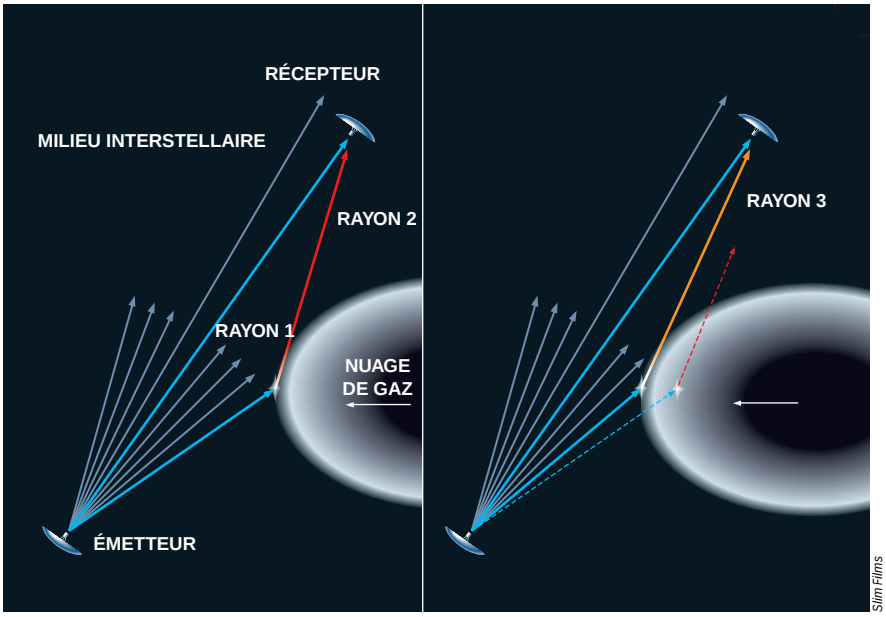
passante plus étroite, les besoins en puissance seraient réduits.

De surcroît, une très grande antenne de réception, constituée d'un assemblage d'antennes et de récepteurs individuels, peut être programmée de manière à recevoir simultanément des faisceaux de nombreuses directions différentes, ce qui faciliterait la recherche des émetteurs inconnus. L'utilisation simultanée de nombreux canaux de réception, un multiplexage en fréquence, déjà en œuvre dans les programmes SETI actuels, améliore l'efficacité de la réception. Toutefois, les avantages de ce multiplexage ne peuvent s'appliquer à l'émission sans une perte de puissance de chaque faisceau ou de chaque canal de fréquence, car la puissance totale est fixée.

À travers le milieu

Jusqu'ici, nous avons examiné la conception des deux extrémités de la communication, l'émetteur et le récepteur. Cependant, entre les deux, le rayonnement peut se propager difficilement dans l'immensité interstellaire. Dans le vide, les ondes électromagnétiques se propagent en ligne droite, à moins de rencontrer un obstacle matériel qui les absorbe, les réfléchit ou les réfracte. Or, le milieu interstellaire n'est pas vide : il contient des gaz, des poussières et des champs magnétiques qui perturbent la propagation de ces ondes. Le long des trajets immenses, ces éléments dévient les ondes de leur chemin en ligne droite, changent leur polarisation et font fluctuer l'intensité du signal reçu. Ces phénomènes plaident contre l'utilisation de faisceaux d'émission et de réception trop étroits, et rendent plus criants encore le besoin d'émetteurs puissants.

Lorsque les ondes traversent un nuage de gaz, où la vitesse de propagation est inférieure à celle dans le vide, les ondes sont réfractées. Cette réfraction dévie les ondes et superpose parfois deux ondes qui ont la même source. Lorsque l'onde pénètre dans le nuage, l'une de ses composantes peut être plus ralentie qu'une autre : un déphasage apparaît entre les composantes de l'onde. Selon l'amplitude du déphasage et selon la différence de chemin entre les composantes de l'onde, les portions déphasées peuvent se renforcer, s'annuler ou donner toutes les combinaisons intermédiaires.



2. LES ONDES sont perturbées quand elles se propagent dans la Galaxie. Notamment, les nuages de gaz réfractent, c'est-à-dire dévient la lumière *(en rouge et en orange)*, de sorte qu'elle interfère avec un autre rayon *(en bleu)* issu du même émetteur. Quand le nuage se déplace, la différence de longueur de trajectoire entre le rayon direct et le rayon réfracté change. Ainsi, les rayons reçus passent continûment et continuellement d'un renforcement constructif à une annulation réciproque : le signal reçu (la somme des signaux propagés selon les divers chemins) scintille sur le récepteur, et une partie de sa cohérence est perdue.

