Analyse informatique des données moléculaires

F. CORPET, C. CHEVALET

INRA, Laboratoire de Génétique Cellulaire, BP 27, 31326 Castanet-Tolosan cedex

e-mail : chevalet@toulouse.inra.fr

Résumé. *Les données biologiques, en particulier les séquences d'ADN, s'accumulent extrêmement rapidement. Pour exploiter toutes ces données, une nouvelle science est née, la bioinformatique. Accéder de manière rapide et fiable aux données disponibles dans les banques internationales et analyser les données expérimentales produites à grande échelle nécessitent des outils informatiques puissants et en perpétuel développement. Assembler les séquences brutes, trouver les unités fonctionnelles des séquences génomiques, comparer les séquences entre elles, prédire les structures et les fonctions des macromolécules, comprendre les interactions entre les gènes et leurs produits en termes de réseaux métaboliques mais aussi d'évolution des espèces : toutes ces questions nécessitent l'utilisation de la bioinformatique et son développement.*

La bioinformatique est la science de l'utilisation de l'ordinateur dans l'acquisition, le traitement et l'analyse de l'information biologique. Le terme, très vague au départ, tend maintenant à se limiter à la biologie moléculaire. Les données traitées par la bioinformatique sont toutes celles qui intéressent le biologiste : séquences d'ADN ou de protéine mais aussi références bibliographiques, images, résultats expérimentaux bruts, logiciels, etc.

Le premier travail de l'informaticien est de représenter ces données biologiques sous une forme assimilable par l'ordinateur. La biologie expérimentale à haut débit nécessite une acquisition et une conversion de données analogiques en données symboliques sans intervention humaine : le processus d'interprétation des signaux d'un séquenceur à fluorescence comme un flux de nucléotides passant devant le détecteur en est un exemple. Pour être de portée générale, la modélisation des données doit être compatible avec un grand afflux de données et compatible avec des outils développés par d'autres. Très souvent, le développement de nouveaux outils se heurte à un manque de standard ou à l'existence de trop nombreux 'standards' : l'incompatibilité des formats est le pain quotidien de l'utilisateur comme du développeur en bioinformatique.

Une fois les données représentées, il faut développer des méthodes de traitement et d'analyse qui répondent aux demandes des biologistes. Ils ont besoin d'outils simples, puissants, intuitifs dans leur manipulation et combinables à l'infini de manière imprévisible… Il y a bien là un défi pour des scientifiques, à l'interface de plusieurs domaines : informatique, mathématiques, statistiques et biologie moléculaire. Peu de personnes peuvent à la fois appréhender le problème biologique, créer l'algorithme mathématique qui propose une solution, concevoir l'outil informatique, le réaliser et faire un modèle statistique pour aider à l'analyse des résultats.

Les outils à la disposition du biologiste moléculaire

Internet offre au biologiste une quantité écrasante d'information et d'outils pour analyser ses données et on trouve assez facilement des listes de sites intéressants, par exemple sur le site d'Infobiogen (tableau 1). Certains serveurs proposent d'analyser les données en direct (réponse sur une page Web) ou en différé (réponse par e-mail). D'autres permettent de télécharger leurs programmes pour les installer localement ce qui nécessite plus ou moins de connaissances en informatique. On peut enfin ouvrir un compte sur des gros centres serveurs qui possèdent de nombreux logiciels accessibles par les utilisateurs enregistrés : par exemple le paquet GCG, qui est le plus complet et le plus connu des ensembles de programmes pour le biologiste moléculaire, est disponible à l'INRA de Jouy et de Toulouse ou à Infobiogen.

Cet article fait le point sur l'analyse informatique des données moléculaires, et principalement des séquences d'acides nucléiques ou d'acides aminés. Un premier paragraphe traite de l'accès à l'information, les suivants présentent le travail d'analyse des données depuis le séquençage de l'ADN jusqu'à la recherche de fonction des protéines ; enfin, quelques pistes de développement nécessaires à l'analyse des données de grands séquençages sont proposées.

INRA Prod. Anim., 2000, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 191-195.

