Outils informatiques de biologie moléculaire appliquée

1. Définiton de la bioinformatique

C'est répondre à des problématiques biologiques en utilisant des méthodes informatiques. A partir d'une problématique biologique et éventuellement de données expérimentales (séquençage, puces à ADN, etc ….), la bioinformatique permet un traitement massif et rapide du problème afin de réduire les champs d'investigation à venir et/ou de formuler des prédictions. Les prédictions établies sur la base d'une méthodologie bioinformatique sont ensuite validées (ou invalidées) expérimentalement. Rien n'empêche que la bioinformatique soit l'élément déclenchant du questionnement.

Lors de sa création, la bioinformatique correspondait à l'utilisation de l'informatique pour stocker et analyser les données de la biologie moléculaire. Cette définition originale a maintenant été étendue et le terme bioinformatique est souvent associé à l'utilisation de l'informatique pour résoudre les problèmes scientifiques posés par la biologie dans son ensemble. Il s'agit dans tous les cas d'un champ de recherche multidisciplinaire qui associe informaticiens, mathématiciens, physiciens et biologistes.

La bioinformatique est une:

- discipline récente (quelques dizaines d'années).
- discipline hybride: elle est fondée sur des concepts et des formalismes issus de la biologie, de l'informatique, des mathématiques et de la physique, de la chimie (techniques de séquençage, ...).
- discipline qui utilise tout le potentiel de traitement de l'informatique: modèles théoriques, algorithmes et programmes, bases de données, ordinateurs, réseau Internet, protocoles de communication, langages,...

2. Objectifs de la bioinformatique

Les champs d'investigation de la bioinformatique sont vastes et variés.

Quelques exemples:

- Analyse de séquences (comparaisons, recherche de motifs/domaines, recherche de répétitions, recherche de biais du contenu, etc.)
- Prédictions de structures tri-dimensionnelles (protéines, ARNs)
- Bases de données pour stocker et mettre à disposition les données (séquences) ou répertorier des plans expérimentaux
- Analyses phylogénétiques et évolutives (classification, arbre, étude des pressions évolutives)

La démarche de la bioinformatique peut être résumée selon les étapes suivantes:

1) Compilation et organisation des données biologiques dans des bases de données:
   - bases de données généralistes (elles contiennent le plus d'information possible sans expertise très poussée de l'information déposée).
   - bases de données spécialisées autour de thèmes précis.
2) Traitements systématiques des données: l'un des objectifs est de repérer et de caractériser une fonction et/ou une structure biologique importante.
3) Elaboration de stratégies:

a. apporter des connaissances biologiques supplémentaires en combinant les données biologiques initiales et les données biologiques obtenues "*in silico*" (recherche ou un essai effectué au moyen de calculs complexes informatisés ou de modèles informatiques. Cette expression est surtout utilisée dans les domaines de la génomique et la bioinformatique**)**.

b. ces connaissances permettent, à leur tour, de développer de nouveaux concepts en biologie.
c. concepts qui, pour être validés, peuvent nécessiter le développement de nouvelles théories et outils en mathématiques et en informatique.

3. Banques de données

3.1. Banques généralistes

Nucléotides (N): Genbank (banque américaine créée en 1982)
Protéines (P): EMBL (banque européenne qui existe depuis 1980)
sont les grandes banques de séquences généralistes. Leur mission est de rendre publiques les séquences qui ont été déterminées. On trouve également une expertise biologique directement liées aux séquences traitées.

3.2. Banques spécialisées

De nombreuses bases de données spéciques ont été créées pour des besoins spéciques liés à l'activité d'un groupe de personnes. Elles ont pour but de recenser des familles de séquences autour de caractéristiques biologiques comme les gènes identiques issus d'espèces différentes. Elles peuvent aussi régrouper des classes spécifiques de séquences comme les vecteurs de clonage ou toutes les séquences d'un même génome.

3.3."Genome browsers"

Ils correspondent à différentes bases de données qui permettent d'accéder aux données du génome humain (et de celui d'autres espèces) à l'aide d'une interface graphique. En plus des données de séquence, ces navigateurs permettent d'accéder à de nombreuses données d'annotation (gènes avec exons et introns, sites de fixation, régions d'homologie).

Les plus populaires sont:
● Ensembl (European Bioinformatics Institute / Wellcome Trust Sanger Institute)
● NCBI (National Cancer for Biology Information)
● UCSC (University of California Santa Cruz)
D'autres méritent également le détour:
● Vista (University of California)
● Argo (BROAD Institute)
● Genatlas (Université René Descartes - Paris)

3.4. "Proteins"

Les plus populaires sont:
● Uniprot/Swiss Prot/Expasy (Uniprot Consortium)
● Protein Data Bank (Research Collaboratory for Structural Bioinformatics)
● NCBI (National Cancer for Biology Information)
D'autres bases de données sont particulièrement utiles pour identifier des domaines protéiques présents chez plusieurs protéines et ainsi définir des familles et des superfamilles de protéines:
● CATH protein structure classification (University College London)
● Protein Information Resource (University of Delaware / Georgetown University Medical Center)
● Structure Function Linkage Database (University of California, San Francisco)

4. Champs d'application de la bioinformatique

4.1. Acquisition des données biologiques

- les séquences nucléotidiques et les séquences polypeptidiques
- les gels bidimensionnels et les différentes méthodes de spectromètrie de masse (protéomique)
- les données de puces à ADN
- les données de structures tridimensionnelles
- l'uniformisation - standardisation des (formats de) données
- la recherche de phase de lecture ouverte (gène) et de signaux de régulation de la transcription et de la traduction, détection de bornes introns/exons
- la recherche de régions transcrites (EST) - profil d'expression des gènes (puces à ADN, analyse d'images)
- la détection de polymorphismes de nucléotide simple ou d'insertion / délétion
- la reconstruction d'arbres phylogèniques
- l'analyse de génomes entiers (génomique structurale, synténie) - réseaux de gènes
- l'ontologie: l'organisation hiérarchique de la connaissance sur un ensemble d'objets par leur regroupement en sous-catégories suivant leurs caractéristiques essentielle

4.2. Séquençage

La bioinformatique intervient aussi dans le séquençage, avec par exemple l'utilisation de puces à ADN ou biopuces.

4.3. Modélisation moléculaire

Les macromolécules biologiques sont en général de dimensions trop petites pour être accessibles à des moyens d'observation directs tel 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.

4.4. 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 orthologues s'ils se retrouvent dans des espèces différentes (spéciation sans duplication), ou qu'ils sont paralogues s'ils se retrouvent chez la même espèce (duplication à l'intérieur du génome).