Supposons maintenant que la nappe de gaz qui se trouve sur le chemin de la deuxième onde soit en mouvement par rapport à la trajectoire de l'onde, de sorte que le déphasage varie dans le temps *(voir la figure 2)*. Dans ce cas, l'onde qui résulte de la superposition des deux composantes variera dans le temps, tantôt s'annulant, tantôt se renforçant. Des effets du même type résultent des réflexions, des décalages Doppler, etc. Toutes ces perturbations transforment un signal initialement régulier en un signal fortement modulé à son arrivée, sur un lointain récepteur.

L'utilisation des ondes radio pour les contacts interstellaires est décourageante. La taille gigantesque de la Galaxie impose des systèmes rédhibitoires : des puissances d'émission considérables ou des antennes immenses, avec des faisceaux si fins qu'ils risquent de ne servir à rien. Le système nécessaire pour envoyer un signal vers un grand échantillon d'étoiles est certainement hors de portée de notre technique. De surcroît, même si l'on parvenait à établir le contact, plusieurs siècles pourraient s'écouler avant que nous recevions une réponse à notre message. Même si nous réussissions à venir à bout des formidables contraintes physiques, le projet de communication avec des civilisations extraterrestres nécessiterait une ténacité dont l'humanité n'a pas encore fait preuve.

George SWENSON est professeur émérite d'électrotechnique et d'astronomie à l'Université de l'Illinois.

Extraterrestrials, Where Are They? sous la direction de Ben Zuckerman et Michael H. Hart, Cambridge University Press, 1995.

Christian DE DUVE, *Vital Dust : Life as a Cosmic Imperative*, Basic Books, 1995.

James M. CORDES, T. Joseph W. LAZIO et Carl SAGAN, *Scintillation-Induced Intermittency in SETI*, in *Astrophysical Journal*, vol. 487, pp. 782-808, 1er octobre 1997.

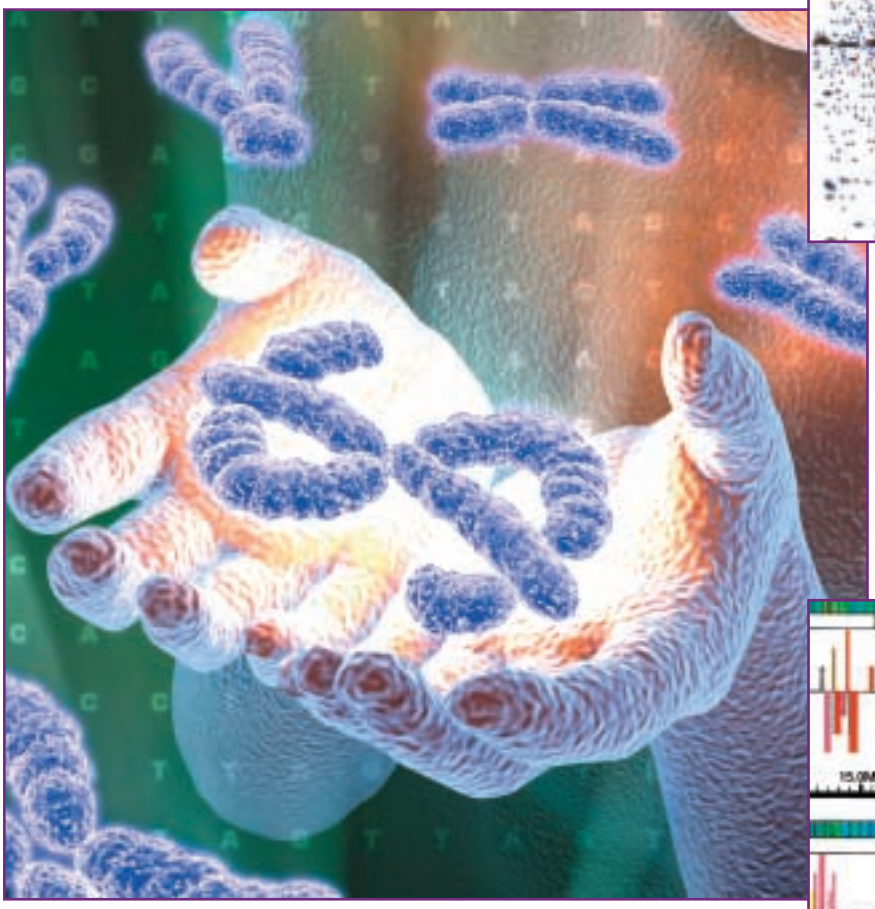
Andrew J.H. Clark et David H. CLARK, *Aliens : Can We Make Contact with Extraterrestrial Intelligence?* Fromm International, 1999.

