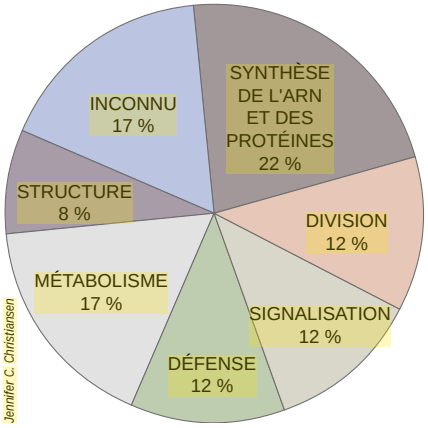
ger et non des molécules entières, longues parfois de plusieurs milliers de bases. Différentes parties d'un même gène donnent des ADN complémentaires dont l'origine commune n'est pas toujours évidente, mais un ADN complémentaire, même s'il n'est composé que de quelques milliers de bases, conserve des traces de son origine : la probabilité que deux gènes différents aient en commun la même séquence de plusieurs milliers de bases est très faible ; une molécule d'ADN complémentaire identifie spécifiquement un gène, tout comme un chapitre d'un livre ne laisse pas de doutes sur sa provenance.

Une fois que nous avons obtenu une molécule d'ADN complémentaire, nous savons en fabriquer des copies à volonté, de sorte que nous pouvons déterminer l'ordre des bases. Cet ordre des bases révèle la séquence des acides aminés dans la partie de la protéine qui serait codée par l'ARN messager. Quand nous comparons ensuite cette séquence à celles de protéines déjà connues, nous découvrons parfois la fonction de la protéine complète, car des protéines dont les fonctions sont voisines ont souvent des séquences d'acides aminés qui se ressemblent.

L'analyse des séquences des ADN complémentaires était très longue, mais elle a récemment été automatisée. À l'aide d'ordinateurs puissants et de programmes spécialisés, nous traiterons les considérables quantités de données produites.

L'assemblage du puzzle

Pour identifier les gènes activés dans une cellule, nous analysons une séquence de 300 à 500 bases à l'une des extrémités de chaque molécule d'ADN complémentaire ; ces séquences servent de marqueurs des gènes. Nous avons choisi ce nombre de bases pour que la séquence soit analysée suffisamment rapidement et que le gène soit identifié sans ambiguïté. Si l'on compare une molécule d'ADN complémentaire à un chapitre d'un livre, une séquence de 300 à 500 bases est comparable à la première page du chapitre : elle permet d'identifier le livre et donne même une idée de son contenu. De même, une séquence d'ADN complémentaire nous renseigne sur le gène dont elle provient. Nous élucidons ainsi, chaque jour, une longueur totale d'ADN complémentaire qui atteint un million de bases.



1. LA PROPORTION DES GÈNES participant aux principales activités cellulaires a été déduite de l'étude de 150 000 séquences partielles. D'après les ressemblances entre les gènes humains et d'autres gènes dont on connaissait le rôle, on a attribué une fonction probable aux séquences nouvelles.

Les séquences produites sont pour la plupart issues de l'un des 6 000 gènes déjà identifiés par diverses méthodes. Quand nous ne pouvons pas associer une nouvelle séquence d'ADN complémentaire à un gène connu, nous consultons diverses bases de données afin de chercher si cette nouvelle séquence s'identifie à une séquence déjà répertoriée. Quand deux séquences ont un long segment commun, nous construisons des segments plus longs, par réunion des deux. Nous supposons alors que ces séquences incomplètes existent quelque part dans le gène d'origine. Ce procédé s'apparente à celui qui consiste à rassembler des morceaux de phrase tels que «La cigale ayant chanté tout l'été», «tout l'été se trouva fort dépourvue» et «fort dépourvue quand la bise fut venue», et à les combiner pour retrouver la fable de La Fontaine.

Simultanément, nous tentons de prévoir la fonction probable des protéines en fonction de leur séquence. Ayant calculé la structure possible, nous la comparons avec celle d'autres protéines connues. Nous trouvons parfois une ressemblance avec une protéine humaine, mais, souvent, c'est avec une protéine de bactérie, de champignon, de plante ou d'insecte : beaucoup d'organismes produisent des protéines proches des protéines humaines. Nos ordinateurs réactualisent constamment ces classifications provisoires.

