n° 2, février 2018, page 161

Il reste du chemin à faire

Deux cents Mo, c'était, il y a quelques années, une capacité respectable pour un disque dur. Rappelons que les disquettes des années 1990 avaient une capa-

[2] Un octet (Byte dans l'usage anglais) correspond à huit bits d'information. Ne pas confondre bytes et bits, ni mégabases d'ADN (Mb) et mégabytes (ou mégaoctets) de données (MB ou Mo)...

[3] Une base, qui peut avoir quatre valeurs (T, A, G ou C) correspond à deux bits d'information. Deux milliards de bases portent donc quatre milliards de bits, donc un demi-milliard de Bytes ou octets : soit 2,5 fois 200 Mo.

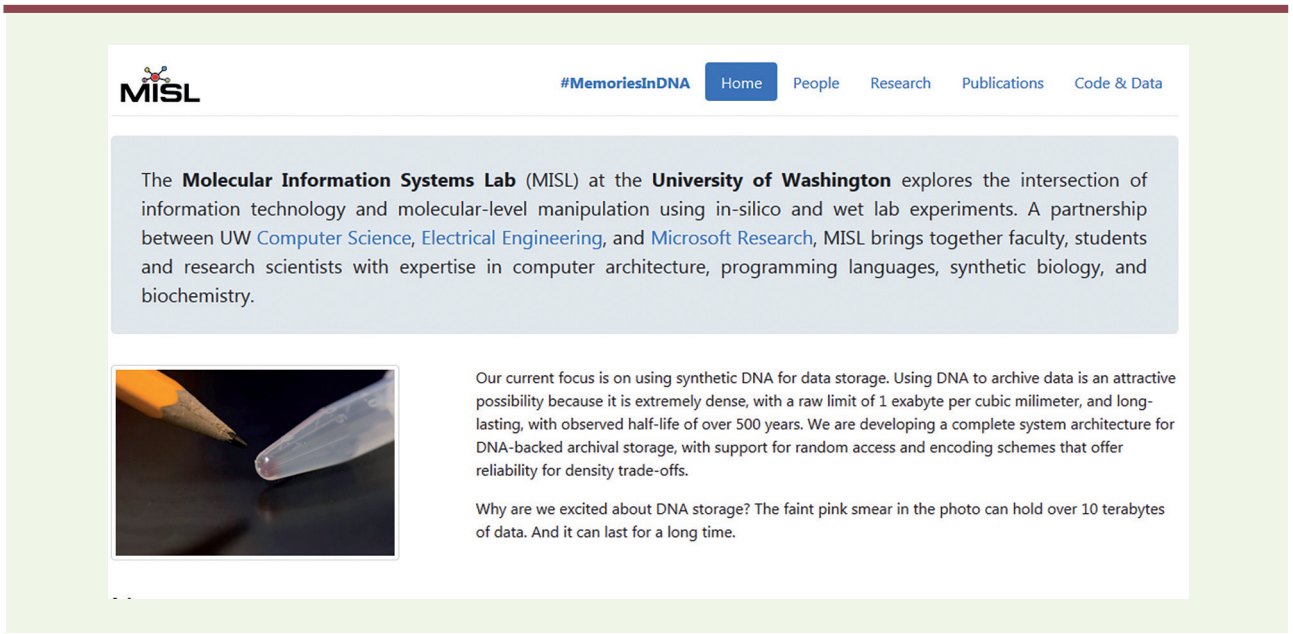


Figure 2. ***Site du MISL.*** Le site du *Molecular information systems lab* (MISL), responsable avec Microsoft du travail rapporté ici, et promoteur du projet *Memories in DNA* (site du MISL, misl.cs.washington.edu).

cité de 1,44 Mo, les CD 800 Mo et les DVD, vers 2000, 4 500 Mo ; mentionnons aussi que le tout premier disque dur, commercialisé en 1956 par l'entreprise IBM, occupait la place d'une grosse armoire et pesait une tonne pour une capacité de 3,75 Mo *(Figure 1)*.

