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.

références aux revues scientifiques relatives à chaque information des banques de données.

L'utilité de la bio-informatique s'est illustrée dès 1993, quand des biologistes ont isolé des «ostéoclastes» de personnes souffrant de tumeurs osseuses. Les ostéoclastes sont des cellules qui participent à l'équilibre des os ; tandis que des cellules nommées ostéoblastes font croître les os, les ostéoclastes les dégradent. On les pense hyperactives chez les personnes qui souffrent d'ostéoporose. Les généticiens isolèrent des ADN complémentaires d'ostéoclastes de tumeurs, déterminèrent la séquence de ces ADN et la comparèrent à celle

Ne dites plus *in vitro*. Dites plutôt *in silico*.

Après le séquençage du génome humain, les spécialistes de prospective imaginent la modélisation bio-informatique de toutes les réactions biochimiques qui s'enchaînent jusqu'à l'apparition de la vie humaine. Stuart Kauffman, directeur du Groupe Bios de l'Institut de Santa Fe, prévoit des applications inédites de la bio-informatique.

Pour la Science : Quel est l'avenir de la bio-informatique et de la biologie *in silico*?

S. Kauffman : On peut imaginer que les gènes d'une cellule humaine fonctionnent à la manière d'un ordinateur chimique, parallèle, où des gènes sont en permanence activés ou inactivés, et ce, à l'intérieur d'un réseau complexe d'interactions. Des voies de signalisation intracellulaires (voir *La communication dans les cellules*, par John Scott et Tony Pawson, *Pour la Science*, août 2000) sont reliées à des voies de contrôle génétique selon des modalités que nous entrevoyons à peine. Le plus énorme projet bio-informatique consiste à identifier les mécanismes de ce réseau, qui commande le développement cellulaire, de la fécondation de l'œuf jusqu'à l'âge adulte.

PLS : Quelles en seront les retombées?

S. K. : Nous saurons sur quels gènes agir, ou sur quelles séquences de gènes et dans quel ordre, afin de redonner un comportement normal aux cellules cancéreuses ou, au contraire, de provoquer leur apoptose (la mort cellulaire programmée). On pourra aussi activer la régénération des tissus. Par exemple, on apprendra à régénérer la moitié du pancréas qu'un malade aura perdue ou des cellules bêta, productrices d'insuline, chez des diabétiques.

PLS : Comment atteindre un tel objectif?

S. K. : La bio-informatique ne suffira pas. Elle devra être épaulée par les mathématiques, qui mettront au point les outils nécessaires pour trouver des circuits de rechange dans des réseaux cellulaires défectueux. On devra également confronter les nouveaux outils à des études expérimentales pour mettre en évidence ce que sont réellement les réseaux biochimiques dans les cellules. La bio-informatique doit s'ouvrir et inclure l'idée d'expérience. Chacun des secteurs de la bio-informatique fournira des hypothèses qui seront à vérifier.

PLS : Quels sont les principales difficultés?

S. K. : Imaginons que je choisisse dix gènes qui interagissent par les protéines qu'ils codent, et que j'essaie de construire le circuit de leurs interactions. Ces dix gènes sont probablement reliés à d'autres gènes, situés hors du circuit considéré : je n'aurais ainsi qu'un petit morceau de circuit prélevé à un circuit bien plus grand, qui inclut des milliers de gènes. Comment représenter le comportement d'un circuit sans connaître les gènes externes auxquels il est relié?

Par exemple, nous connaissons depuis longtemps chaque neurone des ganglions gastriques du homard (un faisceau de nerfs reliés au système digestif de l'animal), ainsi que toutes les connexions synaptiques et les neuromédiateurs de ces neurones. Chaque ganglion n'a que 13 à 20 neurones, mais on ne parvient pas à représenter son comportement.

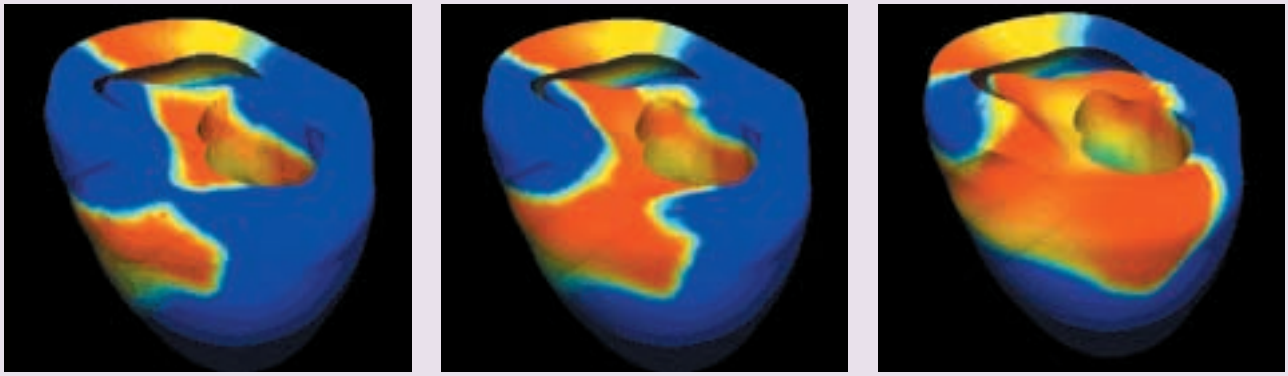
Certes, la compréhension d'un système à 13 variables est difficile, mais le génome humain est un système à 100 000 variables ! Si chaque variable a deux états possibles, combien y a-t-il de situations possibles? $2^{100\,000}$, soit approximativement $10^{30\,000}$. Ainsi, même en ne traitant les gènes que comme des systèmes à deux états (activé ou désactivé, ce qui est inexact puisqu'ils ont des degrés d'activité intermédiaires), on a $10^{30\,000}$ situations possibles. Rappelons que le nombre de particules dans l'Univers connu est égal à 10^{80}...

