LA BIO-INFORMATIQUE

Bio-informatique*

La bio-informatique est un champs de recherche multi-disciplinaire où travaillent de concert **biologistes**, médecins, **informaticiens**, mathématitiens, physiciens et **bio-informatitiens,** dans le but de résoudre un problème posé par la **biologie.**

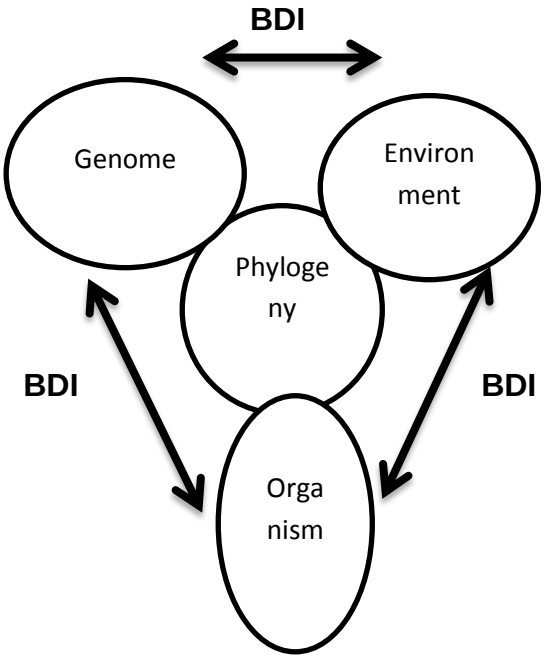
Le terme bio-informatique peut également décrire (par abus de langage) toutes les **applications informatiques** résultant de ces recherches. L'utilisation du terme bio-informatique est documentée pour la première fois en 1970 dans une application de Poulien Hogeweg et Ben Hesper (Université d'Utrecht Pays-bas), en référence à l'étude des processus d'information dans les systèmes biotiques.

Cela va de l'analyse du **génome** à la **modélisation** de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'image, l'assemblage de génomes et la construction d'arbre phylogénétiques (phylogénie).

Cette discipline constitue la « biologie en silico » par analogie avec in vitro ou in vivo.

Depuis quelques années les progrés de l'informatique (et en particulier de la bio-informatique mise au service de la biodiversité (« Biodiversity informatics » pour les anglophones) dopent la biologie évolutive en offrant aux chercheurs un acces à un nombre croissant de données sur la diversité et les variations des génes, ainsi que des génomes, des organismes et de l'environnement en général.

Tout cela peut être relié à la phylogénie (de l'étude des populations à celle de clades entiers), via de nouveaux protocoles et de réseaux dans le domaine de l'informatique de la biodiversité (**DBI** : **B**io **D**iversity **I**nformatics).



Définitions et champs d'applications

La bio-informatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique de l'information biologique. Plusieurs champs d'application ou sous-disciplines de la bio-informatique se sont constitués.

- La bio-informatique des séquences, qui traite de l'analyse de données issues de l'information génétique contenue dans la séquence de l'ADN ou dans celle des protéines qu'il code. Cette branche s'intéresse en particulier à l'identification des ressemblances entre les séquences, à l'identification des gènes ou de régions biologiquement pertinentes dans l'ADN ou dans les protéines, en se basant sur l'enchainement ou séquence de leurs composants élémentaires (nucléotides, acides aminés).
- La bio-informatique structurale, qui traite de la reconstruction, de la prédiction ou de l'analyse de la structure 3D ou du repliement des macromolécules biologiques (protéines, acides nucléiques), au moyen d'outils informatiques.
- La bio-informatique des réseaux, qui s'intéresse aux interactions entre gènes, protéines, cellules, organismes, en essayant d'analyser et de modéliser les comportements collectifs d'ensembles de briques élémentaires du vivant. Cette partie de la bio-informatique se nourrit en particulier des données issues de technologies d'analyse à haut débit comme la **protéomique** ou la **transcriptomique** pour analyser des flux génétiques ou métaboliques.
- La bio-informatique statique et la bio-informatique des populations.