Tableau 1. *Principaux serveurs généralistes de bioinformatique.*

European Bioinformatics Institute (EBI) : http://www.ebi.ac.uk
ExPASy Molecular Biology Server : http://www.expasy.ch
Informatique appliquée à l'étude des Biomolécules et des Génomes : http://www.infobiogen.fr
Institute for Genomic Research : http://www.tigr.org
National Center for Biotechnology Information (NCBI) : http://www.ncbi.nlm.nih.gov

1 / Recherche d'information

Alors que la quantité de données biologiques augmente si rapidement, il est essentiel de savoir accéder à cette information. Des outils de recherches contextuelles sont disponibles pour de nombreuses banques de biologie moléculaire, qu'elles soient généralistes ou spécialisées. Les méthodes employées s'apparentent à la commande Edition /Rechercher d'un traitement de texte, ou à la fonction « grep » avec une efficacité plus ou moins grande selon la structure des données (texte quelconque ou champs structurés). La syntaxe de la requête varie beaucoup d'un outil à l'autre, ce qui nécessite un apprentissage avant de pouvoir faire des requêtes élaborées. Les deux outils principaux sont Entrez et SRS. Ils permettent non seulement de trouver des entrées correspondant à une requête, mais procurent aussi des liens vers d'autres informations dans la même banque ou dans une autre. Entrez permet, par exemple, de faire des recherches dans la partie publique de la base bibliographique Medline. SRS permet de croiser des recherches entre plusieurs banques (un plus ou moins grand nombre selon les serveurs). Ces recherches basées sur le texte sont dépendantes de la qualité des annotations présentes dans les banques (erreurs de frappe, fantaisie des auteurs…).

2 / Séquençage

Un séquenceur peut produire jusqu'à 80 kb de séquences brutes par jour. Il faut ensuite rassembler les morceaux produits (d'environ 500 bases) en contigs les plus longs possible, en tenant compte d'erreurs (substitution, insertion ou délétion) et de la redondance (chaque base de la séquence finale est présente plusieurs fois dans les séquences brutes). Les algorithmes d'assemblage de séquences les plus sophistiqués tiennent compte du niveau de confiance de chaque détermination de base, acceptent des erreurs de séquençage, donnent une probabilité pour chaque base dans la séquence consensus finale, utilise des informations auxiliaires comme la longueur des clones. Les zones de fortes répétitions gênent beaucoup cette reconstitution. L'idéal serait que le processus soit totalement automatique depuis le séquençage brut jusqu'à la séquence finale assemblée, mais des allers et retours entre informaticiens et biologistes sont encore nécessaires.

3 / Prédiction de domaines fonctionnels

L'avancée des technologies de séquençage rend plus économique l'acquisition de longues séquences anonymes que la détermination et le séquençage des seules régions codantes. Localiser des gènes ou d'autres régions fonctionnelles sur les séquences génomiques devient primordial si on veut tirer parti des grands projets de séquençage.

Il y a deux manières de trouver un gène. La plus évidente est de comparer la séquence génomique aux banques de séquences connues. Cette approche donne à la fois la position du gène dans la séquence et des éléments pour déterminer sa fonction, mais elle repose sur l'existence de séquences homologues, correctement annotées. L'autre approche repose sur des propriétés intrinsèques des gènes. Depuis la simple détection de cadres de lecture ouverts, les méthodes ont évolué pour proposer la détection d'un gène entier, avec les limites exon-intron. Ces méthodes utilisent des modèles de Markov cachés et nécessitent un apprentissage pour chaque espèce.

La détection de signaux fonctionnels non codants, comme les régions de régulation génétique, peut être beaucoup plus difficile que la détection des zones codantes. Certains motifs, comme les sites de reconnaissance des enzymes de restriction, sont si bien définis qu'un simple algorithme de recherche de mots suffit (stringsearch). D'autres ne sont définis que par un petit nombre de positions invariantes séparées par un nombre variable de positions dégénérées. Les méthodes utilisées peuvent soit partir de l'expertise humaine, soit utiliser un système d'apprentissage.

