

**Gilbert Deléage
Manolo Gouy**

Bioinformatique

Cours et cas pratique

DUNOD

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements

d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour

les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée.

Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du

droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).



© Dunod, Paris, 2013
ISBN 978-2-10-058751-3

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2^o et 3^o a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

TABLE DES MATIÈRES

Comment utiliser cet ouvrage	VI
Avant-propos	IX
Chapitre 1 • La composition en acides aminés	1
1.1 Acides aminés et séquence	1
1.2 Informations déduites de la composition en acides aminés	4
Chapitre 2 • Bases de données pour données de bases	7
2.1 Les banques de données généralistes	7
2.2 Une entrée SWISS-PROT	14
2.3 Les interrogations Entrez, ACNUC, SRS	17
Chapitre 3 • La comparaison de deux séquences	21
3.1 Matrice de points	21
3.2 Matrice de substitution	26
Chapitre 4 • Recherche dans les banques	33
4.1 Score de similitude entre séquences	33
4.2 Recherche globale ou locale	36
4.3 FASTA	37
4.4 BLAST	41
Chapitre 5 • Alignement de séquences	47
5.1 Introduction	47
5.2 Comparaison de protéines homologues (algorithme global)	49
5.3 Meilleur chevauchement entre séquences (algorithme local)	52
5.4 Alignements multiples	54
5.5 Représentation « logo »	57
Chapitre 6 • Bases théoriques de la phylogénie moléculaire	59
6.1 Arbres phylogénétiques	59
6.1.1 Arbres racinés et arbres non racinés	61
6.1.2 Le format Newick d'arbres phylogénétiques	62
6.2 Arbre des espèces – arbres de gènes	63
6.2.1 Nombre d'arbres binaires possibles	64
6.3 Modèle markovien de l'évolution moléculaire	65

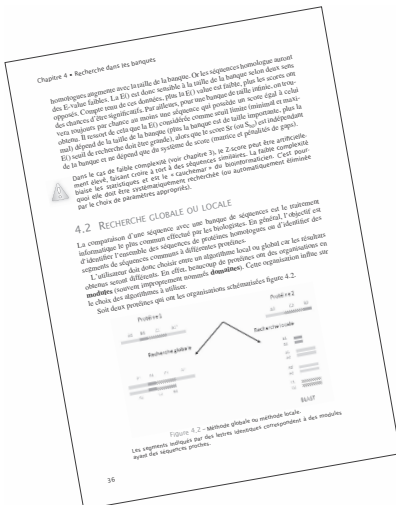
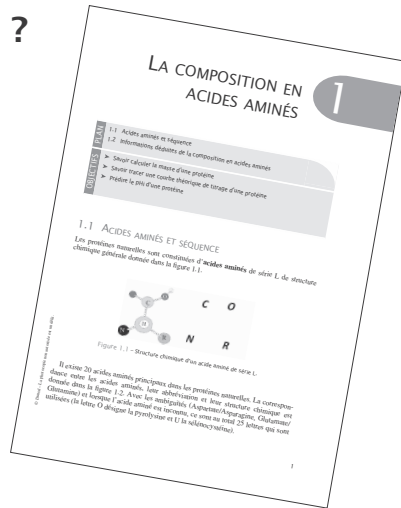
Table des matières

