

LOGIQUE & CALCUL

Le vivant, plus fort que l'électronique

Des systèmes biologiques sont capables d'enregistrer de l'information et de la manipuler, comme le font les ordinateurs. Fondées sur ce constat, de nouvelles technologies informatiques pourraient concurrencer celles d'aujourd'hui.

Jean-Paul DELAHAYE

L'ADN est comme un programme d'ordinateur, mais bien plus avancé que n'importe quel logiciel jamais créé.

Bill Gates

L'idée réductionniste que tout n'est que du calcul dans l'Univers procède de l'idée (passée de mode) que l'Univers n'est qu'une grande horloge. On aurait seulement changé le modèle de référence, qui, de mécanique, serait devenue informatique.

En passant d'un objet matériel et rigide, l'horloge, à un objet idéal plus abstrait et polyvalent, l'ordinateur, nous avons trouvé un moyen de nous représenter le monde d'une manière plus générale et efficace. La métaphore informatique de l'Univers n'est peut-être pas la dernière, mais l'ordinateur est incontestablement bien plus commode que l'horloge pour modéliser la variété et la complexité des objets et des situations du monde. De plus, s'appuyer sur les ordinateurs pour penser l'Univers, c'est considérer que les concepts physiques classiques doivent être complétés : à côté du temps, de la longueur et de la masse, il faut, pour comprendre ce qui se passe, utiliser les concepts d'information et de calcul.

Or on trouve dans les mécanismes biologiques des composants qui, bien utilisés, réalisent des calculs à la demande quand on a maîtrisé leur programmation. L'idée que la biologie, aussi éloignée qu'elle paraisse de l'informatique, est aussi une science du

calcul et de l'information devient année après année plus concrète.

La coopération entre informatique et biologie se développe selon deux axes complémentaires. D'une part, en s'inspirant du monde biologique, on a conçu de nouvelles façons de programmer nos ordinateurs. Avec ces méthodes informatiques inattendues, on résout parfois mieux qu'auparavant des problèmes d'ingénierie.

Imiter le vivant et le faire calculer

Ainsi, les techniques de réseaux de neurones artificiels ont été mises en œuvre pour faire gagner au jeu de go, contre un champion humain, un système informatique (voir Yoshua Bengio, « La révolution de l'apprentissage profond », *Pour la Science* n° 465, juillet 2016).

Il y a aussi les algorithmes génétiques, où l'on mime des mécanismes évolutifs qui sélectionnent des objets informatiques, les systèmes immunitaires artificiels, où l'on copie les principes adaptatifs de l'immunité des vertébrés, les systèmes d'intelligence distribuée, où l'on s'inspire des insectes sociaux qui, collectivement, réalisent d'étonnantes constructions telles que ruches ou termitières, les méthodes de calcul par les membranes, où l'on copie les échanges cellulaires pour imaginer des systèmes algorithmiques nouveaux.

L'autre axe pour relier biologie et informatique, qui sera le sujet de la suite de

l'article, consiste à identifier et à développer des méthodes autorisant une programmation du vivant et l'exploitation de ses capacités à mémoriser de l'information, cela en utilisant uniquement les molécules et les structures biologiques (sans électronique). Deux domaines aujourd'hui ont réussi cet exploit : le calcul et le stockage d'informations fondés sur l'ADN (*DNA computing* et *DNA data storage*) et le calcul avec des cellules (*Synthetic biological circuit* et *Bacterial computing*).

L'informaticien américain Leonard Adleman est bien connu pour sa découverte, avec Adi Shamir et Ronald Rivest, de l'algorithme de chiffrement asymétrique RSA : c'est lui le « A ». Il a été inspiré par la machine de Turing, un mécanisme de calcul imaginé en 1936 dont la mémoire est constituée par un ruban potentiellement infini qu'un automate parcourt, lit et sur lequel il écrit. Or l'ADN polymérase est une molécule qui s'accroche à un brin d'ADN, puis se déplace dessus en « lisant » chaque base rencontrée et en « écrivant » son complément, ce qui fabrique un nouveau brin d'ADN. Leonard Adleman y reconnut un mécanisme proche de celui des machines de Turing, qui peuvent calculer toute fonction mathématique calculable par algorithme. L'analogie l'incita à utiliser des brins d'ADN et les mécanismes moléculaires qui les manipulent pour calculer.