PLS : Où en sommes-nous, face à ce problème?

S. K. : Nous commençons seulement son étude, mais le jour viendra où un cancer ne sera plus seulement diagnostiqué d'après la morphologie de la cellule cancéreuse, mais aussi à partir des modèles détaillés de l'activité des gènes et des protéines dans cette cellule.

PLS : À quelle échéance : un an, ou 200 ans?

S. K. : Dans 10 à 12 ans, nos outils seront au point. Nous progresserons alors dans la compréhension des circuits commandés par des pans entiers du génome. Nous devrions avoir compris la majeure partie du génome dans 30 à 40 ans.



Raimond L. Winslow et David Scollan (Johns Hopkins University) et Physionx Sciences, Princeton, N.J.

CE MODÈLE INFORMATIQUE D'UN CŒUR en fibrillation (les contractions sont désordonnées et rapides) montre l'activité électrique déréglée qui parcourt l'organe. Le modèle a été établi à partir des modifications de l'expression de quatre gènes dont le fonctionnement est perturbé lorsque se produit une lésion cardiaque.

des banques de données. Ces comparaisons spécifiques de l'échantillon ont permis de découvrir un gène surexprimé dans les ostéoclastes : celui d'une classe de molécules déjà identifiées et nommées cathepsines. Sans la bio-informatique, qui a ainsi fourni une cible thérapeutique prometteuse en quelques semaines, des années d'expériences de laboratoire classiques et un peu de chance auraient été nécessaires.

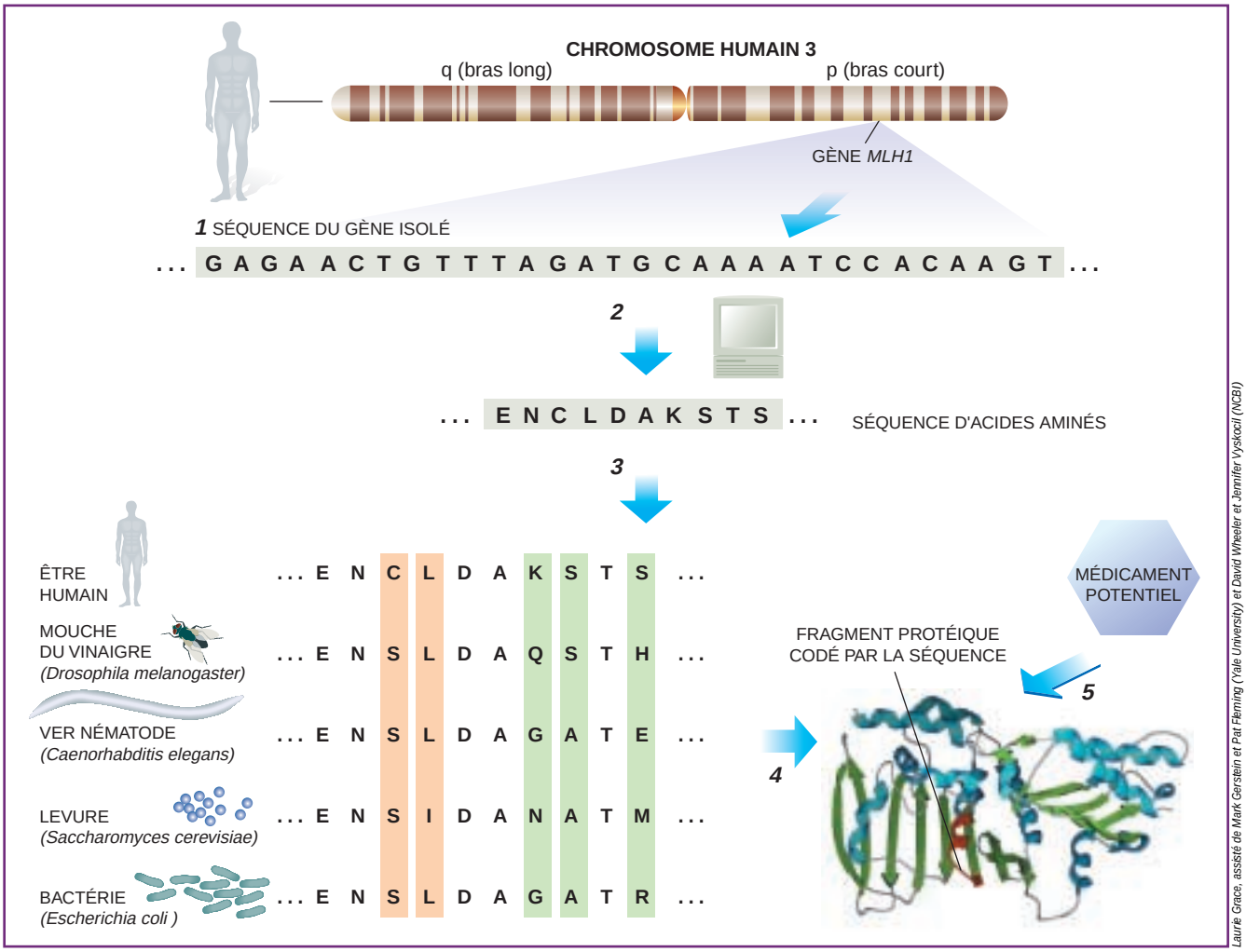
Aujourd'hui, les biologistes cherchent un médicament pour bloquer la cible de la cathepsine K. Les recherches de principes actifs, ainsi que les évaluations de l'activité, de la toxicité et de l'absorption, se font encore dans un laboratoire de biochimie classique, c'est-à-dire sur une paillasse avec de la verrerie. Ce travail long et fastidieux sera progressivement confié aux ordinateurs, grâce aux nouveaux outils bio-informatiques et à la quantité croissante de données sur les structures des protéines et sur les voies métaboliques de l'organisme *(voir l'encadré de la page 48)* : après la biologie *in vivo* est venue la biologie *in vitro* ; bientôt, elle sera *in silico*.

