Qu'est-ce que la bioinformatique?

Victor Jongeneel

Qu'est-ce que la **bioinformatique** ?

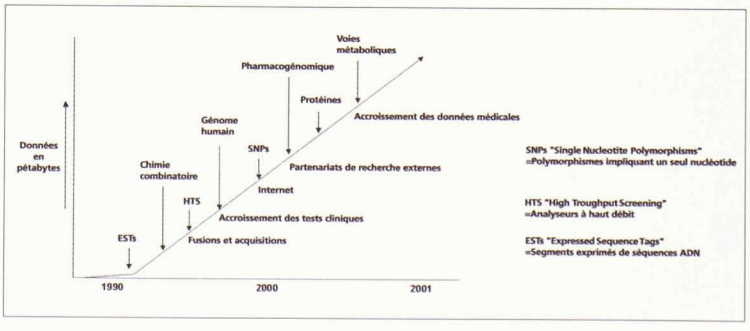
Le terme « bio-informatique »[1] n'est apparu dans la littérature scientifique qu'au tout début des années 90. De façon très générale, on peut inclure dans une définition de la bioinformatique toutes les applications de l'informatique à la biologie.

Celles-ci sont extrêmement nombreuses et couvrent des domaines aussi différents que l'étude *in silico* de la connectique des neurones, le traitement quantitatif et qualitatif d'images microscopiques, la gestion d'échantillons et de données expérimentales dans les grands laboratoires industriels ou la modélisation de l'évolution de populations animales dans des conditions écologiques spécifiques. Dans tous ces cas, l'informatique apporte des outils indispensables à l'analyse de phénomènes biologiques, à la formulation de nouvelles hypothèses ou à la gestion de données expérimentales. Typiquement, on voit émerger des projets pluridisciplinaires où collaborent biologistes et informaticiens (souvent assistés par des ingénieurs, physiciens, mathématiciens, etc.). Cette évolution a créé une demande croissante de scientifiques bénéficiant de connaissances approfondies dans les deux domaines, et capables de communiquer avec des professionnels des deux bords.

De plus, l'émergence en biologie de méthodes générant des quantités très importantes de données qui ne sont pas nécessairement liées à des hypothèses de travail précises a créé un besoin aigu de traitement et d'interprétation de ces données. Les projets les plus visibles sont ceux qui visent à déterminer la séquence complète des lettres qui composent les génomes de différents êtres vivants, dont, bien sûr, l'Homme. En complément indispensable de ceux-ci, on trouve les projets visant à cerner la partie du génome exprimée en ARN messager, le transcriptome, et les protéines synthétisées à partir de ces messagers, le protéome. Récemment, on a appris à mesurer simultanément les niveaux d'expression de dizaines de milliers de gènes dans des populations cellulaires différentes sur des puces à ADN, générant une fois de plus des données en quantités massives (fig. 1). Étroitement couplée aux projets de génomique, une forme spécifique de bioinformatique est ainsi apparue, qui est dérivée d'une discipline plus ancienne connue sous le nom d'analyse de séquences, et cette spécialité reçoit actuellement le plus d'attention parmi les biologistes et connaît le développement le plus rapide. C'est dans cette bioinformatique post-génomique que se profile l'Institut Suisse de Bioinformatique (ISB), né en 1998 de la volonté de cinq groupes de recherche travaillant dans la région lémanique de collaborer étroitement et de mettre en commun une partie de leurs ressources. Il est actuellement composé de huit groupes localisés à Genève, Lausanne et Bâle.

Etude des génomes

Malgré ce que la presse a tendance à dire, la détermination de la séquence d'un génome n'est de loin pas synonyme de décodage. En fait, un immense travail d'interprétation des données ne fait que commencer. Pour donner un exemple simple, il n'y a pas dans une séquence génomique de signaux précis indiquant la localisation des gènes : celle-ci doit être déduite en combinant des outils prédictifs encore très imparfaits, des données expérimentales sur le transcriptome et des comparaisons avec des gènes connus venant d'autres organismes. Encore beaucoup plus complexe est la déduction des signaux de contrôle, des interrupteurs et régulateurs génétiques, dont on sait qu'ils sont encodés dans le génome mais que l'on ne sait pas encore reconnaître de façon fiable, et encore moins analyser globalement. Le géno-



