Bioinformatique : un domaine pluridisciplinaire

» **Olivier DELGRANGE**, Service d'Informatique théorique, FS – contact : olivier.delgrange@umons.ac.be

Les biotechnologies et le génie génétique sont des technologies de pointe dont les applications dans le domaine des sciences biomédicales ne sont plus à démontrer. Ces technologies utilisent intensivement les outils issus de la bioinformatique. Cette discipline a la particularité de regrouper, autour d'un même thème, beaucoup de disciplines des sciences ou de la médecine: la biologie, l'informatique, les mathématiques, la physique, la chimie,... Les informaticiens ne sont pas de simples prestataires au service de la biologie mais des interactions fortes entre les disciplines sont nécessaires pour résoudre les problèmes posés par le traitement de l'information biologique. Le bioinformaticien actuel doit savoir jongler avec beaucoup de spécialisations : les probabilités, les propriétés énergétiques des repliements de molécules, l'algorithmique, la génétique, ...

Le terme bioinformatique évoque souvent une simple interaction entre la biologie et l'informatique. Il s'agirait donc, à l'instar de la bureautique, de l'utilisation de l'ordinateur pour résoudre des problèmes ou pour répondre à des questions de nature biologique. La réalité scientifique de cette discipline est toute autre. Il s'agit d'un domaine de recherche pluridisciplinaire à part entière où se mêlent les compétences des biologistes, des informaticiens, des physiciens et des mathématiciens pour résoudre des problèmes scientifiques posés par la biologie. Rappelons que le terme *informatique* provient du traitement de l'*informa*tion de manière automa*tique*. Admettons pour l'instant qu'il existe une *bioinformation* à traiter. Le terme bioinformatique serait alors le juste terme pour désigner le traitement automatique de cette bioinformation.

Il s'agit d'une discipline récente, qui prend ses origines au début des années 1980 lorsque le laboratoire européen de biologie moléculaire (EMBL : European Molecular Biology Laboratory) et le département américain de la santé (NIH : National Institute of Health) ont créé les banques de données EMBL et GENBANK pour répertorier les séquences d'ADN découvertes par les biologistes[1].

La bioinformation représente l'ensemble de toute les informations biologiques qui concernent les organismes vivants. Avant d'en préciser les divers composants, il convient de préciser certaines notions de biologie moléculaire.

Théorie fondamentale de la biologie moléculaire

L'ADN *(Acide DésoryriboNucléique)* est le support de l'information génétique qui définit les fonctions biologiques d'un organisme (nutrition, croissance, reproduction, respiration, communication, etc.). Chaque cellule d'un organisme contient une copie des molécules d'ADN de l'individu. L'ensemble des molécules d'ADN d'une cellule est appelé le *génome*, il caractérise l'individu. Comme l'ont découvert Francis Crick et James Watson en 1953, l'ADN se présente sous la forme de deux longs enchaînements moléculaires dont chaque élément est appelé un *nucléotide*. Tous les nucléotides ont la même structure chimique à l'exception de la base azotée qui peut être soit *l'Adénine*, la *Cytosine*, la *Guanine* ou la *Thymine*. Nous utiliserons donc uniquement les lettres A,C,G et T pour différencier les nucléotides. Par abus de langage, il est d'ailleurs courant de dire que l'ADN est composé d'une succession de bases. Les deux enchaînements nucléotidiques (on parle de deux *brins* d'ADN) forment une longue structure en double hélice dans laquelle chaque nucléotide d'un brin s'apparie[2] avec son nucléotide complémentaire sur l'autre brin : un A sera toujours en face d'un T et un C en face d'un G. Les deux brins sont

1. Ces banques de séquences sont accessibles aux adresses http://www.ebi.ac.uk/embl/ et http://www.ncbi.nlm.nih.gov/Genbank/index.html
2. Deux bases azotées s'apparient en créant un lien hydrogène entre elles.