Beaucoup pensent que les promesses réelles de la génomique ne seront tenues que si la bio-informatique se développe : ce n'est pas la génomique qui est révolutionnaire, mais l'utilisation qui peut en être faite.

La bio-informatique à domicile

Cette révolution bio-informatique se fonde sur des groupes de toutes tailles. Certaines entreprises de bio-informatique pourvoient de gros clients, en ciblant leurs produits : elles créent des programmes personnalisés et des services de consultation adaptés à l'ensemble de l'entreprise. En revanche, d'autres visent des clientèles plus petites ou des organismes universitaires : elles proposent, par le réseau *Internet*, des services d'achat à la carte et de consultation de banque de données.

Enfin, de grands groupes pharmaceutiques valorisent leurs travaux en génomique en créant des unités internes de bio-informatique. Parmi eux, certains ont organisé de vastes services afin de mettre en œuvre les logiciels d'accès aux bases de données : ces groupes veulent ainsi rapprocher des services classiquement séparés, tels ceux qui sont chargés de la recherche de nouveaux



3. LA COMPARAISON DE GÈNES HUMAINS avec ceux d'animaux modèles offre parfois des pistes pour la recherche de nouveaux médicaments. Par exemple, quand on connaît la séquence du gène humain *MLH1* (1), associé au cancer du côlon, un programme informatique peut traduire cette séquence d'ADN en une séquence d'acides aminés, qui compose la protéine codée (2). On cherche ensuite des séquences similaires dans des banques de données de protéines synthétisées par des organismes modèles (3). Les bandes vertes indiquent des zones de grande variabilité (les acides aminés sont peu conservés d'un organisme à un autre), et les bandes orange, des zones de faible variabilité. On modélise la protéine humaine à partir des structures des protéines similaires des organismes modèles (4) et, enfin, on crée une molécule qui se fixe à la protéine modélisée (5).

produits, de la formulation, de la toxicologie et des essais cliniques.

Naguère, ces services étaient excessivement cloisonnés, et les données d'un service restaient inaccessibles aux autres équipes. La bio-informatique donnera la même information à tous, et chacun fera un usage individuel des données.

En outre, les services internes communs de bio-informatique réduisent les investissements informatiques. Par exemple, les moyens logiciels individuels, utilisés par les différents chercheurs d'une même entreprise pour l'accès aux bases de données et le traitement de ces données, sont remplacés par une plate-forme logicielle unique : les économies se comptent en millions de francs.

Afin d'intégrer pleinement la bio-informatique à leurs sociétés, des groupes pharmaceutiques concluent parfois des alliances stratégiques ou même acquièrent des sociétés de génie génétique plus petites ; ainsi munis de partenaires et de réseaux commerciaux, ils augmentent leurs compétences bio-informatiques et s'adaptent aux nouvelles techniques sans réorganiser leurs propres structures.

Toutefois, le regroupement des banques de données est notoirement périlleux. On cherche à éviter les mauvaises communications à l'aide de systèmes de nomenclature des données comprise aussi bien des diverses banques de données que des systèmes d'appellation. Imaginons qu'une banque *A* communique avec une banque *B*, celle-ci avec une banque *C*, *C* avec *D*, et ainsi de suite. Si *A* change ses données, les références de la banque *D* risquent d'être périmées, surtout quand le renouvellement des données est permanent. Ce problème s'aggrave, car les connaissances en biologie et les analyses par ordinateur sont de plus en plus complexes.

Des améliorations systématiques seront d'une grande aide, mais le vrai progrès semble reposer sur l'ingéniosité des utilisateurs : le succès de la bio-informatique n'est pas une affaire de matériel, ni de logiciel, mais bien d'intelligence.

Ken HOWARD est journaliste scientifique.

D. B. SEARLS, *Using Bioinformatics in Gene and Drug Discovery*, in *Drug Discovery Today*, vol. 5, n° 4, pp. 135-143, avril 2000.

Au-delà du génome

CAROL EZZELL

Le séquençage de l'ADN d'un être humain n'est qu'une première étape dans l'étude du fonctionnement d'une cellule : on doit aujourd'hui explorer les protéines codées par ces gènes.

Au moment où les gènes sont d'actualité, ils sont supplantés par les protéines. Un savoir est acquis : les séquences des milliers de gènes humains. Les biologistes doivent en conquérir un autre : le fonctionnement de ces gènes ; on cherche où et quand ils s'expriment, et quelles sont les propriétés des protéines qu'ils codent. Ce nouveau champ du savoir, nommé protéomique concentre la majorité des efforts financiers et humains des organismes publics et des sociétés privées.

Pour comprendre la protéomique, examinons comment un gène code une protéine. Dans le noyau d'une cellule, un segment d'ADN qui correspond au gène est d'abord transcrit en une molécule d'ARN, également formée de nucléotides. Cet ARN est composé de régions codantes, nommées exons, et de régions non codantes, les introns. L'ARN est débarrassé des introns par des enzymes d'«épissage» : ainsi se forme l'ARN messager. Si l'ADN est la série de plans qu'une cellule utilise pour synthétiser les protéines, l'ARN messager est la copie d'une partie du plan que l'entrepreneur emmène sur le chantier chaque jour. Le chantier est à l'extérieur du noyau, dans le cytoplasme, où la séquence en nucléotides de l'ARN messager est traduite en une séquence en acides aminés, qui composent la protéine.