6.3.1	Matrice de transition	67
6.3.2	Quelques modèles nucléotidiques de Markov	68
6.3.3	Longueur d'une branche	70
6.3.4	Modélisation de la variation des taux d'évolution entre sites	70
6.4	Choix des sites	72
6.5	Matrices de taux de substitution entre séquences protéiques	73
6.6	Distances évolutives entre paires de séquences	74
Chapitre 7 • Algorithmes pour la phylogénie moléculaire		77
7.1	Parcimonie	78
7.1.1	Algorithme	78
7.1.2	Heuristiques	82
7.1.3	Propriétés	82
7.1.4	Implémentations	83
7.1.5	Longueurs de branches des arbres de parcimonie	83
7.1.6	Traitement des indels	84
7.2	Méthodes de distances	85
7.2.1	Méthode d'évolution minimale	86
7.2.2	Méthode Neighbor-Joining	86
7.3	Maximum de vraisemblance	89
7.3.1	Modèle probabiliste utilisé au maximum de vraisemblance	90
7.3.2	Calcul de la vraisemblance	90
7.3.3	L'algorithme de Felsenstein	91
7.3.4	Prise en compte de la variabilité des vitesses d'évolution entre sites	92
7.3.5	Optimisation de la vraisemblance	93
7.3.6	Implémentation	94
7.4	Estimation de la fiabilité d'un arbre par bootstrap	94
7.5	Choix des méthodes de calcul d'arbres	97
Chapitre 8 • Recherche de fonctions		99
8.1	Définitions	99
8.2	Détection de signatures de séquence (PROSITE)	100
8.3	Recherche de fonction avec pondération par la fréquence	103
8.4	Méthodes à base de profils	106
Chapitre 9 • Profils physico-chimiques		111
9.1	Pourquoi les profils physico-chimiques ?	111
9.2	Hydrophobie-paramètres-construction du profil - interprétation	111
9.3	Amphiphilie	114
9.4	Accessibilité au solvant	115
Chapitre 10 • Prédictions de structures secondaires		117
10.1	Méthode « statistique empirique »	120
10.2	Méthode information directionnelle (GOR)	123
10.3	Méthode de recherche des plus proches voisins (NNM)	127

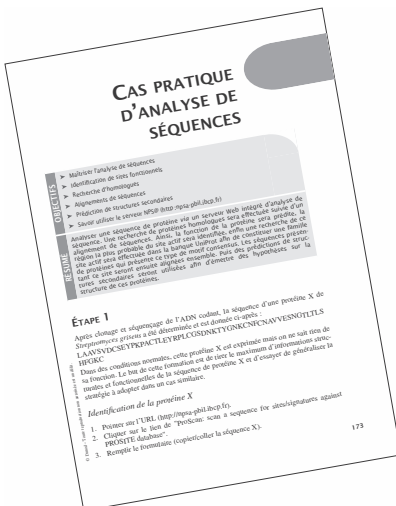
10.4 Méthode auto-optimisée (SOPM)	130
10.5 Méthode auto-optimisée avec alignements (SOPMA)	131
10.6 Méthodes neuronales	132
10.7 Autres méthodes	134
10.7.1 Méthode statistique discriminante (DSC)	134
10.7.2 Méthode neuronale (PREDATOR)	134
10.7.3 Méthode hiérarchisée réseaux de neurones (HNN)	134
10.7.4 Méthodes utilisant les chaînes de Markov	135
10.7.5 Combinaison de méthodes	135
10.8 Critères de qualité prédictive	137
Chapitre 11 • Prédiction de structures 3D	139
11.1 Principe des méthodes de détermination expérimentale	139
11.2 Le format PDB	140
11.3 Les différents modes de représentations	142
11.4 Classification de structures 3D	145
11.5 Comparaison de structures 3D	146
11.6 Énergétique moléculaire	148
11.7 Optimisation de structures 3D	151
11.8 Modélisation de structures 3D	152
11.8.1 Les méthodes d'enfilage des repliements (<i>threading</i>)	153
11.8.2 Modélisation par homologie	154
11.8.3 Les alphabets structuraux	163
11.8.4 Les méthodes de novo	166
Chapitre 12 • Détection de sites 3D dans les protéines	169
12.1 Problématique	169
12.2 Méthode SuMO	170
Cas pratique d'analyse de séquences	173
Conclusion	183
Glossaire	185
Bibliographie	187
Index	193

Comment utiliser cet ouvrage ?

La page d'entrée de chapitre
Elle donne le plan du cours ainsi qu'un rappel des objectifs pédagogiques du chapitre.



Le cours
Le cours, concis et structuré, expose le programme.



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

Les rubriques



Une erreur à éviter.