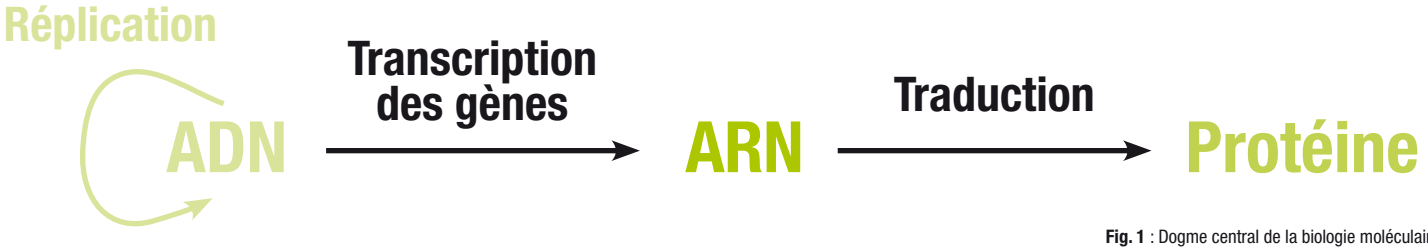

Fig. 1 : Dogme central de la biologie moléculaire

donc complémentaires : la connaissance de la séquence de bases d'un brin est suffisante pour connaître l'autre brin. Cette propriété est très importante car lors de la division cellulaire, les deux brins sont séparés, chacun sert de matrice pour synthétiser un nouveau brin complémentaire de telle sorte que les deux cellules issues de la division cellulaire possèdent toutes deux une copie en double hélice de chaque molécule d'ADN. Cette duplication porte le nom de *réplication* de l'ADN et est responsable de la transmission, de cellule en cellule et d'individu parent en individu enfant, du génome. Chaque molécule d'ADN est appelée un *chromosome*. Une *séquence* d'ADN est un fragment contigu d'un brin d'ADN. Une séquence est généralement représentée par la suite ordonnée des caractères représentant ses bases. Par exemple, ATCGGATCG est une séquence d'ADN.

Si l'ADN est responsable du transport et de la transmission du génome, c'est un autre type de molécules, *les protéines*, qui assurent les fonctions cellulaires de l'organisme. Ainsi, l'hémoglobine, les hormones, etc sont des protéines. Elles sont synthétisées à partir de séquences d'ADN spécifiquement localisées dans le génome de l'organisme : *les gènes*. Un gène est tout d'abord transcrit en une autre molécule : l'ARN (*Acide RiboNucléique*, qui est une chaîne linéaire mono-brin constituée de nucléotides de nature chimique légèrement différente des nucléotides de l'ADN). Durant cette transcription, la séquence d'ADN sert de modèle et l'ARN est synthétisé, nucléotide par nucléotide, comme le complémentaire du nucléotide lu sur l'ADN. Ainsi un un A produit un U (*Uracile*[3]), un C produit un G, un G produit un C et un T produit un A. La molécule d'ARN sort ensuite du noyau de la cellule (chez les organismes évolués dont les cellules contiennent un noyau) et est *traduite* en une succession linéaire d'*acides aminés* qui sont les constituants de base des protéines. Il existe 20 acides aminés différents, chacun d'entre eux étant synthétisé sur base d'un triplet de 3 nucléotides successifs de la molécule d'ARN. Notons donc qu'une protéine peut être représentée par la suite ordonnée des lettres qui représentent chacune un acide aminé. Par exemple MALPTPSDST est une séquence protéique. La protéine se replie sur elle même pour adopter *une forme tridimensionnelle qui détermine sa fonction*. L'ensemble de toutes les protéines est appelé protéome. Il convient de remarquer que seulement 1% du genome humain est finalement traduit en protéines., le surplus d'ADN est appelé *l'ADN poubelle*. Les connaissances à propos de cet ADN poubelle sont limitées mais on sait maintenant que certaines zones sont impliquées dans la machinerie biologique ou même parfois dans certaines maladies.

Pour conclure ces quelques rappels de biologie, le lecteur trouvera, représenté, sur la figure 1, l*e dogme central* de la biologie moléculaire. Notons que certains

3. L'ARN ne contient pas de Thymine, celle-ci est remplacée par l'Uracile.

ARN ne sont pas traduits en protéines, ce sont des ARN fonctionnels dont la fonction est dictée par la manière avec laquelle ils se replient dans l'espace. Enfin, remarquons que chez les organismes évolués dont les cellules possèdent un noyau, la totalité d'un gène n'est pas traduite en protéine : certains segments (les *introns*) sont éliminés au niveau de l'ARN. Les segments qui sont effectivement traduits sont les *exons*.

