Stocker de l'information dans l'ADN

Christophe Dessimoz

Grâce à l'avancée des techniques de la génomique, il devient possible d'utiliser l'ADN comme support numérique pour archiver des données sur le long terme.

Deux piscines olympiques : tel est le volume qu'il faudrait pour contenir l'ensemble des données produites dans le monde en 2012 (environ 3×10^{21} octets), si on les stockait dans les toutes récentes clefs USB de un téraoctet (10^{12} octets). Au fil des ans, ce volume croît de façon exponentielle et, avec lui, le besoin de supports numériques fiables pour archiver l'ensemble des informations sur le long terme. De plus, la durée de vie des supports actuels étant limitée (pas plus de quelques dizaines d'années), la pérennisation des données nécessite une maintenance constante et lourde, fondée sur des copies multiples et régulières sur divers supports *(voir l'article de Francis Daumas et Marion Massol dans ce numéro)*.

Pourtant, il existe depuis plus de trois milliards d'années un support numérique universel, compact et stable : l'ADN, la molécule porteuse de l'information génétique des organismes vivants. Codé sous la forme d'une succession de quatre molécules – les nucléotides A, T, G, C –, le programme de la vie est numérique : les nucléotides codent de l'information, au même titre que les bits (0 ou 1) de l'informatique. L'ADN est de plus très stable – on a retrouvé, dans des carottes de glace, des gènes datant de 450 000 à 800 000 ans – et dense : si l'on codait sur de l'ADN toute l'information produite en 2012, elle tiendrait dans le coffre d'une voiture...

En outre, le fait que l'ADN soit universel limite les risques d'obsolescence. Tant qu'il y aura de la vie sur Terre et un intérêt pour la comprendre, la technique pour lire l'ADN existera. Aussi les progrès récents de la génomique nous ont-ils conduits, à l'Institut européen de bio-informatique, à étudier la possibilité d'utiliser l'ADN comme support d'archives numériques.

Tous les sonnets de Shakespeare sur ADN

Peu après la caractérisation de l'ADN, publiée en 1953 par James Watson et Francis Crick, l'idée d'utiliser cette molécule comme support d'information a été envisagée. En 1964, un physicien et officier radio russe, Mikhail Neiman, voit dans les systèmes et processus d'information biophysiques un moyen de miniaturiser les dispositifs de stockage et de traitement de l'information. L'idée a été mise en œuvre à la fin des années 1990, avec le développement du génie génétique. Plusieurs équipes ont su encoder de l'information sur l'ADN, mais les informations étaient courtes (un message secret annonçant le débarquement du 6 juin 1945, une comptine anglaise avec paroles, musique et petite image, la Déclaration des droits de l'homme…). Surtout, les manipulations génétiques étaient laborieuses et *ad hoc*.

Ces dernières années, cependant, les progrès des techniques de séquençage ont entraîné un tournant important. En 2010, le biologiste américain Craig Venter, de

L'AUTEUR

Christophe DESSIMOZ est professeur de bio-informatique au Collège universitaire de Londres. Il est aussi affilié à l'Institut européen de bio-informatique (EMBL-EBI). Article écrit en collaboration avec Marie-Neige Cordonnier.

BIBLIOGRAPHIE

N. Goldman *et al.*, **Towards practical, high-capacity, low-maintenance information storage in synthesized DNA**, *Nature*, vol. 494, pp. 77-80, 2013.

G. M. Church *et al.*, **Next-generation digital information storage in DNA**, *Science*, vol. 337, pp. 1628-1629, 2012.

Entretien avec Ch. Dessimoz à écouter sur : bit.ly/pls_433_dessimoz

l'Institut du même nom, et son équipe ont synthétisé un génome complet d'un million de paires de bases (ou nucléotides) à partir du code du génome d'une bactérie, qu'ils ont introduit dans une autre espèce de bactérie dont ils avaient ôté le génome. Si l'opération était à nouveau *ad hoc* et fort coûteuse (estimée à plusieurs dizaines de millions d'euros), elle montrait qu'il était désormais possible de coder une grande quantité d'information sur de l'ADN en un temps raisonnable.

Depuis, le coût du séquençage a baissé de façon drastique, ce qui ouvre encore de nouvelles perspectives. Encoder de l'information sur l'ADN n'est pas compliqué. Il suffit de passer d'un système binaire (0, 1) à un système codé en base 4 en convertissant chaque couple binaire – 00, 01, 10, 11 – en un des quatre nucléotides A, T, G, C. Toutefois, deux aspects fondamentaux sont à prendre en compte. D'une part, coder une longue séquence coûte cher. Nous avons donc pris le parti de synthétiser des fragments courts (150 paires de bases), meilleur marché.

