Gilbert Deléage
Manolo Gouy
Alexandre de Brevern

Bioinformatique

De la séquence à la structure des protéines

Cours et cas pratiques

3e ÉDITION

DUNOD

11, rue Paul Bert, 92240 Malakoff
www.dunod.com

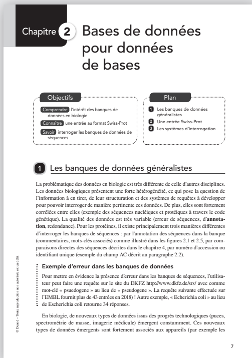
ISBN 978-2-10-081515-9

Table des matières

Comment utiliser cet ouvrage ?

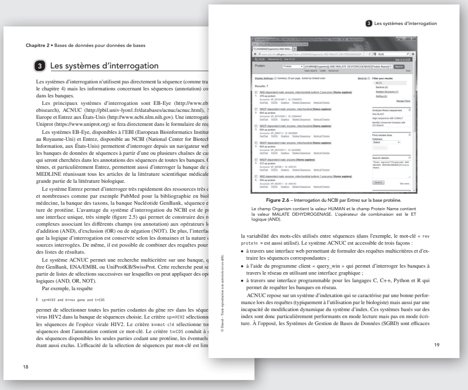
1 La page d'entrée de chapitre

Elle donne le plan du cours ainsi qu'un rappel des objectifs pédagogiques du chapitre.



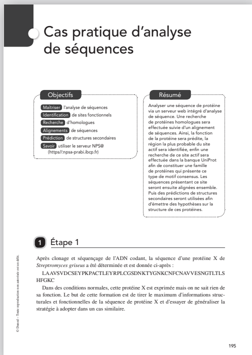
2 Le cours

Le cours, concis et structuré, expose le programme.



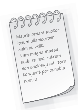
3 Cas pratique

Proposé en fin d'ouvrage, avec les réponses aux questions, pour appliquer les notions du cours.



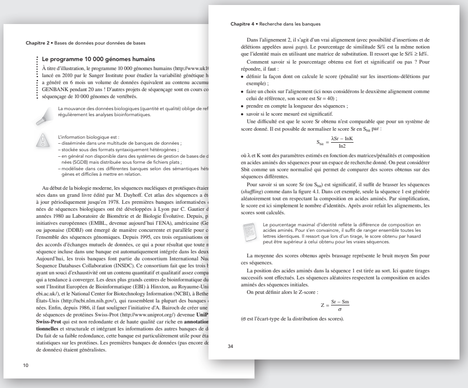
4 Les rubriques

Une erreur à éviter.



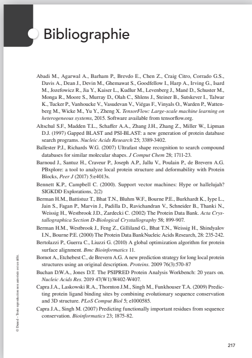
Une note, un complément d'information.

Un éclairage sur une notion (résumé, conseil, etc.)



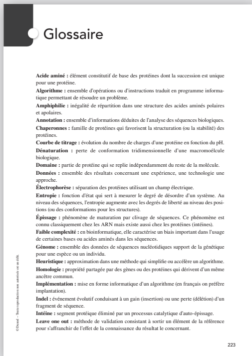
5 La bibliographie

Elle regroupe les articles fondateurs de la discipline.



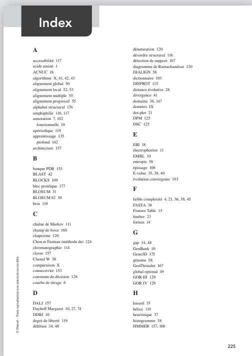
6 Le glossaire

Vous y trouverez les définitions des principales notions développées.



7 L'index

Outil indispensable pour trouver rapidement ce que l'on cherche.