4 / Comparaison de séquences

« Y a-t-il dans la banque une ou plusieurs séquences qui ressemblent à la mienne ? » est la première question que se pose le biologiste lorsqu'il a obtenu une séquence. La réponse à cette question nécessite de définir la ressemblance entre séquences. L'alignement de deux séquences est la base de cette comparaison.

4.1 / Alignement de deux séquences

Un alignement permet de décrire une deuxième séquence comme le résultat d'une série de mutations sur la première (insertion, délétion ou substitution d'un résidu). Insertion et délétion sont représentées par un caractère nul inséré dans l'une ou l'autre séquence, permettant d'aligner les résidus qui se correspondent (substitution ou conservation). L'alignement peut être global (sur toute la longueur de la séquence) ou local (sur les parties les mieux conservées), selon la relation présumée entre les

séquences. On définit un score d'alignement qui permet de définir le meilleur alignement de deux séquences et de quantifier leur ressemblance.

Pour les modèles classiques de score, il existe une solution mathématique au problème de l'alignement de deux séquences : des algorithmes de programmation dynamiques permettent de trouver un alignement global ou un alignement local de meilleur score. Cependant, c'est au biologiste de juger si l'alignement de meilleur score est le meilleur alignement biologique.

4.2 / Recherche dans les banques

Théoriquement, la recherche des séquences semblables à une séquence donnée nécessite la comparaison de toutes les séquences de la banque avec la séquence requête. Ceci n'est possible, dans un temps raisonnable, qu'avec des ordinateurs spécialement conçus pour cela (carte « Bioaccelerator ») : ce n'est pas une méthode praticable pour des recherches ordinaires. On utilise donc des heuristiques qui permettent de réduire l'espace de recherche et le temps de calcul, sans trop de risque de passer à côté des solutions. Les méthodes couramment employées appartiennent aux familles « Fasta » et « Blast ». Des modèles statistiques accompagnent ces méthodes pour permettre d'évaluer la pertinence des résultats. Cependant, on ne doit ni avoir une foi excessive dans les résultats, ni les négliger systématiquement : ce sont des indicateurs fiables de similarité, mais la responsabilité de la décision incombe toujours au biologiste.

4.3 / Alignement multiple

L'alignement multiple est la base de l'étude de familles de protéines et de domaines fonctionnels. Son but est de révéler des similarités de séquence ou de structure dans une famille de séquences voisines dans l'évolution ou par la fonction (figure 1). L'extension mathématique du problème de l'alignement de deux séquences à n séquences n'a pas de solution simple (c'est ce que les mathématiciens appelle un problème NP complet). On utilise un algorithme approximatif connu sous le nom d'alignement progressif par paires. Les solutions proposées par cette méthode ne sont plus des solutions mathématiquement optimales. Elles apportent cependant beaucoup plus d'information que l'alignement de deux séquences (détection de zones plus ou moins conservées, motifs propres à des sous-familles, etc).

La comparaison d'une séquence et d'un alignement multiple (ou profil) est plus sensible que la comparaison de deux séquences. Elle est utilisée pour rechercher toutes les séquences d'une famille dans une banque (profil contre banque de séquences) ou pour prédire la fonction d'une séquence (séquence contre banque de profils).

