

Bioinformatique : Chronique d'une révolution annoncée

(traduction allemande publiée dans la NZZ du 23 août 2000)

A l'heure où le génome humain est sur le point d'être entièrement déchiffré, une jeune discipline fait beaucoup parler d'elle: la bioinformatique. Totalement inconnue il y a encore quelques années, elle est aujourd'hui incontournable en science et en économie. Dans cette ascension, la Suisse n'est pas en reste. Aujourd'hui, elle occupe même le devant de la scène.

C'est une belle revanche pour les bioinformaticiens de la première heure. En quelques années, leur discipline, d'abord considérée par beaucoup de biologistes comme un gentil passe-temps pour chercheurs ratés, s'est imposée au point de devenir essentielle à la recherche biomédicale. Aujourd'hui, des compagnies de bioinformatique fleurissent dans le monde entier, ainsi que des programmes de formation qui peinent à répondre à une demande toujours croissante. Aux sarcasmes a succédé l'enthousiasme. Que s'est-il passé ? Pour comprendre cette révolution, un petit retour dans le temps s'impose.

C'est au milieu des années 80 que le terme bioinformatique fait son apparition pour décrire l'ensemble des applications de l'informatique aux sciences de la vie. Définition vague qui englobe une multitude de domaines très variés, de la robotique à l'intelligence artificielle. Mais dans le contexte qui nous intéresse, ce terme s'applique à l'ensemble des outils informatiques qui permettent de stocker, analyser et visualiser les informations contenues dans les séquences des gènes et des protéines de tous les êtres vivants. Par conséquent, l'histoire de la bioinformatique est étroitement liée à celle de la biologie moléculaire, l'étude des molécules du vivant.

Des gènes aux protéines

Au fin fond de nos cellules se trouve notre patrimoine génétique : l'ensemble de nos gènes ou génome. Véritable base de données de la vie, ces gènes sont des « recettes » permettant à nos cellules de produire les protéines, l'un des matériaux de base pour la construction de notre organisme. Chaque gène contient les instructions pour fabriquer au moins une protéine, dans un « langage » particulier utilisant un alphabet de 4 lettres (A, C, G et T). D'où l'importance de connaître l'ordre de ces lettres, c'est-à-dire la séquence des gènes. Les protéines, elles, peuvent être comparées à de longs « colliers de perles ». Il existe 20 sortes de perles : les acides aminés. La séquence d'une protéine, ou la succession des acides aminés qui la composent, détermine sa forme et donc sa fonction dans l'organisme.

Connaître la séquence est donc essentielle pour comprendre la fonction. Tous les mécanismes du vivant font appel aux protéines. Par exemple, le transport de l'oxygène dans le sang, l'élasticité de la peau ou la digestion des aliments sont assurés par autant de protéines différentes. Si l'on veut espérer pouvoir un jour comprendre comment les protéines interagissent entre elles - ce qui s'avère d'ores et déjà d'une complexité gigantesque - il est nécessaire de connaître leur fonction individuelle. D'autre part, quand on sait que la modification d'un seul acide aminé peut aboutir à une protéine anormale, donc à une maladie, on comprend facilement tout l'enjeu : connaître la séquence « saine » et « malade » permettra peut-être de réparer ou, du moins, corriger la fonction correspondante.

La naissance de la bioinformatique

Dans l'histoire de la biologie moléculaire, l'année 1953 marque un tournant décisif : pour la première fois, la séquence complète d'une protéine, l'insuline, est obtenue par le biochimiste Frederick Sanger. Au début des

années 60, on découvre que cette séquence est déterminée génétiquement et en 1966 le code génétique - c'est-à-dire la correspondance entre les « lettres » des gènes et les « perles » des protéines - est élucidé. A cette époque, seules vingt séquences de protéines étaient connues. La première base de données les réunissant date de 1965. En réalité, il s'agissait d'un livre, publié aux Etats-Unis sous le nom d'*Atlas of protein sequence and structure*. En 1981, cet atlas ne contenait encore que 1660 séquences qui n'étaient disponibles sur aucun support informatique. Aujourd'hui le nombre de séquences de protéines répertoriées, tout organisme confondu, avoisine les 400'000 ! Et avec les nombreux séquençages de génomes en cours, dont celui de l'Homme, ce nombre va encore se multiplier.

