CHAPITRE I.

INTRODUCTION A LA BIOINFORMATIQUE

1. Introduction

Ces dernières années, une croissance massive de l'information biologique recueillie par les communautés scientifiques a vu le jour. Le déluge de ce type d'informations sous la forme de génomes, de séquences protéiques, de données d'expressions génétiques, …a conduit à la nécessité de concevoir des outils informatiques performants pour stocker, analyser et interpréter ces données, ce qui a engendré une science nouvelle, appelée Bioinformatique… Le terme bioinformatique signifie littéralement la science de l'informatique appliquée à la recherche biologique.

D'autre part, l'informatique est la gestion et l'analyse de données en utilisant diverses techniques computationnelles de pointe. Ainsi, en d'autres termes, la bioinformatique peut être décrite comme l'application de méthodes computationnelles pour la découverte de connaissances biologiques. Elle représente une symbiose de plusieurs domaines différents de la science, notamment, l'informatique, la biologie, les mathématiques et les statistiques.

Les tâches de la bioinformatique ne cessent de s'accroître avec les données biologiques. Les premiers temps, elles consistaient en une simple analyse des séquences à travers des comparaisons, aujourd'hui, une séquence est analysée sous toutes les coutures. A partir d'un gène, il est possible de construire une puce ADN, de réaliser une transcription de ce gène et d'obtenir la protéine correspondante ou de chercher des séquences similaires afin d'étudier l'homologie1 d'une espèce donnée. Cependant, la tâche majeure en bioinformatique demeure la prédiction de fonction de protéine, qui consiste en l'identification de son rôle cellulaire et biologique dans l'organisme afin de mieux comprendre son comportement.

2. Historique de la bioinformatique.

La bioinformatique a gagné une importance programmatique significative au cours du programme « European Commission's Fifth Framework Programme (FP5 2007) » de 1998 jusqu'en 2002, et dans la communauté scientifique, où l'activité dans ce domaine est principalement liée au développement du stockage et l'organisation de quantités croissantes

de données produites par des technologies génétiques de plus en plus sophistiquées, en liaison avec les besoins en infrastructures qui accompagnent la recherche génétique de base.

Tableau 1. Historique de la bioinformatique.

Années	Evénement
1965	Première compilation de protéines (Atlas of Protein Sequences) : Margaret Dayhoff et al.
1967	Article : Construction of Phylogenetic Trees – Fitch & Margoliash
1970	Algorithme pour l'alignement global des séquences : Needleman & Wunsch
1974	Programme de prédiction de structures secondaires des protéines : Prediction of Protein Conformation‖ – Chou & Fasman
1978	Premières bases de données : EMBL, GenBank, PIR
1981	370.000 nucléotides et 270 séquences à GenBank Programme d'alignement local de séquences : Smith & Waterman
1985	FASTA: Programme d'alignement local de séquences – Pearson & Lipman
1990	BLAST: Programme d'alignement local de séquences – Altschul et al.
1991	Grail: Programme pour la localisation de gènes – Mural et al.
1992	Fondation du Centre de recherche SANGER: il réalise la moitié de la « production‖ » mondiale. Publication de la 2e carte génétique du génome humain, établie par le Généthon à partir de 814 fragments génomiques
1993	SRS: logiciel d'interrogation multi-banques accessible sur le web – Etzold & Argos
1998	Séquençage de 2 millions de nucléotides par jour
2001	Séquence 'premier jet' complète du génome humain
Fevrier 2015	Plus de 1.143.000.000.000 nucléotides !

3. Les objectifs de la bioinformatique

Le rôle de la bioinformatique est d'aider les biologistes dans la collecte et le traitement des données génomiques afin d'étudier la fonction des gènes et des protéines. Un autre rôle important de la bioinformatique est d'aider les chercheurs des compagnies pharmaceutiques à élaborer des études détaillées des fonctions des protéines afin de faciliter la conception de médicaments.

