Introduction

Le cours intitulé « Bioinformatique» est destiné aux étudiants de première année Master spécialité « Parasitologie ». Il contient l'essentiel du cours avec des exemples.

La bioinformatique est devenue un outil incontournable pour comprendre le fonctionnement des génomes et des organismes vivants. L'objectif du module est de présenter les méthodes informatiques issues de l'algorithmique ou de la modélisation de systèmes complexes mises en œuvre pour l'analyse de données de biologie moléculaire. Cela comprend l'analyse de séquences (ADN, ARN, protéines) et la modélisation 3D.

Objectifs du cours

- Apprendre à :
  - Interroger les bases de données,
  - Rechercher des séquences similaires entre-elles,
  - Réaliser des comparaisons et des alignements de séquences nucléiques ou protéiques.
- Acquérir une expertise dans l'analyse des données biologiques grâce à l'utilisation des outils de bioinformatique.

Le polycopié est composé de :

- **Chapitre 1 : Les séquences de Nucléotide et de peptides :** qui définit la bioinformatique, ses différentes facettes, ses types et ses objectifs ainsi que les principaux formats de séquence.
- **Chapitre 2 : Les bases de données en biologie :** qui explique la notion de base de données, les différents types de bases de données biologiques, et donne un aperçu sur bases de données bioinformatique les plus utilisées.
- **Chapitre 3 : Base de données sur les structures de protéines :** qui présente les structures des protéines et l'ensemble des bases de données de structure les plus utilisées en bioinformatique.

- **Chapitre 4 : Algorithmes utilisés en bioinformatique :** ce chapitre décrit l'intérêt et les types de l'alignement de séquences ainsi que quelques algorithmes d'alignement utilisés en bioinformatique.
- **Chapitre 5 : Prédiction sur les structures et fonction des protéines :** ce chapitre donne une idée sur les algorithmes de prédiction 3D des protéines.
- **Chapitre 6 : Utilisation des données en génomique et protéomique :** ce chapitre présente les différentes fonctionnalités de la plateforme NCBI pour pouvoir utiliser les données biologiques.
- **Chapitre 7 : Analyse de génome et génomique fonctionnelle :** ce chapitre est consacré au logiciel BLAST ainsi que les matrices de similarités.
- **Chapitre 8 : Analyse en 3D des protéines et acides nucléiques :** énumère quelques logiciels de modélisation 3D.

Chapitre 1

Les séquences de Nucléotide et de peptides

Introduction

Le développement de la bioinformatique suit l'augmentation exponentielle de la quantité de données provenant, entre autres, des programmes de séquençage systématique des génomes. Si, dans un premier temps, la priorité fut de stocker le flot d'informations, le rôle de la bioinformatique a rapidement évolué vers la transformation de ces données brutes en connaissances.

La bioinformatique se définie actuellement comme un domaine de recherche qui analyse et interprète des données biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie. Cette discipline étudie l'information contenue dans les séquences des gènes et des protéines.

Définition de la bioinformatique

La bioinformatique est un champ de recherche multidisciplinaire où travaillent des biologistes, médecins, informaticiens, mathématiciens, physiciens et bio-informaticiens, dans le but de résoudre un problème scientifique posé par la biologie.

La bioinformatique est l'application de la statistique et de l'informatique à la science biologique. Cette discipline constitue la « biologie in silico ». Elle applique des algorithmes, modèles statistiques dans l'objectif d'interpréter, classer et comprendre des données biologiques.

La bioinformatique est la discipline de l'analyse de l'information biologique, en majorité sous la forme de séquences génétiques et de structures de protéines. Les trois activités principales de la bioinformatique sont :

- Acquisition et organisation des données biologiques ;
- Conception de logiciels pour l'analyse, la comparaison et la modélisation des données biologiques ;
- Analyse des résultats produits par les logiciels.

Les domaines respectifs de la bioinformatique et de l'informatique peuvent être décrits comme suit (voir Figure 1.1) :

La bioinformatique est une discipline relativement nouvelle, qui évolue en fonction des nouveaux problèmes posés par la biologie moléculaire.

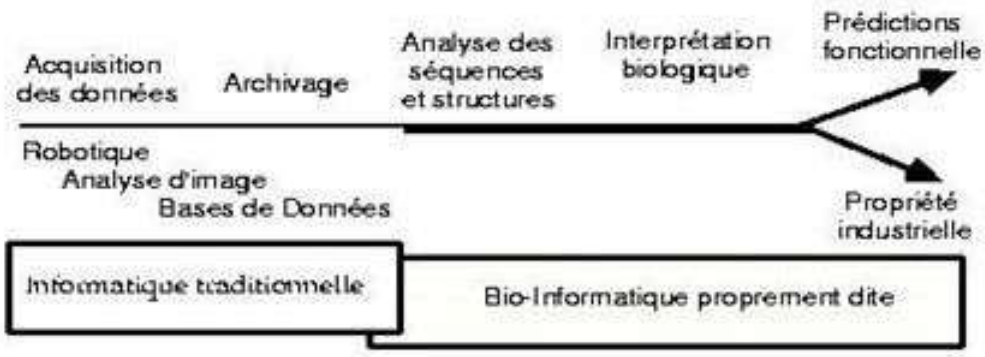

FIGURE 1.1: L'informatique traditionnelle et la bioinformatique

Les différentes facettes de la bioinformatique

Pour l'analyse des données expérimentales que représentent les séquences biologiques, l'apport informatique concerne principalement quatre aspects :

- Compilation et organisation des données : Cet aspect concerne essentiellement la création de bases de données. Certaines ont pour vocation de réunir le plus d'informations possible sans expertise particulière de l'information déposée alors que d'autres sont spécialisées dans un domaine considéré avec l'intervention d'experts. Ces dernières bases sont généralement construites autour de thèmes précis comme l'ensemble des séquences d'une même espèce ou les facteurs de transcription. Incontestablement, toutes ces banques de données constituent une source de connaissance d'une grande richesse que l'on peut exploiter dans le développement de méthodes d'analyse ou de prédiction.
- Traitements systématiques des séquences : L'objectif principal est de repérer ou de caractériser une fonctionnalité ou un élément biologique intéressant. Ces programmes représentent les traitements couramment utilisés dans l'analyse des séquences comme l'identification de phases codantes sur une molécule d'ADN ou la recherche de similitudes d'une séquence avec l'ensemble des séquences d'une base de données.
- Élaboration de stratégies : Le but est d'apporter des connaissances biologiques supplémentaires que l'on pourra ensuite intégrer dans des traitements standards. On peut donner comme exemples la mise au point de nouvelles matrices de substitution des acides aminés, la détermination de l'angle de courbure d'un segment d'ADN en fonction de sa séquence primaire, ou encore la détermination de critères spécifiques dans la définition de séquences régulatrices.

- Évaluation des différentes approches dans le but de les valider : Très souvent, tous ces aspects se confondent ou sont étroitement imbriqués pour donner naissance à un ensemble d'outils, d'études ou de méthodes qui convergent vers un but commun que l'on appelle l'analyse informatique des séquences.

Il est maintenant facile et courant d'effectuer certaines opérations plus ou moins complexes à l'aide de logiciels plutôt que manuellement. Pourtant, ces pratiques ne sont pas toujours systématiques car il est souvent difficile pour certains utilisateurs de savoir quel programme utiliser en fonction d'une situation biologique déterminée ou d'exploiter les résultats fournis par une méthode. C'est pourquoi ce cours contient la présentation d'un certain nombre d'outils ou de méthodes couramment utilisés et reconnus dans l'analyse informatique des séquences. Cependant, cette présentation ne constitue en aucun cas un exposé exhaustif de tout ce qui existe.

Historique de la bioinformatique

L'historique de la bioinformatique est synthétisé dans le Tableau 1.1.