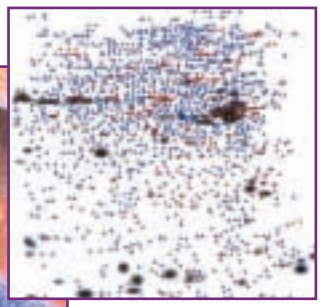
Peter Douglas WARD et Donald BROWNLEE, *Rare Earth : Why Complex Life Is Uncommon in the Universe*, Copernicus Books, 2000.

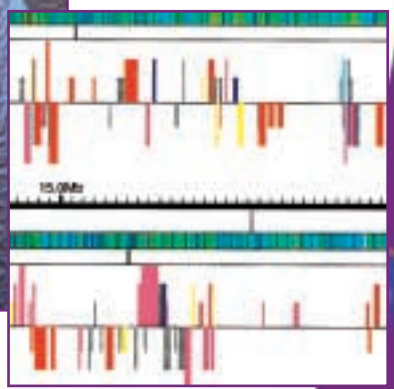
Introduction

Il y a seulement dix ans, les mots «génome» ou «génomique» n'étaient connus que des généticiens. Aujourd'hui, ils sont dans tous les journaux, évoqués par toutes les télévisions ou les radios, connus de l'ensemble du public. Entre-temps, plusieurs milliards de francs ont été alloués au *Projet génome humain* : des laboratoires de plusieurs pays ont déchiffré ensemble les trois milliards de lettres qui composent le «mode d'emploi» d'un être humain. Aujourd'hui, ce projet est quasi achevé, et la concurrence fait rage entre des organismes publics et des sociétés privées. De nombreux groupes pharmaceutiques cherchent à exploiter les résultats du travail effectué.

Avec les articles de ce dossier, *Pour la Science* fait le point sur la génomique et présente deux disciplines – la bio-informatique et la protéomique – qui sont les filles naturelles du décryptage du génome humain.







Fond : Laurie Grace. À gauche : Jeff Johnson. En haut à droite : Julio E. Celis, Danish Center for Human Genome Research, Aarhus. En bas à droite : Science, vol. 287, n° 5461, 24 mars 2000

La lecture du génome humain

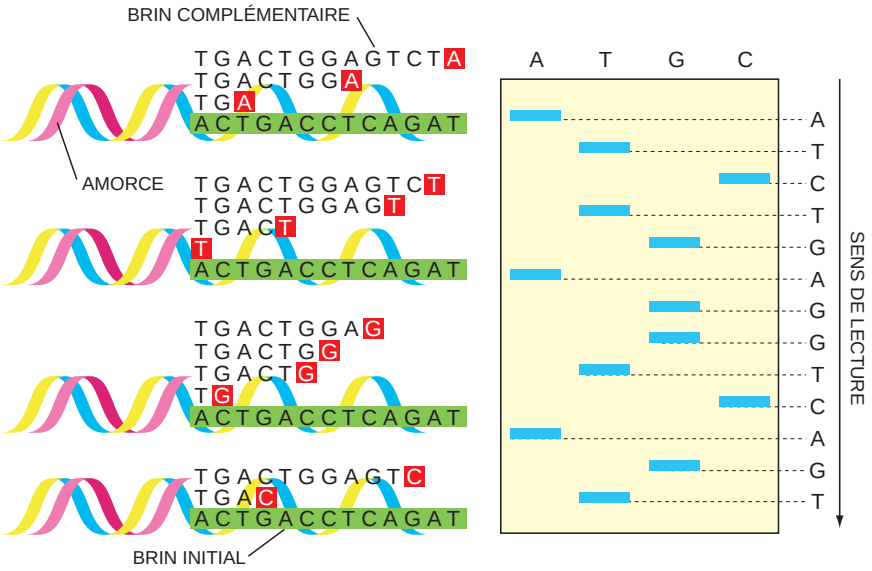
KATHRYN BROWN

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.