La Bioinformation

La bioinformation est donc constituée, entre autres, de toutes les séquences d'ADN, d'ARN et de protéines connues, de l'influence connue de l'environnement sur ces différentes molécules, de l'impact de la présence de certaines motifs spécifiques dans l'ADN sur le métabolisme de l'individu, des informations concernant la proximité évolutive des espèces, des interactions entre gènes, etc. Lors de ces dernières décennies, le volume connu de ces informations a grandi de manière exponentielle. Ainsi, les séquences d'ADN actuellement connues, toutes espèces confondues, totalisent environ 262 milliards de caractères contre 80 milliards en 2005 ou encore 10 milliards en 2000 (voir figure 2). De plus, les séquences sont annotées, c'est-à-dire que toutes les connaissances issues de la recherche sont associées aux séquences concernées. Le volume d'informations à traiter est dès lors gigantesque !

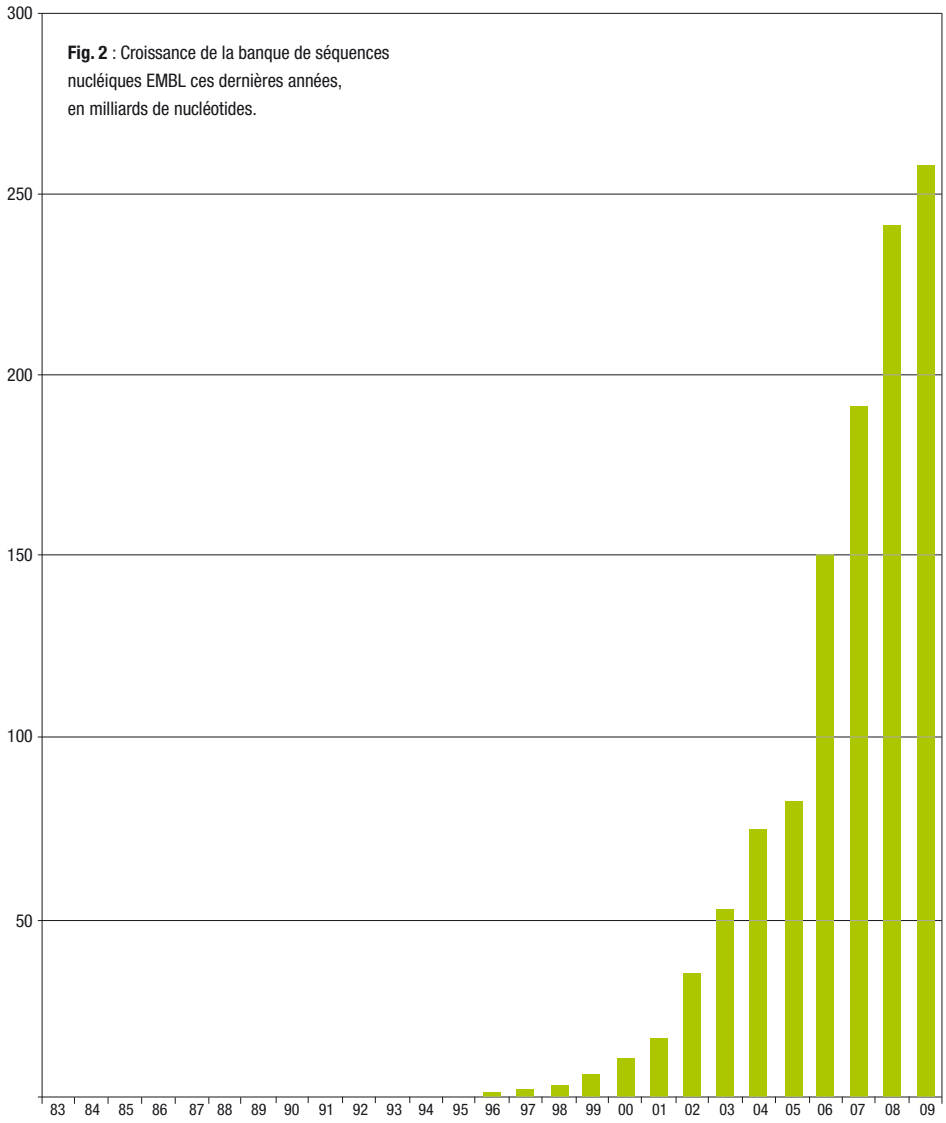

Fig. 2 : Croissance de la banque de séquences nucléiques EMBL ces dernières années, en milliards de nucléotides.