< 1980 :	• Première banque de séquences protéiques (PIR) • Algorithme de comparaison de séquences (Needleman)
1980 :	• Banques de données (EMBL, GENBANK) • Début de la micro-informatique
1990 :	• Développement de l'Internet et de réseaux • Apparition des logiciels d'alignement (FASTA et BLAST) • Projets de séquençage de génomes complets
2000 :	• Séquençage du génome humain (Première ébauche)

TABLE 1.1: Historique de la bioinformatique

Types de la bioinformatique

- **La bioinformatique des séquences :** Sert à analyser les données issues de l'information génétique contenue dans les trois types de séquences. La bioinformatique des séquences s'intéresse en particulier à l'analyse et la comparaison multiples des séquences, l'identification de séquences à partir de données expérimentales, l'identification des ressemblances entre les séquences, la classification et la régression des séquences, l'identification des gènes ou de régions biologiquement pertinentes dans l'ADN ou dans les protéines, en se basant sur les composants de bases (nucléotides, acides aminés).
- **La bioinformatique structurale :** Traite la reconstruction, la prédiction ou l'analyse de la structure au moyen d'outils informatiques.
- **La bioinformatique des réseaux :** S'intéresse aux interactions entre gènes, protéines, cellules et organismes.

Objectifs de la bioinformatique

La bioinformatique s'applique à tout type de données biologiques, en particulier moléculaires :

- Les séquences d'ADN et de protéines
- Les structures de protéines
- Les puces à ADN (microarrays)
- Les réseaux d'interactions entre protéines
- Les réseaux métaboliques
- Les arbres de phylogénie

Parmi les objectifs de la bioinformatique, nous citons :

- Faire avancer les connaissances en biologie, en génétique humaine, en théorie de l'évolution, etc.,
- Aider à la conception des médicaments,
- Comprendre les maladies complexes,
- Développement de logiciels pour la biologie,
- Recherche dans un laboratoire,
- Aide à la création d'organismes génétiquement modifiés (bactéries, plantes, etc.).

Définition d'une séquence

Définition 1 : Alphabet

Un alphabet Σ est un ensemble fini de symboles distincts. Dans le cas de séquences d'ADN ou d'acides aminés on définit le symbole vide ou gap par -. L'alphabet de l'ADN est composé par les symboles suivants : -, A, C, G, T (voir le Tableau 1.2).

La base	**Abréviation**
Adénine	A
Cytosine	C
Guanine	G
Thymine	T

TABLE 1.2: L'alphabet de l'ADN

L'alphabet de l'ARN est composé par les symboles suivants : -, A, C, G, U (voir le Tableau 1.3).

La base	**Abréviation**
Adénine	A
Cytosine	C
Guanine	G
Uracile	U

TABLE 1.3: L'alphabet de l'ARN

L'alphabet des acides aminés est composé des symboles : -, A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y (voir la Figure 1.2).

Les 20 acides aminés

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

FIGURE 1.2: L'alphabet des acides aminés

Définition 2 : Séquence, segment

On appelle séquence S une suite ordonnée de caractères pris dans un alphabet : S = $\{x_1, x_2, ..., x_n\}$ (voir la Figure 1.3).

On note |S| = n la longueur de la séquence.

- On appelle segment toute séquence ou partie d'une séquence.
- Une séquence contient les informations sur le rôle biologique d'une macromolécule : fonction, relation avec les autres molécules, etc.
- Une séquence reflète les contraintes physico-chimiques imposées par la fonction, l'environnement (aqueux, lipidiques, intra- ou extracellulaire), l'évolution moléculaire.

```
>gi|15805103|ref|NP_293788.1| ABC transporter, ATP-binding
protein [Deinococcus radiodurans R1]
MTAAAPALSLRGLSKAFGAVQAVGDVSLEVQAGETLALLGPSGCGKSTVLRSVAGLERP
DAGQVLVGGRDVTALPPEARHLGLVFQDYALFPHLSVLDNVAYGPRRRGSSRPDAAQQA
REALALVGLSEHERRLPAQLSGGQQQRVALARALATRSPLLLLDEPLSNLDEKLRSELR
HDLRGLFGQLGAGVLLVTHDQREALALAHRVAVMRAGHVVQEGAAADLFARPATAWVAE
FLGWTNVFAHPQVSGQALLVPESAVQLGAGGELLRVLSRQRSETGETVTLAHPLGPLTL
SLSPREAAAASGDELRLTVPSAALLQVPDDREG
```

FIGURE 1.3: Une séquence protéique

Formats de séquences (Structuration des données)

Les séquences sont stockées en général sous forme de fichiers texte qui peuvent être soit des fichiers personnels, soit des fichiers publics (séquences des banques) accessibles par des programmes interfaces.

Le format correspond à l'ensemble des règles de présentation auxquelles sont soumises la ou les séquences dans un fichier donné. Le format permet :

- Une mise en forme automatisée ;
- Un stockage homogène de l'information ;
- Un traitement informatique ultérieur de l'information.

Il y a :

- **Formats issus des banques de séquences :** GenBank, EMBL, SwissProt, PDB, PIR, Prosite, etc.
- **Formats liés aux outils :** Staden, Fasta, Phylip, Stanford/IG, Fitch, DNAStrider, etc.
- **Formats enrichis :** GFF (Génome).
- **Nouveaux formats liés aux NGS :** BED, SAMs.

Conclusion

La bioinformatique est l'application de la statistique, l'informatique, la chimie et la physique à la science biologique pour résoudre un problème scientifique posé par la biologie. Le spécialiste qui travaille à mi-chemin entre ces sciences et l'informatique est appelé bio-informaticien.

Les données bioinformatiques sont très diverses et en évolution constante, ce qui nécessite l'utilisation des logiciels et des convertisseurs affin d'analyser l'information biologique.

Chapitre 2

Les bases de données en biologie

Introduction

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.

Définition d'une base de données

En informatique, une base de données est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter l'exploitation (ajout, mise à jour, recherche de données). Ces informations sont en rapport avec une activité donnée et peuvent être utilisées par des programmes ou des utilisateurs. Une Base de données est :

- Un ensemble de données relatives à un domaine, organisées par traitement informatique, généralement accessibles en ligne et à distance,
- Souvent, les données sont stockées sous la forme d'un fichier texte formaté (respectant une disposition particulière),
- Besoin de développer des logiciels spécifiques pour interroger les données contenues dans ces bases.

Définition d'une base de données biologique

Les bases de données biologiques sont des bases de données informatiques collectant des données biologiques très variées, il en existe différentes catégories selon le type de données stockées, celles-ci sont complétées par des annotations.

Les données y sont stockées sous la forme de fichiers textes, en relation entre eux. Ce sont donc des bases de données relationnelles. Leur conception est complexe et évolue rapidement avec l'augmentation des données et des outils d'étude.

Rôle des bases de données biologique

Le rôle des bases de données biologique est (voir Tableau 2.1) :

Collecter les informations auprès de biologistes, littératures et d'autres bases de données	• Séquences, cartographie physique,etc. • Données structurales, relationnelle,etc.
Stocker et organiser	• Logique cohérente
Distribuer l'information	• Large diffusion
Faciliter l'exploitation	• Interface conviviale • Définition des critères de recherche • Comparaison de données

TABLE 2.1: Rôle des bases de données biologique

Type des bases de données biologiques

Il existe un grand nombre de bases de données d'intérêt biologique, que ce soit d'intérêt biochimique, génétique, pharmaceutique ou génomique.

- **Les 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 banques de données généralistes contiennent des données hétérogènes :
  - Collecte la plus exhaustive possible,
  - Banques de séquences nucléiques,
  - Banques de séquences protéiques,
  - Banques de structure 3D de macromolécules,
  - Banques d'articles scientifiques.
- **Les bases de données spécialisées :** elles correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée.

Les banques de données spécialisées contiennent des données homogènes (une collecte établie autour d'une thématique particulière).