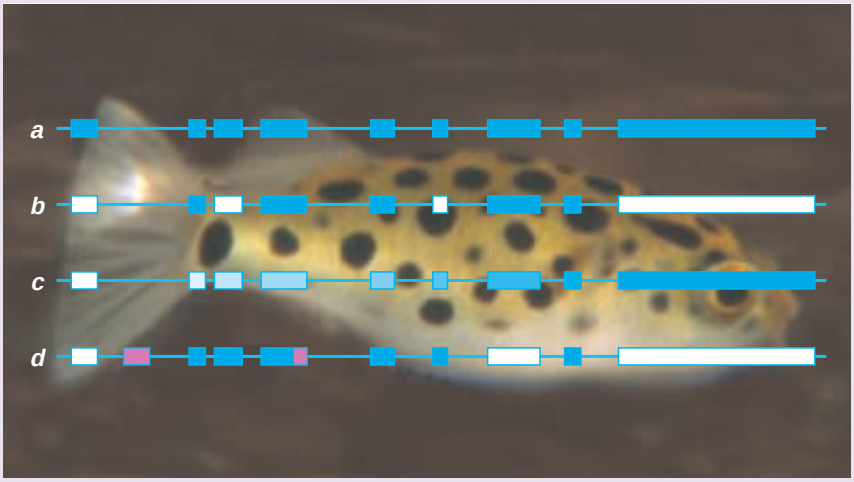
Certains auteurs soulignent l'intérêt des comparaisons avec l'ADN d'autres génomes pour identifier les gènes et pour essayer de connaître les fonctions de leurs protéines. C'est précisément afin d'identifier des gènes humains au moyen de telles comparaisons, que nous poursuivons, au *Genoscope*, le séquençage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

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


A. Syred SPL/Photo Researchers, Inc.

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.

Le ver


S. Stammers SPL/Photo Researchers, Inc.

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.

La drosophile


A. Pasjeka, Peter Arnold, Inc

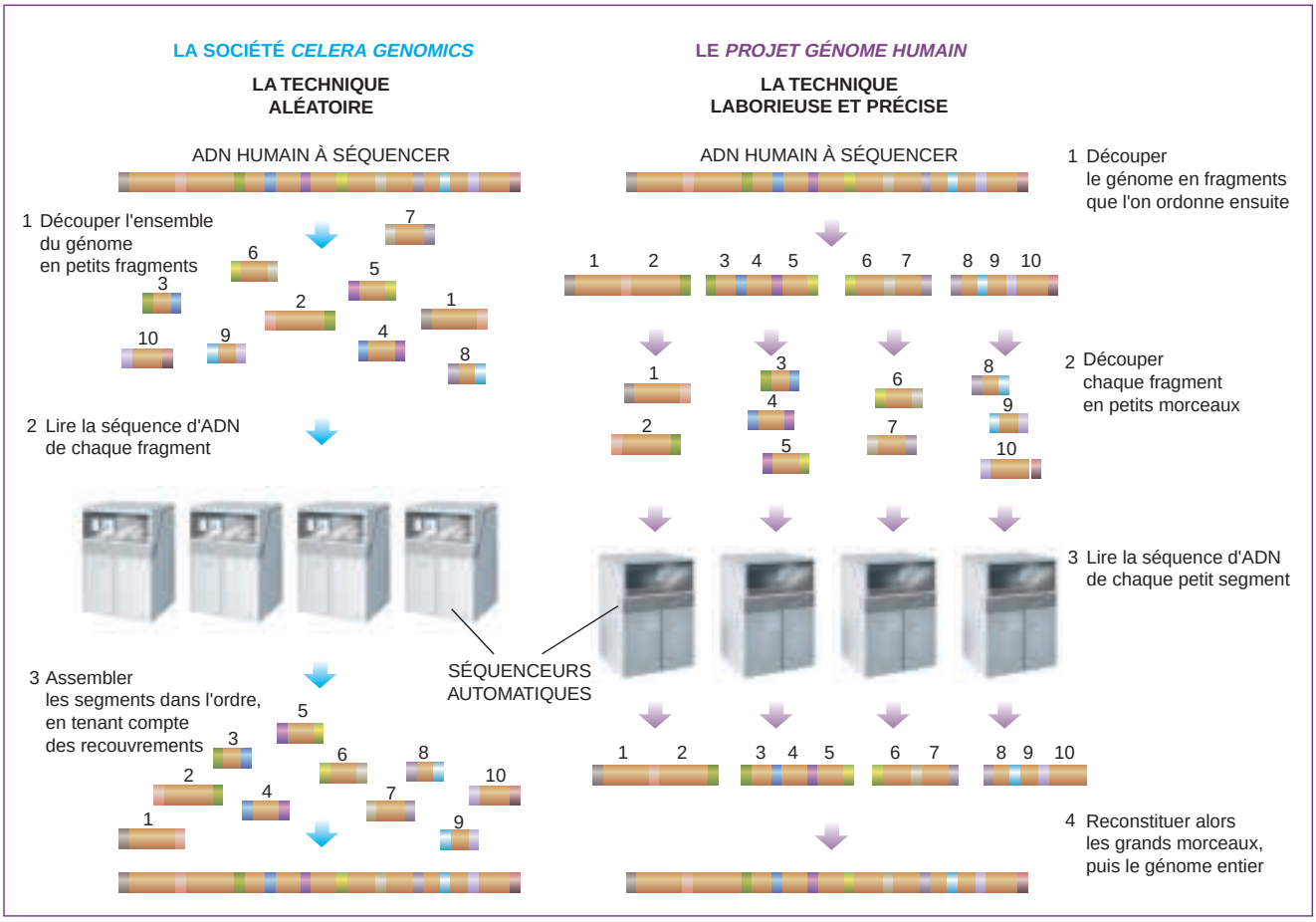
Le séquençage du génome de la drosophile s'est terminé en mars 2000. Il a été effectué par des organismes publics et par la Société *Celera Genomics*. Environ 60 pour cent des 289 gènes connus de maladies humaines ont un gène homologue chez cette mouche, et 50 pour cent des protéines de drosophiles ressemblent 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 les 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.