Les objectifs de la bioinformatique peuvent se résumer dans ce qui suit :

- Collecter et stocker des informations dans des bases de données, accessibles en ligne.

- Explosion de la quantité de données biologiques nécessitant des outils de stockage adaptés.

❖ Fournir des outils de comparaison de séquences protéiques et nucléotidiques.

- Identifier une séquence en la comparant aux séquences d'une base de données.
- Déterminer le degré de similitude entre deux séquences.
- Repérer des motifs structuraux.

❖ Fournir des outils de traduction de séquences.

- Simplifier les tâches de traduction.
- Proposer plusieurs possibilités de protéines pour une même séquence.
- Repérer les exons2 /introns.

❖ Fournir des outils de prédiction physiologique et fonctionnelle et de prédiction expérimentale.

- La recherche de similarité est au centre de la bioinformatique.

4. Les fondements biologiques de la Bioinformatique

Il est indispensable d'avoir des pré-requis biologiques nécessaires pour le développement et l'évaluation de modèles ou de techniques en bioinformatique. Ceci est effectué en ayant les connaissances basiques sur les principes fondamentaux de la biologie moléculaire tels que les structures des gènes et des protéines...

4.1. De la génomique vers la protéomique

La compréhension du fonctionnement d'une cellule vivante suppose celle des mécanismes moléculaires complexes qui sous-entendent les diverses activités cellulaires. Tous les gènes d'un organisme, ou son génome, constituent une base de données statique et spécifique de l'être vivant.

A partir d'un génome unique, chaque type cellulaire d'un organisme exprimera un ensemble de protéines (voir Figure 1), ou protéome, qui variera en fonction de l'environnement des cellules. La synthèse des protéines comprend deux étapes - La

transcription permet de copier l'ADN en ARN messager (ARNm), elle se déroule dans le noyau - La traduction correspond au décodage de l'information portée par l'ARNm en polypeptides reliés en protéines.

La génomique et la protéomique sont intrinsèquement « globales », dans le sens où des centaines, si ce n'est des milliers de bases de données, de bases de connaissances, de programmes informatiques et de bibliothèques de documents sont disponibles via Internet et sont utilisés par des chercheurs et des développeurs à travers le monde dans le cadre de leurs travaux.

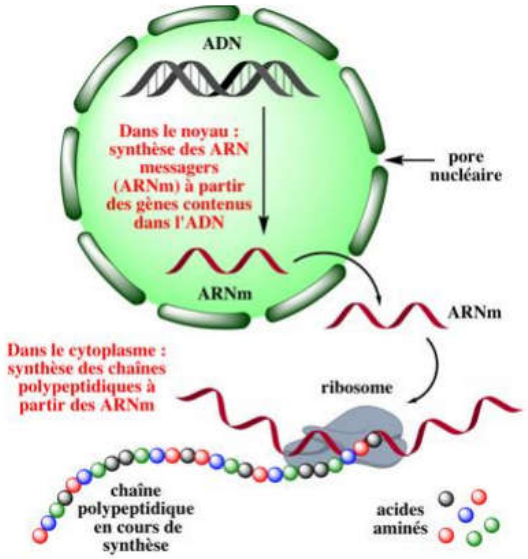

Figure 1. Synthèse des protéines.

Comme les protéines sont les principaux acteurs finaux des processus biologiques, leurs études peuvent offrir la vision la plus pertinente du fonctionnement d'une cellule vivante. La protéomique désigne la science qui étudie les protéomes, c'est-à-dire l'ensemble des protéines d'une cellule, organite, tissu, organe ou organisme à un moment donné et sous des conditions données.