Quelques conseils

- Méfiez-vous des résultats donnés par les logiciels :
  - La qualité des résultats est parfois diminuée au profit de la rapidité
  - Certains problèmes admettent un ensemble infini de possibilités
- Ce n'est pas toujours la solution la meilleure qui est trouvée
  - Beaucoup de logiciels ne font que de la prédiction
- Méfiez-vous des banques de données :
  - Les données ne sont pas toujours fiables
  - La mise à jour n'est pas toujours récente

La réalité mathématique n'est pas la réalité biologique. Les ordinateurs ne font pas de biologie, ils calculent vite !

Comment s'assurer de la qualité de l'information ?

- Autorité :
  - Source de l'information, auteurs, statut, ...
- Péremption :
  - Date de création, de mise à jour, ...
  - Attention, ce qui est validé un jour peut être démenti par la suite !
- Transparence :
  - Documentation disponible
- Règles valables aussi bien pour une banque de données, que pour un logiciel, un site web, etc.

Conclusion

L'origine des banques de données biologiques remonte à l'utilisation des premiers ordinateurs par des cristallographes ou des biochimistes. Parmi ceux-ci, Margaret Dayhoff, biochimiste américaine, fut la première à voir l'intérêt de rassembler toutes les données

sur les séquences des protéines afin d'étudier leurs relations évolutives et de les classer en familles.

Pour permettre le stockage et l'organisation des données biologiques à différents niveaux, de nombreuses bases de données ont été mise en place telles que :

- GenBank, EMBL, DDBJ.
- UniProt, PDB.

Chapitre 3

Base de données sur les structures de protéines

Introduction

La structure des protéines est la composition en acides aminés et la conformation en trois dimensions des protéines. Elle décrit la position relative des différents atomes qui composent une protéine donnée.

Les protéines sont des macromolécules de la cellule, dont elles constituent la « boîte à outils », lui permettant de digérer sa nourriture, produire son énergie, de fabriquer ses constituants, de se déplacer, etc. Elles se composent d'un enchaînement linéaire d'acides aminés liés par des liaisons peptidiques. Cet enchaînement possède une organisation tridimensionnelle (ou repliement) qui lui est propre. De la séquence au repliement, il existe quatre niveaux de structuration de la protéine.

Les différents outils et bases de données permettent de collecter les informations en relation avec les protéines à ces différents niveaux (lorsque des informations sont disponibles ce qui est toujours vrai pour la séquence primaire mais peu fréquent pour la séquence tertiaire et encore plus rare pour la séquence quaternaire). Parallèlement à ces données classiques, des annotations complémentaires sont de plus en plus fréquemment disponibles (domaines protéiques en relation avec une structure ou une fonction particulières, structure de protéines mutantes . . .).

Dans ce chapitre, nous allons présenter quelques bases de données qui sont en effet indissociables dans le cas des structures puisque les données brutes ne sont pas directement interprétables par l'homme et nécessitent l'utilisation d'outils de visualisation.

Structure des protéines

Nous pouvons distinguer plusieurs niveaux dans la description de la structure des protéines (voir Figure 3.1) :

- La structure primaire : elle correspond à la séquence des acides aminés constituant la protéine. Il s'agit d'un assemblage linéaire des acides aminés codés par l'ARN messager.
- La structure secondaire : elle décrit un niveau structural plus complexe : les structures secondaires qui sont représentées par les repliements locaux de la protéine. Elle comporte les structures en hélices (α, 310, π, type II) et les feuillets (β parallèles et antiparallèles) et enfin les coudes (types I, II, III et γ).

- La structure tertiaire : décrit la structure tridimensionnelle de la protéine ou plus précisément d'une forme particulière que peut prendre dans l'espace la protéine d'intérêt dans des conditions expérimentales données et ceci à un tempst.
- La structure quaternaire : permet de décrire les interactions entreprotéines.

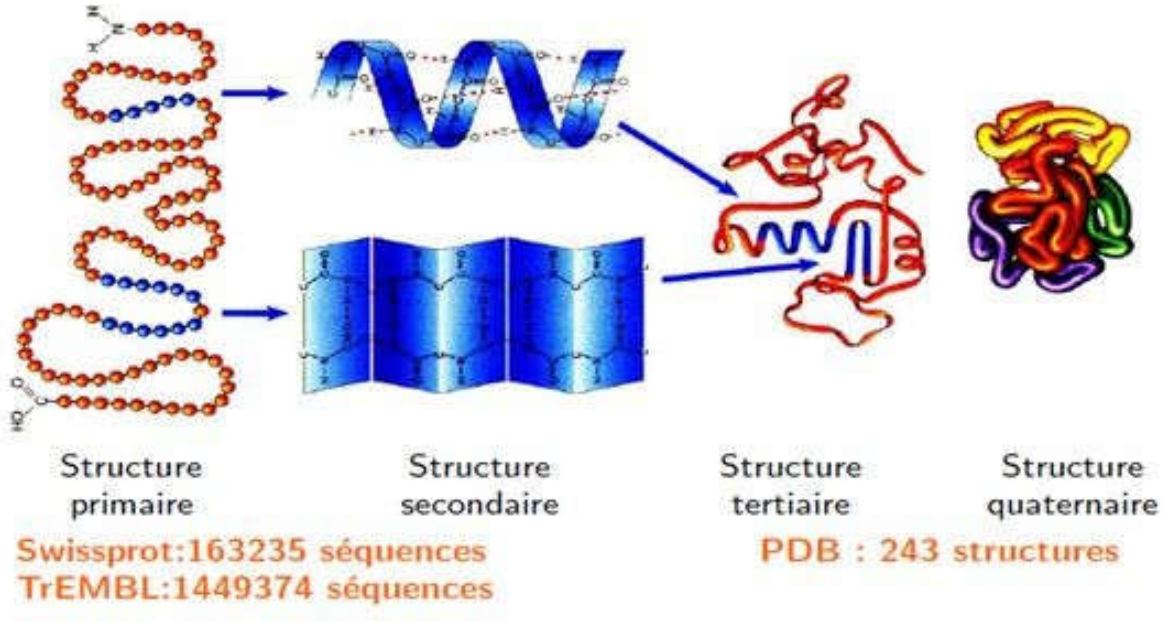

FIGURE 3.1: Structures de protéine

Bases de données de structure

- **PDB (Protein Data Bank) :** La banque de données sur les protéines du Research Collaboratory for Structural Bioinformatics, plus communément appelée PDB, est une collection mondiale de données sur la structure tridimensionnelle (ou structure 3D) de macromolécules biologiques : protéines, essentiellement, et acides nucléiques. Ces structures sont essentiellement déterminées par cristallographie aux rayons X ou par spectroscopie RMN. Ces données expérimentales sont déposées dans la PDB par des biologistes et des biochimistes du monde entier et appartiennent au domaine public. Leur consultation est gratuite et peut se faire directement depuis les sites web de la banque. La PDB est la principale source de données de biologie structurale et permet en particulier d'accéder à des structures 3D de protéines d'intérêt pharmaceutique.
  Comme c'est le cas pour GenBank, les coordonnées atomiques doivent être déposées avant publication. Les structures sont vérifiées lors de leur dépôt afin de s'assurer que les coordonnées déposées soient conformes aux standards établis.

- **E-MSD (European Macromolecular Structure Database) :** C'est la banque Européenne de structures tridimensionnelles de macromolécules biologiques, maintenue par l'EBI. Elle dérive de la PDB mais contrairement à cette dernière c'est une banque relationnelle. Il y a donc pour chaque entrée des liens croisés vers les bases de données (structurales, modulaires ou de séquences) à information ajoutée. Comme dans le cas de la PDB, il est possible d'y déposer de nouvelles structures. Un des atouts majeurs de E-MSD par rapport à la PDB est de générer une banque de ligands (chempdb) à partir des structures résolues en complexes. En plus des informations générales et structurales propres au ligand, elle fournit des détails assez précis de l'environnement chimique des sites de liaison de ce dernier.
- **NDB (Nucleic acid structures Database) :** Les structures disponibles dans la NDB incluent des structures d'ARN et d'oligonucléotides d'ADN composés d'au moins deux bases. Ces molécules peuvent être seules ou complexées avec des protéines ou de petits ligands. Les archives stockent des informations primaires et dérivées des structures. Les données primaires incluent les coordonnées atomiques, les facteurs de structure pour les structures aux rayons X ou les contraintes pour les structures RMN, et le détail des expériences (la condition de cristallisation, l'empilement cristallin, la collecte de données, et les statistiques d'affinement). L'information dérivée correspond à l'analyse de chaque structure. On retrouve des informations telles que la géométrie de valence, des angles de torsion et des contacts intermoléculaires. La NDB est partiellement redondante avec la PDB lorsqu'il s'agit des complexes avec des protéines.

