L'ordinateur biologique, pour demain ?

S'appuyant sur les concepts et les outils de la biologie moléculaire, le calcul par ADN pourrait, un jour, s'imposer comme outil mathématique de référence…

par Lila Kari et Laura Landweber *

Que ce soit pour mesurer la durée des mois ou des saisons, pour les besoins du commerce ou de l'architecture, les peuples ont éprouvé, depuis les temps les plus reculés, le besoin de compter et de calculer. Ils ont à cette fin utilisé tous les moyens à leur disposition : on est ainsi passé progressivement d'outils manuels (les doigts) à des outils mécaniques (l'abaque, par exemple), puis à des dispositifs électroniques. L'ordinateur marque seulement l'avancée la plus récente dans l'histoire des instruments de calcul : leur apparition n'en constitue ni la première, ni la dernière étape. Ils ont, d'ailleurs, eux aussi leurs limites : limites des quantités de données qu'ils peuvent stocker ; limites des vitesses d'exécution, qui sont fixées par les lois de la physique et seront prochainement atteintes. Pour forcer ces barrages, la plus récente tentative propose de remplacer – une fois de plus – les outils de calcul : d'utiliser, en lieu et place des instruments électroniques, des instruments biologiques.

Exposés pour la première fois en 1994 par Léonard Adleman, les calculs par ADN (ou encore biomoléculaires, voire moléculaires) constituent un nouveau paradigme de calcul. S'appuyant sur la manipulation de (bio) molécules pour résoudre des problèmes mathématiques, ils considèrent les processus naturels comme des modèles. Ils ont rapidement soulevé l'enthousiasme de la communauté scientifique. En quoi consistent-ils ?

En codant des données sur des brins d'ADN et en utilisant les outils de la biologie moléculaire, on va pouvoir effectuer des opérations mathématiques. Cela requiert un milliard de fois moins d'énergie que les calculs électroniques traditionnels, tout en stockant des données sur mille milliards de fois moins d'espace. Qui plus est, le calcul avec l'ADN est hautement parallèle : en théorie, des milliards de molécules sont soumises à des réactions chimiques – c'est-à-dire effectuent des calculs – simultanément.

En dépit de sa complexité, cette technologie repose sur une analogie simple entre deux processus, l'un biologique, l'autre mathématique. En effet, la structure complexe d'un organisme vivant résulte, en dernière analyse, de l'application d'une petite série d'opérations simples (copie, épissage, insertion, délétion, etc.) aux informations initialement encodées dans une séquence d'ADN. D'un autre côté, tout calcul – si complexe soit-il – provient de la combinaison d'opérations logiques et arithmétiques très simples.

Pour Leonard Adleman, ces deux processus ne sont pas seulement similaires : grâce aux progrès de la biologie moléculaire, la biologie peut être utilisée pour faire des mathématiques. C'est donc ce qu'il a fait lui-même pour résoudre un difficile problème de calcul. Plus précisément, le cas à sept points du problème du « chemin hamiltonien orienté ». Explications…

Soit un graphique G orienté, c'est-à-dire constitué d'un ensemble de points v reliés les uns aux autres par des flèches de sens déterminé. On dit d'un tel graphique qu'il a un « chemin hamiltonien » s'il existe une suite de flèches compatibles e_1, e_2,…, e_n (c'est-à-dire un chemin) partant du point v_{in} et arrivant en v_{out}, et passant par chacun des autres points une fois et une seule. Le problème « du voyageur de commerce » en constitue une version simplifiée : étant donné un ensemble arbitraire de villes par lesquelles un vendeur doit passer (voir la

* Cet article constitue une adaptation d'un texte publié par les mêmes auteurs dans « Bioinformatics Methods and Protocols », *Methods in Molecular Biology*, ed. Misener & Krawetz, Humana Press.