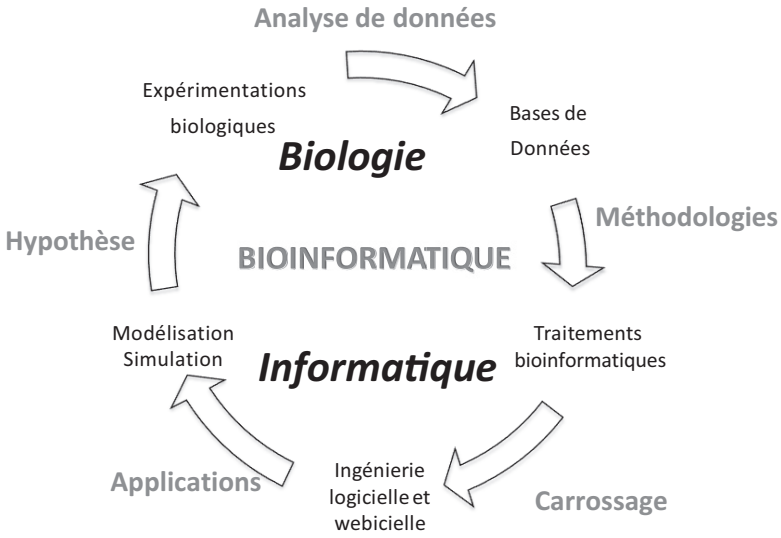
Avant-propos

La bioinformatique est une « interdiscipline » à la frontière de la biologie, de l'informatique et des mathématiques. Les systèmes biologiques sont très complexes et les techniques modernes d'investigation du monde biologique fournissent une vaste quantité de **données** expérimentales. Le but ultime de la bioinformatique est d'intégrer ces données d'origines très diverses pour modéliser les systèmes vivants afin de comprendre et prédire leurs comportements (biologie systémique ou biologie des systèmes) dans des conditions de fonctionnement normales ou pathologiques. Ainsi, à titre d'exemple, le séquençage à très haut débit offre la possibilité de connaître de manière personnalisée le **génome** de chacun. Pour tirer le bénéfice de cette connaissance, il faut développer et appliquer de nouvelles méthodes d'analyse bioinformatique qui permettent d'extraire l'information utile cachée dans la séquence du génome et, de manière plus générale, des données biologiques à grande échelle issues des progrès de l'expérimentation et des technologies de l'automatique. La bioinformatique est donc étroitement couplée à ses applications. Bon nombre de bioinformaticiens ne travaillent pas dans des laboratoires formellement estampillés « bioinformatique ». La bioinformatique et la modélisation procèdent selon un cercle vertueux (schématisé page suivante) dans lequel le point de départ est l'expérimentation biologique (un séquençage par exemple), les données produites sont ensuite organisées dans des dépôts de données (banques ou bases de données). Les méthodes d'analyse qui utilisent ces données sont développées par les bioinformaticiens souvent en association avec des informaticiens et mathématiciens. Pour que ces méthodes permettent le traitement ultérieur des données, il est nécessaire de « carrosser » ces méthodes (sous forme de logiciels ou serveurs web) afin de permettre au biologiste de les utiliser pour émettre de nouvelles hypothèses qui seront testées et qui généreront de nouvelles données.

Aujourd'hui tout projet de biologie comporte une étape d'analyse bioinformatique des données. Par conséquent, un biologiste passe environ 20-30 % de son temps à utiliser des outils bioinformatiques. D'ailleurs, il est remarquable de voir que les deux articles les plus cités en biologie sont ceux décrivant BLAST et CLUSTAL, deux méthodes bioinformatiques largement utilisées par les biologistes.

Ce livre décrit de manière simple les tâches courantes de la bioinformatique qu'un biologiste/biochimiste doit savoir traiter par lui-même sans avoir recours au spécialiste afin de répondre à des questions usuelles comme :

- Comment extraire des informations pertinentes dans les banques de données biologiques ?
- Est-ce qu'une nouvelle séquence a déjà été complètement ou partiellement répertoriée ?
- Est-ce que ce gène appartient à une famille connue ?

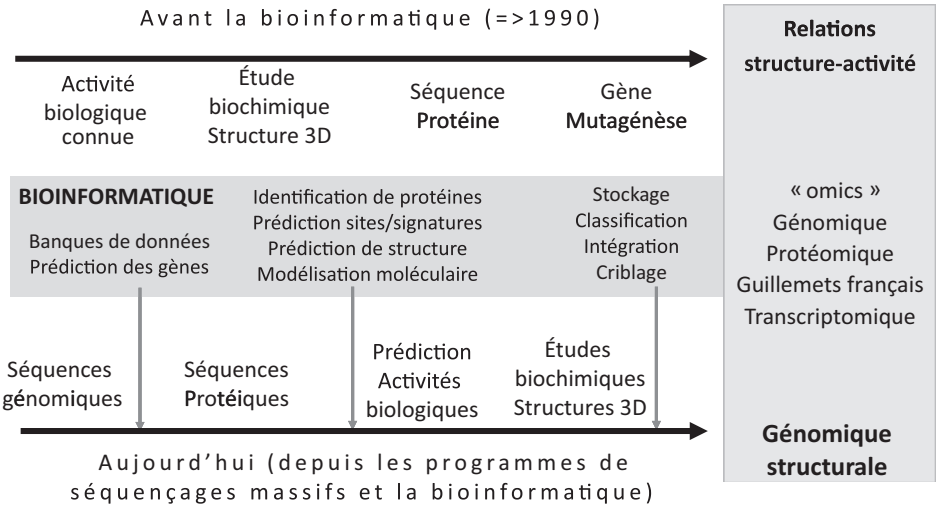


- Existe-t-il d'autres gènes homologues ?
- Est-ce que deux séquences correspondent à deux gènes homologues ?
- Existe-t-il des résidus essentiels à la fonction ?
- Alignement multiple, quel outil ? Pour quoi faire ? Établissement de consensus.
- Quelle peut être la fonction d'une protéine (prédit d'après sa séquence, sa structure...) ?
- Recherche de sous-motifs communs à un ensemble de séquences.
- Recherche de régions contenant des séquences répétées.
- Recherche d'hélices ou de brins dans les protéines.
- Comment construire un modèle tridimensionnel de protéine ?
- Optimisation et comparaison de structures 3D.
- Quelle est la charge globale d'une protéine à un pH donné ?

Ce livre n'a pas la prétention d'être exhaustif (il se limite d'une manière générale aux protéines, mais les **algorithmes** sont souvent très proches de ceux développés pour les acides nucléiques). Il a été rédigé afin de faciliter la compréhension des approches, méthodes, algorithmes et **implémentations** les plus courantes en bioinformatique moléculaire et structurale. À ce titre, il est parfois simplificateur et doit être considéré comme une introduction à la bioinformatique moléculaire et structurale. Il s'adresse donc aux étudiants de biologie/biochimie, de niveau licence, master ou classes préparatoires, ou bien aux biologistes qui souhaitent s'initier et comprendre les méthodes sous-jacentes aux programmes afin d'estimer la qualité de leurs analyses.