Bien que chaque cellule de l'organisme contienne l'ensemble de l'ADN utile à la fabrication et au fonctionnement d'un être humain, beaucoup de ces gènes ne s'expriment jamais, c'est-à-dire ne sont jamais transcrits en ARN messagers, lorsque le développement embryonnaire est achevé. En outre, les gènes ne s'expriment qu'en fonction de leur situation dans les tissus et de leur rôle dans l'organisme. Par exemple, dans une cellule bêta du pancréas, le gène de l'insuline s'exprime intensément (la cellule est pleine d'ARN messagers qui codent l'insuline), alors qu'il reste muet dans un neurone.

Jusqu'au début des années 1980, les biologistes pensaient qu'à un gène correspondait un seul ARN messager et une seule protéine. Il n'en est rien : un gène peut être transcrit par segments, qui sont ensuite coupés et regroupés en une variété d'ARN messagers ; en outre, les protéines ainsi synthétisées sont parfois modifiées : ajout de groupes méthyle ($-CH_3$), liaison à des sucres, etc. Ainsi la connaissance du génome humain n'est pas celle du fonctionnement cellulaire : les biologistes doivent étudier le «transcriptome», c'est-à-dire l'ensemble des ARN messagers produits par les cellules, et le «protéome», c'est-à-dire l'ensemble des protéines synthétisées à partir des ARN messagers.

Des puces biologiques

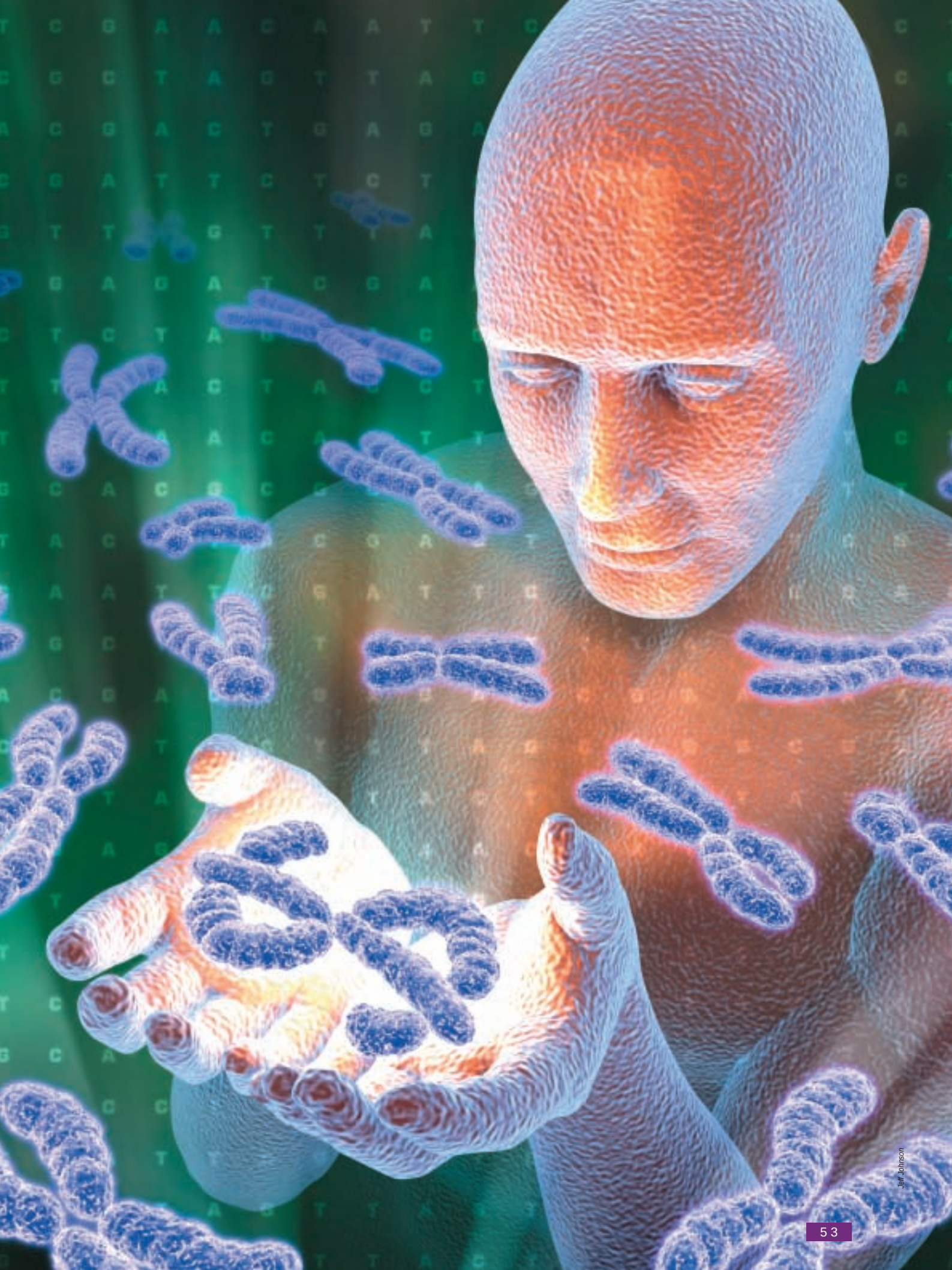
L'étude du transcriptome bénéficie d'un progrès technique récent : la mise au point de biopuces. Sur des plaques de verre d'environ un centimètre carré, on ancre les molécules d'ADN complémentaires des ARN messagers que l'on recherche. Ces molécules d'ADN complémentaires sont obtenues à partir des ARN messagers. Ces derniers s'associent spécifiquement à leurs ADN complémentaires selon le même principe que les deux brins de la double hélice d'ADN. L'ARN messager d'un gène ne s'associera qu'au seul ADN qui lui est complémentaire. Celui-ci est une sorte de copie du gène sans intron.

