TE Bio-Informatique

BOUKADIDA Jawer DENIS Julien

juin 2004

Table des matières

Introduction

Ces dernières années, la recherche en biologie et tout particulierèment en génétique a connu un formidable essor et continue sur sa lancée. Les avancées réalisées sont considérables. Cependant si à l'époque de Gregor Mendel[1], il suffisait de faire certains croisements entre divers espèces de pois comestibles pour faire de grandes avancées dans le domaine de la génétique, de nos jours les chercheurs utilisent d'autres techniques : techniques qui demandent de traiter une très importante somme d'informations que nous ne pouvons traiter sans l'aide de l'informatique.
A tel point, qu'il y a un peu plus d'une dizaine d'années, une nouvelle discipline a été créée : la bio-informatique. Située au carrefour de la biologie, des mathématiques, des statistiques et de l'informatique, elle consiste à utiliser les possibilités offertes par l'informatique afin d'acquérir, traiter, organiser et interpréter l'information concernant la vie comme l'indique son préfixe «bio-» qui a pour racine «$\beta\iota o\varsigma$» signifiant «la vie» en grec ancien.
Pour la suite de ce travail, nous nous limiterons à l'étude de la bio-informaqtiue du point de vue de l'information génétique (séquences) et structurale (repliement 3-D).

Nous allons tout d'abord vous expliquer les raisons pour lesquelles la bio-informatique est devenue aussi indispensable à la génétique et quel est son rôle.

Nous vous montrerons ensuite quels domaines de l'informatique sont plus particulièrement utilisés et de quelles manières.

Enfin, nous vous présenterons les principales retombées engendrées par la bio-informatique ainsi que toutes celles à venir et les progrès qu'il reste à faire dans ce domaine.

[1]moine et botaniste autrichien (1822-1884), communément reconnu comme le père fondateur de la génétique.

1 Qu'est-ce que la bio-informatique ?

1.1 Historique

1.1.1 Les origines

Même si certains chercheurs travaillaient en «biomathématiques» dès 1965 lorsqu'il fallait étudier des séquences afin de rendre compte du degré de parenté de certaines molécules, ce n'est que dans les années 1980 que la bio-informatique est née.
Le terme bio-informatique lui-même n'est apparu dans le langage scientifique qu'au début des années 1990 [6].
Cette naissance a concordé avec le début des premiers séquençages d'ADN[2]. Grossièrement, ce qui en résulte sont de longues chaines constituées uniquement de quatre caractères : A, C, G ou T[3]. Ces données se sont très vite mises à croitre et il a fallu trouver un moyen de les stocker.
Les biologistes ont donc eu l'idée de faire appel à l'informatique et ses bases de données. Les premières bases de données de séquences ont donc été créées par des informaticiens qui se sont aussi attelés à d'autres tâches telles que le fait de regrouper, analyser et trier ces données. Les données grandissantes et les méthodes nécessitant d'être améliorées, ont petit à petit entrainé la spécialisation d'informaticiens ou de biologistes dans la bio-informatique qui est devenue une dicipline à part entière et une branche de la biologie.

1.1.2 Evolution de la Bio-informatique

L'évolution de la bio-informatique s'est faite à l'image de l'évolution de l'informatique, qui est de nos jours nécessaire au développement de domaines tels que la génétique.
Sans l'informatique, il ne serait plus possible de stocker et d'analyser toutes les données récoltées.
Ainsi, plus la capacité de stockage des ordinateurs et les méthodes de séquençage progressèrent, plus le nombre de données stockées grandissait, et plus les processeurs augmentèrent leur capacité et les méthodes d'analyse gagnèrent en efficacité, plus le nombre de données pouvant être analysées s'accroissait.

De la même façon, la formation des bio-informaticiens évolua, alors qu'au début ce furent des informaticiens ou biologistes qui se convertirent en apprenant sur le tard les connaissances qui leur manquaient, petit à petit il s'est construit des structures au sein des facultés ou autres organismes permettant de former des scientifiques ayant une double compétence et destinés pour la plupart à devenir bio-informaticiens.

[2] Acide désoxyribonucléique : Support biochimique de l'information génétique chez tous les êtres vivants
[3] Adénine, Cytosine, Guanine et Thymine

1.1.3 Principaux acteurs

Les trois principaux acteurs dans le domaine de la bio-informatique sont les Etats-Unis, le Japon et le Royaume-uni. Pour prendre un exemple, en 2001, on pouvait compter dans le monde près de 400 entrerpises spécialisées en bio-informatique, et elles se situaient presque toutes dans ces trois pays. [11]
Les autres pays :

- ***Le Canada*** : on peut citer qu'en 2001, le canada a inauguré son premier programe d'étude sur la bio-informatique en premier cycle du pays. Il a aussi créé les organismes Génome Québec et Génome Canada, dotés d'un budget de 200 M$ sur cinq ans. Ce pays dispose aussi de chercheurs de renoms tels que Franz Lang, Steven Michnick, Nadia El-Mabrouk, François Major, ... [8]

- ***La France*** : qui malgré son retard, dû entre autre au manque de personel qualifié, ayant la double spécialisation requise, est en train de devenir un acteur majeur dans ce domaine.
  Ce revirement est le fruit notamment de l'apparition de plusieurs génépoles telles que celles de Marseille, Lille, Toulouse ou Clermont-Ferrand, ou encore avec la création de nombreux DESS et DEA de bio-informatique (tels qu'à Lille, Clermont-Ferrand, Monptellier ou encore Toulouse).