5 / Classification des protéines

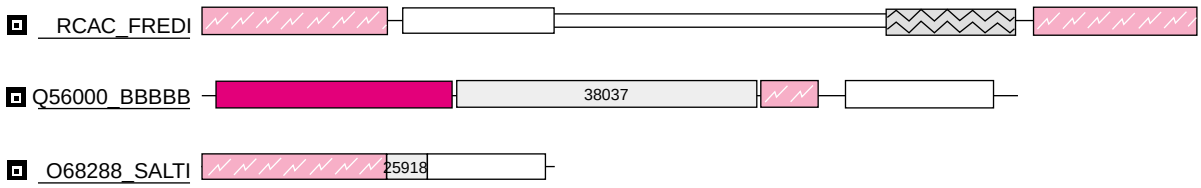
Dans le meilleur des cas, la comparaison entre une nouvelle séquence et une banque montre une similarité claire avec une seule protéine sur toute sa longueur. Dans le pire des cas, il n'y aura aucun résultat significatif. Le plus souvent, on obtient une liste de séquences qui ont des ressemblances partielles avec la séquence requête et qui, pour la plupart, n'ont pas de fonction bien définie, ou bien des fonctions contradictoires. Une grande partie de cette confusion est due à la modularité des protéines : les protéines sont généralement faites d'un assemblage de modules structurellement et fonctionnellement indépendants. Une protéine multidomaines a donc plus d'une fonction et peut appartenir à plusieurs classes fonctionnelles. Des banques de motifs protéiques ou de profils de domaines, comme PROSITE, permettent de déceler la présence d'acides aminés caractéristiques d'une fonction, comme un site catalytique, ou d'un domaine de fonction connue. D'autres banques de domaines permettent de faire le même genre d'analyse, comme Pfam ou ProDom (cf figure 2) : ces banques sont moins fiables mais plus exhaustives.

Figure 1. *Alignement de 5 séquences de cytochrome C - Une séquence consensus donne pour chaque position le résidu, ou la classe de résidus, le plus présent (les symboles représentent l'un des acides aminés voisins). Les positions conservées à plus de 90 % sont en rouge, et à plus de 50% en rose.*

```
            1                                                                   70
   CCPC50   QDGDAAKGEK  EFN-KCKACH  MIQAPDGTDI  I-KGGKTGPN  LYGVVGRKIA  SEEGFK-YGE  GILEVAEKNP
   CCRF2C     GDAAKGEK  EFN-KCKTCH  SIIAPDGTEI  V-KGAKTGPN  LYGVVGRTAG  TYPEFK-YKD  SIVALGASG-
   CCRF2S   QEGDPEAGAK  AFN-QCQTCH  VIVDDSGTTI  AGRNAKTGPN  LYGVVGRTAG  TQADFKGYGE  GMKEAGAKG-
   CCQF2R     EGDAAAGEK  VSK-KCLACH  TFDQGGAN--  -----KVGPN  LFGVFENTAA  HKDNYA-YSE  SYTEMKAKG-
   CCQF2P     AGDAAVGEK  IAKAKCTACH  DLNKGGPI--  -----KVGPP  LFGVFGRTTG  TFAGYS-YSP  GYTVMGQKG-
 Consensus  ..GDaa.GeK  .fn.kC.aCH  .i....gt.i  .....KtGPn  L%GVvgrtag  t...%k.Y.e  g..e.gakg.

            71                                                          134
   CCPC50   DLTWTEADLI  EYVTDPKPWL  VKMTDDKGAK  TKMTFKMGKN  QA--DVVAFL  AQNSPDAGGD  GEAA
   CCRF2C   -FAWTEEDIA  TYVKDPGAFL  KEKLDDKKAK  TGMAFKLAKG  GE--DVAAYL  ASVVK
   CCRF2S   -LAWDEEHFV  QYVQDPTKFL  KEYTGDAKAK  GKMTFKLKKE  ADAHNIWAYL  QQVAVRP
   CCQF2R   -LTWTEANLA  AYVKNPKAFV  LEKSGDPKAK  SKMTFKLTKD  DEIENVIAYL  KTLK
   CCQF2P   -HTWDDNALK  AYLLDPKGYV  QAKSGDPKAN  SKMIFRLEKD  DDVANVIAYL  HTMK
 Consensus  .1tWt#..1.  .Yv.#Pk.fl  .ek.gD.kAk  .kMtFk$.K.  ....#!.A%L  ..........  ....
```

Figure 2. *Décomposition en domaines de trois protéines de régulation (système bactérien à deux composants) selon la banque ProDom (version 99.2).*



6 / Prédiction de structure

6.1 / Structure de protéine

Déterminer la structure tridimensionnelle d'une protéine peut être une étape importante dans la recherche de sa fonction. Cependant, il est actuellement impossible de déterminer une structure 3D à partir de la seule donnée de la séquence. Les meilleures méthodes actuelles s'appuient sur la similarité de séquences avec des protéines de structures connues (modélisation par homologie).

