Cour 04 : Alignement et structure 3D de protéines

I. Alignement

Introduction

Au cours de l'évolution naturelle, les mutations causent des erreurs au moment de la réplication de l'ADN car l'évolution se fait par mutations successives. Ces erreurs peuvent être :

- Des substitutions (changement ponctuel d'un nucléotide par un autre). On parle de transition ou de transversion,
- Des insertions (ajout d'un ou plusieurs nucléotides),
- Des délétions (suppression d'une base ou d'un segment d'ADN).

Il en découle alors des différences, plus ou moins importantes, dans les structures (primaire, secondaire, …) de ces séquences, d'où la divergence et la biodiversité des espèces.

En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines…) repose essentiellement sur la notion de l'alignement1, et permet de déterminer le degré de ressemblance entre celles-ci (similitude ou identité en révélant des régions proches dans leurs séquences primaires). Cela peut alors indiquer que :

- La structure (primaire, secondaire ou tertiaire) des deux séquences est semblable,

- La fonction biologique est proche ou différente (dans le cas de la dissémilarité),

- L'origine des séquences alignées est commune ou éloignée (notion d'homologie), …

Cependant, la comparaison pour l'obtention d'un alignement optimal entre deux séquences biologiques, nécessite néanmoins la mise en œuvre de procédures de calcul (algorithmes) et de modèles biologiques permettant de quantifier la notion de ressemblance entre ces séquences.

1. Définitions

Alignement : processus par lequel deux (ou n) séquences sont comparées afin d'obtenir le plus de correspondances (identités ou substitutions conservatives) possibles entre les lettres qui les composent.

Alignement local : alignement des séquences sur une partie de leur longueur

Alignement global : alignement des séquences sur toute leur longueur

Alignement optimal : alignement des séquences qui produit le plus haut score possible

Alignement multiple : alignement global de trois séquences ou plus Brèches ou "gap" : espace artificiel introduit dans une séquence pour contre-balancer et matérialiser une insertion dans une autre séquence. Il permet d'optimiser l'alignement entre les séquences.

indel : "in" = insertion **"del"** = délétion

Similarité : c'est le pourcentage d'identités et/ou de substitutions conservatives entre des séquences. Le degré de similarité est quantifié par un score. Le résultat de la recherche d'une similarité peut être utilisé pour inférer l'homologie de séquences.

Homologie : 2 séquences sont homologues si elles ont un ancêtre commun. mésappariement : non correspondance entre deux lettres. Un mésappariement peut être : soit la substitution d'un caractère par un autre, c'est-à-dire une mutation soi l'introduction d'un "gap"

Score : un score global permet de quantifier l'homologie. Il résulte de la somme des scores élémentaires calculés sur chacune des positions en vis à vis des deux séquences dans leur appariement optimal. C'est le nombre total de "bons appariements" pénalisé par le nombre de mésappariements.

2. TRAITEMENT DES SEQUENCES NUCLEIQUES (ADN ou ARN)

Notion de score : Le score élémentaire (noté "s") est une entité numérique que l'on attribue à chaque couple de nucléotides des deux séquences à comparer. Il prend la valeur de 1 lorsque les deux nucléotides des deux séquences sont identiques, et la valeur de zéro sinon. Exemple :

Séquence1	A	G	C	T	A	C	C	T	G	T	Score global : Total des scores
Séquence1	A	A	G	T	A	G	C	T	T	T	
Point de comparaison	1	2	3	4	5	6	7	8	9	10	
Score élémentaire (s)	1	0	0	1	1	0	1	1	0	1	1+0+0+1+1+0+1+1+0+1=6

Dans cet exemple, constatez qu'au niveau du premier point de comparaison (ou site de comparaison), les deux séquences contiennent le même nucléotide A, donc le score élémentaire (s) à ce point prend la valeur de 1 (s = 1).

Au deuxième point de comparaison, la séquence 1 contient un G et la séquence 2 contient un A. Elles sont donc différentes en ce point d'où un score élémentaire de zéro (s = 0)…

Au 10ème point de comparaison, les deux séquences contiennent le même nucléotide T donc un score élémentaire de 1.

Constatons que la somme des scores élémentaires est égale à six (s = 6). Donc il y a six points identiques entre les deux séquences ; soit 60% d'identité entre les deux séquences ([6/10] x100). On dit alors que le score global entre les deux séquences est égal à six. Le score a donc permis de quantifier la ressemblance entre les deux séquences.

La relation entre le score global (S) et les scores élémentaires (s) pour deux séquences est de la forme :

$$S = \sum si\ n\ i{=}1$$

3. Alignement pair

Si une nouvelle séquence est obtenue à partir du séquençage génomique, la première étape est la recherche de similarités avec des séquences connues dans d'autres organismes. Si la fonction/structure des séquences similaires/protéines est connue, très probablement (highly likely) la nouvelle séquence correspond à une protéine avec la même fonction/structure. En effet, il a été trouvé que seulement à peu près 1% des gènes humains n'ont pas de contrepartie dans le génome de souris et que la moyenne de similarité entre les gènes de la souris et de l'homme est de 85%.