Format de fichier PDB

Un fichier au format PDB est un fichier texte composé de caractères ASCII (voir Figure 3.2 et 3.3). Il est donc possible d'accéder à l'information brute contenue dans ces fichiers en les ouvrants avec un éditeur de texte.

Les fichiers PDB contiennent les coordonnées cartésiennes des atomes qui constituent la molécule ainsi que des métadonnées. Ces métadonnées peuvent par exemple être la structure primaire de la molécule, ses éventuelles structures secondaires, la méthode expérimentale qui a permis d'obtenir les coordonnées des atomes, etc. Un fichier PDB est composé de :

- 1ère Partie : appelé en-tête Contient des informations bibliographiques attachées à la structure, sur la résolution et les paramètres cristallographiques, la séquence et parfois la structure secondaire.
- 2ème partie : Elle contient les coordonnées atomiques Dans cette partie les atomes désignés par ATOM se situent sur la chaine protéique, tandis que les atomes désignés par HETATM (HETeroAToM group) font partie des molécules cofacteurs, substrats, ions ou autres groupes qui sont liés par une liaison covalente à la chaîne protéique.

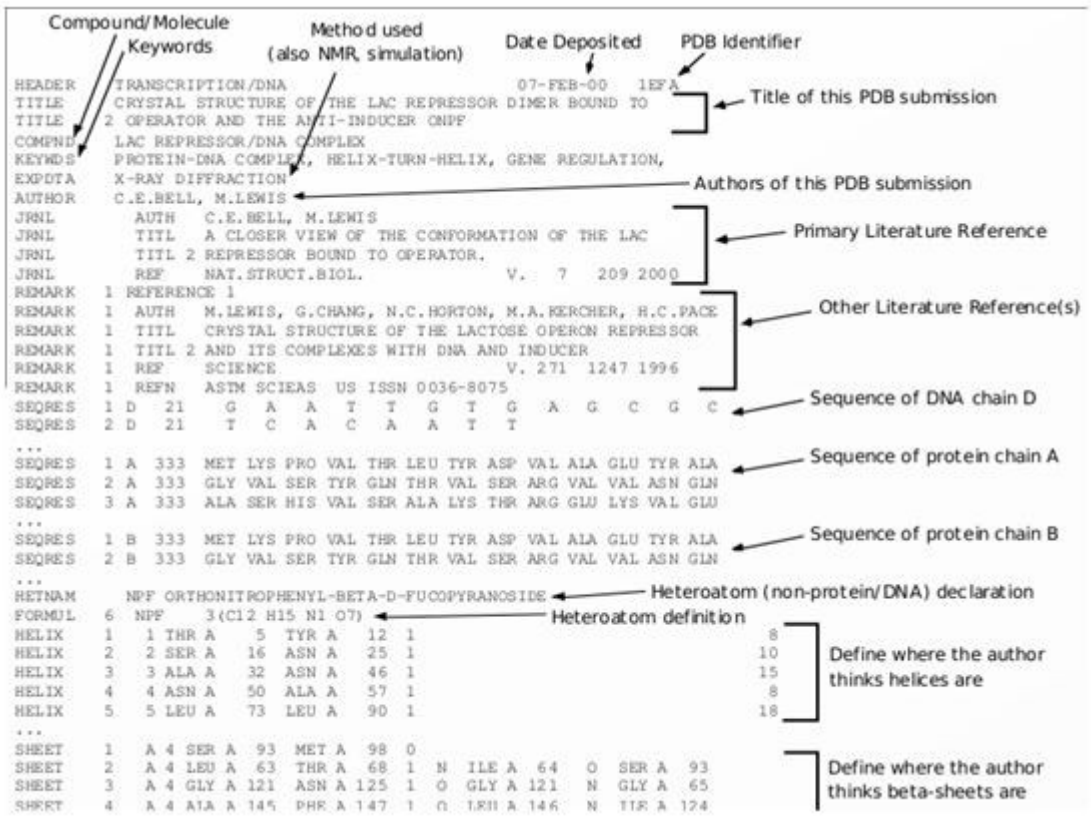

FIGURE 3.2: Format de fichier PDB (Partie 1)

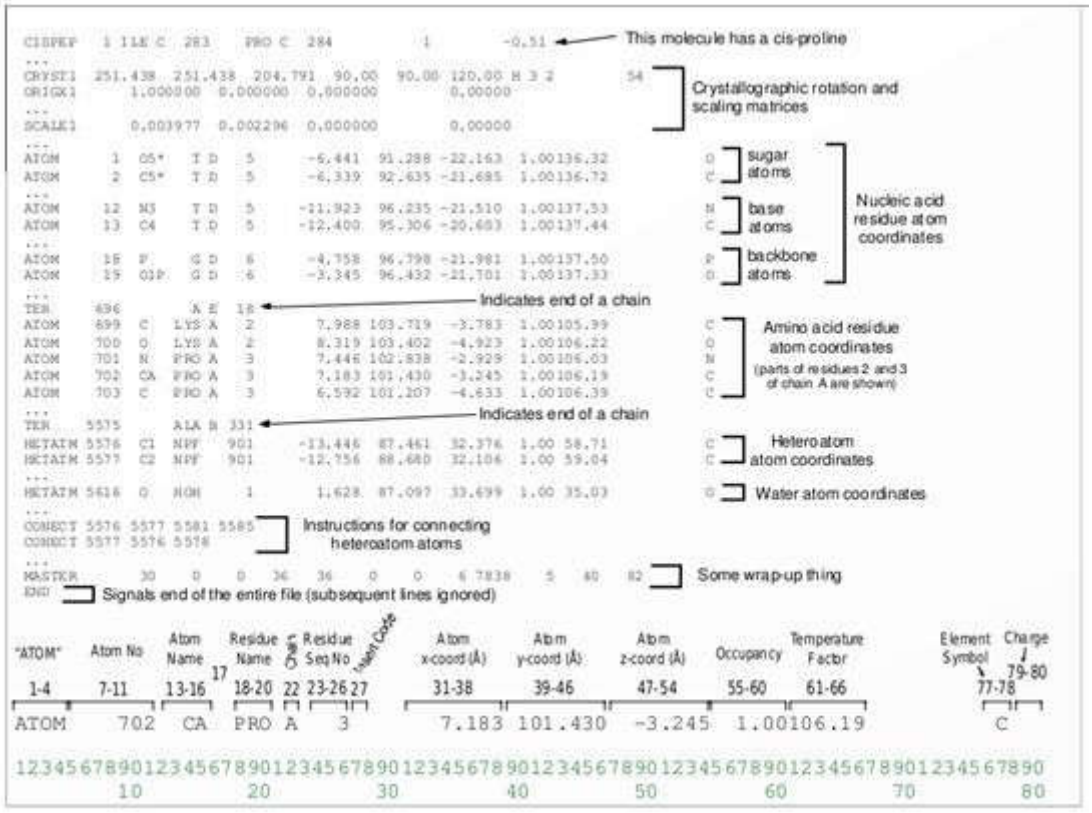

FIGURE 3.3: Format de fichier PDB (Partie 2)

Exemple

Dans la base de données PDB (https://www.rcsb.org/), nous allons effectuer une recherche pour trouver la structure tridimensionnelle de l'insuline chez l'être humain. Quelques informations obtenues sont présenté dans le tableau suivant (voir Tableau 3.1) :

Method	X-RAY DIFFRACTION
Resolution	1.6 Å
Classification	HORMONE
Organism	Homo sapiens
Deposited	2009-07-01
Deposition Author(s)	Timofeev, V.I., Bezuglov, V.V., Miroshnikov, K.A., Cuprov-Netochin, R.N., Samigina, V.R., Kuranova, I.P.

