La lecture du génome humain

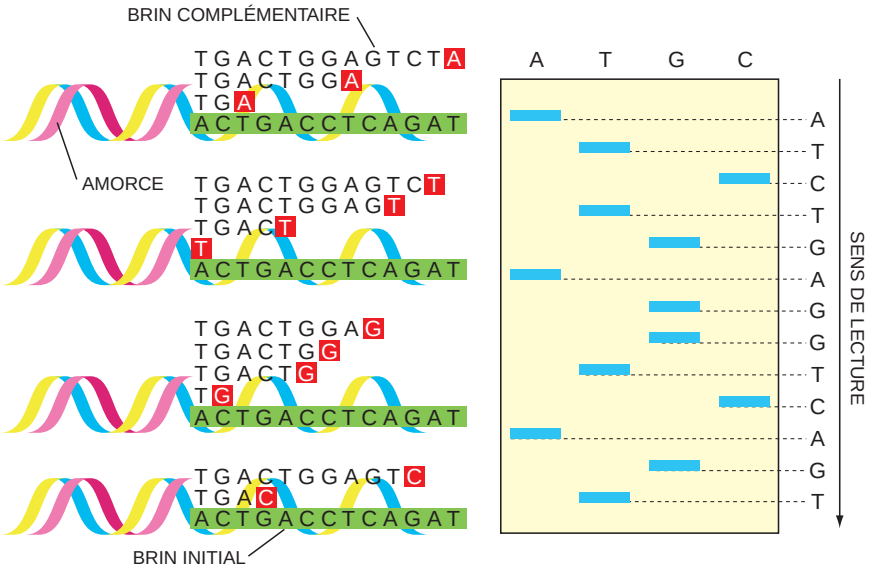
KATHRYN BROWN

Le décryptage de l'information génétique humaine a été une rude épopée... Pourtant l'aventure ne fait que commencer.

À l'heure où vous lisez cet article, l'essentiel de l'information génétique d'un être humain est disponible sur le réseau *Internet* : le profane trouvera, répétées, les lettres *A*, *T*, *C*, *G* qui font un texte qui remplirait plus de 200 annuaires téléphoniques. Les biologistes y lisent un enthousiasmant roman : celui de l'être humain.

En effet, les lettres *A*, *T*, *C* et *G* désignent les nucléotides, c'est-à-dire les sous-unités chimiques de l'ADN, la molécule qui porte les gènes et commande le fonctionnement de toutes nos cellules. Le *Projet génome humain*, qui a coûté plus de 1,8 milliard de francs, a établi la carte complète de nos gènes, composée de six milliards de nucléotides. Ce consortium public, qui réunit depuis dix ans plus de 1 100 scientifiques, fédère quatre grands centres de séquençage aux États-Unis, le Centre Sanger, en Angleterre, et divers laboratoires au Japon, en France, en Allemagne et en Chine.

Ce travail de fourmi est resté discret jusqu'en avril 2000, lorsque la Société privée *Celera Genomics* a annoncé sa propre évaluation du génome humain. Cette rivalité du privé contre le public a mis en vedette le génome de l'être humain, son séquençage et son utilisation. Détaillons l'histoire de ce séquençage. En quoi peut-il améliorer notre quotidien, et notamment la médecine? Quelles contraintes ses applications devront-elles respecter?



1. LE SÉQUENÇAGE DE L'ADN selon la méthode de F. Sanger s'effectue avec des molécules fluorescentes *(en rouge)* qui bloquent la synthèse d'un brin complémentaire à partir d'un brin initial *(en vert)*. À partir d'une amorce *(en rose)*, une enzyme greffe les nucléotides complémentaires : en ajoutant la molécule fluorescente, on obtient des fragments de longueurs différentes, que l'on sépare sur un gel d'électrophorèse selon leur taille *(à droite)*. Lorsque cette opération est répétée pour chaque type de nucléotide, on lit la séquence de l'ADN directement sur le gel.

De l'ADN au séquençage

En 1953, Francis Crick et James Watson découvrent la structure de l'ADN, la molécule que l'on savait être le support de l'hérédité de tous les organismes vivants : deux chaînes d'éléments nommés nucléotides (représentés par les lettres *A*, *T*, *C* et *G*) s'apparient de façon qu'un *A* soit toujours face à un *T*, et un *C* face à un *G* ; les deux chaînes forment une double hélice. Puis, en 1966, Marshall Nirenberg et H. Gobind Khorana déchiffrent le code génétique, c'est-à-dire qu'ils découvrent comment une succession de nucléotides code la synthèse d'une protéine. L'ADN est une gigantesque bibliothèque où sont stockés tous les programmes de synthèse des protéines, dont une cellule a besoin pour fonctionner. Chaque programme est un gène : déchiffrer, ou séquencer l'ADN d'une cellule, consiste à «lire» toutes les lettres de l'ADN, afin, ensuite, d'identifier tous les gènes. Les biologistes ont d'abord séquencé le génome de divers organismes simples : la levure, la mouche... *(voir l'encadré de la page 43)*, puis ils ont cherché à identifier celui d'un être humain, mais comment lire les trois milliards de lettres qui le compose?