On se rapproche donc de valeurs réalistes, et on peut espérer des progrès rapides. Néanmoins, dans les conditions actuelles, ce mode de stockage implique de nombreuses manipulations et un coût élevé. On peut, par exemple, estimer que les oligonucléotides fournis par *Twist Bioscience* ont dû coûter environ 200 000 dollars (au tarif avantageux de 0,0001 dollar par base pour deux milliards de bases) ; quant à la lecture de ces deux mégabases (Mb), elle représente moins d'un génome humain (soit 6 Mb lus à une redondance de 30x), donc *a priori* moins de 1 000 dollars. Reste à évaluer le coût des différentes manipulations impliquées – disons qu'au total le coût du stockage sous forme d'ADN et de la lecture d'un ensemble de fichiers représentant 200 Mo est sans doute de l'ordre du demi-million de dollars – soit environ 2 500 dollars par Mo. D'ailleurs l'entreprise *Twist Bioscience* propose, d'après la *MIT Technology Review* [10], de stocker le fichier de votre choix (12 Mo) sous forme d'ADN pour 100 000 dollars – offre dont l'intérêt pratique semble assez limité mais qui corrobore le chiffrage tenté ci-dessus[4].

En tout état de cause, l'obstacle principal à franchir est celui du coût de la synthèse d'ADN – il faudrait qu'il baisse de plusieurs ordres de grandeur pour que ce type de mémoire devienne envisageable en pratique, sans doute dans un premier temps pour des données très précieuses et que l'on souhaite stocker à très long terme.

Ainsi ce qui était une idée folle il y a deux décennies s'approche peu à peu de la réalisation concrète. L'article qui fait l'objet de cette chronique [7] est une démonstration convaincante de sa faisabilité technique ; la faisabilité financière et pratique reste à atteindre, mais elle n'est pas exclue à moyen terme. Dès à présent, un projet intitulé *Memories in DNA* (http://memoriesindna.com/) et impliquant pour l'essentiel les structures responsables du travail rapporté ici [7], vous suggère de fournir une photo de votre choix qui sera transcrite en ADN dans le cadre d'un projet visant à rassembler 10 000 images, à les coder sous forme d'ADN et à explorer les manières de les classer en explorant directement leur séquence nucléotidique *(Figure 2)*. Au-delà d'un indéniable effet de mode, ces travaux pourraient avoir des retombées dans de multiples domaines : après l'époque du séquençage tous azimuts, l'ADN rentre dans l'époque de la synthèse à grande échelle, ouvrant à terme la voie à de nombreuses applications. ◊

SUMMARY

DNA for information storage?

The very high information density of DNA has prompted speculations on its use for information storage. The high costs of DNA synthesis and sequencing made this highly unpractical; however recent progress (notably array oligonucleotide synthesis) is changing the situation. A recent paper shows encoding and decoding of significant amounts of data (200 MB) with random access to individual files and faithful

[4] Ce coût correspond à 8 000 dollars par Mo contre 2 500 estimés pour le travail décrit ici [7], mais il s'agit là d'une proposition commerciale intégrant une certaine marge.

retrieval of content, at a cost that is still high but not extreme. Much progress remains to be achieved, but this use of DNA in now technically achievable and may eventually become practical. ◊