La souris


The Jackson Laboratory

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*



Laurie Grace, assisté de Eric Lander, Whitehead Institute

3. LES STRATÉGIES DE SÉQUENÇAGE.

autre version du même médicament. D'autres sociétés mettent au point des tests sanguins qui détecteront des gènes pathogènes et prévoiront, par exemple, l'apparition de la maladie de Huntington. Enfin, ultime Graal, la thérapie génique, qui vise à réintroduire directement des gènes fonctionnels dans l'organisme du malade, suscite de nombreux espoirs.

Après s'être uniquement intéressés aux 99,9 pour cent de nos gènes que nous partageons avec nos voisins, les généticiens et les sociétés pharmaceutiques ont compris l'intérêt des 0,1 pour cent de nos gènes qui varient : la modification, ou polymorphisme, d'un seul nucléotide, tel un *T* à la place d'un *C* dans la séquence d'un gène, peut entraîner un dysfonctionnement.

Ces variations génétiques ténues expliquent les variations d'efficacité d'un grand nombre de médicaments : certains ne sont efficaces que sur 30 à 50 pour cent de la population ; dans des cas extrêmes, un produit qui sauve une personne peut en intoxiquer une autre. Par exemple, la Rezuline, administrée contre certains diabètes, a été associée à plus de 60 pour cent de décès par empoisonnement hépatique. Un test génétique individualisé serait utile en déterminant si un tel médicament peut vous soigner efficacement ou risque de vous tuer.

Des géants de la pharmacie s'associent à de plus petites sociétés spécialistes de génomique : aujourd'hui, la médecine personnalisée n'en est qu'au stade de l'éprouvette, mais certains experts lui prédisent un avenir florissant.

Des brevets...

La connaissance du génome modifiera, sans doute aussi, la conduite des essais thérapeutiques et inaugurera une nouvelle médecine. Pourtant, les difficultés restent de taille. Certaines sont d'ordre technique : on devra chercher la fonction réelle des gènes qui seront trouvés dans le génome. Des problèmes sociaux et politiques se posent aussi : le diagnostic d'une maladie incurable qui se manifestera dans 20 ans est-il souhaitable? Enfin, l'un des points délicats est celui des brevets : comment peut-on breveter un gène?

Personne ne s'alarme lorsqu'une marque de voitures dépose un brevet pour un nouveau modèle ou quand une société de micro-électronique protège un nouveau système. En revanche, beaucoup s'indignent quand les groupes industriels revendiquent des droits sur l'ADN humain. Le problème est complexe, car de tels brevets sont parfois, pour les sociétés, le seul moyen de mettre au point des tests ; ainsi, sans brevets, la Société *Myriad Genetics* n'aurait pu mettre au point les tests de dépistage de mutations dans les gènes *BRCA1* et *BRCA2*, associées au cancer du sein et de l'ovaire.