1.2 Stocker les données

1.2.1 Pourquoi et comment ?

Permettre de stocker des données afin de pouvoir y accèder aisément dans le futur est un des buts premiers de la bio-informatique.

Ces données sont essentiellement le résultat du séquençage des génomes, duquel découle une énorme quantité de données qu'on ne peut conserver sur papier d'une part à cause de la place que cela prendrait d'autre part, du fait qu'il serait difficile d'y chercher et de traiter les enregistrements.

Les bases de données sont donc l'outil idéal.

1.2.2 Y accéder

Alors qu'à une certaine époque, pour pouvoir accéder à ces données, on recevait par courrier, des bandes mangnétiques puis des cd-roms, aujourd'hui avec la venue d'internet et des lignes à haut débit, le monde entier peut accèder à toutes ces informations.

Pour faciliter la recherche d'informations, des logiciels ont été créés permettant de lire, d'interroger ces bases de données et d'en extraire une ou des informations.
De plus, afin que l'accès aux diverses banques de données se fasse d'une manière assez similaire, des normes ont été définies [3].

1.2.3 Quelques exemples

De nos jours, il existe donc de très nombreuses bases de données publiques réparties dans le monde, contenant les résultats de nombreuses années de recherche en génomique. Pour information, début 2004, il existait environ 400 bases de données dans le domaine des sciences de la vie.

Voici le noms de quelques unes des plus importantes dans le domaine de la génétique :

- ***EMBL***, banque européenne qui existe depuis 1980.
- ***Genbank***, banque américaine crée en 1982.
- ***DDBJ***, située au Japon, elle fonctionne depuis 1987.

1.3 Analyser ces données

C'est une chose de pouvoir stocker une grande quantité de données, cependant, il faut encore pourvoir l'utiliser et en tirer profit.

Tout comme un fichier au format mp3 qui est une suite de 0 et de 1, la séquence d'un génome est une succession de 'A', 'C', 'G' et 'T'. Ni l'une ni l'autre n'a de signification particulière si on ne les décode pas.
Si pour le format mp3, il existe un algorithme de décryptage, ainsi que de nombreux programmes permettant de l'éxécuter, pour le génome c'est autre chose.

Une séquence de génome peut aussi être comparer à un livre écrit dans une langue inconnue, et dont les mots ne seraient pas discernables. Pour le traduire, il faudrait alors commencer par identifier les mots (gênes pour la séquence) et ensuite comprendre le sens de chaque mot (fonctions des gênes).

Afin de comprendre le fonctionnement des cellules et en tirer profit, il faut donc commencer par décoder les séquences de nucléotides (les 'A', 'C', 'G' et 'T') afin d'identifier les gênes et leurs fonctions biologiques.

2 Comment ça fonctionne ?

La bio-informatique est le traitement automatique de l'information biologique sous forme de données accessibles aisément et exploitables.
Comme pour toute discipline, on part d'une matière première (ici c'est les bases de données) ainsi que des connaissances primaires et on déploie un ensemble d'outils (algorithmes et techniques) en cherchant à produire d'autres données (dites données secondaires) et à construire de nouvelles connaissances.

2.1 Bases de données

2.1.1 Terminologie

Il y a deux sortes de bases de données :
- celles qui offrent des informations plutôt hétérogènes
- celles qui correspondent à des données plus homogènes d'espèces précises.

Il est fréquent de parler de "banques de données" pour les premières et de "bases de données" pour les secondes, mais cette distinction n'est pas très connue par les non biologistes. Pour éviter les équivoques nous appellerons les premières banques de données ou bases de données généralistes et les secondes spécialisées.

Face à la croissance rapide et phénoménale et à la diversité des séquences des bases généralistes, les banques spécialisées ont vu le jour pour recenser des familles de séquences réparties autour de caractéristiques biologiques précises en vue de lever les ambiguïtés laissées par les grandes banques publiques parfois confuses.

2.1.2 Création

Il existe plusieurs outils, logiciels et systèmes de gestion de bases de données.

Il y en a certains qui sont gratuits, très performants et utilisés dans de très nombreux projets de base de données et d'autres qui le sont moins. La plupart de ces logiciels sont disponibles pour les plate-formes courantes (Windows, Linux, ...).

En gros, pour créer une base de données, il nous faut disposer de :

- ***Un système de gestion de bases de données*** : il s'agit du coeur des bases de données. On compte trois types de bases de données : les bases de données relationnelles, les bases de données purement objet (ozone), et les bases de données relationnel-objet. Un quatrième type de base de données est néanmoins en train d'émerger, il s'agit de bases permettant de stocker des données au format XML de manière native.

- ***La plate-forme ou système d'exploitation***

- ***Le serveur web*** : utile si cette base de données doit être accessible au travers d'une interface web (soumission de données ou interrogation)

- ***Le serveur d'application*** : dont le role est de permettre l'extraction de données à partir de la base en fonction des requêtes effectuées par l'utilisateur, puis la construction automatique de pages HTML à partir de ces mêmes données.