Cette formidable accélération est due à l'émergence dans les années 80 de techniques de séquençage automatiques, rapides et efficaces. En 1995, la première séquence d'un génome entier, celui de la bactérie *Haemophilus influenzae* (2 millions de lettres), est obtenue. En 99, c'est le génome du premier organisme pluricellulaire, le ver *Caenorhabditis elegans* (100 millions de lettres), qui est déchiffré. Face à ce déluge d'informations, la bioinformatique se développe. Dès le début des années 90, on assiste à une véritable explosion du nombre de bases de données qui permettent de gérer ce flot d'informations. Très vite, l'apparition de serveurs web, qui assurent la diffusion de ces données à l'échelle mondiale, va permettre à la bioinformatique de littéralement décoller.

De la séquence à la fonction

Aujourd'hui, les 3 milliards de lettres du génome humain sont sur le point d'être entièrement connues. Une étape cruciale, certes, mais pour l'instant cette séquence ne représente qu'un texte gigantesque dans lequel on aurait tout bonnement oublié d'indiquer la ponctuation. Reste à définir le début et la fin de chaque phrase (les gènes) sachant que celles-ci ne représentent que 5 % de l'ensemble du texte ! En résumé, dans l'état actuel il n'a pas de sens. Le but étant à terme de découvrir la fonction de toutes ces séquences, le déchiffrement du protéome (ensemble des protéines) est une tâche mille fois plus colossale. Certaines protéines ne seront exprimées qu'en réponse à des conditions particulières ou dans un seul tissu ou encore à un moment précis de notre développement. Une fois produites, elles subiront de multiples modifications qui ne sont pas indiquées dans la recette de base (le gène) et ne peuvent être déterminées qu'expérimentalement. Or, ces informations sont essentielles pour comprendre le rôle et le fonctionnement des protéines. En conséquence, le nombre de protéines humaines après modifications est probablement plus proche du million que des 30'000 à 40'000 estimées à partir du génome. A l'ère génomique succède celle de la protéomique. Là encore la bioinformatique vient au secours des chercheurs.

Sortes d'encyclopédies informatisées, les bases de données concentrent toutes les informations amassées par les chercheurs du monde entier, contenues dans les génomes et les protéomes des organismes vivants, de la bactérie à l'Homme en passant par la levure, la mouche ou la souris. Au niveau des protéines, ces informations concernent leur séquence, mais aussi leur structure, leur fonction, les modifications subies une fois produites ou encore leur relation avec une maladie. Mais la bioinformatique ne s'arrête pas là. Elle fournit aussi des logiciels d'analyse de plus en plus performants qui permettent de prédire la fonction et la structure d'une protéine inconnue. En quelques années, un changement de paradigme s'est opéré: hier encore, on connaissait la fonction d'une protéine et l'on cherchait la séquence d'acides aminés qui lui correspondait. Aujourd'hui, on a déchiffré des milliers de séquences dont il reste à découvrir la fonction. Pour cette tâche, le chercheur possède à disposition différents outils qui lui permettent de faire des recherches de similarité. Nombre de protéines sont très ressemblantes d'une espèce à l'autre. En clair, il s'agit pour lui de comparer une séquence inconnue à celles d'autres protéines dont on connaît la fonction, la forme et même les modifications, pour trouver des similarités, c'est-à-dire des régions ou des domaines conservés. Toutes ces informations aideront le chercheur à poser des hypothèses concernant la fonction voire la structure de cette protéine. Cela fait, celui-ci retournera à ses éprouvettes pour vérifier ses hypothèses de manière expérimentale. Mais cette démarche lui aura fait gagner un temps considérable.