1

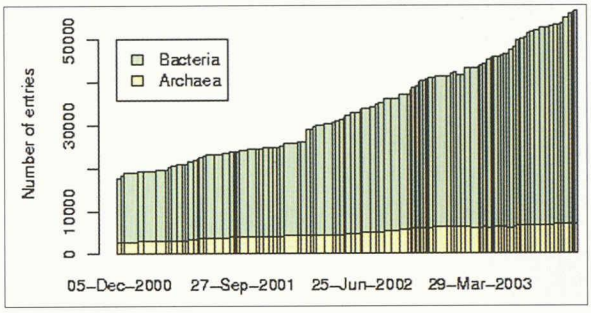
[1] D'abord publié dans le *Flash Informatique* N° 10 du 19 décembre 2000, le présent article a été repris et adapté avec l'accord de son auteur.

Fig. 1 : Les nouvelles technologies, dont la protéomique et les recherches pharmacologiques et médicales qui en découlent, ont produit une explosion des volumes de données (Source Jeff Augen, IBM)

Fig. 2 : Evolution du nombre de références concernant le protéome microbien dans SWISS-PROT *(www.expasy.org)*

Fig. 3 : Exemple de représentation d'une structure tridimensionnelle de la protéine dj1A de la bactérie Escherichia coli à gauche et l'équivalent humain à droite (Document ISB)

Fig. 4 : Principe du scanner moléculaire (Document ISB)



2

me est certainement le plan directeur permettant à un organisme de se former et de fonctionner, mais nous sommes encore très loin de comprendre comment cela est mis en œuvre. La différence entre une chenille et un papillon en constitue un exemple extrême : ces deux états de développement d'un même être vivant déclinent deux manières très distinctes d'utiliser les informations codées dans le même génome. Il faudra développer des outils aussi bien expérimentaux qu'informatiques pour réellement décoder les génomes et c'est à mon avis dans ce sens qu'il faut comprendre la génomique.

Un autre niveau d'analyse des génomes est celui du polymorphisme. On sait qu'il y a sur Terre peu d'individus génétiquement identiques, et que des différences suffisamment grandes définissent des espèces incapables de se reproduire entre elles. Les différences génétiques entre individus ne se traduisent pas seulement par des dissemblances visibles (physionomie, couleur de peau, etc.), mais aussi par des altérités physiologiques et métaboliques à l'origine de propensions à développer certaines maladies plutôt que d'autres, de sensibilités à certains médicaments, ou d'aptitudes et de goûts personnels divergents. La découverte des polymorphismes génétiques, et leur association avec des caractéristiques phénotypiques, est dans une grande mesure un problème d'analyse (bio)informatique de données expérimentales complexes; la bioinformatique est aussi indispensable à la formulation d'une stratégie cohérente de découverte de ces polymorphismes.

Etude des transcriptomes

Le transcriptome est la partie du génome transcrite en ARN, et en particulier en ARN messager pouvant encoder des protéines. Les ARN messagers sont les molécules qui serviront de matrice pour la synthèse de ces protéines, et représentent le potentiel de codage du génome. Une caractérisation complète du transcriptome est donc nécessaire pour obtenir un catalogue représentatif de tous les gènes. Les techniques actuelles produisent de petits fragments de séquences d'ARN messager, qu'il s'agit d'attribuer à un gène particulier, puis de ré-assembler pour produire la séquence complète de l'ARN correspondant. Il faut aussi les placer sur la séquence génomique. Tout cela est fait *in silico*, par le biais de techniques informatiques perfectionnées. Plusieurs projets de séquençage de transcriptomes à grande échelle sont actuellement en cours, qui couvrent la plupart des organismes pour lesquels des séquences génomiques sont disponibles. L'ISB et l'Institut Ludwig de recherches sur le cancer, en collaboration avec le National Cancer Institute américain, travaillent actuellement à une telle reconstitution, en utilisant les séquences génomiques comme matrices sur lesquelles faire l'assemblage.