TABLE 3.1: Quelques informations de l'insuline obtenues de la PDB

La figure 3.4 représente la structure 3D de l'insuline.

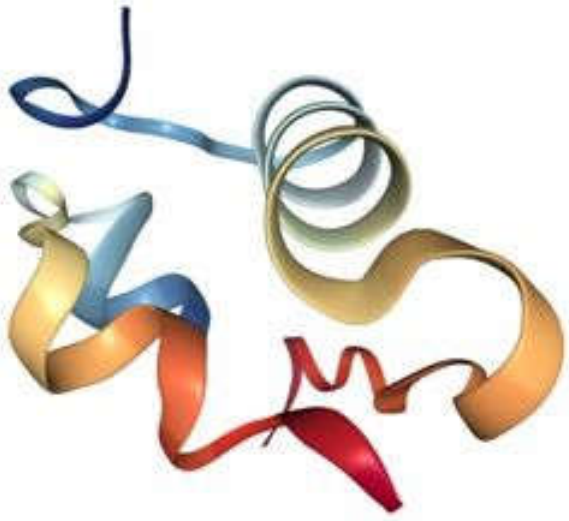

FIGURE 3.4: Structure 3D de l'insuline

Conclusion

Les structures 3D des protéines sont longues et coûteuses à déterminer que ce soit par cristallographie ou par résonance magnétique nucléaire.

Grâce à l'analyse de l'information disponible sur les structures 3D des protéines, il a été possible de mettre en évidence des relations étroites entre séquence et structure, en particulier une conservation de la structure 3D pour des séquences similaires mais aussi pour des séquences très différentes en particulier celles qui dérivent d'un ancêtre commun (séquences homologues).

Chapitre 4

Algorithmes utilisés en bioinformatique

Introduction

L'avènement de la biologie moléculaire a entrainé l'apparition de l'outil informatique pour appliquer la méthode comparative aux données de séquences. La comparaison des séquences passe par la comparaison de chaines de caractères. L'alignement de séquences qui résulte de la comparaison permet ainsi d'identifier les structures (séquences ou sous-séquences) conservées et donc les structures importantes.

L'alignement de séquences permet de trouver des similarités (similitudes) entre les séquences analysées. Ces similarités sont dues à une origine évolutive commune (homologie) ou à des fonctions semblables. D'autre part l'alignement de séquences permet de vérifier l'absence de similarité entre séquences afin de vérifier l'unicité d'hybridation d'une séquence donnée et donc sa spécificité.

Alignement de séquences

L'alignement est la mise en correspondance de deux séquences ou plus (ADN ou protéines).

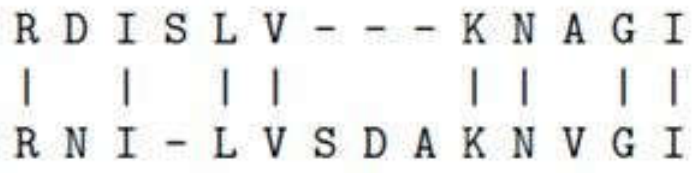

FIGURE 4.1: Alignement de deux séquences

Exemple : soit les séquences S1 = *{ACATT}* et S1 = *{AAGTT}*, par exemple un alignement de S1 et S2 est :

S1	A	C	A	-	T	T
S2	A	-	A	G	T	T
Consensus	A	.	A	.	T	T

TABLE 4.1: Exemple d'alignement

Un alignement peut s'interpréter comme le fruit d'un travail d'édition : trouver le nombre minimum d'opérations élémentaires d'édition qui permettent de transformer une séquence en une autre. On considère les opérations suivantes :

- Identité (Match) :un caractère de la première séquence est mis en regard d'un caractère identique dans la deuxième séquence,
- Substitution (Mismatch) : un caractère de la première séquence est mis en regard d'un caractère différent dans la deuxième séquence (remplacement d'une lettre par une autre),
- Indel (Gaps) : insertion ou suppression d'une ou plusieurs lettres.

Pourquoi comparer des séquences ?

- Les programmes de séquençage fournissent des séquences qu'il faut annoter.
- Recherche d'homologie : La similitude syntaxique est un signe de proximité fonctionnelle.
- Recherche de fonction commune : Les régions conservées correspondent à des régions fonctionnellement importantes.
- Prédiction de gènes

Donc l'alignement de séquences à pour but de :

- Identifier au sein d'une banque une séquence obtenue en laboratoire de biologie ;
- Localiser une séquence au sein du génome d'un organisme ;
- Identifier un rôle à une molécule séquencée par comparaison avec des molécules de fonctions similaires déjà répertoriées ;
- Réaliser une étude phylogénétique ;
- Prédire la structure secondaire (tertiaire) d'une protéine.
- Prédire des informations pertinentes sur la fonction d'une macromolécule à partir seulement de sa séquence.
- Identifier des sites fonctionnels.
- Prédire la/les fonction(s) d'une protéine.

Types d'alignement

L'alignement de séquences est une opération de base en bioinformatique qui a pour but d'identifier des zones conservées entre séquences. Dans la compréhension du fonctionnement de la vie, les protéines jouent un rôle essentiel. On part donc de l'hypothèse que

des protéines comportant des séquences similaires risquent fort de posséder des propriétés physico-chimiques identiques : à partir de l'identification de similarités entre une première séquence dont on connait le mécanisme d'action et une deuxième séquence dont on ne connait pas le mécanisme de fonctionnement, on peut inférer des similarités structurelles ou fonctionnelles sur la séquence non connue et proposer de vérifier de manière expérimentale le comportement d'action supposé.

On distingue différents types d'alignement :

- **Alignement par paires :** consiste à aligner 2 séquences. Il est possible de réaliser un :
  - **Alignement global :** alignement de deux séquences sur la totalité de leur longueur en tenant compte de tous les résidus. Si les longueurs des séquences sont différentes, des insertions ou délétions sont introduites pour aligner les deux extrémités des deux séquences. Il permet de mesurer le degré de similarité entre 2 séquences connues.
  - **Alignement local :** alignement entre une séquence et une partie de l'autre séquence, c'est-à-dire l'alignement de deux séquences portant sur des régions isolées et permettant de trouver des segments qui ont un haut degré de similarité. Outil efficace et rapide de recherche dans les bases de données en comparant une séquence inconnue à celles de la banque.
- **Alignement multiple :** alignement portant sur plusieurs séquences à la fois et dans leur intégralité. Il nécessite un temps de calcul et un espace de stockage exponentiel en fonction de la taille des données.

Algorithmes d'alignement utilisés en bioinformatique

Dans cette section, nous allons présenter les algorithmes les plus utilisés en bioinformatique.

Le dotplot

Le dotplot est l'une des plus anciennes méthodes utilisées en comparaison de séquences. Le dotplot compare deux séquences en les plaçant sur les axes x et y d'un tableau à deux dimensions (Chaque axe du tableau représente l'une des séquences à comparer). On

place un point (ou tout autre symbole) dans la ième rangée et jième colonne du tableau si le iième résidu de la première séquence est identique (ou presque identique) au jième résidu de la deuxième séquence.

Programmation dynamique

La programmation dynamique est appliquée à des problèmes d'optimisation pour lesquels un choix doit être fait entre plusieurs solutions possibles afin d'aboutir à une solution optimale.