Dans la pratique, la protéomique s'attache à identifier les protéines extraites d'une culture cellulaire, d'un tissu ou d'un fluide biologique, leur localisation dans les compartiments cellulaires, leurs modifications post-traductionnelles ainsi que leur quantité. Elle peut également permettre de quantifier les variations de leur taux d'expression en fonction du temps, de leur environnement, de leur état de développement, de leur état physiologique et pathologique, de l'espèce d'origine. Elle étudie aussi les interactions que les

protéines ont avec d'autres protéines, avec l'ADN ou l'ARN ainsi que les fonctions de chaque protéine.

4.2. Les protéines

Les protéines représentent l'une des classes moléculaires les plus importantes dans les organismes vivants. Leurs fonctions incluent la catalyse de processus métaboliques sous la forme d'enzymes, elles jouent un rôle important dans la transmission du signal, les mécanismes de défense, et de transport de molécules, et elles sont utilisées comme matériaux de construction, par exemple dans les cheveux (la protéine de la kératine).

Chaque protéine est une macromolécule produite par un organisme vivant. Les protéines sont formées de chaînes d'acides aminés liées par des liaisons peptidiques. Elles sont généralement représentées sous forme de séquences. Elles se replient en structures tridimensionnelles plus ou moins stables (voir Figure 2). Les protéines ont des tailles de plusieurs centaines d'acides aminés. Plus spécifiquement, les petites chaînes sont appelées peptides, les protéines étant des polypeptides pouvant être réunies par des ponts disulfures.

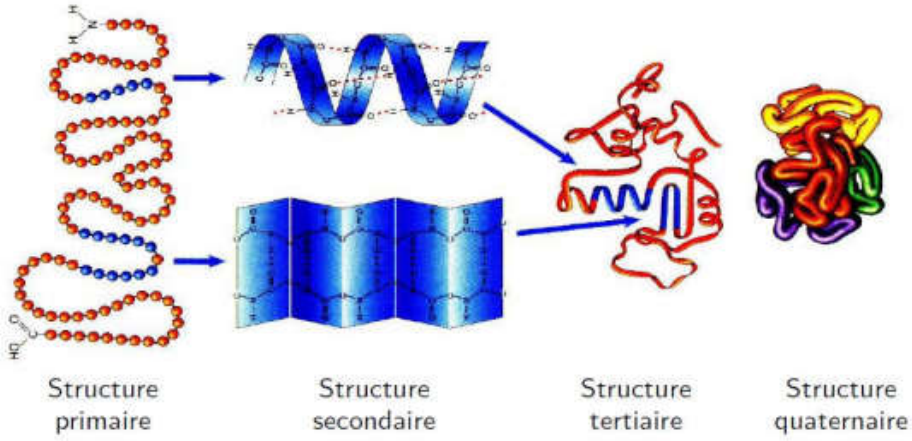

Figure 2. Les différentes structures d'une protéine.

Les protéines se répartissent en trois classes générales, sur la base de leur structure tridimensionnelle globale et sur la base de leur rôle fonctionnel, en fibreuse, membranaire, et globulaire. Les protéines fibreuses ont tendance à être de longues molécules étroites. Elles sont utilisées pour construire les structures macroscopiques, en particulier les structures à l'extérieur des cellules. Elles ont également tendance à avoir un rôle structurel, même si certaines ont ainsi des fonctions plus actives.

Les protéines membranaires comprennent une classe unique de protéines. Pour les protéines membranaires, une zone importante de la protéine doit être stable dans un environnement hydrophobe. Ceci est généralement réalisé en ayant des chaînes latérales non polaires sur des zones de surface spécifiques de la protéine. En raison de cette surface hydrophobe exposée, et parce que de nombreuses protéines membranaires sont déstabilisés par l'élimination de la membrane, la plupart des protéines membranaires sont difficiles à manipuler. Par conséquent, l'information structurelle n'est disponible que pour un nombre relativement faible de ces protéines, bien que de nouvelles techniques aient permis de déterminer la structure tridimensionnelle pour un nombre croissant de protéines membranaires au cours des dernières années.