La logique suivie dans le livre est de partir des séquences de protéines pour aller vers leurs structures secondaires, leurs structures tridimensionnelles et finir par leurs fonctions. Elle suit la stratégie actuelle d'analyse d'une question biologique qui a été revisitée du fait de l'avènement de la bioinformatique et des séquençages massifs. La

bioinformatique moléculaire a pour première mission de « faire parler cette séquence » pour en tirer le maximum d'informations selon le schéma suivant :



La plupart des images des structures 3D présentées ont été générées à l'aide du logiciel AnTheProt pour Windows (http://antheprot-prabi.ibcp.fr).

En fin d'ouvrage, un exercice de mise en pratique de l'analyse de séquence est fourni avec son corrigé. De même, un exercice avec corrigé concerne la modélisation moléculaire par homologie. Un quizz en ligne est disponible à l'adresse suivante : https://www.gdeleage.fr/prof/bio_info.php?choix=BIOINFORMATIQUE.

Les auteurs remercient Christophe Combet et Céline Brochier pour la relecture du livre.

Chapitre 1 La composition en acides aminés

Objectifs

- **Savoir** calculer la masse d'une protéine
- **Savoir** tracer une courbe théorique de titrage d'une protéine
- **Prédire** le pHi d'une protéine

Plan

1 Acides aminés et séquence

Les protéines naturelles sont constituées d'**acides aminés** de série L de structure chimique générale donnée dans la figure 1.1.

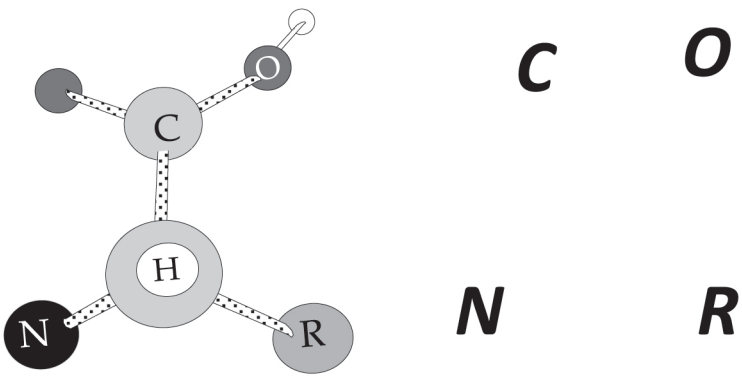


Figure 1.1 – Structure chimique d'un acide aminé de série L.

Il existe 20 acides aminés principaux dans les protéines naturelles. La correspondance entre les acides aminés, leur abbréviation et leur structure chimique est donnée dans la figure 1.2. Avec les ambiguïtés (Aspartate/Asparagine, Glutamate/Glutamine) et lorsque l'acide aminé est inconnu, ce sont au total 25 lettres qui sont utilisées (la lettre O désigne la pyrolysine et U la sélénocystéine).

A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic Acid	Asp
E	Glutamic Acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
O	Pyrrolysine	Pyl
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
U	Sélénocystéine	Sec
V	Valine	Val
W	Tryptophane	Trp
Y	Tyrosine	Tyr
B		
Z		
X	Inconnu	

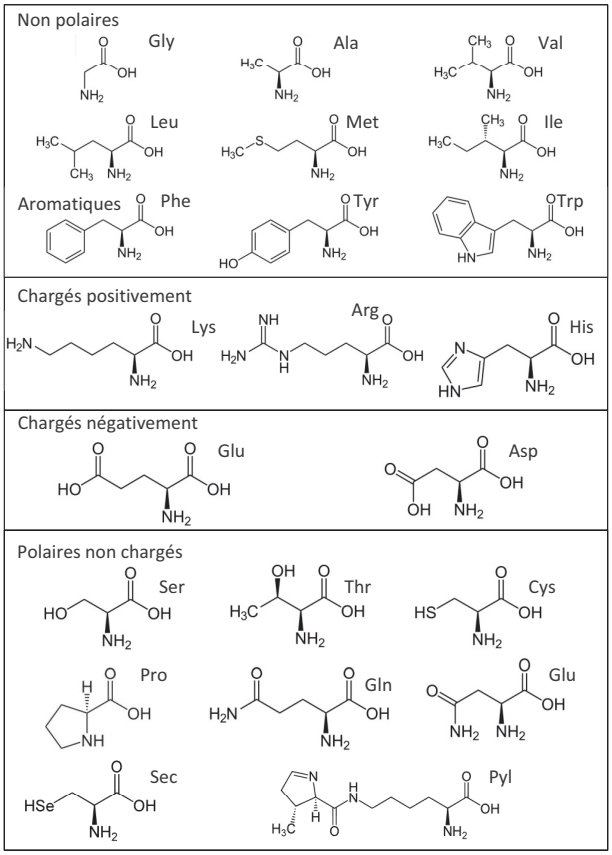


Figure 1.2 – Correspondance entre CODE 1 lettre, CODE 3 lettres et la structure chimique des acides aminés trouvés dans les protéines.

Pour identifier la série d'un acide aminé, il suffit de regarder le Cα avec le H devant les autres atomes. On doit pouvoir lire « CORN » comme illustré dans la figure 1.1.

Certains acides aminés partagent des propriétés physico-chimiques avec d'autres. Cela conduit à une distribution des groupes d'acides aminés selon le diagramme (non exclusif) de Venn schématisé figure 1.3.

Au niveau chimique, les protéines sont obtenues par condensation des acides aminés et élimination d'eau lors de la formation de la liaison peptidique (pour chaque acide aminé ajouté). La suite des lettres indiquant l'enchaînement des acides aminés constitue la **séquence** de la protéine (on parle aussi de **structure primaire**). Chaque séquence caractérise de manière unique une protéine. Une infime partie des séquences théoriquement possibles existe vraiment. Ce sont celles qui ont été sélectionnées par l'évolution et qui sont douées d'une activité biologique (structurelle et/ou fonctionnelle).