FIGURE 1

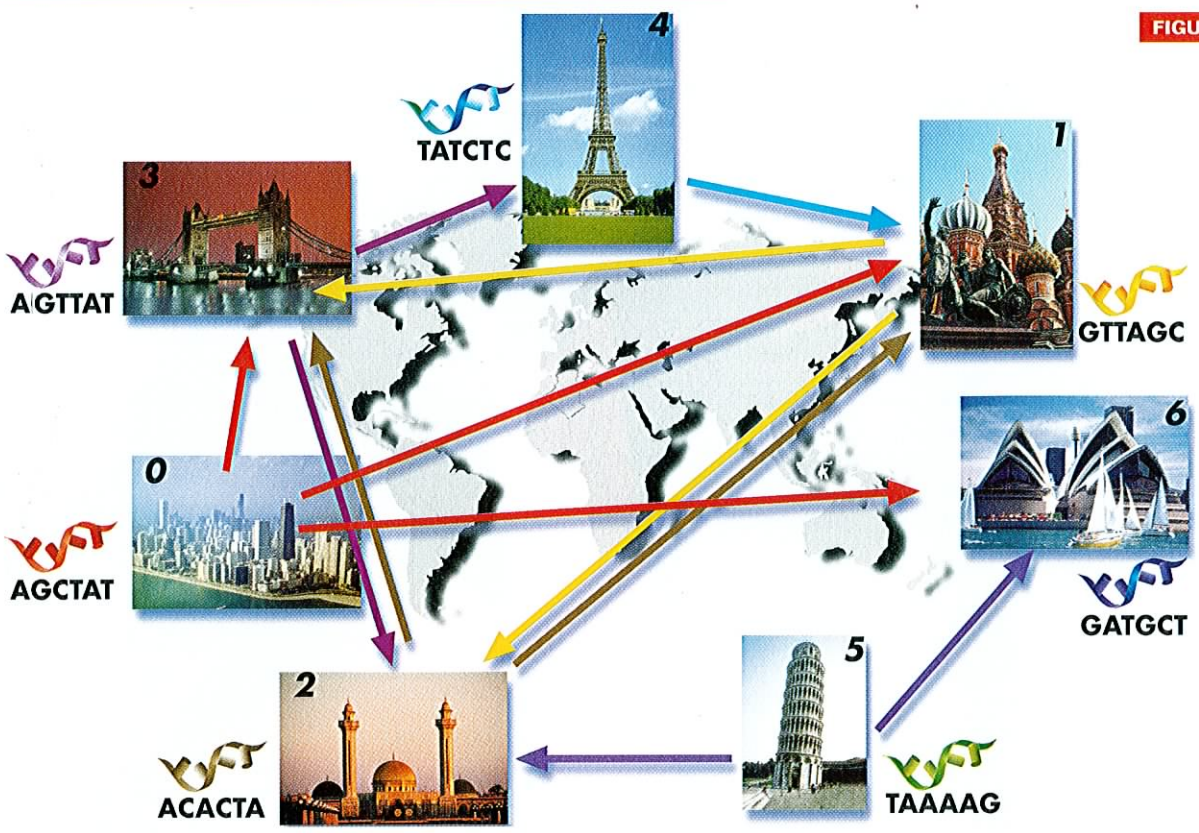


figure 1), quel est le moyen le plus court de les relier ? Le modèle d'Adleman limite le nombre d'itinéraires entre les villes, en spécifiant les destinations de départ et d'arrivée du voyage. Il s'agit alors de découvrir, s'il existe, le chemin continu les reliant toutes ensemble.

Pour résoudre ce problème, Adleman utilise l'algorithme suivant :

Étape 1 : générer des chemins aléatoires sur le graphique ;

Étape 2 : ne retenir que ceux qui commencent en v_{in} et finissent en v_{out} ;

Étape 3 : si n est le nombre de points du graphique, ne retenir que les chemins passant exactement par n points ;

Étape 4 : ne retenir que les chemins passant au moins une fois par chaque point du graphique ;

Étape 5 : s'il reste au moins un chemin, dire « OUI » ; dans le cas contraire, dire « NON ».

