**BIOINFORMATIQUE**

I : Définition

La bioinformation est l'information liée aux molécules biologiques : leurs structures, leurs fonctions, leurs liens de "parenté", leurs interactions et leur intégration dans la cellule.

Divers domaines d'études permettent d'obtenir cette bioinformation : la génomique structurale, la génomique fonctionnelle, la protéomique, la détermination de la structure spatiale des molécules biologiques, la modélisation moléculaire ...

La bioinformatique est l'analyse de la bioinformation.

II : Description

C'est une discipline récente (quelques dizaines d'années).

C'est une discipline "hybride" (au même titre que la biochimie ou la biophysique) : elle est fondée sur des concepts et des formalismes issus de la biologie, de l'informatique, des mathématiques et de la physique.

C'est une discipline qui utilise toutes les potentialités de traitement de l'informatique : modèles théoriques, algorithmes et programmes, ordinateurs, réseau Internet, bases de données ...

Il y à deux types de bioinformation : la séquence des nucléotides et la séquence des acides aminés. Les séquences constituent l'un des principaux types de bioinformation qu'analyse la bioinformatique.

III : Démarche

• Compilation et organisation des données biologiques dans des banques de données : ces banques sont soit généralistes (elles contiennent le plus d'information possible sans expertise particulière de l'information déposée), soit spécialisées dans un domaine autour de thèmes précis.

• Traitements systématiques des données : l'objectif principal est de repérer et de caractériser une fonction et/ou une structure biologique importante. Les résultats de ces traitements constituent de nouvelles données biologiques obtenues "in silico".

• Elaboration de stratégies :

Le but est d'apporter des connaissances biologiques supplémentaires en combinant les données biologiques initiales et les données biologiques obtenues "in silico".

Ces connaissances permettent, à leur tour, de développer de nouveaux concepts en biologie.

Concepts qui nécessitent l'élaboration de nouvelles théories et outils en mathématiques et en informatique.

IV : Deux types de molécules support de la bioinformation : les acides nucléiques et les protéines

1- Les acides nucléiques

ADN : Acide DésoxyriboNucléique

Macromolécule : chaîne nucléotidique constituée par un enchaînement d'unités élémentaires : les désoxyribonucléotides, forme de stockage de l'information génétique.

ARN : Acide RiboNucléique

Macromolécule : chaîne nucléotidique constitué par un enchaînement d'unités élémentaires : les ribonucléotides, forme qui permet de transférer et de traiter l'information dans la cellule le plus souvent formé d'un simple brin.

2 - Les protéines

Protéine : Macromolécule : chaîne polypeptidique constitué par un enchaînement d'unités élémentaires : les acides aminés.

V : Quelques dates importantes

1984 à 1990 : Création de la fondation **HUGO** ("Human Genome Organization") pour coordonner le séquençage au niveau mondial (éviter les doublons). Echec du projet (coût trop élevé).

1990 : Les objectifs du HGP étaient de :

Séquencer les 3 milliards de paires de bases du génome humain avec un taux d'erreur minimal, identifier tous les gènes, développer des méthodes plus rapides et efficaces pour le séquençage de l'ADN et l'analyse des séquences,

transférer ces technologies à l'industrie.

1997 : Le décryptage du génome pose la question de la brevetabilité du vivant, l'**UNESCO** le 11 novembre 1997 a déclaré que le génome humain est un patrimoine de l'humanité, or un patrimoine de l'humanité ne peut pas être la propriété d'un individu. Donc, une séquence d'ADN ne peut pas être brevetée.

1998 : Création de la société "**Celera Genomics**" (USA) par Craig Venter dans le but de séquencer le génome humain en compétition avec l'HGP.

2000 : HGP annonce 90 % du séquençage du génome humain.

"**Celera Genomics**" propose les premiers résultats du séquençage total du génome d'une personne.

Février 2001 : La même semaine, publication du brouillon initial des travaux de séquençage du génome humain complet par le **HGP** et "**Celera Genomics**"

Le génome humain contient 3,2 milliards de nucléotides et il contiendrait environ **30.000 à 40.000** gènes (chiffre nettement inférieur à ce qui était prévu).

VI : Le stockage de la bioinformation : les banques de données

Les fichiers contenant l'information biologique sous la forme de séquences est l'élément central autour duquel les banques de données se sont constituées. Il existe un grand nombre de bases de données d'intérêt biologique. On peut distinguer :

Les banques de données ou bases de données généralistes : elles correspondent à une collecte des données la plus exhaustive possible et qui offrent un ensemble plutôt hétérogène d'informations

Les bases de données ou bases de données spécialisées : elles correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée.

Depuis la première ébauche du séquençage du génome humain, en 2001, les biologistes produisent toujours plus de données bioinformatiques. Pour les stocker, ils **multiplient les bases de données**, mais aucune standardisation n'existe encore dans ce domaine. En conséquence, les données ne sont pas toutes présentées sous la même forme. De plus, les protocoles d'accès sont différents, de même que les algorithmes qui traitent les données.

Le but du projet "**Embrace**" est de standardiser l'accès aux innombrables données issues des projets de génomique, et leurs méthodes d'étude, afin que les chercheurs puissent les consulter et les exploiter facilement.