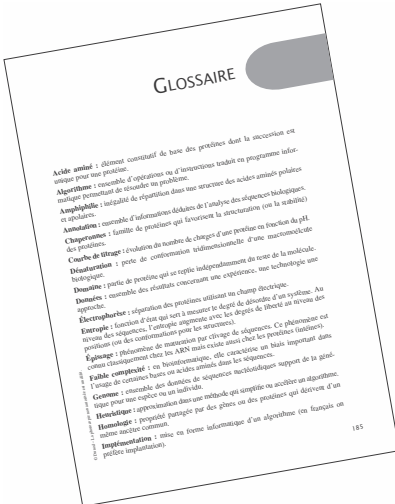
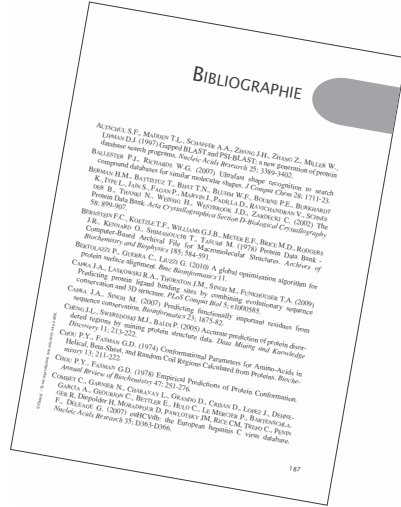
Une note, un complément d'information.



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

La bibliographie

Elle regroupe les articles fondateurs de la discipline.



Le glossaire

Vous y trouverez les définitions de principales notions développées.



L'index

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

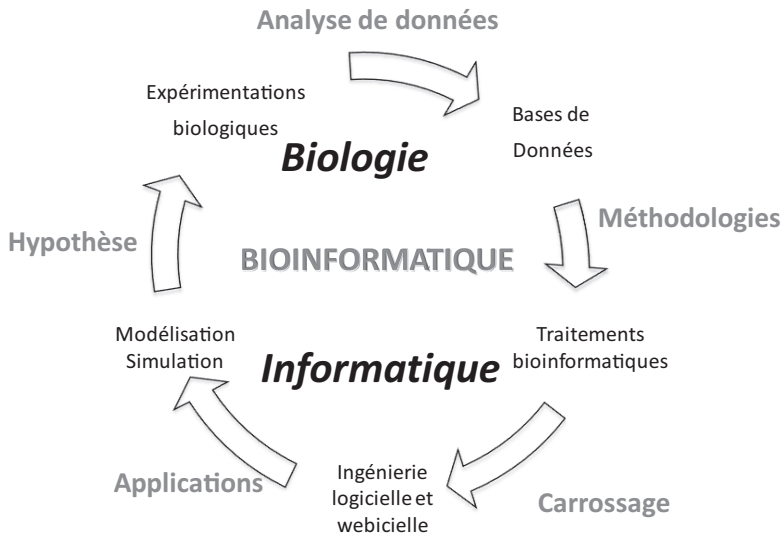
AVANT-PROPOS

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 **génom**e 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.

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 ?

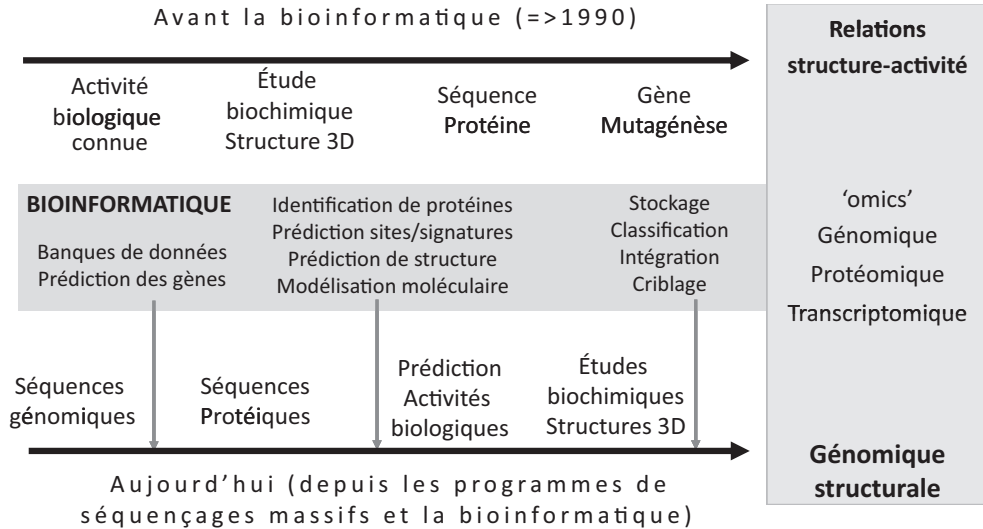


- Existe-t-il d'autres gènes homologues ?
- Est-ce que deux séquences correspondent à deux gènes homologues ?
- Existe-t-il des résidus essentiels à la fonction ?
- Alignement multiple, quel outil ? Pour quoi faire ? Établissement de consensus.
- Quelle peut être la fonction d'une protéine (prédit d'après 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 :



