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. ■

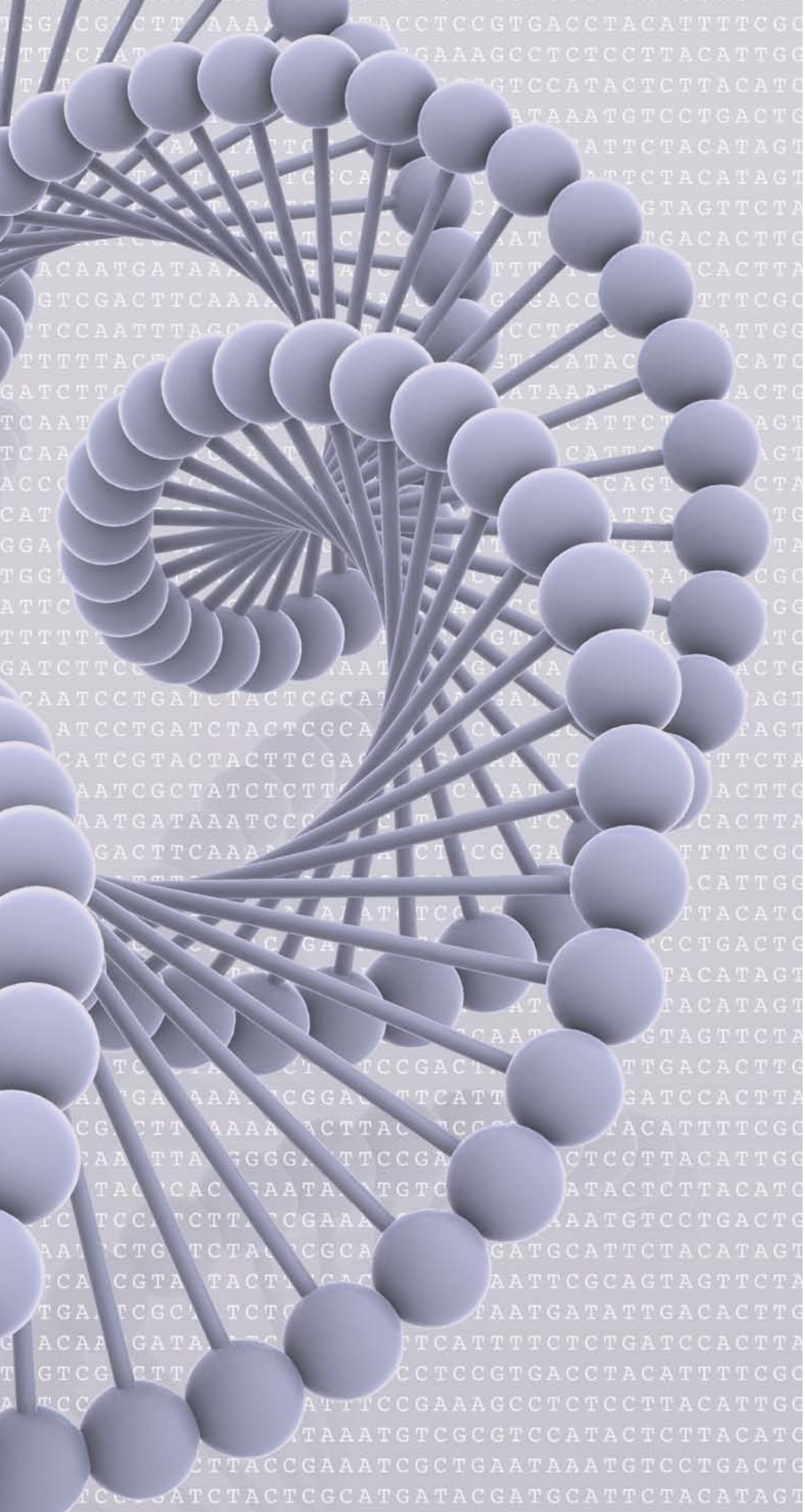
L'ADN
mémoire numérique du vivant

Vincent Daubin, Simon Penel et Éric Tannier

L'ADN, un code à quatre lettres, est une source d'information sur l'histoire du monde vivant. En comparant les génomes actuels, on reconstitue les gènes d'espèces disparues il y a plusieurs centaines de millions d'années.

L'ESSENTIEL

- **Par les mécanismes de l'hérédité, les génomes actuels recèlent des informations sur l'évolution du vivant depuis plusieurs milliards d'années.**
- **L'information est brouillée par le jeu des mutations, de la sélection naturelle et du transfert de gènes entre espèces.**
- **Néanmoins, en modélisant l'histoire des gènes, les phylogénéticiens ont accès à des informations sur des espèces dont on n'a plus trace aujourd'hui.**



© Shutterstock/ Tatiana53

Depuis plus de trois milliards d'années, des informations numériques écrites en langage ADN sont copiées, plus ou moins fidèlement, puis interprétées, conduisant à la formation de nouvelles générations d'organismes. À ce titre, la théorie de l'évolution est une théorie de l'information, analogue à celle qu'on utilise en mathématiques depuis sa formalisation en 1948 par le mathématicien américain Claude Shannon : un ensemble de lois décrivant comment un objet – un texte dans le cas de la théorie mathématique, un organisme dans le cas de la reproduction – peut être reconstitué à partir d'une information sur cet objet, en général codée de façon plus compacte que l'objet lui-même. La biodiversité actuelle résulte de ce processus de duplication – ou réplication – avec erreurs des données génétiques héritées d'une multitude d'ancêtres.

Dans les génomes des organismes actuels figurent les traces de ces milliards d'années de transmission d'informations codées génétiquement. Aussi, chaque lignée ayant enregistré des événements différents depuis sa séparation des autres, il est possible, en comparant ces génomes, de reconstruire l'histoire du vivant. Depuis 50 ans, les phylogénéticiens, qui étudient les relations évolutives entre les organismes, comparent les séquences ADN des espèces actuelles. Avec les progrès du séquençage génomique, ces recherches ont changé d'échelle. Elles permettent aujourd'hui de comprendre à la fois les mécanismes de l'évolution et l'origine des espèces. Quelles informations anciennes peut-on espérer décrypter dans les génomes actuels ? Quelles sont les méthodes employées ? Quelles sont les limites de cette approche ? L'analogie avec la théorie de l'information apporte plusieurs réponses.

L'hérédité : une théorie de l'information

Si, en semant des graines de tomate, on obtient des tomates et non des courgettes, c'est parce que dans les graines figure une information qui dicte la forme de la plante. La nature de cette information, celle qui est transmise des parents à leur progéniture dans tout le monde vivant lors de la reproduction, a été longuement débattue. Selon la théorie de la préformation, par exemple, dont on trouve des traces dès l'Antiquité, chaque graine contenait