LIENS D'INTÉRÊT

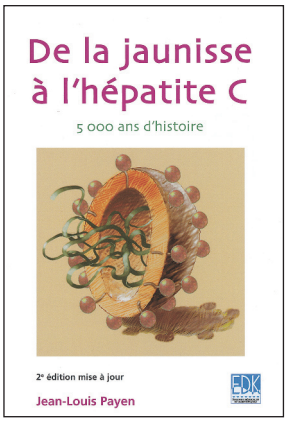
L'auteur déclare n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Cox JPL. Long-term data storage in DNA. *Trends Biotechnol* 2001 ; 19 : 247-50.
2. Heckel R. An archive written in DNA. *Nat Biotechnol* 2018 ; 36 : 236-7.
3. Jordan B. HGP-write : après la lecture, l'écriture ? *Med Sci (Paris)* 2016 ; 32 : 898-901.
4. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods* 2014 ; 11 : 499-507.
5. Carlson R. On DNA and transistors. *Synthesis* March 9, 2016. www.synthesis.cc/synthesis/2016/03/on_dna_and_transistors
6. Carlson R. Guesstimating the size of the global array synthesis market. *Synthesis* August 30, 2017. http://www.synthesis.cc/synthesis/2017/8/guesstimating-the-size-of-the-global-array-synthesis-market
7. Organick L, Ang SD, Chen YJ, *et al*. Random access in large-scale DNA data storage. *Nat Biotechnol* 2018 ; 36 : 242-8.
8. Jordan B. Séquençage d'ADN : l'offensive des nanopores. *Med Sci (Paris)* 2017 ; 33 : 801-4
9. Montel F. Séquençage de l'ADN par nanopores : résultats et perspectives. *Med Sci (Paris)* 2018 ; 34 : 161-5.
10. Walsh B. This company can encode your favorite song in DNA-for $100,000. *MIT Technology Review*, April 6, 2018 (https://www.technologyreview.com/s/610717/this-company-can-encode-your-favorite-song-in-dnafor-100000/).

TIRÉS À PART

B. Jordan



ISBN : 978-2-8425-4136-1 128 pages

La jaunisse est un symptôme facilement identifiable ; il paraissait bien naturel que l'homme, confronté à une modification de la couleur de ses yeux et de sa peau ait de tous temps recherché les causes de cette transformation.

Il n'est donc pas surprenant que le premier traité de médecine, écrit 3 000 ans avant J.C. par un médecin sumérien, décrive déjà la jaunisse. À chaque époque de l'histoire de la médecine, les praticiens, influencés par les concepts médicaux de leur temps, attribuèrent une ou plusieurs explications particulières à ce symptôme. Ainsi, du démon *Ahhâzu* des Sumériens à la sophistication des biotechnologies qui permirent la découverte du virus de l'hépatite C, le lecteur cheminera sur une période de 5 000 ans au travers des différents continents.

Ici encore, l'histoire se révèle une formidable source de réflexion : le foie souvent impliqué dans l'apparition des jaunisses est-il le siège de l'âme ?

Les expérimentations humaines chez des volontaires ou chez des enfants handicapés mentaux étaient-elles justifiées pour permettre la découverte des virus des hépatites ?

Le formidable développement de la transfusion sanguine, des vaccinations, mais aussi de la toxicomanie explique-t-il les épidémies d'hépatites du XXe siècle ?

Autant de questions qui sont abordées dans ce livre passionnant et accessible à tous.

BON DE COMMANDE

À retourner à EDK, 17 avenue du Hoggar, 91944 Les Ulis Cedex A
Tél. : 01 41 17 74 05 - Fax : 01 43 29 32 62 - E-mail : edk@edk.fr

NOM : ..... Prénom : .....
Adresse : .....
Code postal : ..... Ville : .....
Pays : .....
Fonction : .....
Je souhaite recevoir l'ouvrage **De la jaunisse à l'hépatite C, 5 000 ans d'histoire :** 12 € + 3 € de port = **15 € TTC**
en ..... exemplaire, soit un total de ..... €
❒ Par chèque, à l'ordre de **E D K**
❒ Par carte bancaire : ❒ Visa ❒ Eurocard/Mastercard
Carte n° |_|_|_|_| |_|_|_|_| |_|_|_|_| |_|_|_|_| Signature :
Date d'expiration : |_|_| |_|_|
N° de contrôle au dos de la carte : |_|_|_|

Tarifs d'abonnement ***m/s*** - 2018

Abonnez-vous

à ***médecine/sciences***

> Grâce à ***m/s***, vivez en direct les progrès des sciences biologiques et médicales

Bulletin d'abonnement
page 626 dans ce numéro de *m/s*