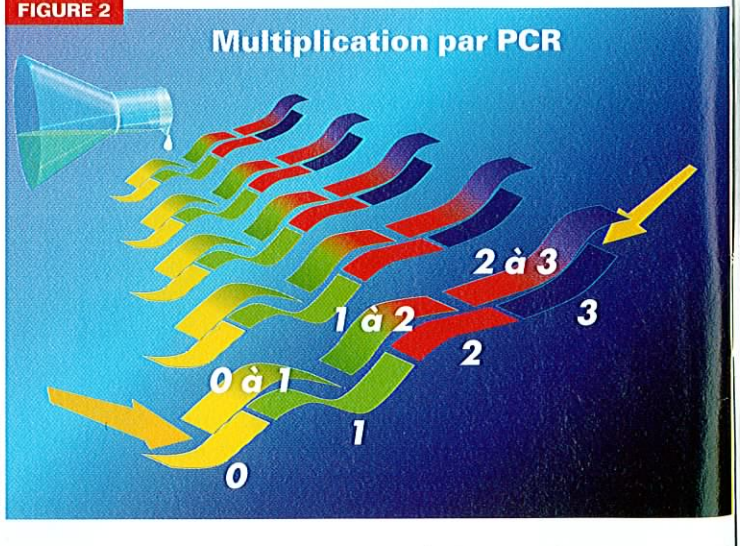
Pour exécuter la première étape (voir la figure 2), chaque point du graphique est codé par un brin d'ADN constitué de vingt bases prises au hasard – un oligonucléotide. Puis, pour chaque flèche du graphique, on génère un oligonucléotide dont les dix premières bases sont complémentaires des dix dernières de son point d'origine et les dix dernières correspondent aux dix premières du point suivant. En utilisant ces séquences complémentaires comme attelles, les séquences d'ADN correspondant à des flèches compatibles devraient s'apparier et être accolées bout à bout par l'enzyme ligase T4. Des molécules d'ADN codant pour des chemins aléatoires sont ainsi produites.

Dans la deuxième étape, le résultat obtenu précédemment est amplifié par PCR (amplification en chaîne par réaction), en utilisant comme amorces des oligonucléotides codant pour v_{in} et v_{out}. La réaction n'amplifie et ne retient ainsi que les molécules