De nombreux biologistes conviennent que les brevets semblent nécessaires, mais certains se plaignent que les sociétés privées profitent des découvertes qui ont demandé un travail de séquençage contraignant avec les fonds publics. Les données accessibles sur le site *GenBank* fournissent au monde entier un aperçu sans

précédent sur ce qui constitue un être humain : en avril 2000, 35 000 personnes ont consulté sur *Internet*, chaque jour, le site *GenBank*. Certaines de ces personnes travaillent dans des sociétés qui établissent des catalogues de gènes, afin d'en breveter les utilisations potentielles. Par exemple, la Société *Incyte* a obtenu plus de 500 brevets de gènes complets, et elle en a demandé 7 000 autres. Pourtant, la fonction des gènes brevetés parfois n'est pas connue. Simultanément, la Société *Incyte* – et d'autres – vendent par *Internet* la séquence de plus de 100 000 gènes ou des exemplaires réels de ces gènes.

De nombreux biologistes craignent que de telles activités ne ralentissent les recherches : si l'accès aux informations est protégé ou si l'utilisation des gènes est limitée par des brevets, l'étude de nombreuses maladies sera entravée.

...et des tests

Alors que le débat fait rage, les États ajoutent à la confusion. Ainsi, en mars 2000, au cours d'une déclaration à la presse ambiguë, Bill Clinton et Tony Blair ont revendiqué un libre accès aux résultats bruts des études de séquençage : les sociétés privées de génomique qui avaient précieusement protégé leurs séquences semblaient visées. Suite à la conférence de presse et, surtout, à la chute des cours des actions en Bourse des sociétés concernées, les dirigeants de certaines d'entre elles ont annoncé la mise à disposition gratuite de leurs résultats bruts sur le génome. Cependant, dans les semaines suivantes, des responsables du gouvernement américain ont précisé que les brevets pour de nouveaux produits de santé d'origine génétique restaient autorisés.

En France, les lois de bioéthique promulguées en 1994 interdisent la commercialisation de tout produit d'origine humaine. Toutefois, au niveau européen, l'interdiction est plus floue : un élément isolé du corps humain, notamment la séquence d'un gène, peut être breveté s'il a été isolé ou produit. Le gouvernement français hésite à incorporer cette directive dans le droit français et le ministre de la Recherche reste opposé aux brevets des séquences d'ADN.

Au cœur du débat figure l'inventivité et la démonstration de l'utilité des séquences brevetées. Aujourd'hui, les brevets découlent principalement des analyses informatiques des gènes : à partir de séquences complètes ou partielles de gènes, on détermine la séquence en acides aminés de la protéine codée par le gène. En comparant cette protéine hypothétique à des protéines connues, on cherche la fonction du gène et son utilité pour la mise au point d'un médicament ou d'un test de diagnostic. Ce résultat est la base de la demande de brevet.

Aujourd'hui, les brevets concernent plus de 740 tests génétiques qui sont déjà commercialisés ou sont en cours de mise au point. Cependant, les lacunes de la génétique sont considérables : par exemple, plusieurs années après la mise au point des tests des gènes *BRCA1* et *BRCA2*, les généticiens ignorent la contribution exacte des mutations de ces gènes au risque de cancer chez les femmes. Ce n'est qu'en août 2000 que des biologistes ont découvert que la protéine codée par le gène *BRCA1* participe, avec d'autres, au repliement de l'ADN dans le noyau des cellules et commande ainsi l'expression d'autres gènes. Pour la maladie de Huntington, le test identifie exactement la modification du gène, mais il n'indique ni l'âge auquel apparaîtront les symptômes ni la gravité et la progression de la maladie.

Les conséquences sociales de ces tests sont également importantes. La présence d'une maladie dans une famille peut inciter une personne à subir un test génétique de dépistage, et les employeurs ou les assureurs pourraient utiliser les résultats à des fins discriminatoires. Une législation pour protéger les individus et bannir la discrimination génétique est nécessaire. Ainsi, le séquençage est porteur d'espoir, mais aussi d'abus : la possession des résultats compte moins que leur utilisation.

Kathryn BROWN est journaliste scientifique.

E. PENNISI, *Are Sequencers Ready to Annotate the Human Genome?*, in *Science*, vol. 287, n° 5461, page 2183, 24 mars 2000.

E. GREEN, *The Human Genome Project and Its Impact on the Study of Human Disease*, in *Metabolic and Molecular Bases of Inherited Disease*, sous la direction de Charles R. Scriver, huitième édition, McGraw-Hill, 2000.