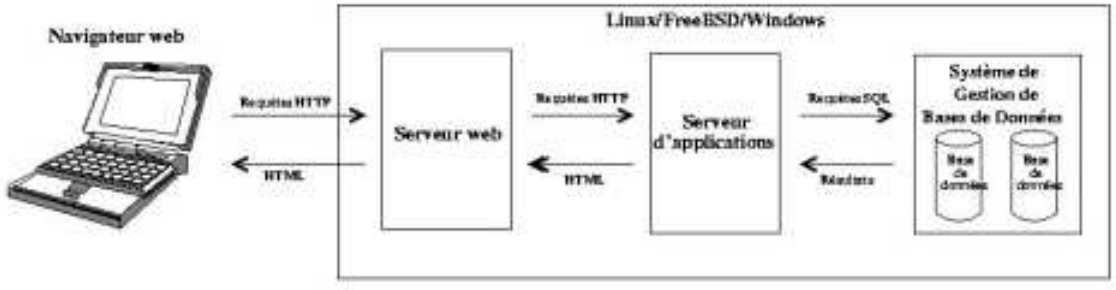


FIG. 1 – Fonctionnement d'une base de données

2.1.3 Interrogation

Afin que les bases de données soient plus facilement exploitables, et que les utilisateurs puissent en extraire les sous-ensembles de séquences qui les intéressent, deux types de logiciels leur sont généralement offerts :

- ***Des logiciels généralistes*** : Systèmes de gestion de bases de données (SGBD) qui utilisent un langage de requête standard (par ex SQL) et un format de données ne dépendant pas du type de l'information stockée.

- ***D'autres plus spécifiques à certaines bases de données seulement*** : Ce sont des systèmes d'interrogation dédiés qui sont programmés exclusivement pour la manipulation de séquences biologiques. On peut en citer le programme Stringsearch (1984) qui permet une interrogation à deux critères. D'autres permettent des interrogations multicritères plus complexes comme les logiciels ACNUC (1985) ou SRS « Sequence Retrieval System » (1993).

2.1.4 Modification et insertion de données

Il y a deux manières de modifier les données contenues dans la base :

- ***Ou bien directement par des requêtes du langage utilisé.***
  Exemple de requête en SQL :
  - *Pour insérer une donnée :*
    INSERT INTO nom_table SET nom='toto', time='2564582', message='.....'

  - *Pour modifier une donnée :*
    UPDATE nom_table SET parcours='dede ;toto' WHERE sessid='DF458SE'

- ***Ou bien en passant par un formulaire HTML***, comme celui-ci :

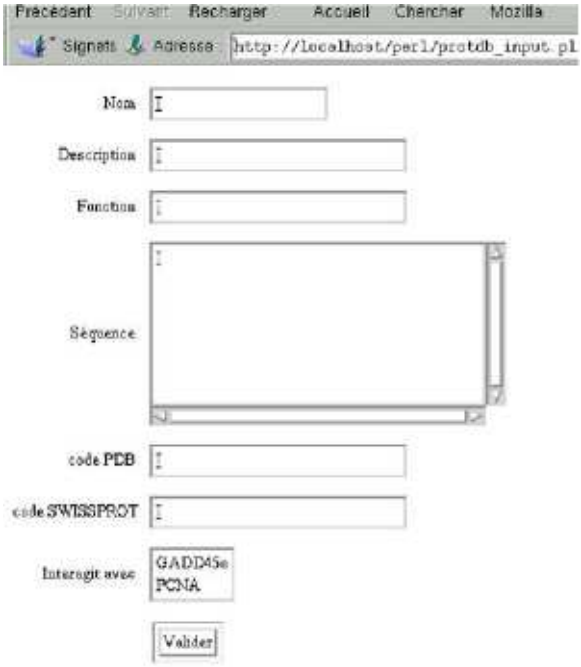


FIG. 2 – Formulaire HTML

2.2 Algorithmes

2.2.1 Intérêt de l'analyse des séquences

Actuellement il y a des milliards de séquences génétiques et protéiques.
Seulement peu d'entres elles ont des structures et des fonctions connues. Les autres de-

meurent encore ambiguës. Mais si jamais on réussi à les aligner, on peut prétendre qu'elles sont similaires par conséquent elles ont les mêmes structures et des fonctions proches. On peut comparer soit des séquences d'ADN soit des séquences de protéines personnelles (qu'on aura trouver et qu'on voudra déterminer) à des séquences dont on connaît déjà l'origine et les fonctions et qu'on trouve dans des bases de données accessibles à tous dans le but d'établir des homologies structurales et fonctionnelles entres elles.