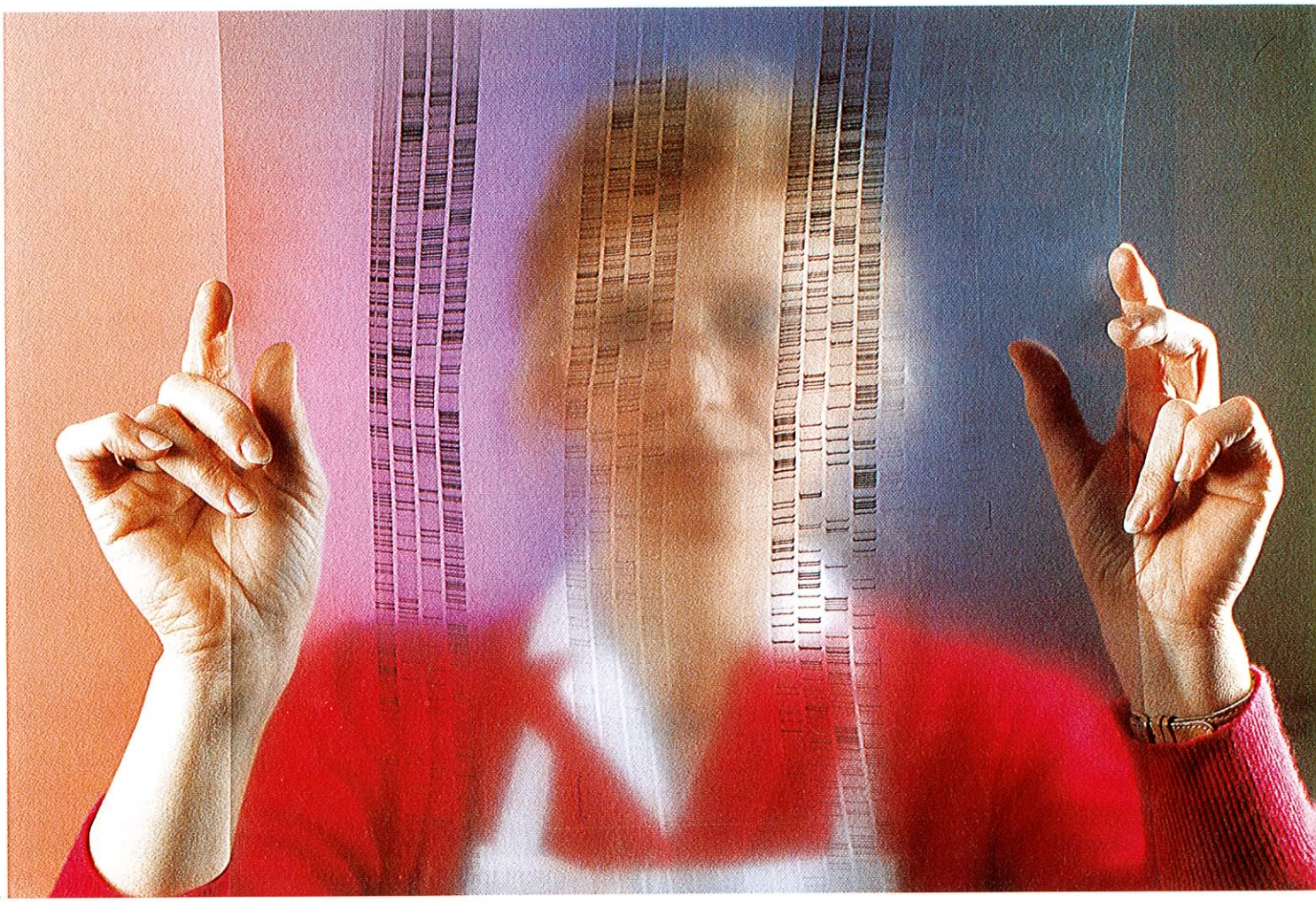
Le meilleur chemin

Soit 7 villes par lesquelles un voyageur de commerce doit passer, quel est l'itinéraire le plus court les reliant toutes ensemble ? En mathématiques, cela correspond au cas à 7 points du problème dit du « chemin hamiltonien orienté ». L. Adleman y répond en s'appuyant sur la biologie moléculaire (à droite, technique classique de décryptage d'un gène). Chaque ville est codée par un brin d'ADN de vingt bases prises au hasard. Pour chaque flèche, on génère des brins d'ADN dont les dix premières bases correspondent à la ville d'origine, et les dix dernières à celle d'arrivée. On en fait ensuite de nombreuses copies par la technique de PCR (ci-dessous).

FIGURE 2



DESSINS A. MEYER



J.C. RÉVY - CNRI

codant pour des chemins ayant v_{in} pour départ et v_{out} pour arrivée.

Lors de la troisième étape, une électrophorèse sur gel d'agarose (2) permet de séparer et recueillir les brins d'ADN de longueur désirée. Le chemin recherché, s'il existe, doit passer par les sept points, chacun d'eux ayant été associé à vingt paires de bases. Les produits de la PCR codant ce chemin devraient donc comporter 7x20 = 140 paires de bases.

La quatrième étape s'effectue par purifications successives, pour chaque point autre que le départ et l'arrivée. Il s'agit, à chaque fois, de sélectionner des brins d'ADN codant pour un point donné. Pour ce faire, des brins complémentaires de la séquence codant pour ce point sont fixés à des billes magnétiques. La solution hétérogène de simples brins d'ADN est alors déposée sur les billes : seuls les brins contenant la séquence recherchée sont retenus. Ainsi, les brins qui ne possèdent pas une seule des séquences correspondant aux points sont éliminés.