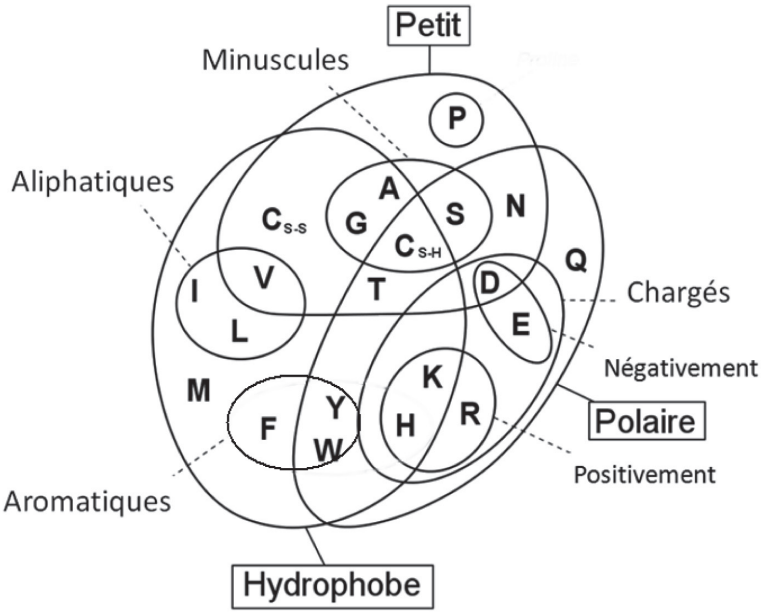


Figure 1.3 – Diagramme de Venn des propriétés des acides aminés.

Le génome humain comprend $3{,}4 \cdot 10^9$ bases et coderait pour 20 563 séquences protéiques.

La bioinformatique s'est emparée très tôt de la comparaison des séquences. En effet, au sens informatique, il s'agit principalement de comparer des mots entre eux, rechercher des mots communs, trouver le plus grand mot commun, aligner les mots en autorisant des « jokers » à certaines positions.

Le nombre de séquences de longueur 100 réalisable à partir de 20 acides aminés différente (20^{100}) est supérieur au nombre d'atomes dans l'Univers ($\sim 10^{80}$).

Encart 1.1 Combien de séquences protéiques différentes peut-on générer en théorie ?

Le nombre de séquences différentes de longueur N qu'il est possible de générer en prenant les 20 acides aminés principaux est 20^N.

Exemples :

Peptide (5 acides aminés) : 20^5

Protéine de taille standard moyenne de 400 acides aminés : 20^{400}

Protéome humain (soit ~20 000 protéines de longueur moyenne 400) : $20^{\,8\,000\,000}$

2 Informations déduites de la composition en acides aminés

La première information dérivable d'une séquence est la composition en acides aminés. Cette composition (nombre et pourcentage de chacun des acides aminés) peut aussi être obtenue expérimentalement par des méthodes d'analyse biochimiques.

Si la composition en acide aminé d'une protéine X est biaisée par rapport à la composition moyenne de l'ensemble des protéines, on dit que la protéine X présente une **faible complexité**. Cette faible complexité peut aussi ne concerner qu'une partie de la séquence. Ainsi, dans certains récepteurs stéroïdiens, on observe jusqu'à 37 glutamines consécutives constituant un cas extrême de faible complexité.

Tableau 1.1 – Les pKa des acides aminés ionisables.

I	pKa i	j	pKa j
His	6,00	**Ser**	13,60
Arg	12,48	**Tyr**	10,10
Lys	10,53	**Glu**	4,20
N_{ter}	9,80	**Thr**	13,60
		Asp	3,86
		C_{ter}	2,10
		Cys	8,33

La composition permet au biochimiste de calculer la masse moléculaire théorique M de la protéine en utilisant la relation suivante :

$$M = \sum_{i=1}^{N} m(i) - 18 \times (N - 1)$$

où $m(i)$ est la masse moléculaire de l'acide aminé i et N le nombre d'acides aminés. Connaissant la composition en acides aminés, le coefficient ε_{280} d'extinction molaire à 280 nm se calcule grâce à la relation suivante :

$$\varepsilon_{280} = [N_{Trp} \times 5\,500] + [N_{Tyr} \times 1\,490] + [N_{Cys} \times 125]$$

Il est alors possible de doser précisément par spectrophotométrie (densité optique) la concentration en protéine grâce à la relation de Beer-Lambert :

$$DO_{280} = \varepsilon_{280}\, L\, C$$

où L est la longueur du trajet optique, C la concentration en g/l.

Enfin, le pI (ou **point isoélectrique** d'une protéine) correspond à la valeur de pH telle que $NC = 0$ dans la relation suivante :

$$NC = \sum_i Ni\left(1 - \frac{10^{-\mathrm{pka}(i)}}{10^{-\mathrm{pka}(i)} + 10^{-\mathrm{pH}}}\right) - \sum_j Nj\left(\frac{10^{-\mathrm{pka}(j)}}{10^{-\mathrm{pka}(j)} + 10^{-\mathrm{pH}}}\right)$$

NC est le nombre de charges théoriques portées par la protéine.

i désigne un résidu qui peut être chargé positivement (Arg, Lys, His) ayant un pKa(i).

j désigne un résidu qui peut être chargé négativement Asp, Glu, Tyr, Cys, Ser, Thr ayant un pKa(j).