La Suisse sur le devant de la scène

Aujourd'hui, plus de 500 bases de données génétiques et protéiques - en général gratuites pour les universitaires et payantes pour les sociétés - sont accessibles sur le net et des centres de bioinformatique ont vu le jour dans le monde entier. En tête de file, des instituts publics, comme aux Etats-Unis, le NCBI (National Center for Biotechnology Information) ou en Europe, l'EBI (European Bioinformatics Institute). Mais la Chine, le Japon ou l'Australie connaissent également un essor important. De leur côté, les groupes pharmaceutiques et biotechnologiques - à l'exception des grandes compagnies qui ont développé leur propre division de bioinformatique - font appel aux services d'entreprises privées qui se multiplient un peu partout. Mais l'utilité

de la bioinformatique pour la biotechnologie moderne est devenue si fondamentale que pour la première fois, le monde académique, les compagnies privées et les consortiums de recherche publics nationaux et internationaux expriment une réelle volonté de coopérer.

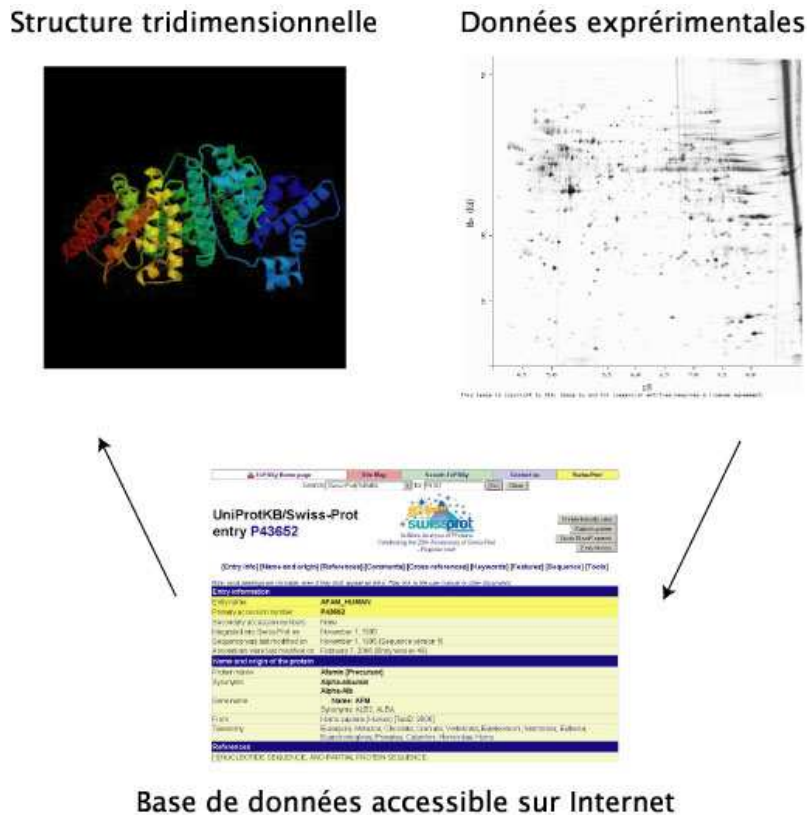


Fig.1 A partir des données expérimentales (ici un gel d'électrophorèse bidimensionnelle qui permet de séparer les protéines), la bioinformatique permet d'analyser et stocker les informations concernant une protéine dans une base de données (ici Swiss-Prot), puis de prédire et visualiser la structure en trois dimensions de cette protéine (ici l'albumine humaine qui assure le transport d'une hormone thyroïdienne du sang au cerveau).
 Mise à jour de la figure, avril 2006

