

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. Connus 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.

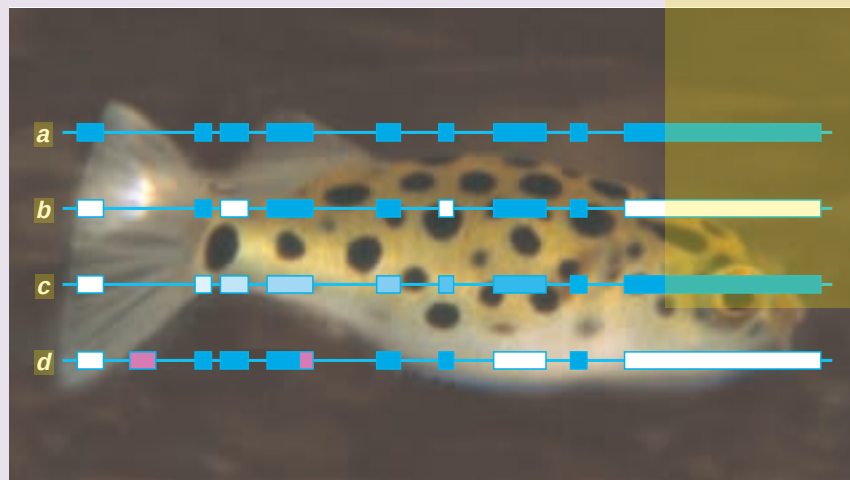
Certains auteurs soulignent l'intérêt des comparaisons avec l'ADN d'autres génomes pour identifier les gènes et pour essayer de connaître les fonctions de leurs protéines. C'est précisément afin d'identifier des gènes humains au moyen de telles comparaisons, que nous poursuivons, au *Genoscope*, le séquençage 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.

résultats du séquençage «approximatif», c'est-à-dire à plus de 90 pour cent, du génome d'une personne anonyme seraient disponibles dans un délai de six semaines. Le consortium public a immédiatement crié au coup bas, arguant que le séquençage approximatif du génome d'une seule personne, et vérifié seulement trois fois, était loin des objectifs annoncés lors de la formation de la société, en 1998 : elle voulait alors séquencer entièrement les génomes de plusieurs personnes, et vérifier le résultat dix fois.

Et maintenant ?

De surcroît, la Société *Celera Genomics* a utilisé les résultats que les équipes du *Projet génome humain*, consortium public, déposaient régulièrement sur une banque de données publique, nommée *GenBank*, accessible par le réseau *Internet*.

Au début du mois de mai 2000, les généticiens du *Projet génome humain* annonçaient à leur tour la fin du séquençage de 90 pour cent du génome humain, ainsi que la séquence complète du chromosome 21, présent en trois copies au lieu de deux chez les personnes atteintes de trisomie 21 et qui contient les gènes de plusieurs autres maladies génétiques (la séquence complète du génome humain s'achèvera en 2003). Avec le séquençage, une première étape est achevée ; disposant des mots du génome, les généticiens cherchent aujourd'hui la grammaire et la syntaxe du message génétique, tandis que les sociétés pharmaceutiques étudient de nouveaux traitements fondés sur l'utilisation de ces informations.

Comment utiliser cette liste de trois milliards de lettres ? Les généticiens ont longtemps pensé qu'après le séquençage de l'ADN ils comprendraient qui nous sommes, pourquoi nous contractons des maladies, pourquoi nous vieillissons... Ils avaient tort : aujourd'hui, ils ne peuvent encore qu'imaginer à quoi servira la séquence du génome (voir l'article de *Caroll Ezzell*).

Ainsi, dans le domaine de la médecine, certaines sociétés pharmaceutiques espèrent fabriquer des médicaments sur mesure : un pharmacien vous délivrera un médicament pour la tension artérielle adapté à votre seul profil génétique, tandis que la personne suivante obtiendra une

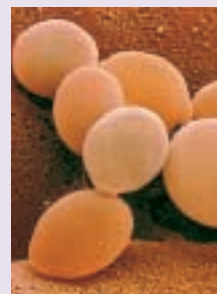
Les «autres» génomes

Qu'avons-nous de commun avec des mouches, des vers, des levures ou des souris ? *A priori*, pas grand-chose. Pourtant les généticiens utilisent les génomes de ces organismes – dits modèles – pour étudier diverses maladies humaines, notamment le cancer et le diabète.