Diverses techniques, dont les puces à ADN, permettent de mesurer simultanément, dans des populations de cellules données, les niveaux d'expression d'un grand nombre d'ARN messagers. Les méthodes de bioinformatique sont absolument centrales pour la mise en place de ces techniques aussi bien que pour l'exploitation des données qui en résultent. Par exemple, le choix des ADN à déposer sur les puces se fait par corrélation croisée entre les données sur le génome et sur le transcriptome, et en fonction de la disponibilité de fragments précis dans des banques d'ADN publiques. D'autre part, l'interprétation des données fournies par ces puces demande des traitements statistiques très semblables à ceux utilisés dans l'exploration de données industrielles (« data mining »). La plupart des biologistes n'ayant pas été formés dans ce domaine, il est actuellement du ressort des (bio)informaticiens.

Etude des protéomes

Le protéome est défini comme l'ensemble des protéines exprimées dans un tissu donné sous des conditions données. Quoique conditionné par le transcriptome, il en est bien distinct tant du point de vue chimique que logique (c'est-à-dire que l'expression d'un gène au niveau de l'ARN messager n'est pas nécessairement corrélée avec celle qui intervient au niveau de la protéine correspondante). Une série d'avances techniques récentes, en particulier dans les séparations par chromatographie ou électrophorèse et l'analyse des macromolécules par spectrométrie de masse, a rendu possible l'analyse directe du protéome dans beaucoup d'échantillons biologiques même extrêmement complexes. Les spectromètres de masse en particulier sont maintenant en mesure d'analyser des échantillons à une cadence très élevée, produisant par là-même des terabytes de données. Le degré de complexité et d'hétérogénéité des protéines ne peut être appréhendé qu'à l'aide d'une multitude de technologies diverses. Ces différentes approches génèrent des données de différents types, spécificité et qualité. Malheureusement, les outils d'analyse informatiques de ces données n'ont pas suivi la cadence, et la protéomique est actuellement une branche qui a cruellement besoin d'outils informatiques nouveaux, et de puissances de calcul rivalisant avec celles utilisées dans les plus pointues des sciences de l'ingénieur. Il est symptomatique que *GeneProt*, une compagnie qui œuvre au décryptage protéique d'échantillons complexes et qui connaît actuellement des difficultés, ait installé à Genève le plus grand parc informatique non-militaire du monde.

3

Banques de données

Un autre aspect important de la bioinformatique est l'organisation du savoir biologique et biochimique. Par exemple, la séquence brute d'un génome, ou d'une protéine, est de très peu d'utilité pour les biologistes. Ce qui compte, c'est de pouvoir corréler ces séquences avec des propriétés biologiques précises, documentées dans la littérature scientifique. C'est à cette tâche que c'est attelé le groupe qui gère *SWISS-PROT*, l'une des banques de données les plus utilisées en recherche biomédicale. *SWISS-PROT* comporte aujourd'hui plus de 138 000 fiches, dont chacune est compilée, vérifiée et régulièrement mise à jour par une équipe de biologistes sous la supervision d'Amos Bairoch de l'UniGe et de l'ISB (fig. 2 et 3). Il existe des milliers de banques de données biologiques de par le monde, certaines très spécialisées et d'autres d'intérêt très général, qui couvrent tous les sujets possibles et imaginables. Les mieux connues sont probablement *GenBank / EMBL / DDBJ*, une collaboration internationale visant à documenter (mais pas à annoter) toutes les séquences générées par les différents centres actifs dans ce domaine. Une banque de données très utilisée par les médecins et biologistes est *OMIM* (Online Mendelian Inheritance in Man), un compendium de connaissances sur les gènes humains et les conséquences médicales de leurs dysfonctionnements, dont l'un des directeurs scientifiques est Stylianos Antonarakis du Département de génétique médicale de l'UniGe.