Dans la cinquième étape, on cherche à vérifier s'il se trouve encore des molécules codant pour un chemin hamiltonien. On y procède par PCR. La première PCR amplifie les résultats de l'étape 4 et confirme ou non l'existence d'une telle molécule, comme dans l'étape 2. Si cette molécule existe, une seconde PCR confirme la présence des points internes du graphique (autres que le départ et l'arrivée) en utilisant comme amorces les séquences d'oligonucléotides complémentaires de chacun de ces points. Fort élégamment, on peut, du même coup, cartographier les liaisons entre les points du graphique, sans recourir à un séquençage d'ADN.

Le problème ainsi résolu par Adleman n'avait rien de trivial. Il ne pouvait être traité que sur un ordinateur non déterministe, c'est-à-dire un ordinateur capable de poursuivre un nombre illimité de calculs indépendants en parallèle. La première expérience de calcul biomoléculaire a donc immédiatement suscité de nombreux espoirs. Depuis, de nombreuses études tant théoriques qu'expérimentales sont venues étendre ce résultat.

2 – L'électrophorèse consiste à soumettre des molécules à un champ électrique, dans un milieu gélifié, afin de les séparer selon leur poids.

Vers un ordinateur à ADN

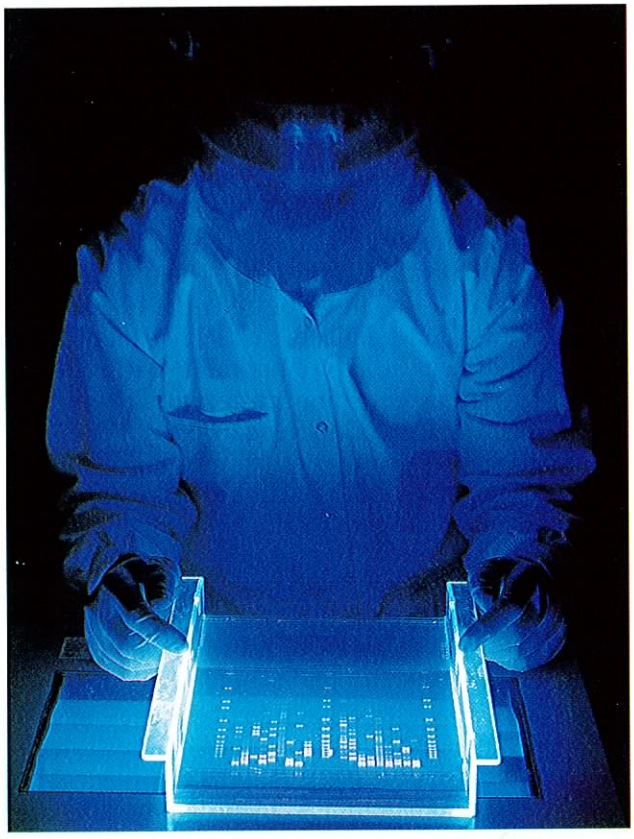
Ces expériences, qui ont utilisé des algorithmes conçus pour résoudre des problèmes mathématiques déterminés, soulèvent deux types d'interrogations. Quelles sont les classes de problèmes mathématiques qui peuvent être résolues par des calculs par ADN ? Est-il possible, au moins en principe, de concevoir un ordinateur à ADN programmable ? Si les modèles proposés diffèrent tous les uns des autres, leurs caractéristiques communes permettent de risquer une réponse à ces questions.

En fait, toute espèce de calculateur – mécanique, électronique ou biologique – doit pouvoir faire deux choses : stocker des données, et effectuer

des opérations sur elles. Il s'agit donc de voir comment des informations peuvent être stockées sur des brins d'ADN, et quelles techniques de biologie moléculaire pourraient servir aux calculs. Pour distinguer clairement les opérations mathématiques des manipulations biomoléculaires effectuées sur des brins d'ADN, on utilisera le terme « bio-opérations » pour désigner ces dernières.

