COURS BIO INFORMATIQUE

I)- Définition :

On trouve un grand nombre de définitions selon l'acception du terme et selon la prépondérance de "bio" sur "informatique" ou l'inverse.

- La bio information est l'information liée aux molécules biologiques : leur séquence, leur nombre, leur(s) structure(s), leur(s) fonction(s), leurs liens de "parenté", leurs interactions et leur intégration dans la cellule ...
- Cette bio information est issue de diverses disciplines : la biochimie, la génétique, la génomique structurale, la génomique fonctionnelle, la transcriptomique, la protéomique, la biologie structurale (structure spatiale des molécules biologiques, modélisation moléculaire ...), ...
- Une définition de la bioinformatique : analyse de la bioinformation par des moyens informatiques.

La définition du NCBI (2001) est : "*Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline*."

De manière générale :

- 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,...

II)- Démarche :

La démarche de la bio informatique 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. Les résultats de ces traitements constituent de nouvelles données biologiques obtenues "*in silico*".

3)- Elaboration de stratégies :

- 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, pour être validés, peuvent nécessiter le développement de nouvelles théories et outils en mathématiques et en informatique.

Quelques étapes clé en biologie moléculaire, en informatique et en bioinformatique (la liste ne peut évidemment pas être exhaustive), sont données dans le tableau suivant, pour nous permettre de comprendre l'évolution :

1965		Margaret Dayhoff *et al.* : Première compilation de protéines ("*Atlas of Protein Sequences*"). Matrices de substitution
1967		Article : "*Construction of Phylogenetic Trees*" - Fitch & Margoliash
1970		Algorithme pour l'alignement global de séquences : Saul Needleman & Christian Wunsch
1971		Premier microprocesseur Intel 4004
1972		Clonage de fragments d'ADN dans un virus, l'ADN recombiné : Paul Berg, David Jackson, Robert Symons
1973		Découverte des enzymes de restriction qui coupe spécifiquement l'ADN.

		Méthode de transfection (introduction d'un ADN étranger) des cellules eucaryotes grâce à un virus (vecteur).
1974		Programme de prédiction de structures secondaires des protéines : "*Prediction of Protein Conformation*" - Chou & Fasman. Vint Cerf et Robert Khan développent le concept des réseaux reliant des ordinateurs au sein d'un « internet » et développent deux protocoles fondamentaux "*Transmission Control Protocol*" (TCP) et "*Internet Protocol*" (IP).
1977		Développement des micro-ordinateurs accessibles à tous Techniques de séquençage d'ADN : Frederick Sanger / Maxam & Gilbert
1978 - 1980		Mutagénèse dirigée : Michael Smith Séquençage du 1er génome à ADN, le bactériophage phiX174 : Frederick Sanger Premières bases de données : EMBL, GenBank, PIR Accès téléphonique à la base de données PIR
1981 : 370.000 nucléotides GenBank : 270 séquences		Micro-ordinateur IBM-PC 8088 Programme d'alignement local de séquences : Temple Smith & Michael Waterman
1983		IBM-XT disque dur (10 Mb)
1984		Amplification de l'ADN : réaction de polymérisation en chaîne (PCR - Karry Mullis) MacIntosh : interface graphique & souris
1985		"FASTA" : Programme d'alignement local de séquences - David Lipman & William Pearson
1987		Nouveau vecteur permettant de cloner des fragments d'ADN 20 fois plus grands : le YAC (*Yeast Artificial Chromosome*) qui rend possible le séquençage de grands génomes.
1988		Taq polymérase, enzyme thermostable pour la PCR. Création du "*National Centre for Biotechnology Information*" (NCBI).
1989		INTERNET succède à ARPANET
1990		Clonage positionnel et premier essai de thérapie génique. "BLAST" : Programme d'alignement local de séquences - Altschul *et al*.
1991		"*Expressed Sequences Tags*" (EST) : méthode rapide d'identification des gènes (C. Venter).
1992		Séquençage complet du chromosome III de levure
1993		"*European Bioinformatics Institute*" (EMBL). Création à terme du "*European Bioinformatics Institute*" (EMBL - EBI).
1995		Analyse du transcriptome : début des puces à ADN
1996		Séquençage complet de la levure (consortium européen).
1997		11 génomes bactériens séquencés Evolutions de BLAST : "Gapped BLAST" et "PSI-BLAST"
1998		Séquençage de 2 millions de nucléotides par jour. Interférence ARN
2000		Séquençage du 1er génome de plante : *Arabidopsis thaliana*
2001		Séquence "premier jet" complète du génome humain

Années 2000		Epigénétique : développement de technologies d'analyse des modifications de l'ADN et des histones. Accès aux revues et journaux scientifiques : développement de *l'"open access".* Montée en puissance de la biologie synthétique. Détermination de structures de systèmes biologiques de plus en plus complexes (ribosomes, spliceosome, virus, ...) - cryo-microscopie électronique et autres techniques ("*femtosecond pulses / X-ray free-electron laser*")
2007 - 2008		Avènement des nouvelles technologies de séquençage à très haut débit, dites de seconde génération et maintenant de 3è génération. Prise de conscience du phénomène "*big data*" (pas seulement en biologie) qui devient peu à peu une discipline scientifique.
Mars 2019 : > 303 milliards de nucléotides > 49 millions séquences d'acides aminés		Plus de 18.900 génomes eucaryotes et procaryotes séquencés et des milliers en cours de séquençage (*Genomes OnLine*).

II)- Quelques champs d'application de la bioinformatique:

1)- L'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 puce 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 essentielles

2)- Le séquençage :

La bio-informatique intervient aussi dans le séquençage, avec par exemple l'utilisation de puces à ADN ou biopuce. 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 les brins d'ADN synthétiques représentatifs de la maladie.

Depuis l'invention du séquençage de l'ADN par Frederick Sanger dans la deuxième moitié des 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'information et de la rendre disponible à l'ensemble de la communauté des chercheurs. Cela a été rendu possible grâce à différentes bases de données, accessibles en lignes. À l'échelon mondial, trois grandes institutions sont chargées de l'archivage de ces données : le NCBI aux États-Unis, l'EBI en Europe et le DDBJ (en) au Japon. Ces institutions se coordonnent pour gérer les grandes bases de données de séquences nucléotidiques comme GenBank ou l'EMBL database, 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 connues. Celle-ci passe par la détermination des cadres de lecture ouverts 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, tblastx et leur dérivés).
- Alignement de séquences : 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.

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. Ces données sont issues principalement d'analyses cristallographiques (étude des figures de diffraction des rayons X par un cristal), de résonance magnétique nucléaire, de cryomicroscopie 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 est 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 la *visualisation moléculaire* (voir les protéines en 3D).

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 dépend 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 tri-dimensionnelle 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

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).

Il est alors possible de quantifier la *distance génétique* entre deux espèces en comparant leurs gènes orthologues. 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 distance. Ils reposent chacun sur des modèles de mécanismes évolutifs différents. Les deux méthodes les plus connues sont la méthode 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 les programmes d'alignements multiples de 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.