Pour utiliser ces biopuces, les biologistes extraient les ARN messagers d'un échantillon, ils les marquent par une molécule fluorescente, puis les versent sur la puce. La lame de verre de la biopuce est divisée en microsurfaces (jusqu'à plusieurs dizaines de milliers). Sur chacune d'elles, on n'ancre qu'un seul ADN complémentaire bien identifié. Chaque ARN fluorescent de l'échantillon s'associe à l'ADN qui lui est complémentaire. À l'aide d'un microscope, on repère les microsurfaces fluorescentes : selon les parties des biopuces auxquelles les ARN messagers fluorescents se lient, les biologistes déduisent, de la séquence des ADN complémentaires, les séquences des ARN messagers de l'échantillon. De telles puces détectent simultanément plus de 60 000 ARN messagers ; par exemple, des biopuces détectent spécifiquement les ARN messagers humains associés au cancer.

Depuis 1998, des centres de recherche publics et privés utilisent de tels systèmes pour identifier les ARN messagers produits par divers types de cellules cancéreuses. Grâce aux biopuces, les généticiens savent, par exemple, que 5 692 gènes sont actifs dans les cellules de cancer du sein et que, parmi eux, 277 sont inactifs dans les cellules des autres tissus. Des composés qui bloqueraient les protéines codées par ces 277 gènes pourraient être des médicaments anticancéreux qui auraient moins d'effets secondaires que les chimiothérapies actuelles. Aussi des millions de francs ont-ils été récemment investis dans l'analyse des protéines impliquées dans le cancer.

L'étude des ARN messagers n'est qu'un moyen : on veut surtout mieux connaître les protéines qui, replacées dans le fonctionnement d'une

1. APRÈS AVOIR ANALYSÉ ses chromosomes, l'espèce humaine doit aujourd'hui déterminer où et quand les gènes sont actifs, et identifier les protéines que ces gènes codent.



Jeff Johnson

cellule, constituent les cibles des médicaments. Seules, mais nombreuses! On suppose que les milliers de gènes humains produisent plus d'un million de protéines différentes. Certains experts estiment qu'au cours de la prochaine décennie l'industrie pharmaceutique disposera de plus de 10 000 protéines humaines contre lesquelles on pourra chercher des principes actifs : c'est 25 fois plus de cibles que l'industrie pharmaceutique n'en a évaluées depuis qu'elle existe. Pour satisfaire leurs actionnaires, les grandes sociétés pharmaceutiques doivent identifier trois à cinq nouveaux médicaments potentiels chaque année, soit deux à cinq fois plus qu'aujourd'hui. Aussi, elles s'associent avec des petites sociétés spécialisées qui leur fournissent des cibles potentielles de médicaments déjà testées.

Des machines à protéines

Des sociétés privées qui ont participé au séquençage du génome s'intéressent aujourd'hui à l'étude de l'expression protéique. La Société *Celera genomics*, en association avec l'Institut suisse de bio-informatique, a créé une entreprise qui veut identifier tout le protéome humain. Simultanément, les techniques d'analyse se perfectionnent : en 1999, une équipe a ainsi publié les plans d'un scanner moléculaire qui automatiserait le processus fastidieux de séparation et d'identification des milliers de protéines d'une cellule.

Aujourd'hui, pour étudier les protéines, on utilise souvent la technique

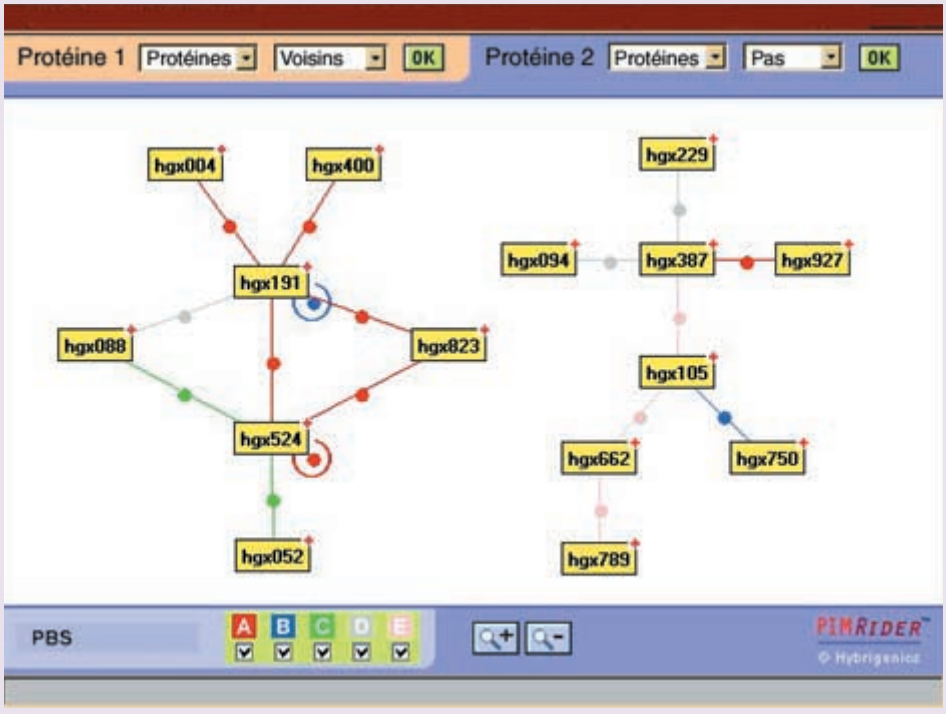
La protéomique fonctionnelle

L'étape suivant le séquençage du génome et l'identification des gènes consiste à étudier les protéines que ces derniers codent. Cependant, les protéines n'interviennent pas de façon isolée dans les processus biologiques cellulaires : elles constituent avec d'autres protéines une ou plusieurs cascades d'interactions, qui sont l'expression de fonction biologiques. La caractérisation de ces interactions entre protéines est le champ d'application de la protéomique fonctionnelle qui met en correspondance les données du génome et les observations biologiques.