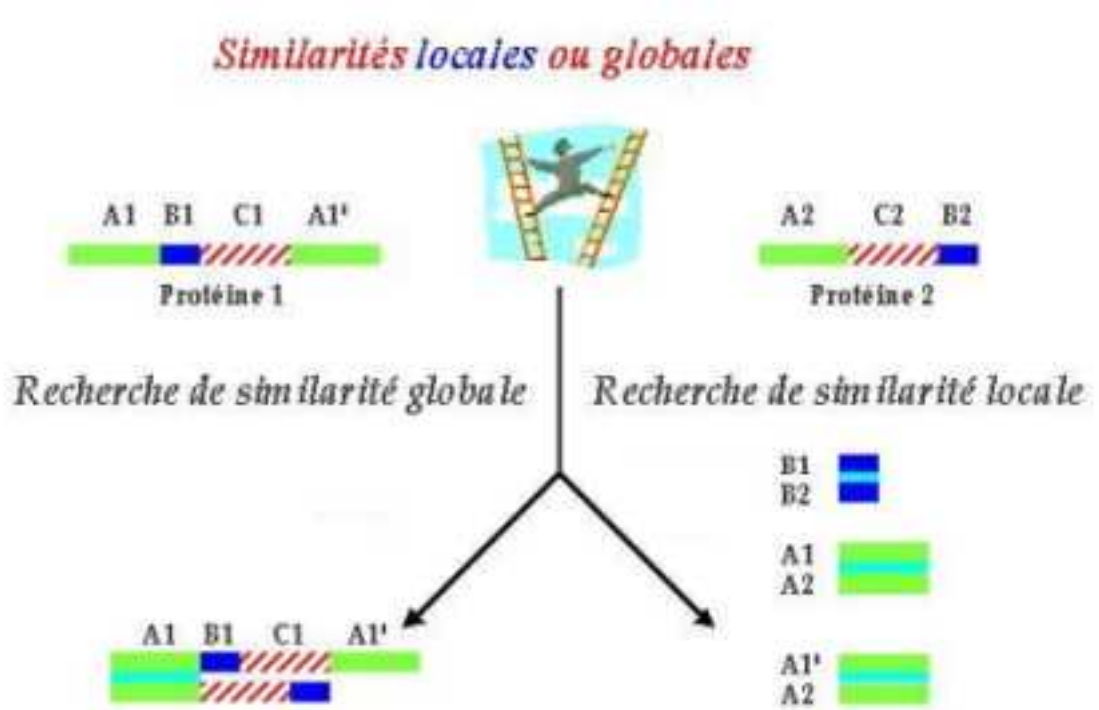
Attention : On sait qu'il peut toujours y avoir des mutations, des adaptations et des évolutions ; ce ne sont donc jamais des comparaison exactes.

2.2.2 Les différents types d'algorithmes

On distingue deux types d'algorithmes se basant sur l'alignement des séquences :

- ***Les algorithmes d'alignement globaux*** : (Needlemann & Wunsch, 1970)
  Ils sont qualifiés de globaux car ils considèrent l'ensemble des éléments des séquences. Des insertions devrant être faites dans la séquence la plus courte pour arriver à les aligner dans le cas où elles auraient des longueurs différentes (exemple de programme basé sur cette méthode : FASTA).

- ***Les algorithmes d'alignement locaux*** : (Smith & Waterman, 1981)
  On cherche à trouver les zones les plus similaires entre deux séquences (c'est à dire deux sous-séquences homologues) sans prédétermination de longueurs. Il ne porte pas sur la totalité des séquences comme le précédent mais seulement une partie de chacune d'elles. (ex : BLAST est un programme qui adopte cet algorithme)



Différence entre ces deux méthodes

Aussi doit on signaler que des fois il peut s'avérer nécessaire de faire des délétions ou des insertions de longueurs variables à des positions différentes au sein des séquences pour en optimiser la comparaison. On parle alors de gap. Une pénalité s'applique suivant que l'alignement s'est fait avec ou sans gap influençant ainsi sur le score qui intervient dans le choix du meilleur alignement. C'est pourquoi il faut introduire un minimum de gap.

Les différentes implémentations de ces algorithmes font appel à la programmation dynamique.

Etudions un exemple d'alignement (adoptant l'algorithme global) pour voir comment s'applique le principe de la programmation dynamique dans notre problème.

Prenons les deux séquences **S** = ACGCTG et **G** = CATGT et voyons comment les aligner.

On va être mener à introduire des « gap » c'est à dire un nombre (minimal) d'opérations de suppression ou d'insertion de caractères. Par exemple :
AC–GCTG
-CTAG-T

Maintenant, il nous faut interpréter ce résultat et donc trouver un critère de jugement du meilleur alignement possible.

Par exemple (c'est pour simplifier. D'autre mécanismes plus compliqués pourraient être utilisés) si :
- Une égalité vaut 2
- Une différence vaut -1
- Un gap vaut -1 (C'est pour ça qu'il faut les éviter. Avec l'algo semi-global, on n'aurait pas compté le gap en tête)

Alors l'alignement :

```
 - A C G  C T  G
 C A T G  - T  -
-1 2 1 2 -1 2 -1 ==> 2
```

Pour avoir l'alignement optimal, on procédera de la telle façon (ça montre un exemple de fonction de score) :
On notera σ(a, b) le score obtenu en alignant le caractère a avec le caractère b. Le score du gap est σ(_ , *) = σ(* , _) = -gap

Soit V un tableau (une matrice) avec par exemple S dans la première colonne et G sur la première rangée.
V(i, j) est le score optimal de l'alignement de $S_1 \ldots S_i$ avec $G_1 \ldots G_j$ avec (0 $\leq$ i $\leq$ n, 0 $\leq$ j $\leq$ m où |S|=n et |G|=m).

ATTENTION : le premier caractère de chaque chaîne est un caractère rajouté, comme par exemple '_'. Donc, le tableau V a comme taille (n+1) x (m+1).