Leonard Adleman comprit cela au cours d'une « nuit de Pascal ». Il raconte : « Je me dressais soudain et fis remarquer à mon

Le chemin hamiltonien codé par des séquences d'ADN

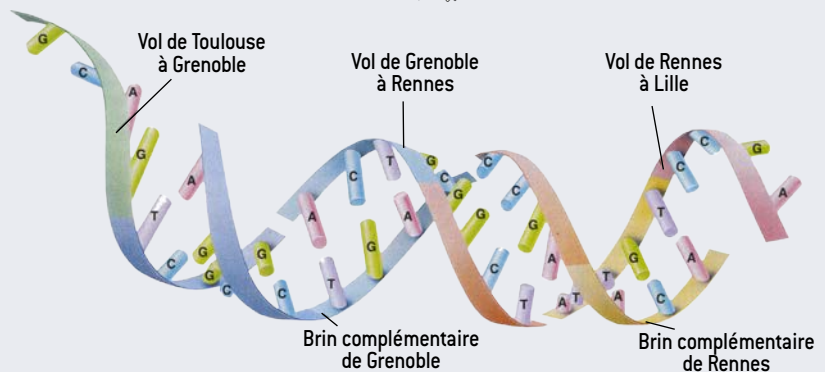
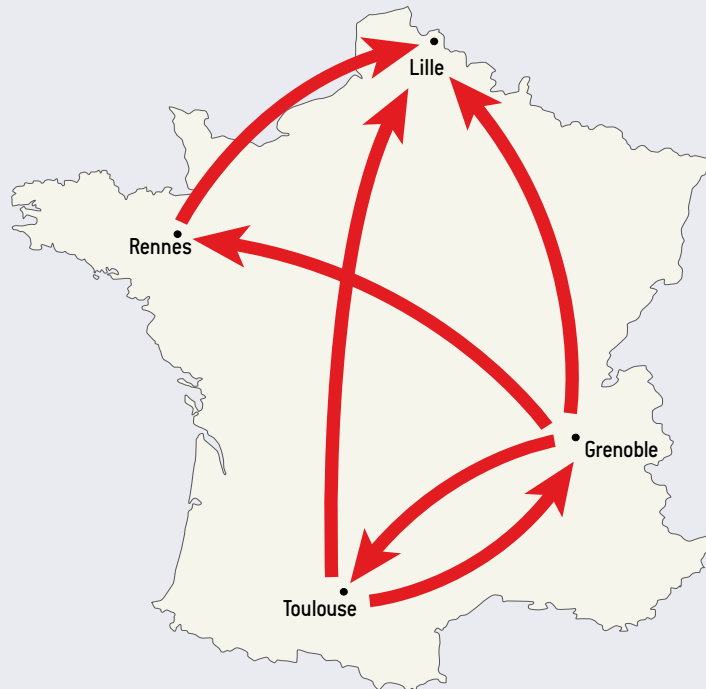
La méthode conçue par l'informaticien américain Leonard Adleman pour résoudre le problème des chemins hamiltoniens est expliquée à partir d'un exemple (emprunté à son article « Calculer avec l'ADN », *Pour la Science*, n° 252, pp. 56-63, octobre 1998).

Imaginons que les vols reliant quatre villes soient ceux dessinés sur le schéma ci-contre. On peut par exemple aller de Grenoble à Lille, mais pas l'inverse. On cherche un chemin partant de Toulouse et arrivant à Lille, qui passe par chacune des quatre villes une fois et une seule.