L'approche des laboratoires du *Projet génome humain* est laborieuse, mais précise. À partir de cellules sanguines et sexuelles (des spermatozoïdes), les généticiens ont isolé les

23 paires de chromosomes. Ils ont alors coupé l'ADN de chaque chromosome en grands fragments d'environ 100 000 nucléotides chacun ; puis après les avoir clonés, c'est-à-dire reproduits chacun en de nombreux exemplaires, ils les ont ordonnés en y détectant de courtes séquences d'ADN dont on connaît la localisation et qui jouent le rôle de balises. Les grands fragments ordonnés ont ensuite été à leur tour découpés en fragments plus petits, dont on a identifié la séquence complète par une méthode qui dérive de celle que le biochimiste britannique Frederick Sanger mit au point en 1977. Dans cette méthode, on utilise une enzyme nommée ADN polymérase, qui, à partir d'une amorce, lit un brin et fabrique un brin complémentaire en additionnant des nucléotides libres qu'elle trouve dans la solution. Lorsque la synthèse est en cours, à l'aide de molécules fluorescentes que l'on ajoute après différents laps de temps, on bloque la synthèse à des stades divers et l'on obtient des fragments de longueurs diverses, que l'on sépare selon leur masse, en les plaçant sur un gel, dans un champ électrique : la vitesse de déplacement de chaque fragment dépend de sa masse et de sa charge électrique. En juxtaposant les quatre résultats (un pour chaque type de nucléotide), un lecteur optique d'un séquenceur automatique localise la fluorescence et recompose finalement la séquence du morceau d'ADN initial *(voir la figure 1)*. Les informations sont stockées dans des banques de données, sur d'énormes disques durs, et gérées par des programmes d'analyse et de récupération de données *(voir l'article de Ken Howard sur la bio-informatique)*.

Ainsi, patiemment, les généticiens ont reconstruit les séquences de tous les chromosomes, puis celles du génome entier. Cette méthode, appliquée à une encyclopédie, consisterait à extraire les pages, successivement, puis à déchirer chaque page en morceaux, à identifier ces derniers, à reconstituer la page et, enfin, l'encyclopédie entière.

En revanche, les généticiens de la Société *Celera Genomics*, après avoir testé

2. L'USINE DE SÉQUENÇAGE de la Société *Celera Genomics* dispose de 300 séquenceurs d'ADN automatisés, capables d'identifier des dizaines de millions de nucléotides par jour.

Kay Chernush

leurs méthodes sur le génome de la drosophile, ont choisi de déchiqueter toute l'encyclopédie en une seule fois ! Avec cette technique, dite aléatoire, tout le génome est découpé en fragments que l'on séquence selon la même technique que celle utilisée par le consortium public. Le travail de reconstitution du génome à partir des fragments est effectué par de puissants ordinateurs : tout repose sur la puissance de calcul et sur l'utilisation d'algorithmes pour analyser les données *(voir la figure 3)*.

Au mois d'avril 2000, les équipes de *Celera Genomics* ont annoncé que les

L'inventaire des gènes

L'interprétation de la séquence du génome humain n'est pas un exercice aussi simple que prévu. Les outils bio-informatiques actuels butent, d'une part, sur des difficultés connues et, d'autre part, ne sont pas adaptés à traiter des données de séquence encore inachevées. À l'exception des deux plus petits chromosomes (21 et 22), dont la séquence est complète, il est aujourd'hui impossible de procéder à un inventaire sérieux des gènes humains. Nous restons limités à des estimations globales parfois discutables. La récente controverse sur le nombre de gènes du génome est un parfait exemple de notre niveau d'ignorance.

L'alternance des régions codantes, nommées exons, et des régions non codantes, les introns, des gènes des organismes multicellulaires est une source de grandes difficultés pour l'interprétation des séquences du génome et, notamment, pour l'identification des gènes. Cette complication est encore accrue par le phénomène d'«épissage alternatif» qui, à partir d'un seul gène, donne naissance à plusieurs ARN messagers, c'est-à-dire à plusieurs protéines. Connu depuis plus de 20 ans, ce mécanisme est plus important que prévu chez les vertébrés.