On peut remarquer (principe de programmation dynamique) que si on veut calculer l'alignement optimal entre $S_1 \ldots S_i$ avec $G_1 \ldots G_j$ alors on sait qu'il provient :

- soit d'un alignement optimal entre $S_1 \ldots S_{i-1}$ et $G_1 \ldots G_{j-1}$ auquel on a rajouté S_i et G_j G_j.
- soit d'un alignement optimal entre $S_1 \ldots S_{i-1}$ et $G_1 \ldots G_j$ auquel on a rajouté S_i et un gap après G_j.
- soit d'un alignement optimal entre $S_1 \ldots S_i$ et $G_1 \ldots G_{j-1}$ auquel on a rajouté gap après S_i et G_j.

Il suffit alors de prendre le meilleur des trois. On aboutit aux formules :

$$V(0,0) = 0$$

$$
\begin{aligned}
V(i,0) &= \sum_{k=1}^{i} \sigma(S_k, \_) \\
V(0,j) &= \sum_{k=1}^{j} \sigma(\_, G_k) \\
V(i,j) &= max \begin{cases} V(i-1,j-1) + \sigma(S_i, G_j) \\ V(i-1,j) + \sigma(S_i, \_) \\ V(i,j-1) + \sigma(\_, G_j) \end{cases}
\end{aligned}
$$

S = ACGCTG et **G** = CATGT
égalité = 2, *diff* = -1, *gap* = -1

		C	**A**	**T**	**G**	**T**
	0	**-1**	-2	-3	-4	-5
A	-1	-1	**1**	0	-1	-2
C	-2	1	0	**0**	-1	-2
G	-3	0	0	-1	**2**	1
C	-4	-1	-1	-1	**1**	1
T	-5	-2	-2	1	0	**3**
G	-6	-3	-3	0	**3**	**2**

La complexité de cet algorithme est O(m x n) en temps et en espace. Il y en a d'autres variantes avec des performances meilleures notamment en espace.
Pour ce qui est des algorithmes d'alignement local et semi-global, on peut reprendre le précédent avec quelques modifications mineures.

En vue d'assimiler, en pratique, la différence entre ces types d'algorithmes, on va se référer à la matrice ci dessus :

- Pour ce qui est de l'alignement global optimal, on veut le meilleurs score qui est dans la dernière rangée ou la dernière colonne (meilleur alignement des deux séquences en entier).
- En ce qui concerne l'alignement local optimal, on veut le meilleur score n'importe où dans la matrice (meilleurs alignement des segments « sous-séquences » sans se fier au reste des deux chaînes).

2.3 Analyse des structures 3D

2.3.1 Distinction

Les séquences sont des structures primaires c'est à dire des mots écrits avec un alphabet de x lettres.

Les structures tertiaires sont des objets 3D.

En parlant de séquences, on a dit que ça concernait à la foi l'ADN et les protéines. Mais quand on parle de structures-3D, c'est la protéine qui est en question.

La structure-3D d'une protéine permet, plus que la séquence, la détermination de sa fonction. En effet, elle va nous renseigner sur les sites actifs et les signatures de la protéine caractéristiques de sa fonction potentielle.
L'analyse de la séquence d'une protéine peut permettre d'émettre des hypothèses fonctionnelles et structurales afin de mieux comprendre le mode d'action de cette protéine. Cependant, seule la connaissance de la structure tridimensionnelle de la protéine peut permettre de comprendre de manière fine son mode d'action.

De plus la prédiction des structure-3D présente un intérêt majeur pour les biologistes car elles sont mieux conservées que les séquences au cours de l'évolution et permettent même d'élucider le mécanisme de certaines mutations et leurs sens.
Elles sont donc très révélatrices.

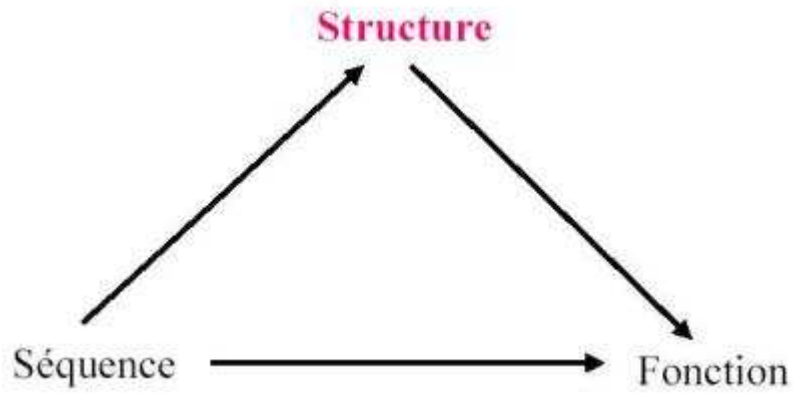


FIG. 3 – Relation entre la séquence, la structure et la fonction [2]

2.3.2 De la structure à la fonction

Afin de prédire les caractéristiques structurales 3D des nouvelles protéines, il est nécessaire d'avoir recours à différentes techniques et outils et il est important que ces derniers soient précis, simples et fiables.

Trois voies différentes sont actuellement suivies pour ces fins :

- ***la modélisation par homologie comparative***, où la structure d'une protéine similaire en séquence (les séquences vues au paragraphe précédent) est utilisée comme empreinte pour définir la structure de la protéine inconnue.

- ***la reconnaissance de repliement***, où une séquence est testée du point de vue de sa compatibilité avec une librairie de repliements. Ce sont les méthodes d'alignement séquences-structures tridimensionnelles.