Les gènes des organismes modèles sont intéressants, car les protéines qu'ils codent ressemblent souvent à celles des êtres humains : un gène humain a souvent son équivalent chez des vers ou chez des drosophiles (ou mouches du vinaigre). Examinons le séquençage génétique des principaux organismes modèles.

La levure

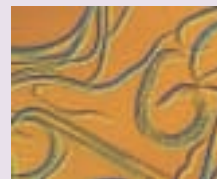
En 1996, la levure de boulangerie *Saccharomyces cerevisiae* a été le premier organisme pourvu d'un noyau dont les secrets génétiques ont été percés. Environ 2 300 (38 pour cent) des protéines de levure ont un équivalent chez les mammifères. Grâce à cet organisme, les biologistes ont découvert les mécanismes fondamentaux du contrôle de la division cellulaire et de la réparation de l'ADN, des processus importants dans le cancer. Puis ils ont élucidé le mode d'action de certains médicaments anticancéreux, tel le cisplatine, qui tue les cellules cancéreuses dotées d'une altération spécifique de leur faculté de réparation de l'ADN.



A. Syred SP/Photo Researchers, Inc.

Le ver

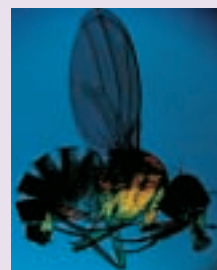
Le génome du ver nématode *Caenorhabditis elegans* a été séquencé en 1998. Les biologistes ont alors découvert qu'un tiers des protéines de ce ver, soit plus de 6 000, ressemblent à celles de mammifères. Aujourd'hui, plusieurs sociétés utilisent ces petits vers – d'environ un millimètre de long – dans des tests de dépistage automatisés, pour rechercher de nouveaux médicaments. Par exemple, pour identifier un principe actif de médicament contre le diabète, on place dans un petit récipient un à dix vers dont une mutation du gène du récepteur de l'insuline bloque la croissance, puis on ajoute des molécules variées et l'on identifie celles qui rétablissent la croissance des vers : elles pallient l'anomalie génétique. Comme certains diabètes résultent également d'une mauvaise utilisation de l'insuline par les cellules, la méthode conduit à de nouveaux traitements contre le diabète.



S. Stammers SP/Photo Researchers, Inc.

La drosophile

Le séquençage du génome de la drosophile s'est terminé en mars 2000. Il a été effectué par des organismes publics et par la Société *Celera Genomics*. Environ 60 pour cent des 289 gènes connus de maladies humaines ont un gène homologue chez cette mouche, et 50 pour cent des protéines de drosophiles ressemblent à des protéines de mammifères. Par exemple, en mars 2000, des généticiens ont identifié chez la drosophile un gène homologue au gène humain *p53* : un gène suppresseur de tumeur qui, muté, laisse les cellules devenir cancéreuses. Le gène *p53* intervient dans une voie métabolique qui mène au suicide des cellules génétiquement endommagées. Les cellules de mouches où la protéine P53 est inactivée ne s'autodétruisent plus, après une altération génétique, et prolifèrent anarchiquement. De telles analogies font de la mouche, moins coûteuse que les souris, un modèle parfait pour l'étude des cancers humains dans les cas où ils sont dus à un défaut dans ce gène.



A. Pasicka, Peter Arnold, Inc.

La souris

Aussi appropriés que soient les autres organismes modèles, tous les nouveaux médicaments sont testés sur des mammifères : généralement la souris. Plus de 90 pour cent des protéines de souris identifiées ressemblent à des protéines humaines. Aujourd'hui, environ trois pour cent du génome de la souris est séquencé, et une séquence approximative est prévue pour 2003. Peut-être la Société *Celera Genomics*, qui a décidé d'y consacrer des moyens exceptionnels, y parviendra-t-elle avant ?

Julia KAROW, journaliste à *Scientific American*



The Jackson Laboratory