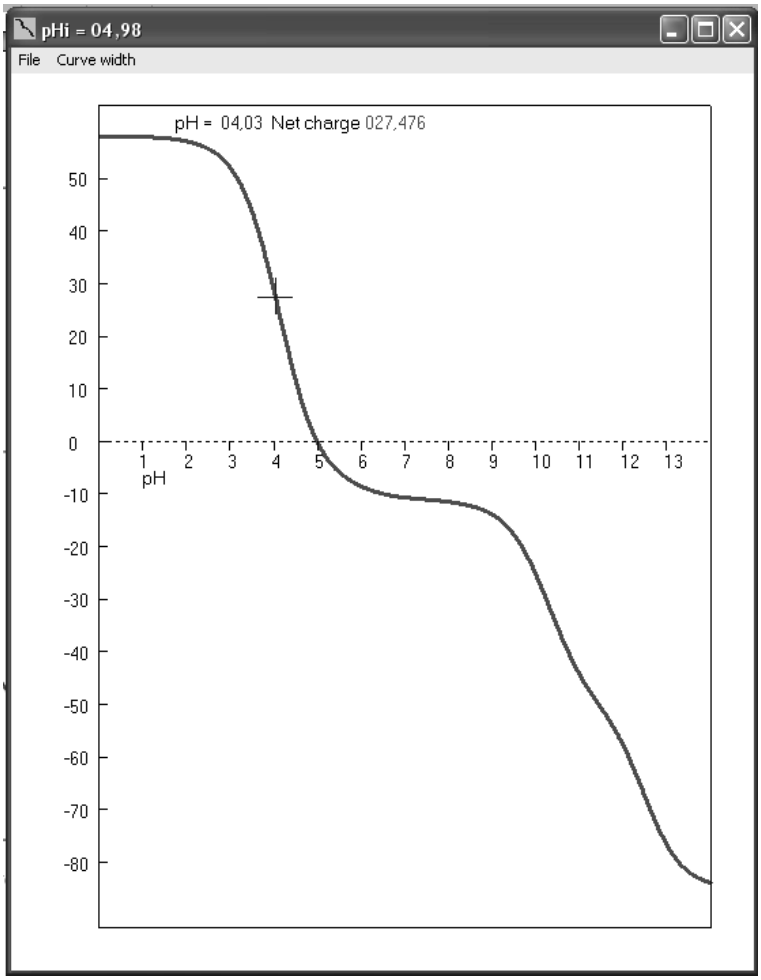


Figure 1.4 – Courbe de titrage théorique d'ATPA_TOBAC.

La courbe représente le nombre de charges théoriques portées par la protéine en fonction du pH. Le point isoélectrique est le pH pour lequel la charge nette est égale à 0 (ici 4,98).

À partir de cette relation, il est possible de calculer la **courbe de titrage** théorique (NC) = f (pH) d'une protéine. Cette information même très approximative est très utile au biochimiste avant de se lancer dans une purification de protéine car la physico-chimie des solutions fait que solubilité d'une protéine est minimale quand le pH de la solution est égal au pHi. Par ailleurs, la connaissance du pHi d'une protéine permet de choisir une colonne de purification de type échangeuse d'ions qui soit adaptée aux conditions de pH utilisées pendant la purification.

Chapitre 2 Bases de données pour données de bases

Objectifs

- **Comprendre** l'intérêt des banques de données en biologie
- **Connaître** une entrée au format Swiss-Prot
- **Savoir** interroger les banques de données de séquences

Plan

1 Les banques de données généralistes

La problématique des données en biologie est très différente de celle d'autres disciplines. Les données biologiques présentent une forte hétérogénéité, ce qui pose la question de l'information à en tirer, de leur structuration et des systèmes de requêtes à développer pour pouvoir interroger de manière pertinente ces données. De plus, elles sont fortement corrélées entre elles (exemple des séquences nucléiques et protéiques à travers le code génétique). La qualité des données est très variable (erreur de séquences, d'**annotation**, redondance). Pour les protéines, il existe principalement trois manières différentes d'interroger les banques de séquences : par l'annotation des séquences dans la banque (commentaires, mots-clés associés) comme illustré dans les figures 2.1 et 2.5, par comparaisons directes des séquences décrites dans le chapitre 4, par numéro d'accession ou identifiant unique (exemple du champ AC décrit au paragraphe 2 de ce chapitre).

Exemple d'erreur dans les banques de données

Pour mettre en évidence la présence d'erreur dans les banques de séquences, l'utilisateur peut faire une requête sur le site de l'EBI https://ebi.ac.uk avec comme mot-clé « psuedogene » au lieu de « pseudogene ». La requête suivante effectuée sur l'EMBL fournit plus de 43 entrées en 2018) ! Autre exemple, « Echerichia coli » au lieu de Escherichia coli retourne 34 réponses.