Un brin simple d'ADN peut être décrit comme une chaîne où se combinent quatre symboles différents : A, G, C, T. En termes mathématiques, cela signifie que nous disposons d'un alphabet à quatre lettres pour coder l'information. De fait, c'est plus qu'il n'est nécessaire : pour faire la même chose, un ordinateur électronique n'a besoin que de deux chiffres, 0 et 1.

Quant aux opérations effectuées sur l'ADN, les modèles de calcul biomoléculaire qui ont été proposés utilisent généralement diverses combinaisons des « bio-opérations » suivantes : synthèse d'un brin d'ADN de longueur convenable ; mélange des contenus de deux éprouvettes ; dissociation d'un double brin d'ADN en ses deux composants complémentaires, par chauffage ; annelage, c'est-à-dire l'effet inverse, par refroidissement ; amplification (copie de brins d'ADN par PCR) ; séparation des brins selon la taille, par électrophorèse ou par d'autres méthodes de fractionnement ; section des doubles brins d'ADN en des sites spécifiques à l'aide d'enzymes de restriction ; collage bout à bout des brins d'ADN avec l'enzyme ADN ligase ; substitution, insertion ou délétion de séquences d'ADN par le biais de la PCR ; marquage de simples brins par hybridation ; sélection par affinité de brins contenant une séquence donnée.



SIMON FRASER / S.P.L. / COSMOS

Étape 3 : électrophorèse

L'électrophorèse sur gel d'agarose permet de séparer des fragments d'ADN de tailles différentes. C'est l'étape 3 de l'algorithme d'Adleman. Son but : sélectionner les brins d'ADN longs de sept fois 20 paires de bases.

Un « bio-calcul » est une séquence de bio-opérations effectuées sur des brins d'ADN que renferme un tube à essai. Les opérations recensées ci-dessus – et éventuellement d'autres – peuvent donc être utilisées pour écrire des « programmes ». Un « programme » reçoit un tube contenant des brins d'ADN codant une information comme *input*, et renvoie comme *output* soit les ordres oui ou non, soit une autre série de tubes.