En revanche, à partir de la séquence seule, on peut prédire certains éléments de structure secondaire, les zones trans-membranaires et les peptides-signaux. Cependant, là encore, les programmes les plus efficaces utilisent un alignement multiple de séquences homologues plutôt qu'une seule séquence.

La cristallographie par rayon X et la RMN sont les méthodes expérimentales qui permettent de définir la structure 3D des protéines. Elles produisent une quantité telle de données qu'elles sont dépendantes d'outils informatiques très puissants pour l'interprétation de ces données. La modélisation moléculaire nécessite des stations de travail informatique dédiées, avec des outils puissants de manipulation d'image 3D.

Les progrès de la cristallographie et de la spectroscopie font croître rapidement les banques de structures de protéines. Des ressemblances de structure (repliement des protéines) sont visibles alors qu'aucune similarité de séquence n'était détectée. Fondée sur ce fait, de nouveaux algorithmes (threading) proposent de comparer une séquence avec une banque de structures connues, en essayant « d'enfiler » la nouvelle séquence dans une structure et en en déterminant la faisabilité. Ainsi, Madej *et al* (1995) ont pu suggérer que la leptine, produit du gène « obese » (ob) chez la souris, avait une structure 3D similaire à celle de l'interleukine 6, deux ans avant que cela soit démontré expérimentalement.

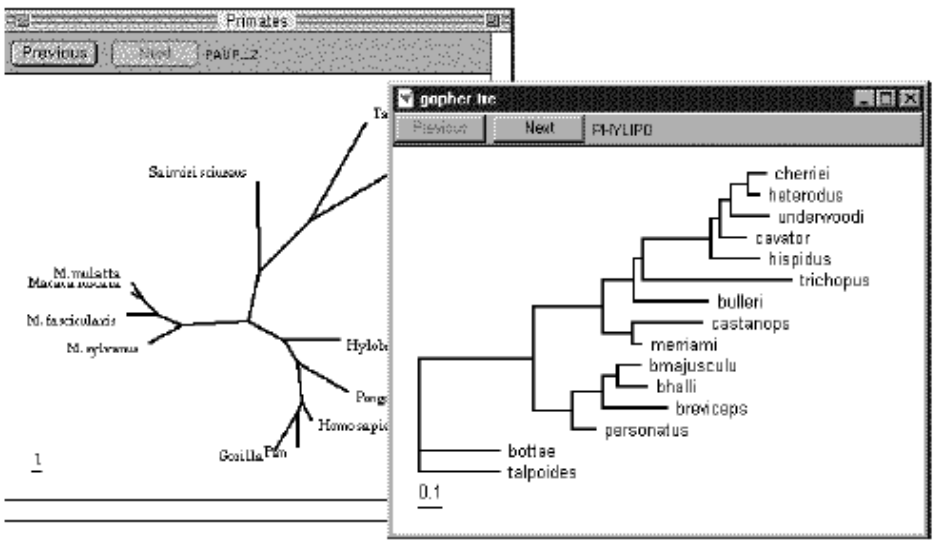
6.2 / Structure d'un ARN

De la même manière, la fonction d'un ARN dépend de sa structure ; cependant les règles de repliement de cette molécule ne sont pas celles des protéines et les programmes utilisés sont différents. La structure secondaire d'un ARN peut être prédite par un modèle thermodynamique ou par comparaison avec des ARN de structures connues. La structure 3D des ARN n'est connue que pour quelques molécules comme l'ARNt ou certains introns. L'expertise humaine est encore largement nécessaire mais les bioinformaticiens ont développé de nombreux outils qui permettent d'avoir un environnement très confortable pour manipuler ces structures *in silico.*

7 / Phylogénie

Les algorithmes de comparaison de séquences recherchent un alignement qui minimise le nombre de changements nécessaires pour passer d'une séquence à l'autre. Cette distance est supposée simuler les mutations naturelles et un alignement multiple peut donc servir à reconstruire un arbre phylogénétique des relations entre séquences et, en principe, entre les espèces d'où proviennent ces séquences (figure 3).