En biologie, de nouveaux types de données issus des progrès technologiques (puces, spectrométrie de masse, imagerie médicale) émergent constamment. Ces nouveaux types de données émergents sont fortement associés aux appareils (par exemple les

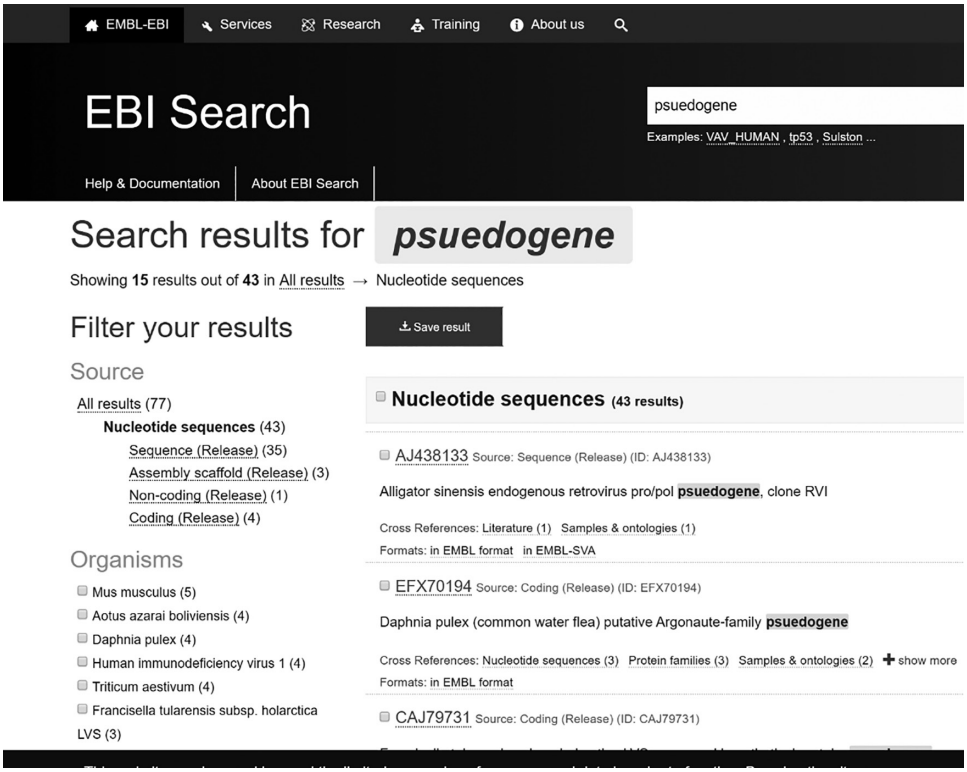


Figure 2.1 – Interrogation de banques nucléiques sur le serveur de l'EBI avec le mot-clé « psuedogene ».

puces Affymetrix ou les appareils de spectrométrie de masse) et aux auteurs qui les produisent, ce qui génère des formats de données différents et le plus souvent incompatibles car souvent liés à des constructeurs d'appareils.

De plus, le volume des données en biologie (en particulier les séquences) croît de manière exponentielle et double tous les 18 mois depuis 1982 imposant au bioinformaticien de refaire périodiquement les analyses. Cette croissance pourtant déjà considérable est bien inférieure à celle liée aux séquençages massifs. Tout d'abord, les programmes de séquençage massif de génomes complets font exploser les volumes acquis. De plus, les capacités d'obtention de séquences par

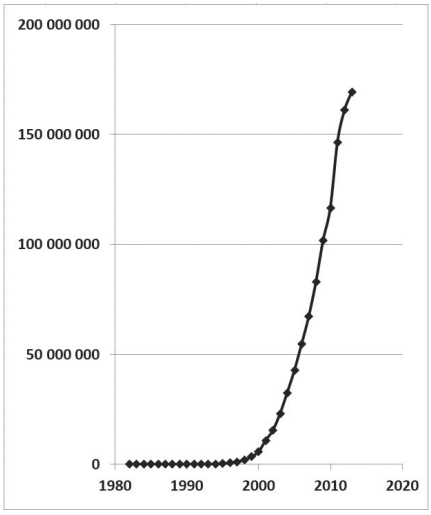


Figure 2.2 – Progression du nombre de séquences dans GENBANK.

les nouvelles techniques de séquençage (NGS ou *Next Generation Sequencing*) sont telles que les coûts des séquençages ont été divisés par 500 000 en vingt ans. À titre de comparaison, en 2001, le coût de 1 génome humain était de 100 M$, en 2019 il est de 700$! (source https://www.genome.gov/). Les nouvelles méthodes de séquençages illustrées dans la figure 2.3 présentent des caractéristiques de taille de séquence, de longueur de lecture (« read »), de temps d'obtention et de degré de parallélisation différents. Toutes ces méthodes permettent de générer un grand nombre de fragments de longueur variable selon la technologie qui seront assemblés par bioinformatique pour finalement donner la séquence (cas d'un petit génome) ou pour positionner la séquence sur un génome de référence.

Technologie	: Pyroséquençage	Fluorescence	Solide
Parallélisation	: $4\ 10^5$	$3\ 10^7$	$5\ 10^7$
Nucléotides par lecture	: ~400	~50	35
Temps d'obtention	: 8 H	144 H	240 H
Longueur de séquence	: $\sim 5\ 10^8$	$\sim 4\ 10^9$	$\sim 2\ 10^{10}$

Figure 2.3 – Next Generation Sequencing (NGS).

Il faut souligner que ces technologies progressent tant sur le plan de la longueur des « read » que sur le degré de parallélisation. Les coûts ont aussi chuté au point que dans un avenir proche, la séquence du génome complet d'un humain coûtera environ 500 €, ce qui ouvre des perspectives de médecine personnalisée mais pose aussi des questions éthiques importantes. Par ailleurs, de nouvelles approches sont en cours de développement (Ion Proton ou GridION™) et permettent une miniaturisation encore plus grande du système (MinIOn USB) et une parallélisation par empilement des unités (comme pour les calculateurs). Dans cette course aux génomes, il ne faut pas perdre de vue que séquencer n'est pas déchiffrer.