4.6 Conclusion

Toutes ces méthodes montrent finalement que le problème de la signification mathématique des similitudes que l'on peut observer entre séquences biologiques est un élément important mais complexe, qui n'est pas encore clairement résolu mathématiquement. Il est vrai que cette signification dépend de nombreux critères eux même complexes comme par exemple l'algorithme utilisé et son paramétrage ou le système de score employé.

Chapitre 5

Prédiction sur les structures et fonction des protéines

Introduction

Les protéines sont des chaînes d'acides aminés réunis par des liaisons peptidiques. De nombreuses conformations de cette chaîne sont possibles du fait de la rotation de la chaîne autour de chaque atome de carbone. Ce sont ces changements de conformation qui sont responsables des différences dans la structure tridimensionnelle des protéines.

La structure protéique peut être considérée comme une séquence d'éléments structurels secondaires, tels que des hélices α et des feuilles β , qui constituent ensemble la configuration tridimensionnelle globale de la chaîne protéique. Dans ces structures secondaires, des motifs réguliers de liaisons H sont formés entre des acides aminés voisins, et les acides aminés ont des angles Φ et ω similaires.

La formation de ces structures neutralise les groupes polaires sur chaque acide aminé. Les structures secondaires sont étroitement emballées dans le cœur de protéine dans un environnement hydrophobe. Chaque groupe latéral d'acides aminés a un volume limité à occuper et un nombre limité d'interactions possibles avec d'autres chaînes latérales proches, une situation qui doit être prise en compte dans la modélisation moléculaire et les alignements.

Dans ce chapitre, quelques algorithmes de prédiction des structures de protéines sont présentés.

Définition de la prédiction des structures de protéines

La prévision de la structure des protéines est l'inférence de la structure tridimensionnelle d'une protéine à partir de sa séquence d'acides aminés, c'est-à-dire la prédiction de son pliage et de sa structure secondaire et tertiaire de sa structure primaire.

La prédiction de la structure est fondamentalement différente du problème inverse de la conception des protéines. La prédiction de la structure protéique est l'un des objectifs les plus importants poursuivis par la bioinformatique et la chimie théorique ; Elle est très importante en médecine (par exemple, dans la conception de médicaments) et en biotechnologie (par exemple, dans la conception de nouvelles enzymes). Tous les

deux ans, la performance des méthodes actuelles est évaluée dans l'expérience CASP (Évaluation critique des techniques de prédiction des protéines).

La structure d'une protéine donne beaucoup d'informations sur sa fonction et sa régulation (voir Figure 5.1). Mais ni la cristallographie aux rayons X ni la RMN n'est la panacée. En effet la première technique nécessite que la protéine étudiée cristallise, ce qui peut prendre plusieurs mois ou même ne jamais se réaliser. La seconde technique, la RMN, est limitée à des fragments d'une centaine d'acides aminés, et l'échantillon doit être soluble à forte concentration. Ainsi il est plus difficile, voire impossible pour l'instant, d'étudier des structures plus grosses, comportant des sous-unités ou flexibles que des plus petites, globulaires et rigides. De plus ces techniques demandent de longues et coûteuses manipulations, ce qui n'est pas conciliable avec le nombre de protéines dont la connaissance de structure présente un intérêt certain. C'est pourquoi, la détermination de structure de protéines par des outils bioinformatiques s'est rapidement révélée une nécessité, et des techniques comme la modélisation moléculaire par homologie ou la prédiction de structure tridimensionnelle de novo se développent.

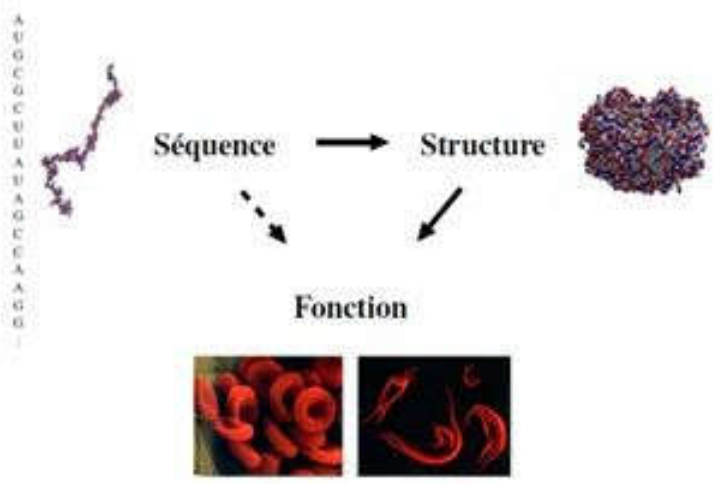

FIGURE 5.1: La structure et la fonction d'une séquence

Méthodes de prédiction de structures 3D

Aujourd'hui, les méthodes de prédiction de la structure tridimensionnelle des protéines peuvent être classées en trois catégories :

- La modélisation par homologie (ou comparative).

- Les méthodes de reconnaissance de repliement (ou d'enfilage, threading en anglais).
- Les méthodes ab initio et de novo.

Le choix de la méthode dépend des informations disponibles pour la réalisation du modèle. Le facteur le plus déterminant dépend de l'existence ou non dans la PDB d'une structure protéique similaire à celle de la protéine à modéliser et du taux d'identité de séquence entre ces protéines.

Conclusion

Dans ce chapitre, nous avons énuméré les différentes méthodes de prédiction des structures protéiniques. Que ces dernières soient faites au moyen de mesures expérimentales ou bien par des programmes plus ou moins évolués, il est essentiel de garder en mémoire que le produit de la méthode utilisée ne sera jamais qu'une prédiction avec son lot d'imprécisions.

La panoplie des méthodes de prédiction de structure in-silico est quant à elle beaucoup plus large. Si plusieurs d'entre elles affichent des performances raisonnables, aucune n'est cependant parvenue à s'imposer. Toutefois, le concours CASP11 (Critical Assessment of Techniques for Protein Structure Prediction) permet de faire régulièrement le point sur l'efficacité des méthodes publiées.

La fiabilité des prédictions est bien entendu fortement reliée au niveau de définition de la prédiction et du raffinement du modèle utilisé.

Chapitre 6

Utilisation des données en génomique et protéomique

Introduction

La bioinformatique des séquences traite les 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'enchaînement ou séquence de leurs composants élémentaires (nucléotides, acides aminés).

La première difficulté consiste à organiser l'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.

Définition de NCBI

Le NCBI (National Center for Biotechnology Information) est un institut national américain pour l'information biologique moléculaire. Cet organisme, fondé en 1988 et situé à Bethesda dans le Maryland, fait partie de la Bibliothèque américaine de médecine, un des Instituts américains de la santé.

Le site américain NCBI est un site hébergeant une banque de gènes, d'ARNm et de protéines. Cette banque est régulièrement complétée par des laboratoires du monde entier et par des centres de séquençage.

Le NCBI conduit des recherches dans la biologie informatique, développe des logiciels pour analyser des données de génome et fournir des informations biomédicales.

Rechercher sur NCBI la séquence d'intérêt

A l'aide du moteur de recherche intégré, rechercher la séquence d'ADN souhaitée (pour rechercher un gène, sélectionner la catégorie Gene dans la barre de recherche (voir Figure 6.1). (ici le gène de la globine B).

FIGURE 6.1: Rechercher sur NCBI la séquence d'intérêt

Conclusion

Le NCBI est une plateforme bioinformatique qui conduit des recherches dans la biologie informatique et développe des logiciels pour analyser des données de génome et fournir des informations biomédicales.

Le système d'interrogation du NCBI, appelé Entrez, permet de faire une recherche globale ou filtrée dans plusieurs bases de données tels que : GenBank, PubMed, OMIM, OMIA, etc.