On choisit pour chaque ville une séquence d'ADN. Pour Toulouse, par exemple, c'est la séquence ACTTGACG. On nomme « début » les quatre premiers nucléotides, ACTT, et « fin » les quatre derniers, GACG. On peut alors coder chaque vol en mettant bout à bout la fin de la séquence associée à la ville de départ et le début de la séquence associée à la ville d'arrivée (voir le tableau) et chaque chemin combinant plusieurs vols sera représenté par l'enchaînement des séquences d'ADN correspondantes.

À chaque ville correspond aussi une séquence complémentaire, obtenue en échangeant C et G, et A et T. L'utilisation de ces séquences complémentaires permet de filtrer, dans une soupe contenant au départ les molécules codant tous les chemins possibles jusqu'à la longueur 3 au moins (correspondant au trajet reliant les quatre villes), les molécules qui correspondent aux chemins passant par toutes les villes.

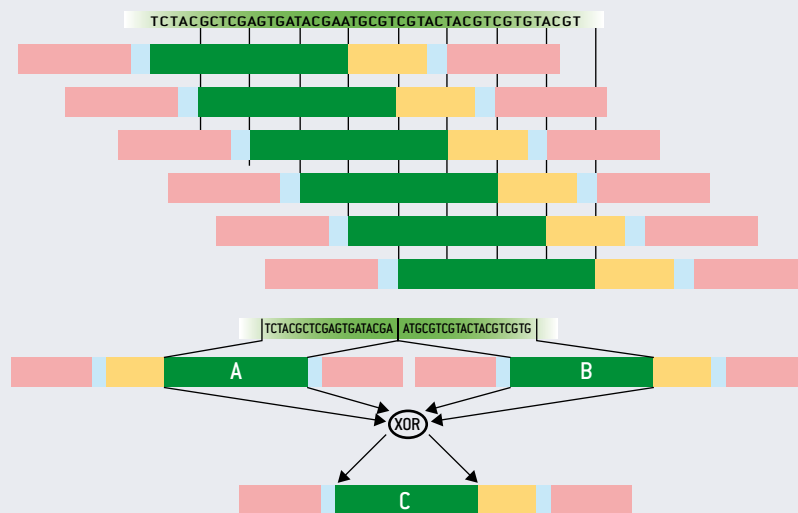
Dans notre problème, il y a un chemin hamiltonien entre Toulouse et Lille, qui sera le seul retenu ayant la longueur voulue et passant par chaque ville. Il passe dans l'ordre par Toulouse, Grenoble, Rennes et Lille. Il est représenté par la séquence GCAGTCGGACTGGGCTA-TGTCCGA, obtenue à l'issue des manipulations de la soupe initiale opérant les filtrages.



Ville	ADN	Complémentaire
TOULOUSE	ACTTGACG	TGAACGTC
GRENOBLE	TCGGACTG	AGCCTGAC
RENNES	GGCTATGT	CCGATACA
LILLE	CCGAGCAA	GGCTCATT
Vol	Code ADN du vol	
TOULOUSE-GRENOBLE	GACGTGG	
TOULOUSE-LILLE	GACCCGA	
GRENOBLE-RENNES	ACTGGGCT	
GRENOBLE-LILLE	ACTGCCGA	
GRENOBLE-TOULOUSE	ACTGACTT	
RENNES-LILLE	ATGTCCGA	

Mémorisation dans l'ADN

James Bornholt et ses collègues ont conçu et mis en application un procédé de mémorisation massive d'informations exploitant les séquences d'ADN. Le procédé se fonde sur l'inscription de couples clé-données, ce qui permet un accès direct aux informations enregistrées analogue à ce qui existe dans la mémoire d'un ordinateur utilisant un système d'adresses : pas besoin, contrairement au cas d'une bande magnétique, de tout parcourir pour retrouver une information.