- ***la prédiction structurelle***, où la structure est directement déduite de la séquence, à partir de règles empiriques et de modèles plus ou moins simplifiés des protéines.

  (Ces méthodes viennent se substituer des techniques physiques expérimentales et qui sont en général lourdes, coûteuses et souvent limitées).

Cette phase de reconnaissance et d'identification doit être complétée par des simulations numériques et des techniques d'ingénierie pour analyser les propriétés physico-chimiques et biologiques de la protéine.

Le fruit de ce travail viendra enrichir, au fur et à mesure, les bases de données de structures protéiques dites banques PDB[4].
C'est pourquoi on vise, en permanence, à produire un ensemble de serveurs et logiciels intégrant des procédures spécialisées pour l'archivage, la communication et la validation de résultats émanant des analyses de structures de protéines.

Ces développements ont donné naissances à divers programmes et logiciels dans le domaine dont on cite [12] :

- ***EVAL123D*** : Ce serveur permet une évaluation synthétique de la validité structurale du modèle tridimensionnel d'une protéine. Outre les outils standards PROSAII et VERIFY3D, EVAL123D intègre les outils locaux EVAL23D et EVTREE qui mesurent la compatibilité entre chaque acide aminé de la protéine et son environnement tridimensionnel local.

- ***CINDY*** : C'est un logiciel de cartographie RMN [5]. Il permet d'analyser les spectres RMN sur une station graphique SGI. Il offre différentes possibilités dont : la superposition de spectres 1D, 2D, tranches 3D, des annotations, l'affichage des systèmes de spins, diverses aides à l'attribution... L'interface graphique permet à l'utilisateur de manipuler et de modifier tous les objets présents sur l'écran de manière instantanée.

[4]Brookhaven Protein Databank

[5]C'est une méthode expérimentale basée sur l'interaction des moments magnétiques nucléaires des atomes

3 Quel avenir ?

Aucun de nous ne pourrait imaginer tout ce que la bio-informatique a pu offrir à la communauté scientifique (aussi bien académique qu'industrielle) dès l'aube de son apparition. Cette discipline en pleine émergence a, en seulement quelques années, bouleversé le monde de la biologie moderne et est venue s'y imposer avec de nouvelles idées et méthodes révolutionnant ainsi tout les domaines qui y touchent de près ou de loin. Au début, son défi majeur n'était que l'annotation des gènes et la prédiction de leur fonction. Mais très vite ça s'est étendue et son champ d'application était allé croissant sous l'impulsion de nouvelles technologies.

Son développement exigeant une mise à jour continuelle des outils informatique nécessaires aux différentes analyses a contribué d'une certaine manière à la révolution des concepts et des méthodes de l'informatique traditionnelle .

Bien qu'à la base, elle était censée concerner la recherche en biologie fondamentale, on la trouve aujourd'hui associée à tous les secteurs où la biologie participe au développement d'une activité économique.

3.1 Les retombées

Voici quelques exemples de retombées :

- ***La santé*** : (avec le traitement des maladies génétiques et des maladies multifactorielles, comme le diabète ou l'obésité) : la sélection de gènes candidats, assistée par l'exploitation intelligente des informations présentes dans les banques de données publiques ou propriétaires, peut accélérer l'identification des gènes responsables de pathologies complexes et par suite le développement de thérapies innovantes et adaptées aux individus [1].

- ***L'industrie pharmaceutique*** : (avec la découverte de nouvelles cibles thérapeutiques et la conception rationnelle de médicaments) : connaissant la structure exacte des protéines, les concepteurs de médicaments élaborent des molécules qui s'adaptent aux sites protéiques et les activeront ou les empêcheront d'agir [13].

- ***L'industrie agro-alimentaire*** : (avec la maîtrise du risque alimentaire ou l'amélioration des espèces) : les connaissances sur la structure tri-dimensionnelle de certaines protéines permettent de concevoir des ligands[6] et autres molécules capables d'interagir avec ces macro-molécules et par exemple d'inhiber une fonction nocive ou réduire sa fréquence d'apparition ou améliorer le rendement ou la qualité... [1]

[6]Molécules capables de s'attacher à un récepteur cellulaire.

- ***L'écologie*** : (avec la lutte contre la pollution des eaux et la pollution chimique ou biologique), les cosmétiques, préservation de la biodiversité, nouvelles sources d'énergie, ...

- ***Biotechnix 3d*** : créée en 1997 sur la technopole de Sophia-Antipolis par le docteur Alaa Khashoggi et développé par la division bio-informatique de Gentech initialement pour les besoins internes de ses chercheurs en biotechnologie végétale. C'est un logiciel regroupant les outils nécessaires à l'analyse de séquences d'ADN et de protéines. Il a permis à Gentech de poursuivre des recherches sur la résistance des plantes à des pathogènes en utilisant une approche originale qui lui permet d'obtenir des plantes résistantes sans introduire de gène étranger dans leur génome. Cette stratégie innovante a fait l'objet d'un dépôt de brevets.