Les similarités existent parce que toutes les cellules possèdent une cellule ancêtre commune (a mother cell). Donc, dans les différents organismes il pourrait avoir des mutations d'acides aminés dans certaines protéines parce que les acides aminés ne sont pas tous importants pour la fonction et peuvent être remplacés par des acides aminés qui ont des caractéristiques chimiques semblables sans changer la structure. Parfois les mutations sont tellement nombreuses qu'il est difficile de trouver des similarités.

La méthode du calcul des fonctions des gènes par similarités est appelée la génomique comparative ou la recherche d'homologie. Deux séquences sont homologues lorsqu'ils ont comme racine un ancêtre commun.

- **Recherche de segments identiques : La matrice de points**

Elle permet une vue (méthode visuelle) englobant les similarités entre les régions des séquences à comparer.

Exemple de réalisation : On donne deux séquences x et y :

x=ACTCGGATT et y=AGCTCGGT

Cette méthode consiste à créer une matrice qui va contenir les deux séquences (la séquence x en horizontal et la séquence y en vertical) et de cocher les cases de cette matrice pour le seul cas où les nucléotides sont identiques (Match). Quand il n'ya pas identité on parle de Mismatch:

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	G					X	X			
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X			
	T			X					X	X

Sur cette matrice, constatons qu'il y a une diagonale formée de cinq cases. Donc le segment identique le plus long entre les deux séquences x et y contient cinq nucléotides identiques et consécutifs qui sont: **CTCGG**

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A									
	G									
	C		X							
	T			X						
	C				X					
	G					X				
	G						X			
	T									

Remarque : Dans le cas où les deux séquences sont complètement identiques, le résultat est une diagonole :

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X	X		
	A	X						X		
	T			X					X	X
	T			X					X	X

- **Méthodes d'alignement par paire des acides nucléiques et aminés**

L'alignement pair est performé par l'utilisation des méthodes suivante :

1. Matrice d'analyse Dot

La matrice d'analyse Dot, est décrit pour la première fois par Gibbs et McIntyre (1970), est premièrement une méthode de comparaison de deux séquences pour trouver des alignements possibles entre ces deux dernières. La méthode est aussi utilisée pour rechercher les répétitions directes ou inversées dans les séquences d'ADN et de protéine, la prédiction des régions auto-complémentaire dans les séquences d'ARN et la prédiction de la structure secondaire par appariement de bases. L'avantage majeur de cette méthode est d'examiner le meilleur alignement qui apparait en diagonal.

2. les programmations dynamiques (dynamic programming or DP algorithm)

La programmation dynamique est une méthode de calcul utilisée pour aligner deux séquences d'ADN ou protéine. La méthode est très importante car elle apporte le meilleur score ou "optimal", les algorithmes fournissent une méthode de calcul fiable pour aligner les séquences. Elle compare chaque paire de nucléotide et génère des résidus appariés et non appariés et aussi des gaps dans les deux séquences, donc le nombre des résidus appariés donne le meilleur score possible, la méthode était prouvé mathématiquement de point de vue optimisation de score entre deux séquences.

3. Méthode Word and K-tuple

La méthode permet d'aligner les séquences rapidement par la recherche des fragment dans les deux séquences qui se ressemblent (nommé words ou K-tuples) puis les rassembler par un programmé dynamique.

4. Alignement multiples

Un alignement multiple consiste à comparer plusieurs séquences. On parvient à ce type d'alignement en considérant successivement tous les alignements possibles deux à deux. En pratique, on compare une séquence à toutes les autres pour tenter de déterminer la voie la plus probable dans l'évolution, étant donné les probabilités de différentes substitutions possibles. Plus on ajoute de séquences à l'alignement multiple, plus le modèle acquiert de la précision sur l'histoire évolutive des familles des séquences comparées. Cet alignement permet de montrer que certains résidus sont identiques dans toutes les séquences. Tout résidu ou toute séquence courte identique dans toutes les séquences d'un groupe donné est dite conservée. Les alignements multiples donnent en général une meilleure évaluation de la similitude que les alignements deux à deux et permettent d'identifier des membres ayant une relation lointaine dans une famille de gènes, qui n'auraient pas été révélés par un alignement deux à deux cet alignement consiste à superposer chaque résidu d'une séquence avec ceux d'une ou plusieurs autres séquences. De plus, pour construire un alignement optimal, on aura souvent besoin d'ajouter des indels (gaps).

5. système de scoring des séquences moléculaires

Le choix de système de Scoring attribue des scores aux résidus identiques (matches), différents (mismatches), substitution, insertion et délétions en séquences d'ADN et protéines. Ce système est appliqué dans les alignements globaux (Needleman- Wunsch algorithm) et les alignements locaux (Smith-Waterman algorithm). Pour l'alignement d'ADN un simple positif score est donné aux résidus identiques et un score négatif est donné aux résidus différents, et les gaps sont pris en considération. D’un autre coté pour scorer les résidus identiques et différents en une séquences de protéines il est important de connaitre comment un acide aminé est substitué par un autre.