La bio-informatique

KEN HOWARD

Une nouvelle discipline issue de la biologie et de l'informatique transforme des données brutes du séquençage du génome en informations utiles aux biologistes.

En 1967, dans le film *Le lauréat*, un ami du héros lui murmure : «le plastique» ; en lui désignant le grand secteur industriel des années 1970, il l'invite à y faire carrière. Si le film était tourné aujourd'hui, le mot «plastique» serait remplacé par «bio-informatique».

Le séquençage du génome humain, c'est-à-dire la lecture des nucléotides *A, C, T* et *G* qui constituent le code génétique humain, a fourni trois milliards de bits d'information : pour les stocker, on devrait utiliser plus de 2 000 disquettes informatiques. Pourtant, aussi impressionnante soit-elle, cette information n'est rien par rapport à ce que l'on vise : on ignore encore où et quand les gènes sont activés; quelles sont les petites différences génétiques entre individus, nommées polymorphismes mononucléotidiques ; quelles sont les structures des protéines codées par les gènes ; comment les protéines interagissent *(voir l'article de Carol Ezzell)* et interviennent dans les maladies... De surcroît, le génome humain ne sera exploré que si on le compare aux génomes d'organismes dits modèles, telles la drosophile et la souris. Au total, les biologistes font face à un «raz de marée» de données. Si la bio-informatique parvient à utiliser ces données, elle changera le visage de la médecine.

Plusieurs sociétés vendent des produits et des services de bio-informatique : elles proposent, par exemple, de collecter et de stocker les données, d'effectuer des recherches dans des banques de données et d'interpréter les données. Leurs clients sont des laboratoires pharmaceutiques et des sociétés de génie génétique, qui déboursent parfois des sommes importantes pour de tels services. Selon des experts financiers, le chiffre d'affaires de la bio-informatique devrait atteindre 15 milliards de francs.

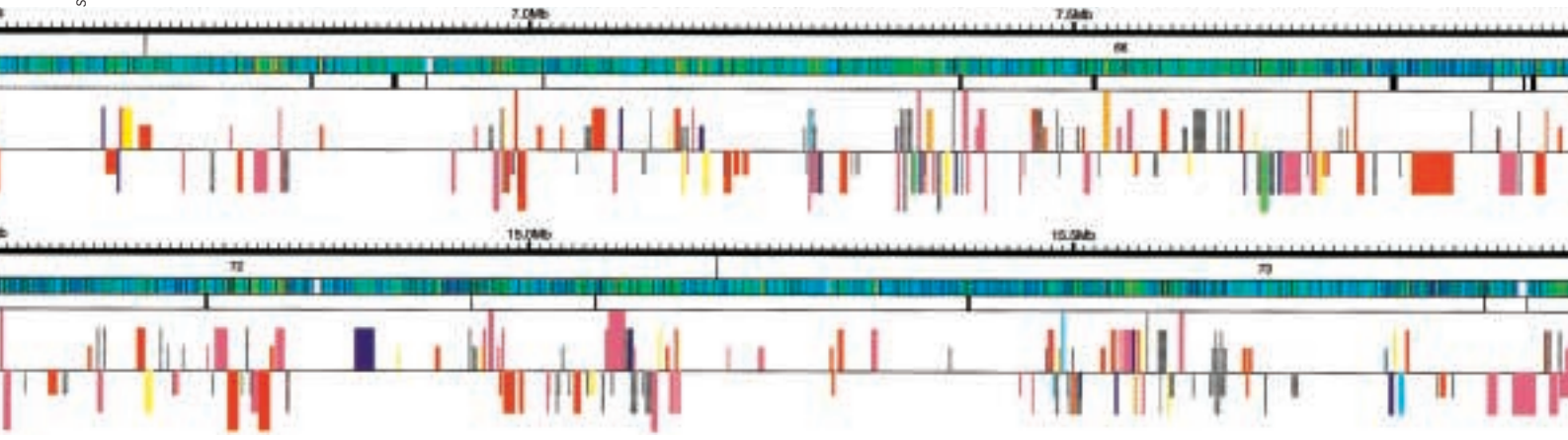
Pourquoi? Parce que la bio-informatique aide à trouver plus efficacement les principes actifs de médicaments que les méthodes classiques : en identifiant les molécules anormales qui causent les maladies, elle facilite le tri des molécules testées. Elle réduit ainsi le temps de recherche et de mise au point des médicaments et augmente donc le temps pendant lequel un médicament est en vente avant l'expiration de son brevet : un an de gagné dans la commercialisation d'un médicament, c'est parfois plusieurs milliards de francs de gains supplémentaires.