Les techniques de la protéomique fonctionnelle s'appliquent à trois niveaux de complexité croissante : l'identification des protéines qui agissent directement avec quelques protéines choisies ; l'identification des cascades d'interactions complètes ; enfin, l'identification des connexions entre les différentes cascades d'interactions.

En France, la Société *Hybrigenics* met au point des méthodes d'étude de la protéomique fonctionnelle. Par exemple, elle a amélioré la technique de double hybride de la levure décrite dans l'article. Ainsi, au lieu de mettre en présence dans une levure deux protéines entières qui déclencheront une réaction observable lorsqu'elles interagissent, on segmente les protéines «proies» en de nombreux fragments, puis on cherche les éventuelles interactions de la protéine «appât» et ces nombreux «fragments proies». Les résultats obtenus sont plus fiables et plus précis. Grâce à cette méthode, on détermine quel fragment d'une «protéine proie» agit avec une protéine «appât» donnée.

En contrepartie, le nombre d'expérimentations et de la complexité du traitement des données issues de ces expérimentations, augmente, en particulier lors de l'identification des réseaux d'interactions. Le nombre d'interactions étudiées atteint souvent plusieurs centaines de millions à quelques milliards : l'exploitation des données requiert alors des systèmes informatiques adaptés. Les systèmes bio-informatiques mis en œuvre effectuent un «criblage à haut débit», c'est-à-dire le tri de plusieurs centaines de millions d'interactions testées en quelques semaines. On obtient ainsi une cartographie des réseaux d'interactions protéiques *(voir la figure)* que les utilisateurs visualisent directement à partir de leur ordinateur via le réseau *Internet*.



Sur son ordinateur, un utilisateur obtient une cartographie d'interactions protéiques. Les rectangles jaunes représentent les protéines et les lignes de couleur, les interactions.

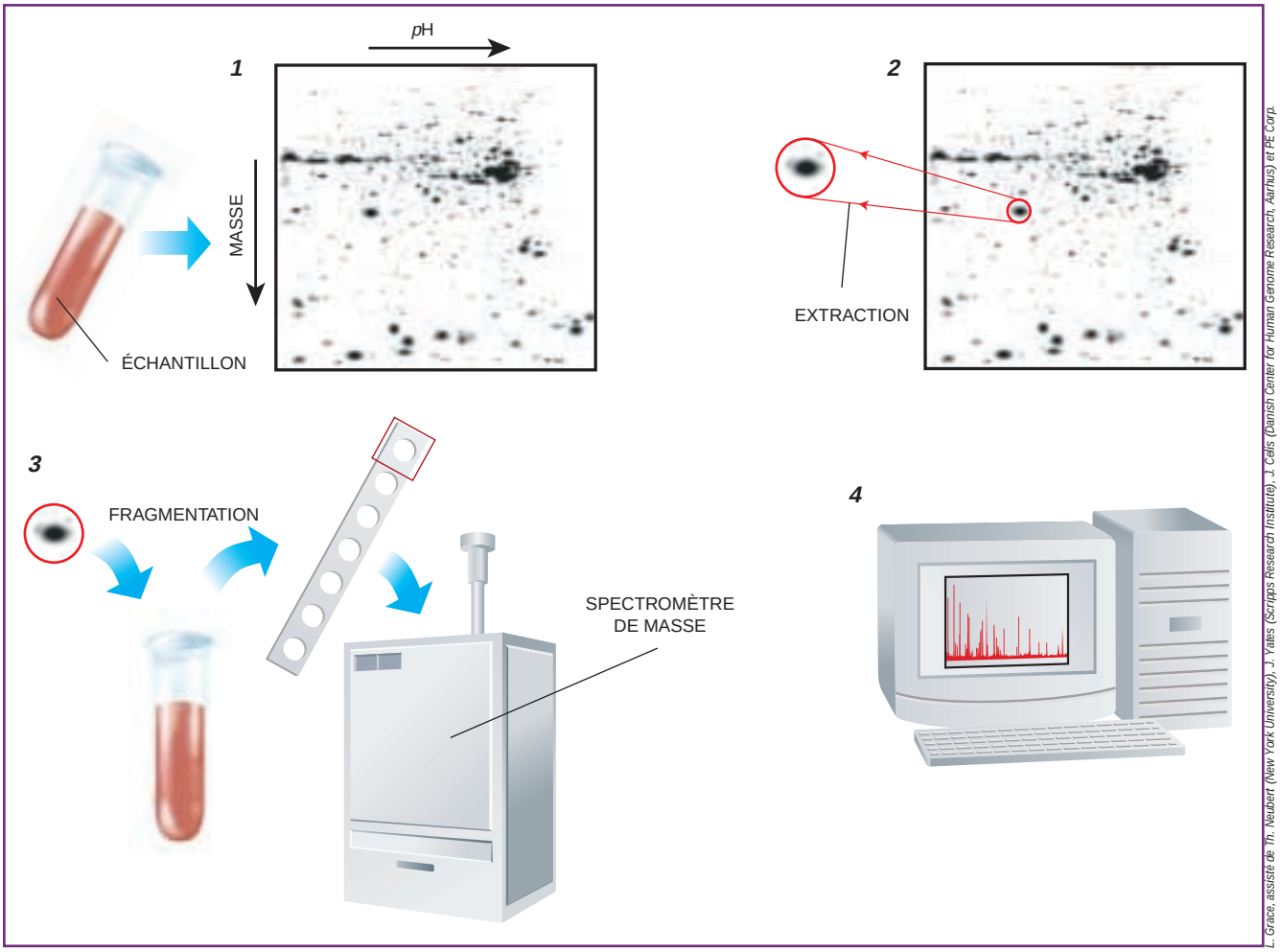
Donny STROSBERG, PDG de la Société *Hybrigenics*