Notre banque de données contient aujourd'hui plus d'un million de petites séquences de gènes, assemblées en 170 000 séquences plus longues. Ces séquences couvrent sans doute la totalité des gènes humains qui s'expriment : lorsque des biologistes trouvent un nouveau gène, nous avons déjà, dans plus de 95 pour cent des cas, une séquence correspondant à ce gène dans notre banque. L'assemblage de plusieurs petites séquences donne souvent la séquence totale des gènes découverts. Enfin, plus de la moitié des gènes que nous avons identifiés ressemblent à des gènes dont la séquence est établie et dont on croit connaître la fonction. Cette proportion s'accroît régulièrement.

Quand un même gène produit un nombre anormalement élevé de séquences d'ADN complémentaires, c'est qu'il synthétise beaucoup d'ARN messager et une quantité élevée de la protéine correspondante ; on en déduit que cette protéine joue un rôle important. De surcroît, les gènes qui s'expriment uniquement dans quelques tissus très limités risquent d'être impliqués dans les maladies de ces tissus. Ainsi, 300 gènes sont susceptibles d'avoir un intérêt médical parmi les milliers de gènes que nous avons découverts.

Nouveaux gènes, nouveaux médicaments

En utilisant la méthode des séquences d'ADN complémentaires pour découvrir de nouveaux gènes, nous avons dénombré les gènes qui participent aux principales fonctions cellulaires, par exemple aux fonctions de défense ou au métabolisme. L'abondance des informations fournies par les séquences d'ADN complémentaires ouvre de nouvelles voies de recherche en pharmacologie. Avec le Laboratoire *SmithKline Beecham* et d'autres laboratoires publics ou privés, nous explorons diverses possibilités.

Ainsi, on détecte le cancer de la prostate en mesurant les concentrations sanguines en une protéine nommée l'antigène spécifique de la prostate (chez les personnes malades, cette concentration est anormalement élevée). Malheureusement ce dosage est ambigu, car des tumeurs assez bénignes et qui progressent très lentement entraînent parfois une augmentation de la concentration en antigène aussi forte que des tumeurs malignes qui nécessitent un traitement d'urgence.

Nous avons analysé des ARN messagers provenant de plusieurs échantillons de tissus prostatiques sains, ainsi que de tumeurs prostatiques bénignes ou malignes. Nous avons trouvé envi-

ron 300 gènes uniquement exprimés dans la prostate ; parmi ces gènes, 100 environ ne sont actifs que dans les tumeurs prostatiques et 20 dans les tumeurs qualifiées de malignes par les spécialistes. Avec ces 20 gènes et les protéines qu'ils produisent, nous cherchons à élaborer des tests qui permettront de dépister spécifiquement les tumeurs malignes de la prostate. Des recherches identiques sont en cours pour les cancers du sein, du poumon, du foie et du cerveau.

Les séquences d'ADN complémentaires servent aussi à la découverte de petites molécules potentiellement thérapeutiques. Les méthodes que les chimistes utilisent pour découvrir et tester de petites molécules ont notablement progressé au cours des dernières années. Des robots testent rapidement les effets de composés naturels et synthétiques sur des protéines humaines associées à diverses maladies. Jusqu'à présent, on manquait de protéines cibles, mais notre méthode en produit beaucoup, qui sont systématiquement tes-

Quatre décennies de progrès

Les avancées actuelles de la génétique reposent sur quatre décennies de progrès continus dans la connaissance de la mécanique génétique qui règle le fonctionnement des cellules.