  Le logiciel Biotechnix 3d mis au point par les bio-informaticiens de la société niçoise est aujourd'hui accessible à l'ensemble des chercheurs en biologie moléculaire du monde entier leur permettant de raccourcir considérablement le chemin du projet à la découverte". En effet, Biotechnix 3d est conçu comme une «boîte à outils» bio-informatique et présente de nombreux avantages par rapport aux logiciels concurrents parmi lesquels :

  - l'accès à la gestion de projets qui permet une structuration claire des fichiers et documents liés aux projets de recherche,
  - la visualisation graphique 3-D des structures protéiques grâce à des lunettes à cristaux liquides infrarouges,
  - un accès rapide et facilité aux bases de données et aux ressources d'analyses disponibles sur Internet,
  - la synchronisation des séquences, structures et résultats d'analyses qui permet de visualiser en temps réel et sur chaque représentation les modifications effectuées sur une zone particulière.

  Grâce à cet outil les scientifiques de Gentech travaillent sur les sites où se déroulent pendant l'infection les interactions entre les protéines de la plante et celles du virus. Après modification de ces sites, la propagation du virus est bloquée et la plante ne présente plus les symptômes de la maladie.

  Les recherches de Gentech sont au coeur d'un véritable débat de société avec à la clé peut-être la création de plantes améliorées de seconde génération, mieux acceptées par les consommateurs et plus respectueuses de l'environnement [4].

- ***Les biopuces*** : le concept de biopuce ou puce à ADN remonte au début des années 1990. Il repose sur une technologie pluridisciplinaire intégrant la micro-électronique, la chimie des acides nucléiques, l'analyse d'images et comme brique de base la bio-

informatique. Le principe de fonctionnement de ces puces repose sur le phénomène de l'appariement par complémentarité des bases de deux séquences d'ADN. Ceci permet d'identifier une séquence de nucléotides, c'est-à-dire l'enchaînement des bases d'un fragment d'ADN en mettant ce dernier en présence d'autres brins d'ADN, dont la séquence est connue. Par exemple, face à des brins d'ADN synthétiques représentatifs d'une maladie, les brins extraits de l'ADN du patient vont s'apparenter si le malade est porteur de l'affection recherchée.

Cette technologie permet de mesurer l'expression de plus de 40 000 gènes dans une cellule ou un tissu et de comparer leur expression entre différentes conditions (normal/pathologique, traité/non traité, cinétiques temporelles, ...) [9].

- ***La guerre bactériologique ou chimique*** : en déterminant par avance les modifications du fonctionnement génétique des cellules immunitaires occasionnées par des agents toxiques, on peut identifier très rapidement et à un coût bas les produits chimiques (mercure, dioxine...) ou bactériologiques (bacille du charbon ou de la diphtérie...) disséminés par un éventuel agresseur [9].

- ***Les molécules darwiniennes*** : faisant appel aux principes de la sélection naturelle, un logiciel engendre des molécules-modèles inédites, futures candidates au statut de médicament.

  Plusieurs modèles chimiques permettent déjà de trouver informatiquement de nouvelles molécules actives. Le dernier logiciel en date propose d'appliquer la sélection naturelle darwinienne sur des modèles de molécules. Il part d'une molécule fournie par un laboratoire, connue pour son action sur certaines maladies. Il en construit ensuite jusqu'à trente « parents », des molécules élaborées automatiquement selon des critères géométriques, physico-chimiques ou autres.

  A chaque molécule est associé un « Chromosome » fictif représentant ses caractéristiques. L'algorithme déclenche alors un processus de reproduction sur cette population de chromosomes, en autorisant les mutations (changement ponctuel et aléatoire) et les échange d'une séquence entre deux chromosomes. À chaque génération, le logiciel choisit les meilleurs reproducteurs, par exemple les molécules ayant la configuration de plus basse énergie.

  En jouant sur les paramètres de sélection, on peut aussi modéliser la fixation de ligands sur une protéine, ou encore prédire la forme exacte d'une molécule d'après sa formule chimique, en un temps record [7].

– ***Le domaine anténatal*** : Le dépistage anténatal[7] se fait par analyse des chromosomes placentaires ou du liquide dans lequel baigne l'embryon et donc indirectement de l'embryon.
Cette analyse va permettre de repérer très tôt toute anomalie grave pouvant toucher le futur bébé. En effet, il est aujourd'hui possible d'étudier les structures des chromosomes en un temps réduit et avec un coût faible et de prédire les éventuels problèmes et malformations pouvant en découler.

En cas d'anomalie, il importe aux parents, avec l'aide du médecin, de prendre la décision de poursuive ou d'interrompre la grossesse [5].

3.2 Les prochains défis

– ***L'ordinateur biologique*** : les premiers prototypes à base d'ADN ont été réalisés à l'institut Weizmann en Isrëal.
Fabriqué à l'aide d'enzymes qui réagissent au contact de l'ADN, il possède une très grande vitesse de calcul. Il a de nombreux intérêts : rapide, fiable et il consomme très peu d'énergie. Cependant, il possède encore de grosses lacunes qui l'empêchent de devenir envisageable. En effet, étant constitué de matériaux vivants, il craint les changements de températures et la déshydratation, de plus il n'est pas programmable.

D'autres chercheurs tentent de créer une machine douée d'auto assemblage à base de protéines virales et de molécules d' ADN : elle serait capable, grâce à un code génétique créé par l'homme, de créer elle même ses cellules-composants [10].

Ceci donne un petit aperçu de ce qui peut nous attendre d'ici quelques années : finit nos bons vieux composants électroniques, bienvenue à l'ADN et autres protéines ...