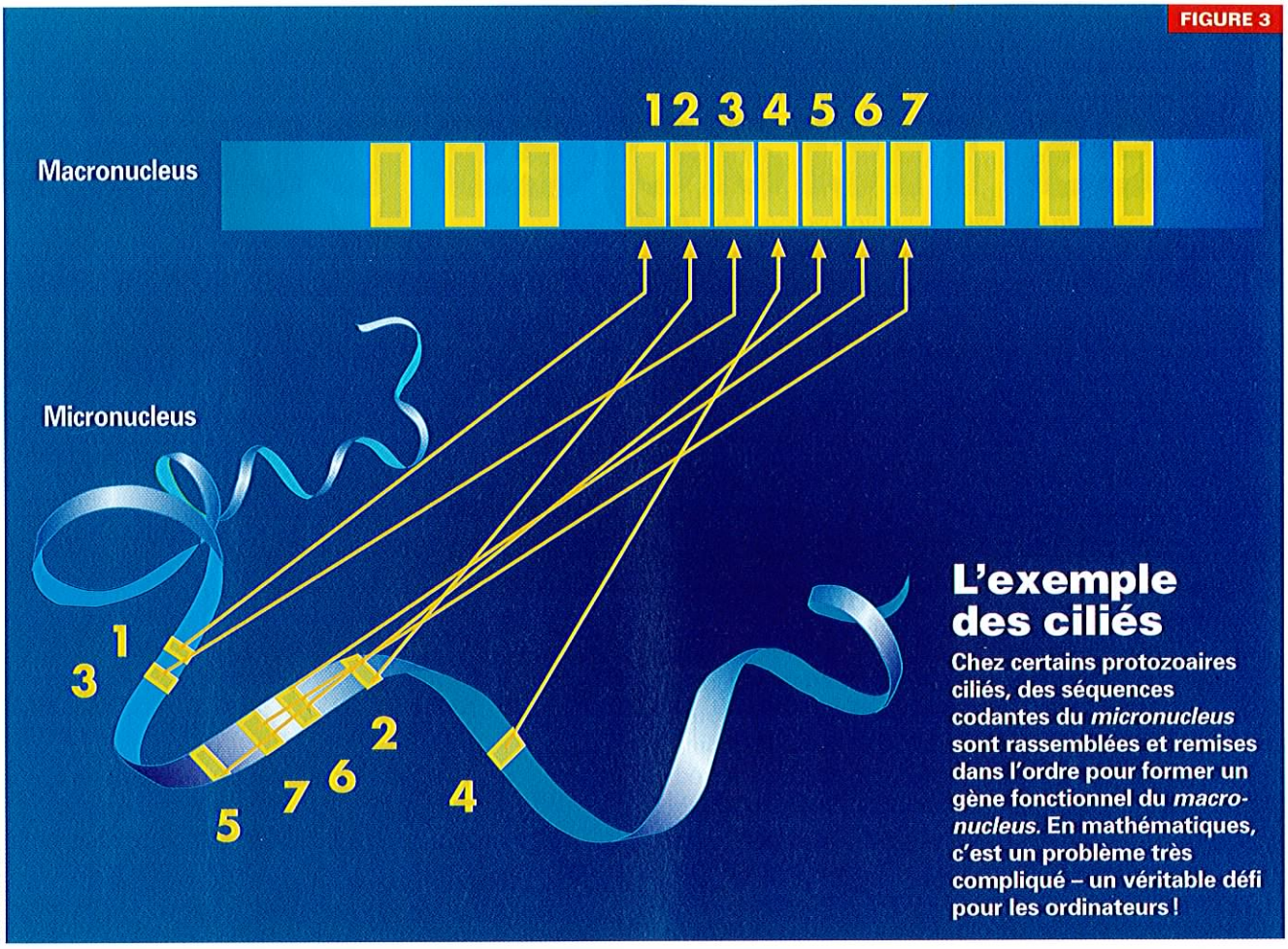
Plusieurs modèles de calcul moléculaire basés sur ces bio-opérations ont été étudiés, tant pour leur puissance de calcul que pour leur faisabilité. S'ils ont tous leurs avantages et leurs inconvénients, leur existence témoigne de la souplesse du calcul par ADN et rend d'autant plus vraisemblable la construction d'un ordinateur à ADN.

Reste cependant à surmonter de nombreux défis techniques, à chaque étape ou presque. Ils viennent principalement de la difficulté à manipuler des systèmes biomoléculaires à grande échelle. On remarquera toutefois que dans les systèmes biologiques, la nature prend en charge des questions telles que le contrôle et l'ajustement des concentrations de molécules, la tolérance à l'erreur... Les cellules, par exemple, doivent ajuster les concentrations de divers composants pour déclencher les réactions qui produiront telle molécule rare ; elles doivent aussi se débrouiller avec les sous-produits indésirables engendrés par leur propre activité. Puisqu'elles parviennent à gérer ces problèmes *in vivo*, on pourrait s'en inspirer pour en faire autant *in vitro*.

Côté théorie, on cherche, entre autres, à obtenir un modèle formel permettant de décrire les calculs par ADN. Cette approche compare fréquemment la puissance de calcul d'un tel modèle à celle d'une machine de Turing, le modèle formel des ordinateurs électroniques actuels. Le système contextuel d'insertion/délétion en constitue une bonne illustration. Réalisable en laboratoire, il possède, en outre, toute la puissance d'une machine de Turing.