Pour certains, la bio-informatique est une branche théorique de la biologie alors que pour d'autre, elle se situe clairement au carrefour des mathématiques, de l'informatique et la biologie. Il s'agit en fait d'analyser, modéliser ou prédire les informations issues de données biologiques expérimentales.

Dans un sens encore plus étendu, on peut aussi inclure sous le concept de bio-informatique le développement d'outils de traitement de l'information basés sur des systèmes biologiques comme, par exemple, l'utilisation des propriétés **combinatoires** du **code génétique** pour la conception **d'ordinateurs à ADN** permettant de résoudre des problèmes **algorithmiques complexes**.

Analyse de séquence

Depuis l'invention du **séquençage de l'ADN** par **Frederick Sanger** dans les années 1970, les progrès technologiques dans ce domaine ont été tels que le volume des séquences d'ADN disponibles a progressé de manière exponentielle, avec un temps de doublement de l'ordre de 15 à 18 mois, c'est-à-dire un peu plus rapidement que la puissance des processeurs des ordinateurs (**Loi de Moore**). Un nombre exponentiellement croissant de séquences de **génomes** ou **d'ADN complémentaires** sont disponibles, dont l'annotation (ou interprétation de leur fonction biologique) reste à effectuer.

La première difficulté consiste à organiser cette énorme masse d'informations et de la rendre disponible à l'ensemble de la communauté de chercheurs. Cela a été rendu possible grâce à différentes bases de données, accessibles en ligne. A l'échelon mondial, trois grandes institutions sont en charge de l'archivage de ces données : Le NCBI aux USA, L'EBI en Europe et le DDBJ au Japon.

Ces institutions se coordonnent pour gérer les grandes bases de données de séquences nuléotidiques comme **GenBank** ou l'**UMBL** databse, ainsi que les bases de données de séquences protéiques comme **UniProt** ou **TrEMBL**.

Il faut ensuite développer des outils d'analyse de séquences afin de pouvoir déterminer leurs propriétés.

- Recherche de protéines à partir de la traduction de séquences nucléiques connue. Celle-ci passe par la détermination des phases ouvertes de lecture d'une séquence nucléique et de sa ou (ses) traduction(s) probables.
- Recherche de séquences dans une banque de données à partir d'une autre séquence ou d'un fragment de séquence. Les logiciels les plus fréquemment utilisés sont de la famille **BLAST** (Blastn, Blastp, Blastx et leurs dérivés).
- **Alignement de séquence** : Pour trouver les ressemblances entre deux séquences et déterminer leurs éventuelles homologies. Les alignements sont à la base de la construction de parentés suivant des critères moléculaires, ou encore de la reconnaissance de motifs particuliers dans une protéine à partir de la séquence de celle-ci.
- Recherche de motifs ou structures consensus pour caractériser les séquences.

La bio-informatique intervient aussi dans **le séquençage**, avec par exemple l'utilisation de puces ADN ou **bio puce**. Le principe d'une telle puce repose sur la particularité de reformer spontanément la double hélice de l'acide désoxyribonucléique face au brin complémentaire. Les quatre molécules de base de l'ADN ont en effet la particularité de s'unir deux à deux. Si un patient est porteur d'une maladie, Les brins extraits de l'ADN d'un patient, vont hybrider avec des brins d'ADN synthétiques représentatifs de la maladie.

Modélisation moléculaire

Les macromolécules biologiques sont en général de dimension trop petite pour être accessibles à des moyens d'observations directs tels que la microscopie. La **biologie structurale** est la discipline qui a pour objet de reconstruire des modèles moléculaires, par l'analyse de données indirectes ou composites. L'objectif est d'obtenir une reconstruction tridimensionnelle présentant la meilleure adéquation avec les résultats expérimentaux. Ces données sont issues principalement d'analyses Cryo microscopie électronique ou de techniques de diffusion aux petits angles (diffusion des rayons X ou diffusion des neutrons). Les données issues de ces expériences constituent des données (ou contraintes) expérimentales qui sont utilisées pour calculer un modèle de la structure 3D. Le modèle moléculaire obtenu peut être un ensemble de coordonnées cartésiennes des atomes composant la molécule, on parle alors de modèle atomique, ou une « enveloppe », c'est-à-dire une surface 3D décrivant la forme de la molécule, à plus basse résolution.