Par ailleurs, la sémantique et la représentation d'un concept ou d'une notion varient selon la culture scientifique, ce qui est une difficulté pour une interdiscipline. Ainsi, la définition même de ce qu'est une protéine est différente pour l'informaticien (qui voit souvent un mot), pour un biologiste (qui voit un intermédiaire dans une chaîne fonctionnelle), pour le biochimiste (qui y associera une activité enzymatique) et pour un chimiste (qui y associera un assemblage d'un grand nombre d'atomes).

Le programme 10 000 génomes humains

À titre d'illustration, le programme 10 000 génomes humains (http://www.uk10k.org/) lancé en 2010 par le Sanger Institute pour étudier la variabilité génétique humaine a généré en 6 mois un volume de données équivalent au contenu accumulé dans GENBANK pendant 20 ans ! D'autres projets de séquençage sont en cours comme le séquençage de 10 000 génomes de vertébrés.

La mouvance des données biologiques (quantité et qualité) oblige de refaire régulièrement les analyses bioinformatiques.

L'information biologique est :
– disséminée dans une multitude de banques de données ;
– stockée sous des formats syntaxiquement hétérogènes ;
– en général non disponible dans des systèmes de gestion de bases de données (SGDB) mais distribuée sous forme de fichiers plats ;
– modélisée dans ces différentes banques selon des sémantiques hétérogènes et difficiles à mettre en relation.

Au début de la biologie moderne, les séquences nucléiques et protéiques étaient déposées dans un grand livre édité par M. Dayhoff. Cet atlas des séquences a été remis à jour périodiquement jusqu'en 1978. Les premières banques informatisées de données de séquences biologiques ont été développées à Lyon par C. Gautier dans les années 1980 au Laboratoire de Biométrie et de Biologie Évolutive. Depuis, plusieurs initiatives européennes (EMBL, devenue aujourd'hui l'ENA), américaine (GenBank) ou japonaise (DDBJ) ont émergé de manière concurrente et parallèle pour collecter l'ensemble des séquences génomiques. Depuis 1995, ces trois organisations ont passé des accords d'échanges mutuels de données, ce qui a pour résultat que toute nouvelle séquence incluse dans une banque est automatiquement intégrée dans les deux autres. Aujourd'hui, les trois banques font partie du consortium International Nucleotide Sequence Databases Collaboration (INSDC). Ce consortium fait que les trois banques ayant un souci d'exhaustivité ont un contenu quantitatif et qualitatif assez comparable et qui a tendance à converger. Les deux plus grands centres de bioinformatique du monde sont l'Institut Européen de Bioinformatique (EBI) à Hinxton, au Royaume-Uni (http://ebi.ac.uk/), et le National Center for Biotechnology Information (NCBI), à Bethesda aux États-Unis (http://ncbi.nlm.nih.gov/), qui rassemblent la plupart des banques de données. Enfin, depuis 1986, il faut souligner l'initiative d'A. Bairoch de créer une banque de séquences de protéines Swiss-Prot (http://www.uniprot.org/) devenue **UniProtKB/Swiss-Prot** qui est non redondante et de haute qualité car riche en **annotations fonctionnelles** et structurale et intégrant les informations des autres banques de données. Du fait de sa faible redondance, cette banque est particulièrement utile pour établir des statistiques sur les protéines. Les premières banques de données (pas encore des bases de données) étaient généralistes.

Différence entre base de données et banque de données

Une banque de données est un ensemble de fichiers textes sans relation entre eux (on parle de fichier « plat »). Une base de données est un ensemble de relations entre des données gérées avec un système de gestion de base de données (SGBD) et interrogeable par SQL (*Structure Query Langage*).

Depuis 25 ans, une explosion des bases de données spécialisées est observée (1641 répertoriées dans NAR).

La revue *Nucleic Acids Research* consacre un numéro spécial « database » chaque année (http://www.oxfordjournals.org/nar/). Avant de se lancer dans un nouveau projet, il convient de vérifier qu'il n'existe pas une banque spécialisée maintenue à jour.

Les bases de données spécialisées présentent l'avantage d'être maintenues par des experts du domaine qui gèrent les problèmes de numérotation, nomenclature, cohérence, annotation. On peut distinguer les bases de données thématiques biologiques (récepteurs couplés aux protéines G comme GPCR, ou immunologie IMGT), par organisme (dont le génome est en général complètement séquencé), par technologie (spectres RMN, cartes de spectrométrie de masse, gels d'**électrophorèse** bidimensionnelle) ou par type (séquence, structure, image, spectre, interaction). Le tableau 2.1 recense quelques ressources notoires en bioinformatique.

L'accès aux génomes se fait grâce à des outils dédiés appelés *genome browser*. Le serveur Ensembl (www.ensembl.org) répertorie les principaux génomes d'organismes modèles. Le serveur offre la possibilité de naviguer depuis le niveau caryotype (figure 2.4) jusqu'au niveau de la séquence nucléique et de sa traduction dans les différentes phases de lecture.