d'électrophorèse bidimensionnelle, qui sépare les protéines selon leur charge électrique et leur taille : on dépose une solution qui contient de nombreuses protéines sur une étroite bande de polymère dont l'acidité varie d'un bord à l'autre. Quand on applique au gel une différence de potentiel électrique, chaque protéine du mélange se déplace à une vitesse qui dépend de sa charge et de l'acidité du milieu. On place ensuite la bande au bord d'un gel plat et on applique une nouvelle différence de potentiel électrique, perpendiculaire au premier ; les protéines migrent sur le gel à une vitesse qui dépend de leur masse moléculaire. Finalement, chacune des taches obtenues (après une réaction chimique de «révélation») contient une protéine différente.

Dans les laboratoires universitaires, les taches sont généralement extraites du gel grâce à un perforateur, et les protéines de chaque tache sont analysées par spectrométrie de masse : à l'aide d'enzymes, on découpe les protéines en peptides (des courtes chaînes d'acides aminés) dont on identifie la masse et la charge électrique ; à partir de ces informations, des programmes d'ordinateur retrouvent dans les banques de données de séquences le gène dont la séquence peut coder pour de tels peptides. Un gène est ainsi relié à une tache : on déduit enfin la séquence de la protéine de la tache analysée *(voir la figure 2)*. Aujourd'hui, les robots remplacent progressivement les biologistes dans cette tâche : des machines entièrement automatiques extraient les taches des gels, utilisent les enzymes pour couper les protéines en morceaux, transmettent les fragments à un spectromètre de masse laser et transfèrent les informations à un ordinateur en vue de leur analyse.

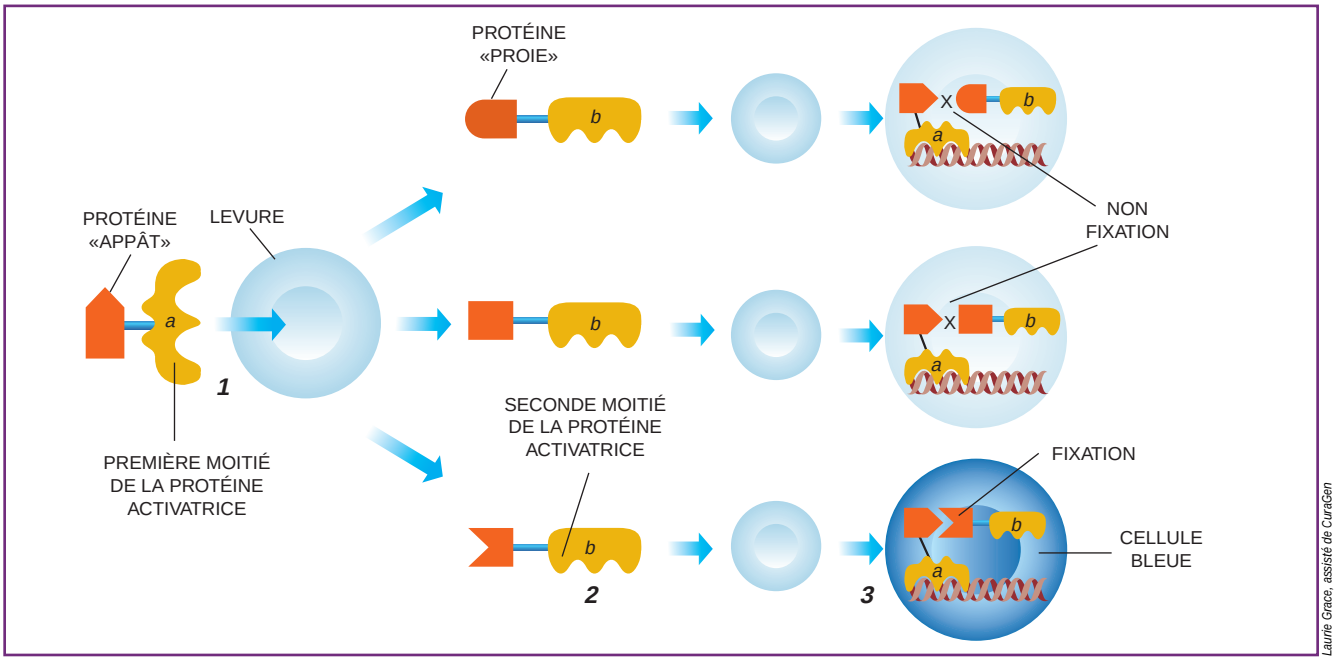
Toutefois, l'automatisation ne pallie pas les inconvénients de l'électrophorèse bidimensionnelle. Notamment, leur préparation est délicate, et les protéines fortement chargées, ou de masse faible, sont mal séparées. Les protéines qui ont des régions hydrophobes, telles les protéines membranaires, sont également difficiles à étudier par cette technique ; or, les récepteurs transmembranaires sont des cibles thérapeutiques importantes.