Au début des années 1960, les pionniers de la biologie moléculaire, les Britanniques Sydney Brenner et Francis Crick, l'Américain James Watson et les Français François Gros, François Jacob et Jacques Monod conçoivent et mettent en évidence, d'abord chez les bactéries, l'existence d'un intermédiaire dans le fonctionnement des gènes, l'ARN messager. À la fin de la décennie, les Américains Howard Temin et David Baltimore montrent que la transcriptase inverse des rétrovirus est une enzyme capable de recopier l'ARN messager des rétrovirus en ADN complémentaire. C'est en 1975 que les Français François Rougeon et Philippe Kourilsky, et le Suisse Bernard Mach, réalisent le premier clonage dans une bactérie d'un ADN complémentaire copie de l'ARN messager contrôlant la synthèse d'une molécule d'hémoglobine. À partir de 1977, les méthodes de séquençage de l'ADN décrites par l'Américain Walter Gilbert et le Britannique Frederik Sanger donnent accès à la structure des gènes et permettent d'en déduire la conformation primaire des protéines qu'ils codent. La capacité de programmer la synthèse de protéines humaines dans des micro-organismes ouvre la voie à l'essor du génie génétique et à la production de nouveaux médicaments.

Au milieu des années 1980, un grand programme international a été conçu, visant à l'analyse complète du génome humain. L'objectif de la première phase du Projet génome humain, démarrée en 1990, a été d'établir des cartes qui permettent de localiser rapidement les régions chromosomiques contenant des gènes impliqués dans telle ou telle fonction biologique ou dans des pathologies. Le Centre d'études du polymorphisme humain (CEPH) a été fondé en 1983 par Jean Dausset. Les familles sélectionnées qui ont accepté de participer à la collecte d'ADN ont joué un rôle central dans ce travail de cartographie. Les équipes françaises de Daniel Cohen et de Jean Weissenbach (au CEPH et au Généthon), soutenues par l'Association française de lutte contre les myopathies, l'AFM, et par les donateurs du Téléthon ont apporté une contribution notable.

La seconde phase du Projet génome humain visant à déterminer la séquence complète des trois milliards de nucléotides dans les 23 paires de chromosomes est en cours d'organisation. Aujourd'hui, moins de deux pour cent de l'ensemble a été déchiffré, et il faudra encore de longues années pour achever le déchiffrage.

Afin d'accéder plus rapidement à la connaissance de la structure fine des gènes, de leur position précise sur les chromosomes, mais aussi de leur lieu de fonctionnement dans l'organisme, plusieurs équipes à travers le monde ont entrepris à partir de 1990 le séquençage systématique de collections de clones d'ADN complémentaires. Cette initiative a été le fait de chercheurs du milieu académique, tels Sydney Brenner et moi-même en Europe, Kenichi Matsubara au Japon, James Sikela et Craig Venter aux États-Unis. Ce dernier a poursuivi son travail au sein d'un institut privé, *TIGR*, financé par la Société *HGS* de William Haseltine et le Laboratoire pharmaceutique *SmithKline Beecham*. Une autre société américaine, *Incyte Pharmaceuticals*, a également développé une importante base de données de séquences d'ADN complémentaires qu'elle met à la disposition de ses clients de l'industrie pharmaceutique.

Poursuivant l'effort que j'ai entrepris en France en 1990 avec le programme Genexpress (il s'agit d'une collaboration entre le CNRS et l'AFM réalisée avec le concours du Laboratoire Généthon et le soutien de l'Union européenne), j'ai fondé en 1993 avec mes collègues Bento Soares, de l'Université Columbia à New York, Greg Lennon, à Livermore, et Mihael Polymeropoulos, à Bethesda, le Consortium international IMAGE. Nous cherchons à collecter le maximum de données biologiques et génétiques pertinentes en favorisant l'étude des mêmes collections de clones d'ADN complémentaires de référence, et la diffusion des résultats dans les bases de données publiques internationales. Ainsi, nos efforts conjugués avec ceux de nombreux laboratoires européens, japonais et américains, notamment le centre génome de l'Université de Washington à Saint Louis, financé par la Société *Merck*, permettent aux chercheurs du monde entier d'avoir accès à des connaissances – au moins partielles – sur la majorité des transcrits des gènes humains. Une collaboration équilibrée entre les pouvoirs publics, les fondations, les associations qui soutiennent la recherche en génétique et les industriels est indispensable pour que l'on puisse déterminer dans un proche avenir la séquence complète de tous les transcrits du génome humain, comprendre la fonction des protéines qu'ils codent, et utiliser ces nouvelles connaissances pour le diagnostic, la prévention et le traitement des maladies.

Charles AUFFRAY, CNRS-UPR 420, Villejuif