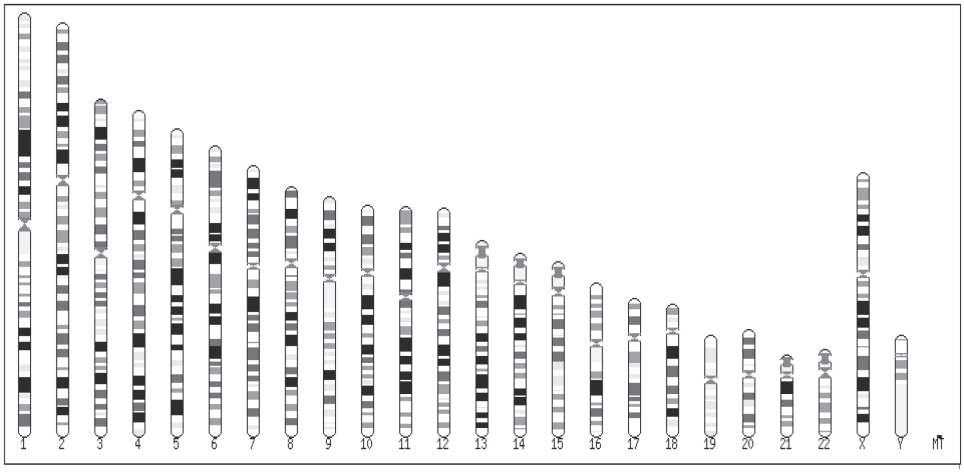


Figure 2.4 – Caryotype humain sur Ensembl (23 chromosomes).

II existe une seule banque de données des structures 3D des macromolécules biologiques appelée historiquement la PDB (*Protein Data Bank*). Cette banque (http://www.rcsb.org/) contient les coordonnées tridimensionnelles atomiques de protéines, d'acides nucléiques, de complexes nucléo-protéiques, de sucres. La croissance de la banque est constante depuis 4 ans et représente environ 10 000 structures/an en moyenne sur la période 2016-2020. En revanche, le nombre de structures présentant une architecture originale (**repliement** ou *fold*) est constant. Ainsi, le nombre de repliements différents connus est d'environ 1 500 et représente la redondance en structures 3D. La redondance en séquence fait qu'on peut distinguer environ 35 707 groupes de séquences qui partagent entre eux moins de 30 % d'identité. Ainsi, il existe des versions de PDB à 95 % (PDB95), 75 % (PDB75) et 25 % (PDB25).

Les données biologiques sont fortement biaisées. À titre d'illustration, même si plus de 13 919 espèces sont représentées dans **UniProtKB/Swiss-Prot (2020 03)**, seulement 20 espèces couvrent 21,6 % des entrées (voir tableau 2.2).

Tableau 2.1 – Quelques bases de données spécialisées.

Acronyme	Description
IMGT	IG, récepteur de cellules T, Complexe Majeur d'Histocompatibilité
HIV	Base de séquences sur le SIDA à Los Alamos
GPCRDB	Récepteurs couplés aux protéines G
euHCVdb	Base de données de séquences du virus de l'hépatite C
OMIM	*Online Mendelian Inheritance in Man*
HGMD	*Human Gene Mutation Database*
KEGG	Kyoto Encyclopedia of Genes and Genomes
ENZYME	Nomenclature des enzymes
BRENDA	Base de connaissance sur les enzymes
NRSub	*Bacillus subtilis*
AceDB	*Caenorhabditis elegans*
FlyBase	*Drosophila melanogaster*
GOLD	Banque des génomes séquencés.
RCSB	Base de données des structures des macromolécules biologiques
IntAct	Base de données d'interactions protéiques
BIND	Base d'interactions
MiMI	Banque d'interactions moléculaires du Michigan
STRING	Banque d'interactions entre protéines
CATH	Banque de classification des structures de protéines
SCOP	Classification structurale des protéines
Ensembl	Explorateur de génomes complets d'organismes modèles.
NucleaRDB	Système d'information pour les récepteurs nucléaires

Tableau 2.2 – Le top 20 des séquences par espèce représentée dans UniProtKB.

N°	Nombre	Nom de l'espèce
1	20 368	*Homo sapiens* (Human)
2	17 042	*Mus musculus* (Mouse)
3	15 983	*Arabidopsis thaliana* (Mouse-ear cress)
4	8 106	*Rattus norvegicus* (Rat)
5	6 721	*Saccharomyces cerevisiae* (Baker's yeast)
6	6 012	*Bos taurus* (Bovine)
7	5 140	*Schizosaccharomyces pombe* (Fission yeast)
8	4 518	*Escherichia coli* (strain K12)
9	4 191	*Bacillus subtilis*
10	4 149	*Dictyostelium discoideum* (Slime mold)
11	4 129	*Caenorhabditis elegans*
12	4 081	*Oryza sativa subsp. japonica* (Rice)
13	3 608	*Drosophila melanogaster* (Fruit fly)
14	3 451	*Xenopus laevis* (African clawed frog)
15	3 158	*Danio rerio* (Zebrafish) (*Brachydanio rerio*)
16	2 295	*Gallus gallus* (Chicken)
17	2 219	*Pongo abelii* (Sumatran orangutan)
18	2 204	*Mycobacterium tuberculosis* (strain ATCC 25618/H37Rv)
19	2 042	*Escherichia coli* O157:H7
20	1 898	*Mycobacterium tuberculosis* (strain CDC 1551/Oshkosh)

Tableau 2.3 – Nombre de structures 3D déposées dans la PDB (25/06/2020).

Méthode	Protéines	Acides nucléiques	Complexes Prot/A.Nuc.	Autres	Total
Cristallographie rayons X	137 970	2 060	6 648	473	147 151
Résonance Magnétique Nucléaire	11 400	1 299	264	49	13 012
Cryo-microscopie électronique	3 956	35	1 101	121	5 213
Méthodes hybrides	159	5	3	1	168
Autres	99	3	0	4	106
Total	**153 584**	**3 402**	**8 016**	**648**	**165 650**

La diversité des banques et des logiciels de traitement de données à conduit à la création de plusieurs formats. Des formats sont adaptés pour les logiciels de traitement de séquences (exemple format Pearson-Fasta) car ils sont économiques en taille