Un exercice de mise en pratique de l'analyse de séquence est fourni avec son corrigé (chapitre 13).

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

Les vidéos fournies dans le complément numérique (www.dunod.com) ont été capturées à l'aide du logiciel CAMSTUDIO (<http://camstudio.org/>). Un quiz en ligne est disponible à l'adresse suivante : https://publi.ibcp.fr/scripts/bio_info.php.

Les auteurs remercient Christophe Combet et Céline Brochier pour leur relecture.

LA COMPOSITION EN ACIDES AMINÉS

1

PLAN

- 1.1 Acides aminés et séquence
- 1.2 Informations déduites de 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

1.1 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.

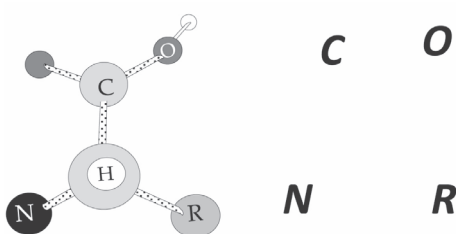


Figure 1.1 – Structure chimique d'un acide aminé de série L.

Il existe 20 acides aminés principaux dans les protéines naturelles. La correspondance entre les acides aminés, leur abréviation 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 pyrrolysine et U la sélénocystéine).

Chapitre 1 • La composition en acides aminés

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	Pyrrrolysine	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		Asn/Asp
Z		Gln/Glu
X	Inconnu	

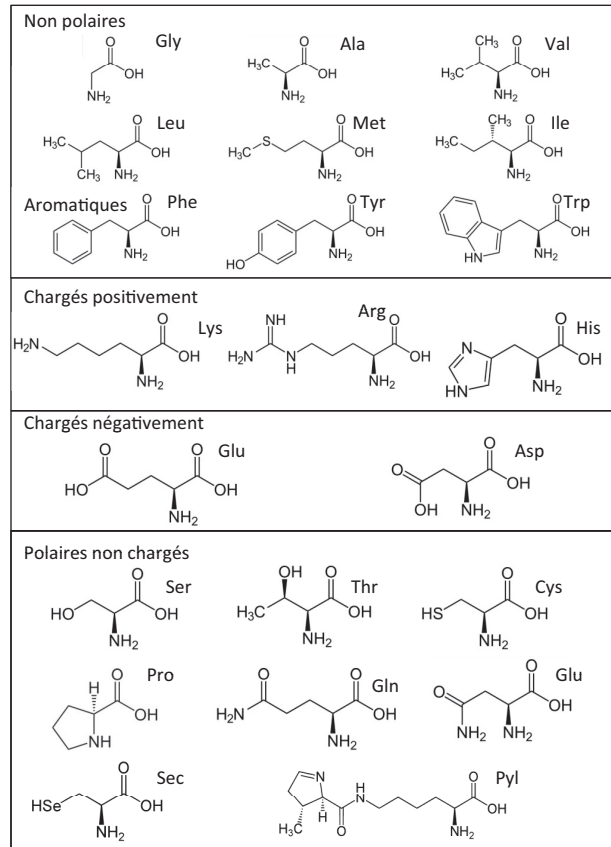


Figure 1.2 - Correspondance entre CODE 1 lettre, CODE 3 lettres et la structure chimique des acides aminés trouvés dans les protéines.

Pour identifier la série d'un acide aminé, il suffit de regarder le C α 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 (structurale et/ou fonctionnelle).