Quelques aspects importants de la bioinformatique dans lesquelles les différentes disciplines interagissent

Mentionnons ici, de manière non exhaustive, différents domaines importants de la bioinformatique.

Organisation et diffusion de l'information génétique

Toutes les informations connues sont mises à disposition des chercheurs du monde entier le plus rapidement possible. Il s'agit non seulement des séquences nucléiques ou protéiques brutes (successions de caractères) mais également de toutes les annotations des séquences et autres informations connexes. Les annotations peuvent porter sur des références à la littérature scientifique, des localisations de zones spécifiques de la séquence avec ses caractéristiques, des informations sur la régulation des gènes, etc. Il s'agit donc là d'un volume très important d'informations qui sont consultables à l'aide d'outils spécifiques de recherche. Il est fondamental de constater que les connaissances de biologie moléculaire et de ses applications ne seraient pas à leur état de connaissance actuel sans l'apport de l'informatique et en particulier du réseau internet et des bases de données. Dans le monde, trois grandes banques centralisent l'information : EMBL, GenBank et DDBJ (Japon). Des accords de partenariat ont étés conclus pour que les mêmes informations soient systématiquement détenues par les trois centres. Certains pays possèdent aussi leurs propres copies des banques de données génétiques, accompagnées de services spécifiques et de ressources humaines pour gérer le tout. Ainsi, en Belgique, l'organisme *BEN (Belgian EMBnet Node)* a géré depuis 1992 les ressources et services centraux de bioinformatique destinés aux chercheurs des communautés néerlandophone et francophone du pays. Malheureusement, la survie de BEN est actuellement fortement menacée : suite à des restrictions budgétaires, le bureau de la politique scientifique fédérale a décidé de ne pas accorder le budget permettant à BEN de continuer ses activités[4]. Les services de BEN sont pourtant intensivement utilisés par la communauté scientifique belge.

Comparaison de séquences

Des séquences de nucléotides proches ont très probablement des rôles proches. Par exemple, si la séquence d'un gène *A* ressemble fortement à la séquence d'un gène *B*, alors séquence de la protéine codée par *A* sera très proche de la séquence de la protéine codée par *B*. Les deux protéines adopteront donc une structure tridimensionnelle proche et auront donc une fonction semblable. Des conclusions fonctionnelles peuvent donc être tirées suite à la comparaison de séquences nucléotidiques ou même protéiques. La comparaison est un outil de base en bioinformatique. On parle plus communément de *l'alignement*

4. Le lecteur intéressé trouvera plus d'informations sur le site de BEN : http://www.be.embnet.org

de séquences qui met en correspondance les parties communes à deux ou à plus de deux séquences. La figure 3 présente un alignement entre deux séquences d'ADN. Des algorithmes[5] spécifiques ont été conçus pour réaliser un alignement. Une problématique différente est celle de retrouver, lorsqu'une nouvelle séquence est identifiée, si elle « ressemble à quelque chose » qui serait déjà connu dans les banques de séquences. Un outil spécifique, nommé BLAST (Basic Local Alignment Search Tool) a été créé pour cela, il utilise des critères empiriques et mathématiques pour identifier des fragments semblables. Il est extrêmement utilisé en pratique pour tirer des conclusions sur une séquence par rapport aux connaissances déjà présentes dans les banques.