Pour identifier les gènes dans des séquences d'ADN d'organismes multicellulaires, les généticiens recourent à deux types de méthodes informatiques. D'une part, ils comparent des séquences à l'aide de programmes, tel BLAST, et, d'autre part, ils utilisent des programmes de prédictions des exons et, dans une moindre mesure des gènes entiers.

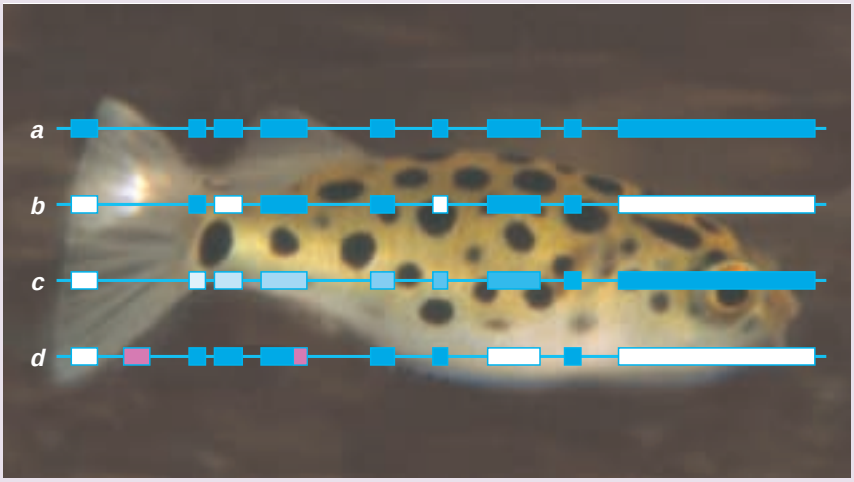
Certains auteurs soulignent l'intérêt des comparaisons avec l'ADN d'autres génomes pour identifier les gènes et pour essayer de connaître les fonctions de leurs protéines. C'est précisément afin d'identifier des gènes humains au moyen de telles comparaisons, que nous poursuivons, au *Genoscope*, le séquençage du génome du poisson Tetraodon *(voir la figure)*. Ces poissons ont un génome compact environ huit fois plus petit que celui des mammifères, mais qui contient les mêmes gènes. Les estimations les plus récentes à partir de ces comparaisons donnent moins de 30 000 gènes chez l'être humain.

Appliquées aux ADN complémentaires, c'est-à-dire des ADN synthétisés à partir d'ARN composés des seuls exons, les comparaisons de séquences, délimitent sans hésitation les introns et les exons d'un gène. Malheureusement, la plupart des séquences d'ADN complémentaires sont incomplètes : au *Genoscope*, nous avons entrepris un programme d'identification des séquences entières d'ADN complémentaires pour remédier à cette situation.

Les programmes informatiques de prédiction utilisent de nombreux critères pour identifier les exons des gènes. Cependant, ils en détectent souvent qui n'existent pas dans la réalité. En outre, les résultats des analyses informatiques sont issus de calculs et ne correspondent pas nécessairement à la réalité : valider ces prédictions par une démarche expérimentale est donc indispensable.

En raison de ces difficultés et de quelques autres, l'inventaire des gènes du génome humain est loin d'être achevé et la controverse actuelle sur le nombre de gènes humains ne se résorbera pas avant des mois voire quelques années. Il est certes capital d'entreprendre dès à présent des programmes à grande échelle pour l'exploration de la fonction des gènes et des protéines, mais il serait grave de négliger de faire un inventaire aussi exhaustif que possible des gènes humains.

Jean WEISSENBACH,
Directeur général du *Genoscope*,
Centre national de séquençage



La composition en exons, des séquences codantes *(rectangles bleus)*, et en introns, des séquences non codantes *(lignes bleues)*, d'un gène *(a)* peut être identifiée par diverses méthodes à partir d'une séquence nouvelle. La première consiste à comparer cette séquence nouvelle à celle de gènes connus. Toutefois, lorsque les gènes ne sont que faiblement apparentés, une fraction des exons *(en blanc)* est ignorée *(b)*. La comparaison à des séquences d'ADN complémentaires est la deuxième méthode. Cependant, les séquences disponibles sont souvent incomplètes : la probabilité de détection des exons varie selon leur position le long du gène *(c, l'intensité du bleu est proportionnelle à la probabilité de détection d'un exon)*. Enfin, la troisième méthode requiert un programme informatique de prédiction *(d)* ; certains exons réels ne sont pas prédits, alors que d'autres, qui n'existent pas, le sont *(en rose)*. D'autres encore ne sont pas prédits intégralement *(en rose et blanc)*. Le recours simultané à plusieurs méthodes augmente la qualité de l'identification d'un gène.