Dans les cas complexes, les biologistes utilisent alors la méthode de «culpabilité par association» : on a des indications sur la fonction d'une protéine en connaissant ses interactions avec une autre protéine déjà connue. Ainsi, en février 2000, des généticiens ont identifié 957 interactions parmi les 1 004 protéines de la levure de boulangerie *Saccharomyces cerevisiae*. Ils ont employé un système de double hybride : une protéine qui active un gène est coupée en deux ; une moitié est fixée à une protéine «appât» dont on connaît la fonction, l'autre est fixée à la protéine «proie» que l'on étudie. Quand les protéines



L. Grace, assisté de Th. Neubert (New York University), J. Yates (Scripps Research Institute), J. Celis (Danish Center for Human Genome Research, Aarhus) et PE Corp.

2. L'IDENTIFICATION DE PROTÉINES INCONNUES s'effectue sur un gel bidimensionnel. On sépare les protéines selon leur charge (*p*H) dans une direction et leur masse dans la direction perpendiculaire *(1)*. Chaque tache obtenue contient une protéine *(2)* que des enzymes découpent en plusieurs fragments *(3)*. L'ordinateur d'un spectromètre de masse établit un diagramme des fragments de protéines en fonction de leur masse *(4)*. Ce diagramme est caractéristique de la protéine initiale.



Laurie Grace, assisté de CuraGen

3. DANS LA TECHNIQUE DE «DOUBLE HYBRIDE» de la levure, on utilise deux moitiés *(a* et *b)* d'une protéine qui active un gène, tel celui qui active la synthèse d'un pigment bleu. Grâce à cette technique, on détermine quelle protéine d'un ensemble de protéines «proies» inconnues se lie à une protéine «appât» connue : l'ADN codant la protéine «appât», lié à l'ADN codant la moitié *(a)* de la protéine activatrice, est introduit dans le noyau de la levure *(1)*. Puis, l'ADN codant l'autre moitié *(b)* de la protéine activante lié à l'ADN codant une des protéines «proies» est également introduit dans le noyau de la levure *(2)*. Lorsque la couleur de la cellule change, la protéine «proie» s'est liée à la protéine «appât» *(3)*.

«proie» et «appât» interagissent, les deux moitiés de la protéine activatrice sont réunies : le gène s'exprime, et la présence de la protéine codée par le gène révèle l'association *(voir la figure 3)*.

Depuis le séquençage du génome de la levure en 1996, les fonctions d'un tiers de ses 6 000 gènes restent mystérieuses. Cependant, grâce au système de double hybride, une partie des protéines codées par ces gènes ont été classées en catégories fonctionnelles, telles la production d'énergie, la réparation de l'ADN, le vieillissement... Le plan des interactions protéiques de la drosophile est également en cours de déchiffrage.

La levure était un prototype ; avec la drosophile, on obtient des informations sur un organisme multicellulaire, et on utilise la protéomique pour trouver de nouveaux médicaments.

Certaines sociétés appliquent une variante du système de double hybride de la levure à l'étude des voies métaboliques spécifiques de maladies. Par exemple, l'une d'elles teste les interactions biochimiques des protéines codées par le gène *MMAC1*, qui, quand il est muté, conduit parfois à un cancer du cerveau et de la prostate.

Enfin, des biopuces d'un type nouveau apparaissent. Certaines, sous la forme de bandelettes, adsorbent, c'est-à-dire retiennent à leur surface, les protéines selon des propriétés telles que leur solubilité dans l'eau, leur affinité pour des ions métalliques, etc. Ensuite placées dans un lecteur de puces, qui inclut un spectromètre de masse, les protéines des bandes sont identifiées.

En mars 2000, une telle biopuce à protéines a été utilisée pour la découverte de 12 marqueurs potentiels d'une maladie bénigne de la prostate et de six marqueurs du cancer de la prostate. On espère que les tests fondés sur les marqueurs ainsi identifiés seront plus fiables et précis que le test utilisé aujourd'hui.

De la structure à la fonction

L'identification des protéines humaines sera un premier travail, qui devra se poursuivre par l'étude de la fonction des protéines, l'élucidation de leur forme et de leur structure. Un groupe de biologistes a récemment proposé un programme d'exploration de la structure des protéines normales et anormales par des techniques de radiocristallographie quasi automatisées.

Classiquement, pour déterminer la structure d'une protéine, on doit utiliser un cristal pur de cette protéine, que l'on irradie avec des rayons X. Ces derniers rebondissent sur les atomes de la molécule et forment une figure de diffraction dont on déduit la forme tridimensionnelle de la molécule entière. Le programme proposé veut perfectionner et automatiser les techniques actuelles.

Connaissant la structure exacte de chacune des protéines du protéome humain, les concepteurs de médicaments élaboreront des molécules qui s'adapteront aux sites protéiques et les activeront ou les empêcheront d'agir. À ce jour, de telles tentatives de conception rationnelle ont eu peu de succès. Cependant, l'espoir subsiste pour les 99 pour cent de protéines humaines dont on ne connaît pas encore la structure.

Carol EZZELL est journaliste à *Scientific American*.

The Chipping Forecast, supplément spécial de *Nature Genetics*, vol. 21, n° 1, janvier 1999.

A. ABBOTT, *A Post-Genomic Challenge : Learning to Read Patterns of Protein Sythesis*, in *Nature*, vol. 402, pp. 715-720, 16 décembre 1999.

R. JAMES, *Differentiating Genomics Companies*, in *Nature Biotechnology*, vol. 18, pp. 153-155, février 2000.

R.F. SERVICE, *Structural Genomics Offers High-Speed Look at Proteins*, in *Science*, vol. 287, pp. 1954-1956, 17 mars 2000.