Les protéines globulaires comprennent le type le plus varié de protéines. Les protéines globulaires sont solubles en solution aqueuse, elles ont généralement des résidus polaires à la surface et des résidus hydrophobes à l'intérieur. Les protéines globulaires ont souvent des structures stables qui se prêtent à la détermination de structure d'elles-mêmes.

4.3. Les acides aminés

Un acide aminé est une petite molécule élémentaire des protéines. Il en existe 20 formes différentes synthétisées par voie ribosomale dans le monde du vivant (Tableau 2).

La configuration générale des acides aminés naturels est caractérisée par un groupe amine et un groupe carboxyle autour d'un atome de carbone central α (Figure 3).

Des liaisons peptidiques connectent des acides aminés individuels dans une chaîne polypeptidique. Chaque acide aminé est lié via la liaison amide de l'acide de son groupe α carboxylique au groupe amine α de l'autre. Par conséquent, ils ont des extrémités N- et C-terminales.

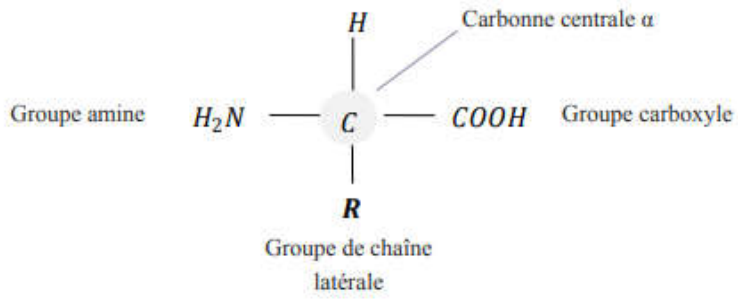

Figure 3. Configuration générale d'un acide aminé.

La structure primaire polypeptidique, c'est-à-dire, la séquence d'acides aminés de l'extrémité N- à l'extrémité C-terminale, peut contenir entre trois et plusieurs centaines d'acides aminés. Chaque acide aminé dans la chaîne polypeptidique est abrégé soit par un code à trois lettres ou une lettre (voir Tableau 2).

Tableau 2. Les vingt acides aminés natifs et leur code officiel.

Code en une lettre	Abréviation	Nom
A	Ala.	Alanine
R	Arg.	Arginine
N	Asn.	Asparagine
D	Asp.	Acide aspartique
C	Cys.	Cystéine
Q	Gln.	Glutamine
E	Glu.	Acide glutamique
G	Gly.	Glycine
H	His.	Histidine
I	Ile.	Isoleucine
L	Len.	Leucine
K	Lys.	Lysine
M	Met.	Méthionine
F	Phe.	Phénylalanine
P	Pro.	Proline
S	Ser.	Sérine
T	Thr.	Thréonine
W	Trp.	Tryptophane
Y	Tyr.	Tyrosine
V	Val.	Valine

Les propriétés structurales et physico-chimiques de chaque acide aminé sont très variées. Cependant, en se basant sur la composition chimique, on peut regrouper les acides aminés en 8 familles :

1. Les acides aminés aliphatiques dont le radical est une chaîne hydrogéno carbonée apolaire.

2. Les acides aminés hydroxylés qui portent un groupe alcool. Ils sont polaires, mais non chargés et neutres.

3. Les acides aminés représentés uniquement par la proline. La chaîne latérale est repliée et établit une liaison covalente avec l'atome d'azote du groupement amine.

4. Les acides aminés soufrés qui comportent un atome de soufre dans la chaîne latérale. L'un d'eux, la cystéine est un thiol, deux molécules de cystéine peuvent établir une liaison covalente entre leurs atomes de soufre et établir une liaison supplémentaire dans la chaîne protéique.

5. Les acides dicarboxyliques portent un groupement acide organique à l'extrémité de leur chaîne latérale. Ils sont donc polaires, chargés négativement (à pH neutre) et également acides.