– ***Les futur vaccins*** : Grâce aux nouvelles connaissances apportées par la bio-informatique, plusieurs laboratoires actifs dans la recherche en immunologie et biotechnologies ont emprunté un grand axe de recherche tournant autour des vaccinations par ADN nu[8].
Cela consiste en l'injection directe des gènes codant la séquence d'un antigène d'un virus ou d'une bactérie dans cellules de l'organisme. Des essais ont montré l'efficacité de ce procédé bien qu'il soit encore limité. Cette nouvelle voie de vaccination est pleine d'avenir car elle évite les problèmes liés à l'utilisation de micro-organismes mais aussi les problèmes de la purification des vaccins atténués.

Par exemple, un vaccin à base d'ADN du VIH est en cours de développement par le Centre de recherche sur les vaccins de l'Institut national sur les allergies et maladies infectieuses (Maryland, Etats-Unis).

[7] avant la naissance
[8] contient la séquence thérapeutique

- ***La correction des anomalies chromosomiques*** :
  C'est le rêve de tout généticien. Cette correction consiste au remplacement d'un chromosome anormal ou l'ajout d'un fragment de chromosome fonctionnant normalement.

  Actuellement et suite à la révolution génomique due au développement de la bioinformatique, il est possible de dépister certaines maladie génétiques et de déterminer les facteurs se cachant derrière en repérant les anomalies dans les structures chromosomiques mais le fait d'y remédier reste encore théorique.

  Avec une meilleure connaissance du génome humain on espère pouvoir y parvenir

- ***La connaissance du génome humain*** : Malgré la réalisation du séquençage complet du génome humain dont le résultat est conservé dans de très grandes banques données, il nous reste encore à découvrir la fonction de chacun des gênes du corps humain, ce qui nous permettrait alors de comprendre le fonctionnement de notre corps et ainsi par exemple de guérir de nombreuses maladies, ...

Conclusion

La bio-informatique est née il y a presque une vingtaine d'années pour subvenir aux biologistes qui avaient besoin d'un support permettant de stocker un nombre de données ne cessant d'augmenter, et d'un outil y facilitant l'accès et en simplifiant le traitement. Certains pays ont un rôle prépondérant dans le dévelopement de cette discipline, alors que d'autres tentent de rattraper le retard qu'ils ont accumulé.

La raison pour laquelle des bio-informaticiens ont été formés, est qu'il était nécessaire de disposer d'individus ayant une double compétence : d'une part biologiste afin d'avoir les connaissances nécessaires pour comprendre les problèmes soulevés par la génétique moderne et autres branches de la biologie et d'autre part informaticien afin de pouvoir créer des bases de donées, mettre au point des logiciels et développer des algorithmes permettant de résoudre les problèmes précédents.

Nous pouvons remarquer que la biologie doit beaucoup à la bio-informatique et qu'aujoud'hui elle lui est indispensable pour continuer son évolution.

Cependant, nous en sommes encore aux prémices de cette discpline qui vient d'émerger, et les retombées commencent seulement à se faire ressentir. Elles sont très prometteuses et ceci explique sans doute les gigantesques sommes d'argent investies par différents pays ainsi que par de grands groupes industriels qui montre que la bio-informatique a encore de beaux jours devant elle.

Enfin même si dans la bio-informatique, c'est plutôt l'informatique qui sert la biologie, il faut noter que les applications résultantes de cette collaboration contribuent aussi au développement de l'informatique avec par exemple les ordinateurs biologiques qui seront peut-être ceux de demain au même titre que les ordinateurs quantiques ....

Références

[1] Bio informatique en lorraine. http ://bioinfo.loria.fr/Bioinfo-Presentation.

[2] Bioinformatique et applications à la génomique. http ://prst-il.loria.fr/conseil-scient/PPT/bioinfo-10-03-p.pdf.

[3] Données et séquençage. http ://www.infobiogen.fr/services/deambulum/fr/bioinfo2.html.

[4] Gentech lance biotechnix 3d. http ://www.gazettelabo.fr/2002archives/prives/2001/61Gentech.htm.

[5] Génome humain et conséquences sur la médecine. http ://www.medecine-et-sante.com/maladiesexplications/genome.html.

[6] Introduction à la bio-informatique. http ://www.infobiogen.fr/services/deambulum/fr/bioinfo1.html.

[7] La chimie combinatoire. http ://www.senat.fr/rap/o99-020/o99-0203.html.

[8] La rentrée des 30 premiers étudiants en bio-informatique. http ://www.bcm.umontreal.ca/Rayonnement/PDF/Burger01.pdf.

[9] Les biopuces. http ://www.senat.fr/rap/o99-020/o99-0202.html.

[10] L'ordinateur biologique. http ://perso.wanadoo.fr/neheven/bio.html.

[11] Naissance de la bio-informatique. http ://www4.ccip.fr/crocis/pdf/enjeux31.pdf.

[12] Plate-forme de bioinformatique. http ://www.lirmm.fr/~w3ifa/MAAS/BioInfo-Montpellier.html.

[13] Carol Ezzell. Au-delà du génome, Septembre 2000. http ://www.fil.univ-lille1.fr/FORMATIONS/BIOINFO/Bioinfo/275_052_056.pdf.