D'autre part, il s'agit de limiter au maximum les risques d'erreurs. Les erreurs de lecture des séquences se concentrent là où se suivent plusieurs nucléotides identiques. Pour éviter ces répétitions, nous avons converti les données binaires en base 3 et non en base 4, et utilisé le quatrième nucléotide comme une sorte d'opérateur de permutation. Le code en base 3 est composé de 0, 1, 2, auxquels on fait correspondre un nucléotide A, T, G ou C. Toute répétition est évitée en modifiant la correspondance chiffres-nucléotides après chaque chiffre codé. Par exemple, si l'on veut coder 22 et que le premier 2 est codé par A, le second 2 sera codé par une nouvelle correspondance ne faisant intervenir que les trois autres nucléotides (T, G ou C).

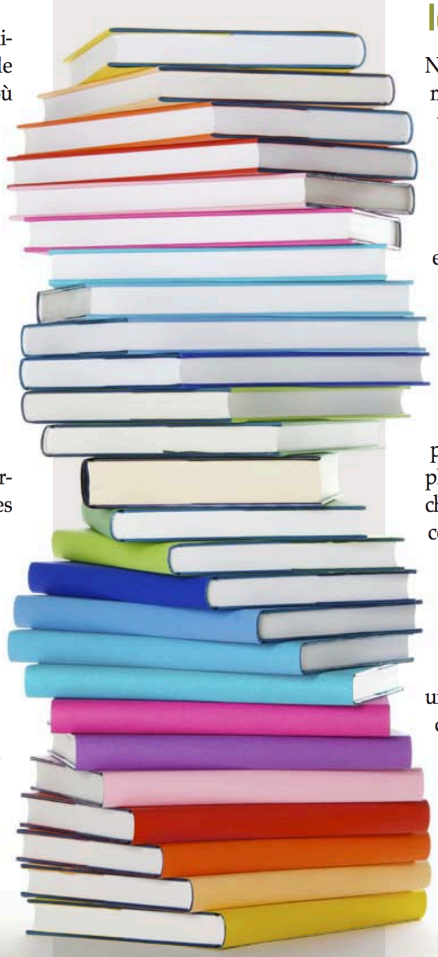
Enfin, pour limiter encore les erreurs, nous avons synthétisé des fragments d'ADN dont les séquences se chevauchent, de façon que chaque nucléotide soit couvert par quatre fragments différents, et nous avons ajouté, au début et à la fin de chaque fragment, des informations indiquant sa position dans la séquence complète.

Nous avons testé notre méthode sur cinq fichiers différents : un fichier texte en code ASCII comportant tous les sonnets de Shakespeare, le PDF d'un article scientifique, un extrait sonore du discours de Martin Luther King (format mp3), une photo (format JPEG 2000) et le texte (ASCII) du code qui nous sert à convertir les données binaires en base 3, pour montrer que l'on peut encoder le codage lui-même…

En deux jours, les données ont été encodées sur 150 000 brins d'ADN. L'ADN a ensuite été lyophilisé et envoyé par avion en Allemagne, où il a été réhydraté et séquencé en deux semaines. Nous avons retrouvé l'intégralité des données, sauf celles contenues dans deux fragments de 25 nucléotides. Ces deux fragments appartenaient à une région répétitive, ce qui nous a permis de les reconstituer manuellement et d'atteindre 100 pour cent de réussite (contre 99,999 pour cent sans ces deux fragments). Les deux fragments avaient chacun la particularité d'avoir des séquences leur permettant de se replier sur eux-même, ce qui a empêché leur séquençage.

Où stocker l'ADN synthétisé ?

Un endroit frais, sombre et sec suffirait à conserver de l'ADN lyophilisé pendant plusieurs milliers d'années. Une telle banque existe déjà : inaugurée en 2008, la Réserve mondiale de semences du Svalbard, une immense enceinte réfrigérée creusée dans l'île norvégienne du Spitzberg pour préserver la diversité des semences, n'a pas de personnel permanent sur place. Une cave à vin conviendrait aussi…



© Shutterstock/ Lichtmeister

Un intérêt pour le stockage à long terme

Nous avons ainsi réussi à stocker de l'information sur de l'ADN et à la restituer selon un procédé déclinable pour toute donnée numérique. Bien sûr, notre technique a ses limites. Avant tout, même s'ils seront sans doute réduits à quelques heures à l'avenir, les temps d'accès, d'écriture et de lecture sont moins bons que ceux d'une clef USB. L'ADN ne peut constituer un support compétitif que pour l'archivage de données à long terme. De plus, le volume de données testé était assez petit – 739 kilooctets, l'équivalent d'une disquette –, mais en théorie, il est possible de passer à une échelle beaucoup plus grande. De même, le procédé reste cher – 7 000 euros par mégaoctet –, mais ce coût diminuera vite si les tendances actuelles se poursuivent.

La technique d'encodage peut encore être améliorée et optimisée. Notamment, le passage à des données plus importantes nécessitera une automatisation et une miniaturisation plus poussées. D'ici quelques années, pour des utilisations sur le long terme, il pourrait devenir plus économique d'investir dans un support certes plus cher à l'achat, mais avec très peu de frais de maintenance, l'ADN, que d'opter pour un support bon marché tel que les bandes magnétiques, mais qui doit être relu tous les cinq ans. ■