Séquençage d'ADN

La détermination de la suite de nucléotides d'une molécule d'ADN porte le nom de *séquençage*. Les appareillages actuels, sont capables d'obtenir uniquement de très courtes séquences de nucléotides (quelques centaines). Or un chromosome humain entier peut comporter plusieurs centaines de millions de nucléotides. L'informatique permet de résoudre le problème d'assemblage de courts fragments en une longue séquence. Notons qu'il s'agit là d'un défi important pour les informaticiens car le problème d'assemblage fait partie de la classe des problèmes *NP-complets*, c'est-à-dire la classe des problèmes « les plus difficiles à résoudre » à l'aide d'un algorithme. Le problème est résolu en utilisant simultanément une grande quantité d'ordinateurs inter-connectés.

Localisation des gènes

Il s'agit de localiser les fragments d'ADN qui codent pour des protéines. Cette localisation s'appuie sur l'utilisation des statistiques et autres méthodologies mathématiques, informatiques, physiques et biologiques. Les gènes localisés par calcul automatique sont ensuite vérifiés en laboratoire car les méthodes de bioinformatique utilisées sont empiriques. Une technique rapide pour localiser un gène dans une séquence *X* est de tirer des conclusions à partir des gènes connus sur des séquences jugées semblables à *X* par BLAST.

Localisation de zones répétitives

Il s'agit de séquences d'ADN, assez mal définies, dans lesquelles la répartition des nucléotides semble fortement biaisée par rapport à une répartition uniforme. Par exemple les *répétitions en tandem approximative (RTA)* qui peuvent être impliquées dans certaines maladies génétiques telles que la maladie de Huntington (maladie d'ordre neurologique à évolution progressive). En collaboration avec E. Rivals du LIRMM (Montpellier), nous avons ainsi développé, dans le Service d'Informatique Théorique de l'Université de Mons, l'algorithme *STAR (Search for Tandem Approximate Repeat)* utilisable, via internet, par la communauté scientifique[6].

```
ACTGTCTACTGAATGT
AC-GTAG---GAACGT
```

Fig. 3 : Alignement entre les deux séquences d'ADN : ACTGTCTACTGAATGT et ACGTAGGAACGT. Les parties rouges identifient les zones communes

Le service de Biologie Moléculaire du Prof. A. Belayew a apporté une contribution très importante en découvrant, dans l'ADN poubelle (inutile a priori), un gène possédant des répétitions locales, produisant une protéine toxique qui provoque une myopathie grave. Ces recherches ont, à de nombreuses reprises, utilisé des techniques de bioinformatique. Un de ces outils a été développé au sein de notre service, il s'agit d'un programme spécifique capable de localiser des zones précises de bégaiements dans l'ADN.

Prédiction de structures 2D et 3D de l'ARN fonctionnel et des protéines

Ces structures aident à déterminer les fonctions des molécules; leurs prédictions reposent souvent sur des critères physicochimiques qui visent à minimiser les énergies libres des molécules qui se replient sur elles-mêmes. Les prédictions sont ensuite vérifiées par des techniques de cristallographie car les algorithmes de prédiction ne peuvent pas prendre tous les paramètres en compte, notamment l'influence de l'environnement de la cellule sur le repliement.

Phylogénie

Il s'agit de l'étude de l'évolution des organismes vivants en vue d'établir leur parenté. Cette problématique peut utiliser l'ensemble de la bioinformation dont on dispose sur les espèces, y compris les résultats d'alignements de séquences.

Modélisation des réseaux de régulation

Cette discipline étudie la manière avec laquelle l'expression des gènes favorise ou inhibe l'expression d'autres gènes. Il s'agit d'un domaine en plein essor en bioinformatique.

Aide à la conception de médicaments

En identifiant, les séquences anormales qui causent les maladies, la bioinformatique rend plus simple le tri des molécules à tester et réduit donc fortement le temps de recherche et de mise au point des médicaments.

Pour conclure, signalons que la bioinformatique est non seulement très présente actuellement en matière de recherche et développement mais également en matière d'enseignement. Ainsi, certains pays ont déjà fait le pas et proposent, dans leur formation universitaire, un cursus complet de « bioinformatique ». Le bioinformaticien, contrairement à d'autres disciplines de recherche, doit être compétent dans les différents domaines qui concernent la bioinformatique. ■

5. Un algorithme est une méthode informatique de résolution d'un problème
6. http://www.atgc-montpellier.fr/star/