Toutefois, avant d'éventuelles retombées financières, les entreprises de bio-informatique doivent apprendre à assimiler la pléthore de données génomiques et affiner leur techniques d'analyse. Dans un souci de prospective, elles doivent aussi se donner des objectifs plus ambitieux, telle la recherche de relations entre les diverses informations et discerner un schéma global.

Du bricolage à l'automatisation

La discipline est née au début des années 1980, quand le Laboratoire européen de biologie moléculaire, puis le Département américain de l'énergie ont créé les banques de données *EMBL* et *GenBank* pour enregistrer les courtes séquences d'ADN que les biologistes commençaient à obtenir. Initialement, la structure était réduite à une seule

Science, vol. 287, n° 5461, 24 mars 2000



1. CES CARTES DE CHROMOSOMES de la drosophile illustrent une utilisation de la bio-informatique : la comparaison des gènes de divers organismes. Les barres montrent les régions où les gènes de la mouche sont semblables aux gènes d'autres espèces.

pièce : sur des claviers ne comportant que les lettres *A*, *C*, *T* et *G*, quelques techniciens saisissaient fastidieusement les séquences d'ADN publiées dans les revues de génétique. Puis les informaticiens mirent au point des protocoles afin que les généticiens contactent EMBL ou *GenBank* et y placent leurs données. Après l'avènement du réseau *Internet*, les données de séquences sont devenues accessibles, gratuitement, de n'importe quel endroit du monde.

Parallèlement, en France, Jean Dausset créait le Centre d'études du polymorphisme humain. À partir de données génétiques recueillies sur les membres de familles à la généalogie connue, les généticiens de ce laboratoire ont cartographié le génome humain, c'est-à-dire identifié et localisé des fragments d'ADN communs à tous les êtres humains. La vocation du CEPH était de mettre ces informations à la disposition de la communauté scientifique.

Après le lancement du *Projet génome humain*, en 1990, le volume des données stockées dans les banques de données a crû exponentiellement. Quelques années plus tard, les techniques de séquençage étaient améliorées grâce à des robots et à des ordinateurs, et les débits de stockage ont encore augmenté. Aujourd'hui, les banques de séquences contiennent les données qui décrivent plus de sept milliards de sous-unités de l'ADN.

Quand le projet public s'est développé, des sociétés privées ont utilisé ses résultats pour effectuer leurs propres séquençages et établir leurs propres bases de données. Certaines d'entre elles identifient plus de 20 millions de nucléotides d'ADN en une seule journée, d'autres détiennent jusqu'à 50 000 milliards de bits de données enregistrées, soit l'équivalent de 80 000 disques compacts.

À titre de comparaison

Que faire d'une telle quantité d'information? La recherche de similitudes, ou d'homologies, entre un fragment d'ADN nouvellement séquencé et des fragments d'ADN séquencés auparavant et provenant d'organismes variés, est l'une des opérations élémentaires de la bio-informatique. Quand ils détectent ainsi une quasi-correspondance, les biologistes prédisent le type de protéine que la nouvelle séquence code. Ainsi, les cibles thérapeutiques pertinentes sont détectées rapidement, et des principes actifs sont plus facilement identifiés.

Dans les années 1990, des informaticiens ont ainsi mis au point un groupe de programmes qui comparent les séquences d'ADN : ce logiciel, nommé BLAST, fait partie d'une série d'outils de recherche de séquences d'ADN ou de protéines qui sont accessibles auprès de nombreux fournisseurs d'accès à des bases de données, tel le Centre de ressources *Infobiogen*, qui coordonne, en France, la recherche, le développement et l'exploitation de l'informatique appliquée à la génomique.

Infobiogen offre aussi l'accès à des banques de données qui contiennent les structures tridimensionnelles des protéines, les génomes complets d'organismes modèles, et des

2. LES DONNÉES GÉNÉTIQUES sont la matière première de la bio-informatique. Celle-ci recherche des informations dans des banques de données gigantesques où sont stockées les séquences génétiques des organismes. L'ADN est décrit par les lettres *A*, *C*, *T* et *G*, qui représentent les quatre constituants de l'ADN. Chercher une séquence particulière dans le génome, c'est comme chercher une aiguille dans une botte de foin… ou les initiales PLS de *Pour la Science* dans le tableau ci-contre.