Un tel système est formé à partir d'un ensemble de lettres ou symboles, un alphabet. Pour le calcul par ADN, notre alphabet comporte quatre lettres, A, C, G, et T. Ces symboles pourront être assemblés en chaînes, de manière à former des mots. Un certain nombre de mots sont pris pour axiomes du système. On s'intéresse alors à l'ensemble des mots qui peuvent être obtenus à partir des axiomes par application répétée de règles d'insertion et de délétion. Ces dernières, spécifiées au départ, indiquent entre quelles chaînes de lettres il est possible d'insérer (ou de supprimer) tel ou tel mot. Un modèle formel de

FIGURE 3



L'exemple des ciliés

Chez certains protozoaires ciliés, des séquences codantes du *micronucleus* sont rassemblées et remises dans l'ordre pour former un gène fonctionnel du *macronucleus*. En mathématiques, c'est un problème très compliqué – un véritable défi pour les ordinateurs !

DESSIN A. MEYER

ce type permet de reproduire les actions de n'importe quelle machine de Turing.

L'intérêt d'un tel modèle n'est pas uniquement théorique. En effet, aux opérations formelles d'insertion et de délétion, on peut faire correspondre des manipulations biomoléculaires : en utilisant des oligonucléotides de synthèse et la technique de mutation génétique localisée par PCR, on insérera ou supprimera des séquences d'oligonucléotides dans des contextes donnés. Le processus de traduction de l'ARN peut servir aux mêmes fins. Des systèmes d'insertion/délétion très simples pourraient suffire : l'opération est réalisable, même avec de fortes restrictions sur la longueur des contextes d'insertion et de délétion, et sur celle des mots insérés ou supprimés.

Les solutions de la nature

Le calcul biomoléculaire éveille ainsi l'espoir d'un nouveau type d'ordinateurs, radicalement différents des ordinateurs actuels. Il jette aussi une lumière neuve sur les phénomènes biologiques étudiés *in vivo*. Sans doute pourra-t-il nous éclairer sur les capacités informationnelles de l'ADN, et sur l'étendue des processus de calcul existant dans la nature. On prendra un exemple : le décryptage des gènes, étudié chez des protozoaires ciliés.

Ces derniers possèdent deux types de noyaux : un *macronucleus* actif, et un *micronucleus* fonctionnellement inerte, qui contribue seulement à la reproduction sexuelle. Le premier se développe à partir du second, après la reproduction. Chez certaines espèces, dans le *micronucleus*, les copies de certains gènes codant pour une protéine sont brouillées par l'insertion de séquences non codantes d'ADN. Qui plus est, elles se présentent parfois dans un autre ordre que dans la réplique macronucléaire. En élucidant le puzzle, et en rassemblant les séquences fonctionnelles, ces protozoaires résolvent un difficile problème de calcul.

Ce processus de décryptage ressemble de près à l'algorithme utilisé par Adleman. Le protozoaire utilise l'information contenue dans des unités de répétition de 2 à 14 nucléotides, pour initier une série de recombinaisons homologues. Ainsi, la séquence d'ADN présente à la jonction entre un gène codant n et le gène non codant qui le suit, est généralement la même que celle qui occupe l'espace entre le gène codant n + 1 et le non-codant qui précède. Ces séquences identiques permettent l'aboutement du gène n au gène n + 1 : dans le graphique d'Adleman, ce processus permet d'assembler, dans le bon ordre, les flèches allant d'un point à un autre. Et *in fine,* on aboutit au chemin hamiltonien recherché. Sauf que dans le cas précis, ce n'est plus un chemin à sept points, mais à environ 50. Ce qui, pour n'importe quel ordinateur, est un véritable défi !

La traduction et le décryptage des gènes constituent un ensemble unique de paradigmes pouvant être utilisés pour les calculs biologiques. Qui plus est, ces processus soulignent la diversité des paradigmes de calcul qui existent dans les systèmes biologiques : il y a bel et bien pléthore de modèles dont pourraient s'inspirer les mathématiques… ■