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.

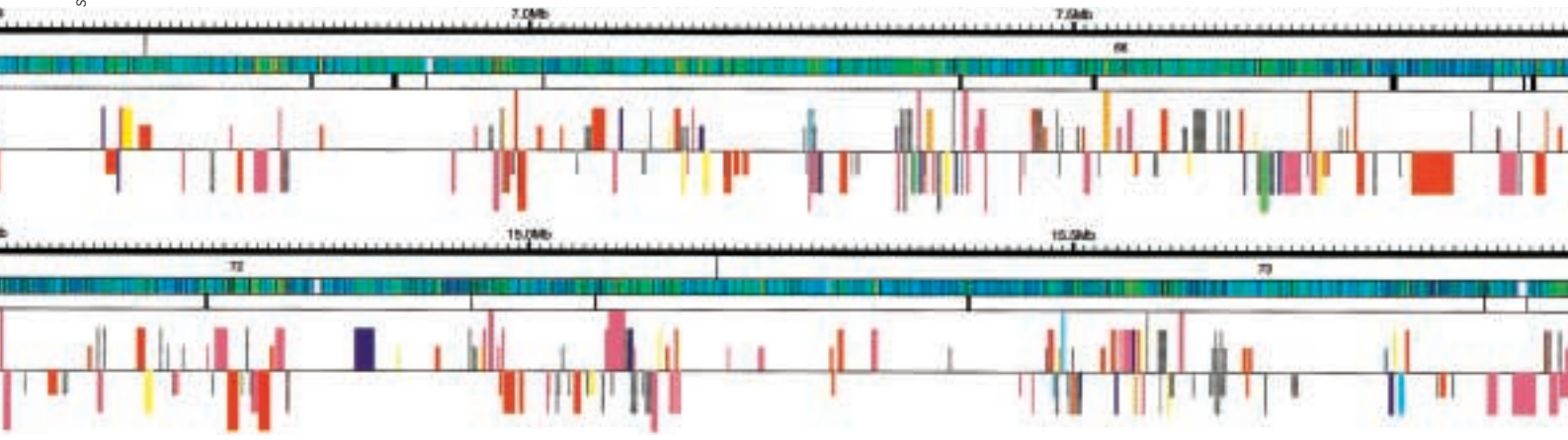
Pourquoi? Parce que la bio-informatique aide à trouver plus efficacement les principes actifs de médicaments que les méthodes classiques : en identifiant les molécules anormales qui causent les maladies, elle facilite le tri des molécules testées. Elle réduit ainsi le temps de recherche et de mise au 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.

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.

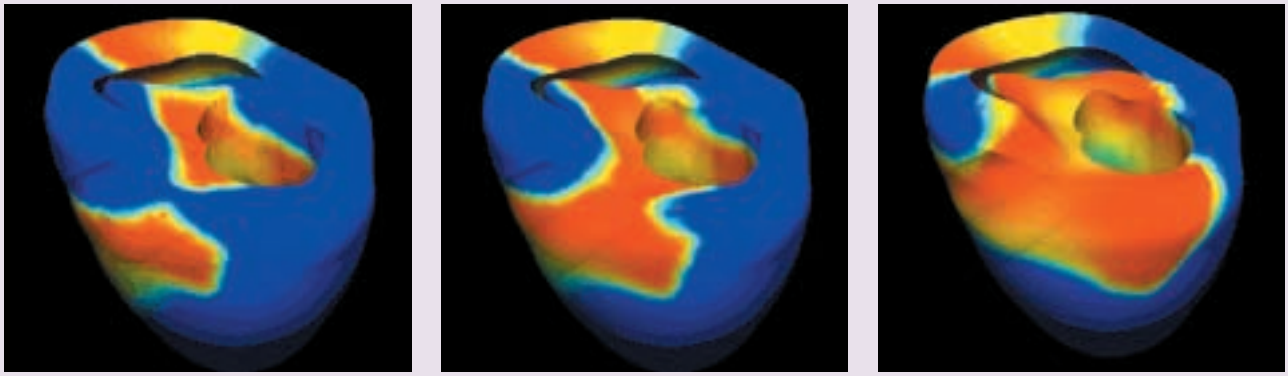
Certes, la compréhension d'un système à 13 variables est difficile, mais le génome humain est un système à 100 000 variables ! Si chaque variable a deux états possibles, combien y a-t-il de situations possibles? $2^{100\,000}$, soit approximativement $10^{30\,000}$. Ainsi, même en ne traitant les gènes 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.