Chapitre 7

Analyse de génome et génomique fonctionnelle

Introduction

La recherche de similitude entre séquences est un élément fondamental qui constitue souvent la première étape des analyses de séquences. Élémentaire, la question de la comparaison et de l'obtention d'un alignement optimal de deux séquences biologiques, nécessite néanmoins la mise en œuvre de procédures de calcul et de modèles biologiques permettant de quantifier la notion de ressemblance entre ces séquences. L'objectif est de révéler des régions proches dans leur séquence primaire en se basant sur le principe de parcimonie, c'est-à-dire en considérant le minimum de changements en insertion, suppression, ou substitution qui séparent deux séquences. On peut apprendre ainsi, par association, des informations importantes sur la structure, la fonction ou l'évolution des biomolécules.

Il faut développer des outils d'analyse de séquences afin de pouvoir déterminer leurs propriétés, comme :

- Recherche de protéines à partir de la traduction de séquences nucléiques connues.
- 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.

Les matrices de similarité ou matrices de substitution

Les matrices de similarité ou matrices de substitution sont des matrices utilisées en bioinformatique pour réaliser des alignements de séquences biologiques. Elles permettent de donner un score de similarité ou de ressemblance entre deux acides aminés.

Ces matrices, M, sont des matrices 20 x 20 pour les 20 acides aminés (ou 4 x 4 pour les 4 bases nucléiques) qui recensent l'ensemble des scores M(a,b) obtenus lorsqu'on substitue l'acide aminé a à l'acide b dans un alignement. Plus le score M(a,b) est élevé,

plus la similarité entre les deux acides aminés a et b est importante. Il existe plusieurs matrices, basées sur des principes de construction différents.

Les principaux logiciels de comparaison avec les banques de séquences

La taille sans cesse croissante des banques de séquences a nécessité l'élaboration d'algorithmes spécifiques pour effectuer la comparaison d'une séquence avec une banque de données car les algorithmes standards de comparaison entre deux séquences sont généralement trop longs sur des machines classiques.

Blast (Basic Local Alignment Search Tool)

Blast est une méthode de recherche heuristique utilisée en bioinformatique. Le logiciel blast permet de comparer une séquence, nucléique ou protéique à une banque de séquences, nucléiques ou protéiques. Blast recherche des régions de similarités locales entre des séquences nucléiques ou protéiques et réalise un alignement local de ces régions homologues. Il existe essentiellement cinq types de comparaison possibles :

- blastn (blast nucléique) : Pour comparer une séquence requête nucléique à une banque de séquences nucléiques.
- blastp (blast protéique) : Pour comparer une séquence requête protéique à une banque de séquences protéiques.
- blastx (blast nucléique vs protéique) : Pour comparer une séquence requête nucléique à une banque de séquences protéiques.
- tblastn (blast protéique vs nucléique) : Pour comparer une séquence requête protéique à une banque de séquences nucléiques.
- tblastx (blast nucléique vs nucléique en passant par un alignement protéique) : Pour comparer une séquence requête nucléique à une banque de séquences nucléiques en alignant les séquences protéiques induites par les séquences nucléiques.

Forces

- L'algorithme de blast permet d'obtenir un logiciel d'alignement très rapide tout en conservant une bonne sensibilité.

- Le logiciel fournit une mesure, la e-value, permettant de tester la significativité statistique d'un score d'alignement

Faiblesses

- L'indexation de la séquence requête puis le parcours des séquences de la banque pour chacune des séquences requêtes se traduit par un temps de calcul qui devient trop important lorsque l'on souhaite comparer des génomes complets ou des données haut débit, issues par exemple des nouvelles technologies de séquençage, avec un génome. Des approches alternatives proposent une indexation des séquences de la banque (voir ci-dessous Autres fonctionalités et Autres logiciels).

7.4 Conclusion

Il existe différentes matrices de scores destinées à aider le biologiste dans ses analyses. L'efficacité de ces matrices dépend du type d'expériences et des résultats utilisés pour l'alignement, et bien que de nombreuses études comparatives aient été menées, il n'y a pas de matrice idéale.

Les programmes de comparaison de séquences ont pour but de repérer les endroits où se trouvent des régions identiques ou très proches entre deux séquences et d'en déduire celles qui sont significatives et qui correspondent à un sens biologique de celles qui sont observées par hasard.

Chapitre 8

Analyse en 3D des protéines et acides nucléiques

Introduction

Malgré la puissance des microscopes modernes, les images des molécules de la matière vivante ne sont pas assez détaillées. Par contre, on parvient, en combinant les résultats des analyses chimiques, cristallographiques et de résonance magnétique nucléaire (RMN) à fournir suffisamment d'informations à des ordinateurs pour qu'ils puissent calculer un modèle moléculaire en 3D que l'on peut ensuite manipuler grâce aux fonctions d'un logiciel de visualisation.

Modélisation moléculaire

La modélisation moléculaire est un ensemble de techniques pour modéliser ou simuler le comportement de molécules. Elle est utilisée pour reconstruire la structure tridimensionnelle de molécules, en particulier en biologie structurale, à partir de données expérimentales comme la cristallographie aux rayons X. Elle permet aussi de simuler le comportement dynamique des molécules et leurs mouvements internes. On l'utilise enfin pour concevoir de nouveaux médicaments. La modélisation moléculaire s'intéresse enfin au rendu visuel des molécules et des simulations en 3D, c'est le domaine du graphisme moléculaire.

Logiciels de graphisme moléculaire

Il existe de nombreux logiciels permettant la représentation de molécules, dont plusieurs sont dans le domaine public ou gratuits pour un usage académique ou pédagogique :

PyMOL

PyMOL est un logiciel libre de visualisation de structures chimiques en 3D créé par Warren DeLano. Il est principalement utilisé par les étudiants, les professeurs et les chercheurs en chimie et en biologie structurale. Il est développé en Python et est multiplateformes. Ainsi, il fonctionne sous Windows, Mac OS X, Linux et les systèmes Unix. Ce logiciel est régulièrement utilisé pour produire des images 3D de grande qualité pour

la publication scientifique. Selon l'auteur, environ un quart des images 3D de protéines publiées dans la littérature scientifique est réalisé avec PyMOL.

L'interface principale du logiciel PyMOL est illustrée dans la Figure 8.1.