L'informatique intervient dans toutes les étapes conduisant de l'expérimentation au modèle, puis dans l'analyse du modèle par visualisation moléculaire.

Un autre volet de la modélisation moléculaire concerne la prédiction de la structure 3D d'une protéine à partir de sa structure primaire (L'enchaînement des acides aminés qui la composent), en prenant en compte les différentes propriétés physico-chimiques des acides aminés. Cela a un grand intérêt car la fonction, l'activité d'une protéine dépondent de sa forme.

De même, la modélisation des structures 3D d'acides nucléiques (à partir de leur séquence nucléotidique) revêt la même importance que pour les protéines, en particulier pour les **structures d'ARN.**

La connaissance de la structure tridimensionnelle permet d'étudier les **sites actifs** d'une **enzyme**, mettre au point informatiquement une série d'**inhibiteurs** potentiels pour cette enzyme, et ne synthétiser et ne tester que ceux qui semblent convenir. Cela permet de réduire les coûts en temps et en argent de ces recherches.

De même la connaissance de cette structure permet de faciliter l'alignement de séquences protéiques.

La visualisation de la **structure tridimensionnelle d'acides nucléiques (ARN et ADN)** fait également partie de la palette des outils bio-informatiques très utilisés.

Construction d'arbres phylogénétiques

On appelle **gènes homologues** des gènes descendant d'un même gène ancestral. De façon plus spécifique, on dit de ces gènes qu'ils sont **ortho logues** s'ils se retrouvent dans des espèces différents, ou qu'ils sont **para logues** s'ils se trouvent chez la même espèce.

Il est alors possible de quantifier la distance génétique entre deux espèces en comparant leurs gènes ortho logues. Cette distance génétique est représentée par le nombre et le type de mutations qui séparent les deux gènes.

Appliquée à un nombre plus important d'êtres vivants, cette méthode permet d'établir une matrice des distances génétiques entre plusieurs espèces. Les **arbres phylogénétiques** rapprochent les espèces qui ont la plus grande proximité. Plusieurs algorithmes différents sont utilisés pour tracer des arbres à partir des matrices de distances. Ils reposent chacun sur des modèles de mécanismes évolutifs différents. Les deux méthodes les plus connues sont la méthode de **UPGMA** et la méthode du **Neighbour joining**, mais il existe d'autres méthodes basées sur le **maximum de vraisemblance** et le **Bayésien Naïf.**

La construction d'arbres phylogénétiques est utilisée par des programmes d'alignement multiples des séquences afin d'éliminer une grande partie des alignements possibles et de limiter ainsi les temps de calcul : il permet ainsi de guider l'alignement total.

Exemples de tâches et débouchés

Voici un exemple de tâches et débouchés réalisé par plusieurs étudiants et professeurs.

- Aide à la création de nouveaux **médicaments** (prédiction de structure d'interaction).
- Développement de **logiciels** pour analyse et **prédiction** de données biologiques (génomiques, transcriptomiques, protéomiques, etc …)
- Développement de **logiciels** pour la biologie (LIMS, interface Web, etc …)
- Recherche dans un laboratoire (entreprise publique, biotechs, pharmaceutique, etc …)
- Modélisation physiologique et simulation informatique d'organes.
- Modélisation d'**écosystèmes** ou de processus écosystémiques (du gène au réseau écologique).
- Informatique pure.
- Aide à la création **d'organismes génétiquement modifiés (bactéries, plantes, etc …).**
- Aide à la création de tests et de systèmes de diagnostics destinés aux laboratoires **d'analyses médicales**, aux centres de **transfusion sanguine** et aux laboratoires de **contrôle industriel.**
- Enseignement.
- Adaptation de technologies informatiques au domaine de la biologie.
- Création, entretien et développement **d'entrepos de données**.

NB : * Source wikipédia.