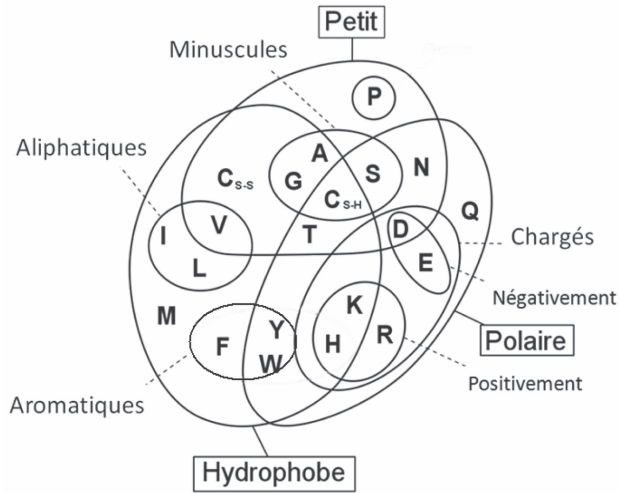


Figure 1.3 - Diagramme de Venn des propriétés des acides aminés.



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 différents (20^{100}) est supérieur au nombre d'atomes dans l'Univers ($\sim 10^{80}$).



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 $\sim 20\,000$ protéines de longueur moyenne 400) : $20^{8\,000\,000}$

1.2 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 :

$$\epsilon_{280} = [N_{\text{Tyr}} \times 5\,500] + [N_{\text{Tyr}} \times 1\,490] + [N_{\text{Cys}} \times 125].$$

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

$$DO_{280} = \epsilon_{280} L C$$

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 :

$$NC = \sum_i N_i \left(1 - \frac{10^{-\text{pKa}(i)}}{10^{-\text{pKa}(i)} + 10^{-\text{pH}}} \right) - \sum_j N_j \left(\frac{10^{-\text{pKa}(j)}}{10^{-\text{pKa}(j)} + 10^{-\text{pH}}} \right)$$

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

1.2 • Informations déduites de la composition en acides aminés

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

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

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 pH_i . Par ailleurs, la connaissance du pH_i 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.

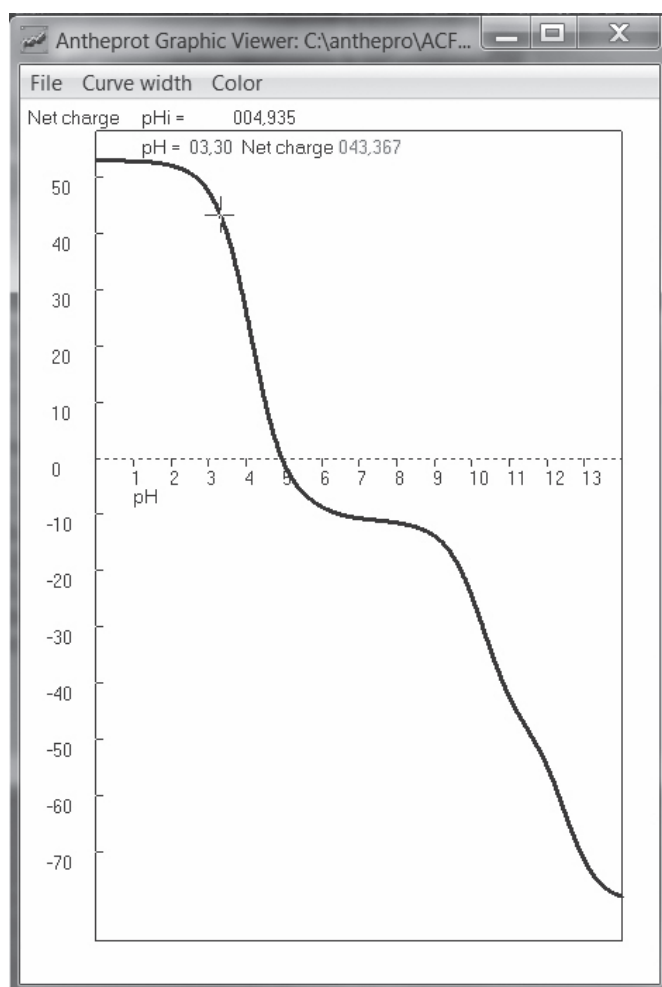


Figure 1.4 - Courbe de titrage théorique d'ATPA_TOBAC.

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