Dans cette aventure, la Suisse n'est pas en reste, loin de là. Ayant même très tôt parié sur la bioinformatique, elle fait aujourd'hui figure de pionnière en la matière. Principalement grâce à une collaboration de longue date entre cinq groupes de recherche qui sont entre autres à l'origine d'ExPASy, le premier serveur web entièrement dédié à l'analyse des gènes et des protéines, ou de Swiss-Prot, la plus importante base de données sur les protéines au monde. Ces groupes - situés à Genève et Lausanne - forment depuis trois ans l'Institut Suisse de Bioinformatique (ISB) qui constitue une concentration unique au monde de spécialistes en la matière. Regroupant une centaine de chercheurs, l'ISB travaille au développement de bases de données et de logiciels dédiés à l'analyse des séquences des gènes et des protéines et assure des programmes de recherche en collaboration avec d'autres groupes en Suisse (BioZentrum de Bâle, ETH de Zürich) et dans le monde entier (NCBI, EBI etc.). L'institut fournit également des prestations de services pour la communauté suisse des chercheurs des sciences de la vie et organise des cours en bioinformatique. Depuis l'automne 99, il propose même un diplôme d'étude approfondie (DEA) co-organisé par les universités de Genève et Lausanne. Ce DEA, qui a pour ambition d'assurer en Suisse la formation de spécialistes dans un domaine où les débouchés se révèlent chaque jour plus nombreux, ne parvient déjà plus à satisfaire toutes les demandes d'inscription. Depuis trois ans, des sociétés ont aussi fait leur apparition, comme Geneva Bioinformatics (GeneBio) qui assure la commercialisation des bases de données développées par l'ISB aux utilisateurs industriels, ou la start-up bâloise Genedata qui fournit des services de conseils ainsi que des logiciels d'analyse génétique et protéique. Récemment, la création à Genève de l'entreprise Proteomics (GeneProt), spécialisée dans la recherche sur les protéomes, a confirmé la place prépondérante que la Suisse occupait déjà dans ce domaine.

A l'aube d'une ère virtuelle

Aujourd'hui, toutes les sciences du vivant utilisent les informations contenues dans les génomes et les protéomes. En dehors de la recherche fondamentale, il existe des applications biologiques, environnementales, agro-alimentaires, mais surtout médicales avec l'espoir de découvrir de nouveaux traitements. A l'horizon se profile une ère virtuelle qui va révolutionner la recherche biomédicale. Les bioinformaticiens mettent au point des outils qui permettent de prédire de plus en plus précisément la structure tri-dimensionnelle des protéines. Un jour, ces modèles informatiques permettront de réaliser des simulations en trois dimensions de l'interaction entre molécules, ce que l'on obtient encore pour l'heure au détriment d'un laborieux travail de laboratoire. Les pharmacologues de demain pourront alors tester virtuellement l'efficacité de molécules chimiques sur une protéine pour stimuler ou au contraire inhiber sa fonction biologique. Fini les tests à l'aveugle, le « drug design » permettra de construire des médicaments pour une maladie précise et peut-être même de mettre au point des médicaments « sur mesure » adaptés à un malade précis. Mais qui sait, peut-être un jour les bioinformaticiens seront-ils capables de modéliser non seulement les protéines mais une cellule entière voire un organisme complet, afin de prédire de manière globale les interactions en jeu. Nous n'en sommes qu'au début. Et la bioinformatique n'a pas fini de faire parler d'elle...

Sylvie Déthiollaz

Sources des Illustrations :

- Fig.1, structure tridimensionnelle: PDB ID: 1GNJ, Petitpas, I., Gruene, T., Bhattacharya, A.A., Curry, S., Crystal structures of human serum albumin complexed with monounsaturated and polyunsaturated fatty acids. *J.Mol.Biol.* **314** pp.955 (2001)
- Fig.1, données expérimentales: [SWISS-2DPAGE](#)

Parution: 23 août 2000

Protéines à la "Une" (ISSN 1660-9824) sur www.prolune.org est une publication électronique du Groupe Swiss-Prot de l'Institut Suisse de Bioinformatique (ISB). L'ISB autorise la photocopie ou reproduction de cet article pour un usage interne ou personnel tant que son contenu n'est pas modifié. Pour tout usage commercial, veuillez vous adresser à prolune@isb-sib.ch