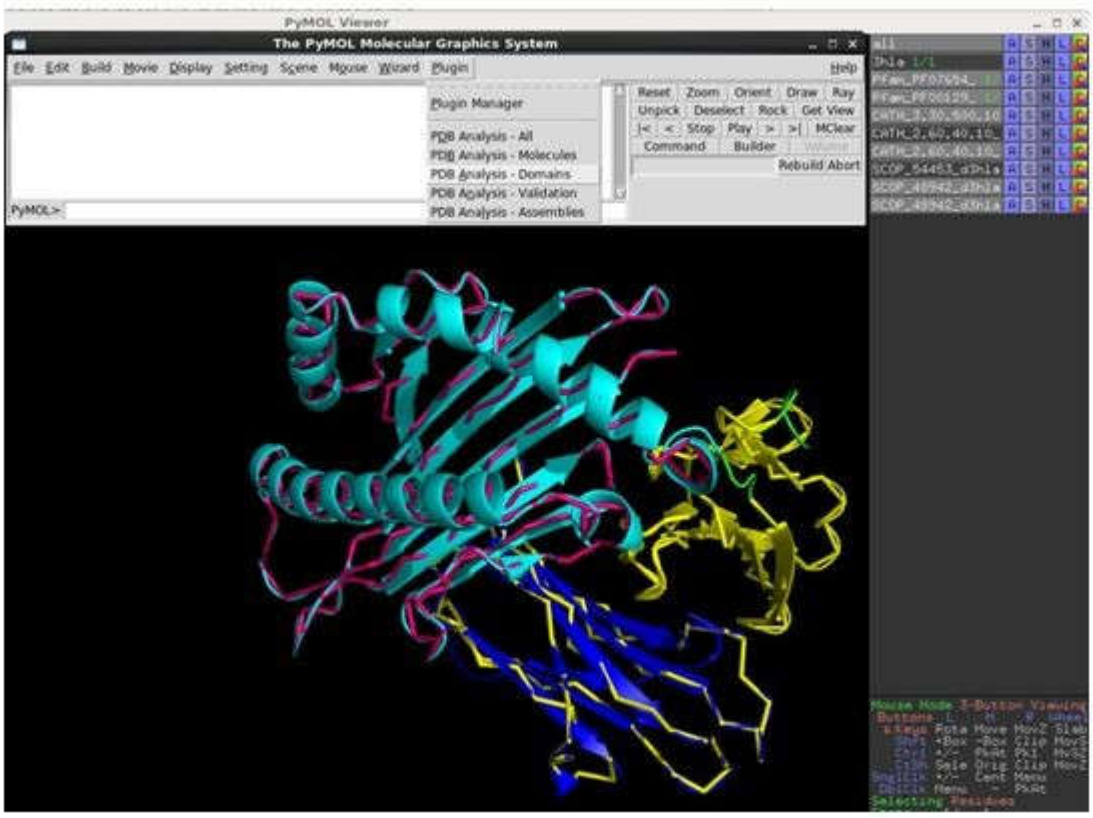

FIGURE 8.1: L'interface du logiciel PyMOL

JMOL

Jmol est un logiciel libre de visualisation de structures chimiques en 3D, gratuit et open source. A l'origine créé pour remplacer un logiciel propriétaire, XMol, qui n'a plus été maintenu, JMol a pour but d'être une interface ouverte de visualisation de molécule libre, ce qui empêchera son abandon à long terme si les développeurs originels venaient à arrêter. Il intègre depuis le projet global OpenScience visant à créer des logiciels libres à but scientifique, et ambitionne de remplacer la solution propriétaire chime. Il est principalement utilisé par les étudiants, les professeurs et les chercheurs en chimie et en biologie structurale. Il est développé en Java et est multi-plateformes. Ainsi, il fonctionne sous Windows, Mac OS X, Linux et les systèmes Unix. Le logiciel est disponible sous trois formes :

- Une application indépendante Jmol qui fonctionne sur le bureau ;
- Un kit logiciel pour intégrer Jmol dans d'autres applications Java ;
- Une applet Java qui peut être intégrée au sein de page Web et qui offre de nombreuses possibilités de visualiser des molécules.

L'interface principale du logiciel JMOL est illustrée dans la Figure 8.2.

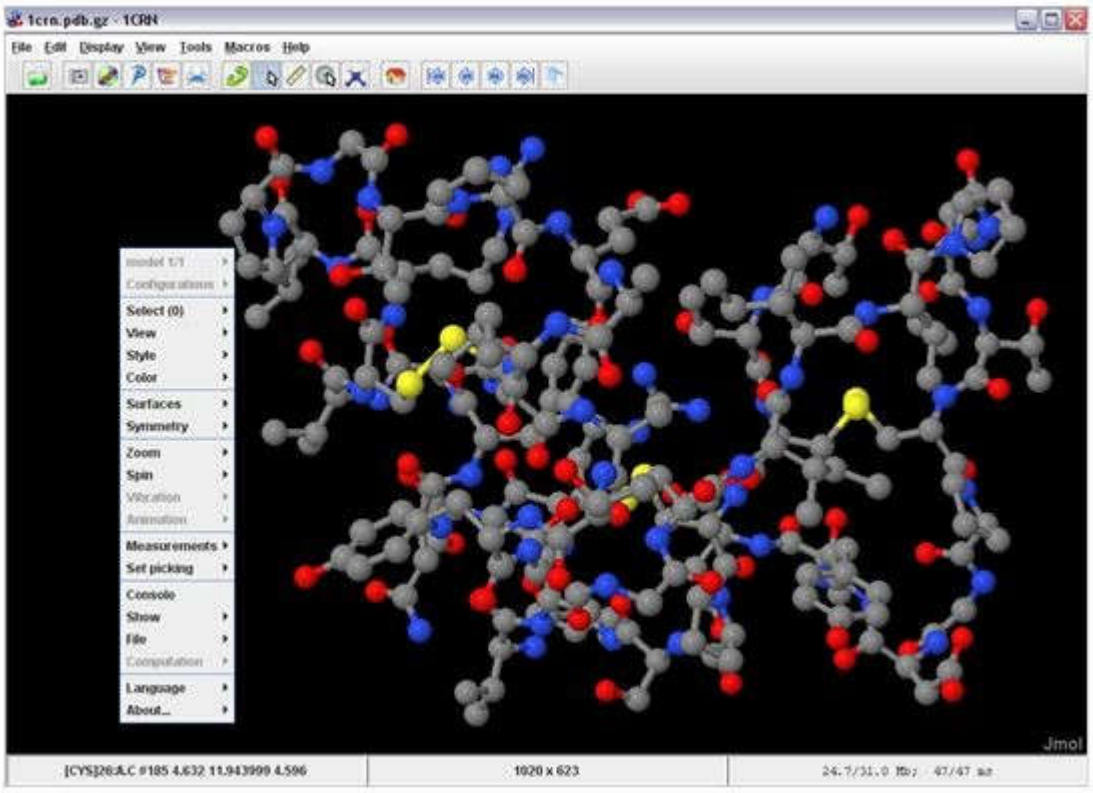

FIGURE 8.2: L'interface du logiciel JMOL

RASMOL

RasMol est un logiciel de graphisme moléculaire gratuit permettant de représenter des molécules en 3D à partir de différents formats standards de fichiers de représentation de molécules. Il fonctionne sur plateformes MacOS, Windows et Linux.

Il utilise les possibilités des cartes graphiques 3D pour faire tourner en temps réel dans l'espace les molécules importées, puis éventuellement le nouvel ensemble de molécules ainsi assemblé.

L'interface principale du logiciel RASMOL est illustrée dans la Figure 8.3.

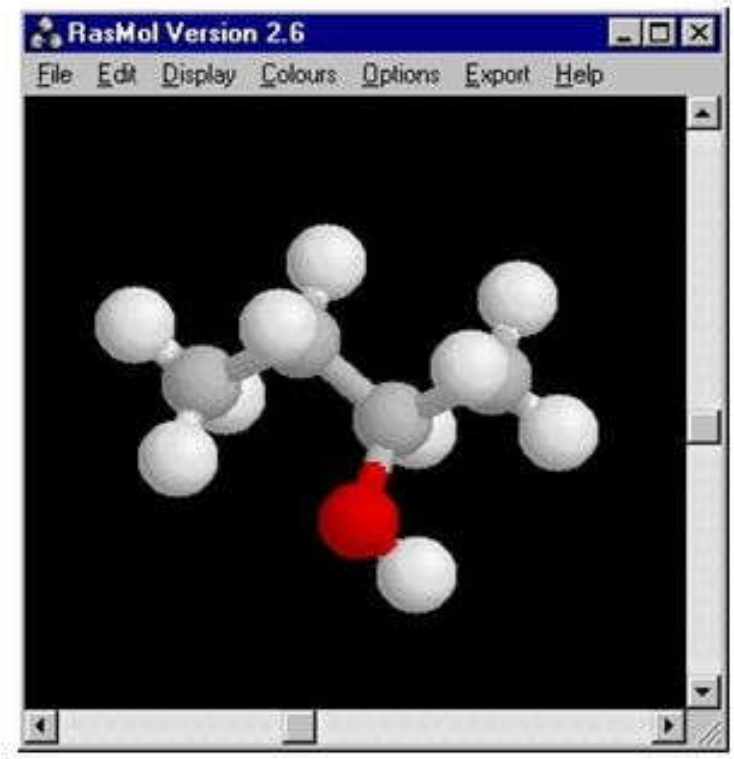

FIGURE 8.3: L'interface du logiciel RASMOL

Conclusion

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'ob- tenir une reconstruction tridimensionnelle présentant la meilleure adéquation avec les résultats expérimentaux. Ces données sont issues principalement d'analyses cristallo- graphiques (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) expé- rimentales 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 mo- dèle, puis dans l'analyse du modèle par la visualisation moléculaire.