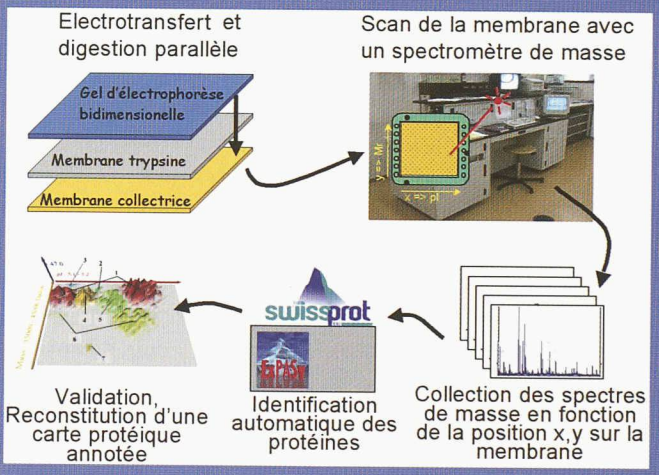
On peut également citer, parmi les bases de données collectant des données expérimentales interprétées, *SWISS-2DPAGE*. Développée en collaboration par le groupe d'informatique protéomique de l'ISB conduit par Ron Appel et le Biomedical Proteomics Research Center de l'UniGe/HUG, cette dernière contient des cartes protéiques annotées pour plus de trente tissus et échantillons biologiques différents. Si l'on peut arguer que la maintenance de banques de données ne relève pas de la bioinformatique à proprement parler, il n'en demeure pas moins qu'il s'agit d'une activité absolument nécessaire aux biologistes et qu'elle demande une infrastructure et un savoir-faire de haut niveau en informatique.

Perspectives

La bioinformatique n'est pas une discipline en soi. Elle résulte du besoin des biologistes d'analyser les données qu'ils produisent en quantités de plus en plus importantes, et d'intégrer celles-ci dans un cadre scientifique rigoureux. Il est donc beaucoup plus facile de définir ce qu'est un bioinformaticien (un scientifique qui maîtrise au moins une discipline biologique et l'informatique nécessaire pour résoudre efficacement les problèmes propres à ce domaine). Dans cette perspective, la bioinformatique est d'ailleurs appelée à s'intégrer toujours davantage aux formations des biologistes et ingénieurs. C'est dans cette optique également que l'ISB organise une formation post-graduée de type DEA d'un an en bioinformatique.

Le scanner moléculaire

A Genève, la collaboration étroite entre le Biomedical Proteomics Research Group de l'HUG et l'ISB illustre l'intégration de la bioinformatique à la protéomique avec notamment le projet de développer conjointement un scanner moléculaire (voir introduction à la protéomique, pages 6 à 10). Cette technique novatrice permet de traiter un échantillon complexe en un minimum d'opérations **(fig. 4)**, la bioinformatique intervenant surtout dans la phase d'analyse des données. Celle-ci comporte les tâches suivantes : détection des signaux livrés par les spectromètres de masse, calibration de ces spectres en utilisant le principe de voisinage, reconnaissance de clusters de masses localisées, soumission de ces ensembles de valeurs de masses à des outils d'identification de protéines, interprétation qualitative des résultats, identification et visualisation des résultats. Le processus de traitement des échantillons peut être soumis à variation et les outils bio-informatiques doivent pouvoir s'adapter à ces modifications. Cela a été rendu possible par une mise au point de la méthode corrélant besoins technologiques et bio-informatiques, expertises de laboratoire et savoir-faire informatique.



4

Victor Jongeneel, ISB, UNIL
Ch. des Boveresses 155, CH - 1066 Epalinges
avec la collaboration de Pierre-Alain Binz, ISB
1, Rue Michel-Servet, CH - 1211 Genève 4

Quelques URL pour en savoir plus :
<http://www.expasy.org>, serveur ExPASy, l'un des premiers serveurs Web de Suisse, et l'un des plus visités dans le domaine de la bio-informatique et de la protéomique
<http://www.ch.embnet.org>, serveur de l'antenne EMBnet suisse, offrant une large gamme de services (dont certains uniques au monde)
<http://www.expasy.org/alinks.html>, liste exhaustive de banques de données biologiques et de serveurs Web en bioinformatique
<http://www3.ncbi.nlm.nih.gov/omim>, banque de données OMIM mentionnée dans le texte
<http://www.isb-sib.ch/DEA>, serveur du DEA en bioinformatique de l'ISB
<http://www.bioinformatik.de/cgi-bin/browse/Catalog/Research_and_Education/Online_Courses_and_Tutorials/>, une bonne liste de cours et didacticiels online sur la bioinformatique
<http://www.ebi.ac.uk>, serveur du European Bioinformatics Institute, producteur de la banque de données EMBL
<http://www.ncbi.nlm.nih.gov>, serveur du National Center for Biotechnology Information, le centre de bioinformatique du Gouvernement américain et producteur de la banque de données GenBank