Figure 3. *Exemple de visualisation d'arbres par TreeView. Les arbres sont calculés avec les programmes les plus répandus, Phylip et Paup.*



Cependant, les pièges sont nombreux : l'alignement peut être incorrect, le taux de mutation peut être variable selon les positions et les séquences évoluent à des vitesses différentes selon les espèces. Pour éviter ces écueils, on peut utiliser des méthodes sophistiquées et supprimer les sites pour lesquels l'alignement est peu fiable ; on perd en rapidité ce qu'on gagne en fiabilité. Des méthodes dites de « bootstrap » permettent d'estimer la signification statistique des résultats.

8 / Les questions liées aux génomes complets

8.1 / Annotation systématique

Lorsque les séquences sortent des ateliers de séquençage, elles sont totalement anonymes et le génome entier ne sera utilisable qu'après le travail des annotateurs. Ceux-ci vont utiliser différents programmes de recherche de gènes par homologie de séquence ou de structure, de prédictions de gènes ou d'autres séquences fonctionnelles, détecter des erreurs de séquençage conduisant à des changements de cadre de lecture… Tout ce travail est grandement facilité par un environnement de travail dédié à l'annotation. Même s'il y a maintenant plus de 20 génomes entièrement séquencés, il n'y a toujours pas un environnement standard utilisable par une nouvelle équipe. Il faut allier efficacité, simplicité et possibilité d'évolution : c'est un vrai défi pour le bioinformaticien.

8.2 / La génomique fonctionnelle

Cette nouvelle approche va permettre de mieux comprendre les réseaux métaboliques auxquels participent des gènes spécifiques et, on l'espère, réduire les lacunes entre séquence et fonction. Les technologies des filtres à haute densité et des puces à ADN, méthodes à grande échelle et haut débit, nécessitent un traitement de l'information et un système d'analyse qui soit à la hauteur. Les logiciels et les systèmes de bases de données développés pour prévoir des expériences de ce type commencent tout juste à apparaître. Les outils que peuvent développer les bioinformaticiens dans ce domaine sont nombreux : définir des procédures statistiques pour établir la justesse et la reproductibilité de données d'expression d'un gène, aider à définir une expérience en spécifiant la redondance et les standards internes nécessaires à une quantification fiable, construire des banques de données d'expression des gènes, développer des algorithmes de classification des gènes en différentes classes de régulations, identifier les réseaux métaboliques auxquels participent les gènes étudiés, etc ; ce domaine est en plein développement. On peut imaginer qu'un jour, les logiciels permettront de pré-interpréter les données en utilisant des bases de connaissances, proposer au biologiste différentes hypothèses ou explications des résultats et aider à définir de nouvelles expériences.

8.3 / La génomique comparée

La connaissance de génomes entiers permet d'approfondir les questions de phylogénie. En particulier, il devient possible de distinguer orthologie (évolution par spéciation) et paralogie (évolution par duplication) des gènes, de comparer l'évolution d'une protéine ou d'un domaine protéique et l'évolution des espèces, de comparer l'organisation des gènes et unités fonctionnelles sur des chromosomes entiers. Les algorithmes de construction d'arbre et l'évaluation statistique de leurs résultats devront progresser pour répondre correctement à ces questions.

Cet article s'inspire du cours, présenté sur Internet, Using Computers in Molecular Biology, NYU Medical Center Course G16.2604.

Références

Andrade M.A., Sander C., 1997. Bioinformatics: from genome data to biological knowledge. Current Opinion in Biotechnology, 8, 675-683.

Le Deambulum, Exploration thématique des BIO-NETosphères : Biologie moléculaire - bioinformatique - Médecine – Biologie, De la Biomolécule à la Biosphère, http://www.infobiogen.fr/services/deambulum/fr/

Madej T., Boguski M.S., Bryant S.H., 1995. Threading analysis suggests that the obese gene product may be a helical cytokine. FEBS Letter, 373, 13-18.

Trends guide to Bioinformatics, Trends Supplement 1998, Elsevier Science.

Using Computers in Molecular Biology, NYU Medical Center Course G16.2604, http://mcrcr0.med.nyu.edu/rcr/course