Le défi relevé était particulièrement difficile pour plusieurs raisons. D'abord, les méthodes de synthèse et de lecture de l'ADN sont imparfaites, avec des taux d'erreurs de l'ordre de 1 % par nucléotide. Les séquences se dégradent aussi progressivement. Il fallait donc concevoir un système fortement redondant. Selon le type de données (image, son, texte), la présence d'une erreur dans les données est plus ou moins grave ; il fallait alors prévoir un système ajustable de redondances.

C'est ce que propose leur procédé. Le point le plus important de leur travail est cependant la mise en place d'un codage qui autorise l'accès à certaines données sans avoir à tout lire. Le système d'adressage conçu utilise la réaction en chaîne par polymérase (PCR, *polymerase chain reaction*) qui amplifie, et donc donne accès, uniquement aux données souhaitées. Bien évidemment, cette lecture choisie accélère l'accès aux données puisqu'elle évite de parcourir tout l'ADN

d'une base de données moléculaire. Notons toutefois que les temps d'accès restent bien supérieurs à ceux de l'électronique et se comptent en heures.

L'information sur les séquences d'ADN doit être inscrite avec des redondances (car la transcription et la lecture amènent des erreurs), ce que plusieurs procédés permettent. On inscrit la longue séquence en s'arrangeant pour que chaque brin d'ADN (plus court que la séquence qu'on veut inscrire) contienne une portion de la séquence à mémoriser et qu'il y ait des chevauchements entre les brins (*schéma du haut*). L'idée utilisée par l'équipe de James Bornholt consiste, pour inscrire la séquence A et la séquence B, à inscrire A sur un brin, B sur un autre et le OU exclusif (XOR) des séquences A et B (après les avoir traduites en binaire), noté C, sur un troisième brin (*schéma du bas*). Les trois séquences A, B, C ont alors pour propriété que la connaissance de deux quelconques d'entre elles (A et B, A et C, B et C) suffit pour retrouver A et B.

épouse : "Bon sang ! Ces choses peuvent calculer." Je ne dormis pas de la nuit, cherchant un moyen d'utiliser l'ADN pour résoudre des problèmes. » La mise au point de mécanismes de calcul fut cependant délicate, car ce que fait la cellule se décompose en opérations élémentaires qui ne permettent pas facilement d'effectuer des opérations complexes. En plus de la copie de brins d'ADN par l'enzyme ADN polymérase, on dispose d'enzymes nommées ligases qui raboutent des molécules d'ADN et d'« enzymes de restriction » qui coupent les molécules d'ADN à des endroits précis. L'enzyme EcoRI, par exemple, coupe l'ADN de la bactérie *Escherichia coli* après le G de la séquence de nucléotides GAATTC.

Des soupes pour mémoriser et calculer

Poursuivant son travail nocturne, dans les jours qui suivirent son eureka, Leonard Adleman conçut et réussit à faire fonctionner un procédé de calcul à base d'ADN et d'enzymes qui résout le problème combinatoire de la recherche des chemins hamiltoniens dans un graphe. Il s'agit de trouver un chemin dans un graphe qui passe par chaque nœud du graphe une fois et une seule. Organiser une tournée visitant des villes fixées à l'avance sur un réseau routier ou aérien revient à résoudre un tel problème.

Ce problème appartient à la classe des problèmes NP-complets. Cela signifie deux choses : d'une part, qu'on ne connaît aucun algorithme rapide qui le résout dès que le nombre de nœuds du graphe devient important (« rapide » signifie que le temps d'exécution est une fonction polynomiale du nombre de nœuds du graphe) ; d'autre part, si l'on connaissait un moyen de résoudre rapidement ce problème, on pourrait aussi résoudre rapidement, c'est-à-dire en temps polynomial, tous les problèmes de la classe NP (problèmes dont on peut vérifier qu'une solution est correcte en temps polynomial).

L'algorithme proposé et mis en œuvre par Leonard Adleman en 1994 avec des soupes moléculaires réelles est le suivant :

- On engendre toutes les séquences possibles correspondant à des chemins