6. Programmes de comparaison avec les banques

Recherches de similitudes dans les banques de séquences, Pourquoi ?

• Savoir si ma séquence ressemble à d'autres déjà connues

• Trouver toutes les séquences d'une même famille

• Rechercher toutes les séquences qui contiennent un motif donné

La taille 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.

La plupart de ces programmes constituent des méthodes heuristiques. Leur but est de filtrer les données de la banque en étapes successives car peu de séquences vont avoir des similitudes avec la séquence comparée.

Ces méthodes utilisent certaines approximations pour éliminer rapidement les situations sans intérêt et ainsi repérer les séquences de la banque susceptibles d'avoir une relation avec la séquence recherchée. Ces programmes permettent de calculer un score pour mettre en évidence les meilleures similitudes locales qu'ils ont observées.

Les deux types de programme les plus utilisés par les biologistes qui sont les logiciels :

1. FASTA

Le logiciel regroupe en fait plusieurs programmes de recherche avec les banques de données :

• Le programme FASTA qui compare respectivement une séquence nucléique avec une base nucléique ou une séquence protéique avec une base protéique.

• Les programmes TFASTA ou TFASTX qui comparent une séquence protéique avec des bases nucléiques traduites.

• Les programmes FASTX ou FASTY qui comparent une séquence nucléique traduite avec des bases protéiques.

2. BLAST : Basic Local Alignment Search Tool

BLAST est l'outil de recherche basique d'alignement local. BLAST, cherche les bases de données des protéines et ADNs pour des séquences (sujets) qui ressemblent à notre séquence (requête) utilisée comme mot clé.

Ce logiciel possède en fait plusieurs programmes de comparaison avec les bases de données :

- BLASTN (pour comparer une séquence nucléique contre base nucléique),

- BLASTP (Pour comparer une séquence protéique contre base protéique),

- BLASTX (comparaison de séquence nucléique (traduite en 6 phases) contre base protéique),

- TBLASTN (comparaison de séquence protéique contre base nucléique (traduite en 6 phases)),

- TBLASTX (comparaison de séquence nucléique (traduite dans les 6 phases) contre base nucléique (traduite dans les 6 phases))

II. Structure des protéines

Si l'ADN est le support physique de l'information biologique, la protéine en est le reflet, et à l'échelle moléculaire c'est déjà une véritable machine fonctionnelle, assurant à la fois des fonctions vitales aussi bien structurales que dynamiques. Ainsi une protéine est un polymère linéaire constitué de différentes unités de base, les acides aminés (aa) ou résidus.

Un aa est constitué d'un carbone central (carbone alpha ou Cα) lié à un groupement carboxyle (COOH), à un groupement amine (NH2), à un atome d'hydrogène (H) et à un radical R. Les vingt aa se différencient par la nature de ce radical qui leur confère différentes propriétés telles que la charge, la flexibilité, l'encombrement stérique ou bien encore l'hydrophobicité. La structure d'une protéine peut être décrite à plusieurs niveaux de structures, chacun apportant un type d'informations spécifiques.

1. La structure primaire

Les aa sont reliés entre eux par une liaison peptidique entre le groupement carboxyle COOH d'un résidu et le groupement amine NH2 du résidu suivant. La chaîne ainsi formée est appelée « la chaîne principale ou squelette », alors que les radicaux sont désignés sous le terme de « chaînes latérales ». On parle de peptide lorsque le nombre de résidus est inférieur à 50 et de protéine au-delà. La principale information apportée par la structure primaire est l'ordre ou la succession des aa formant la molécule protéique.

Secondairement, la structure primaire permet les calculs de valeurs spécifiques intrinsèques telles que le pI, le PM, l'hydrophobicité, etc

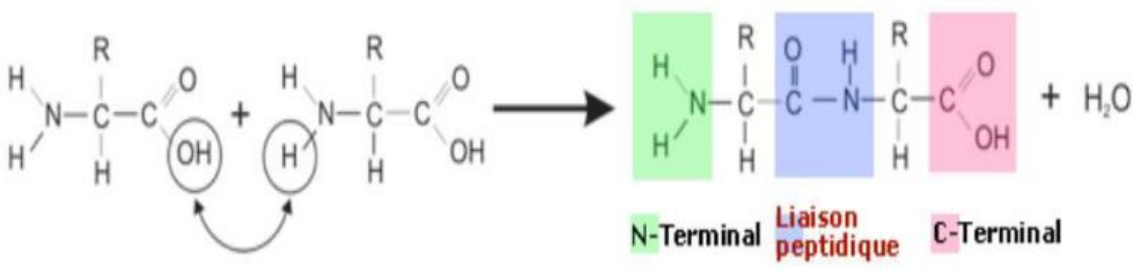

Figure : Formation de la liaison peptidique. La liaison de deux aminoacides est accompagnée de la perte d'une molécule d'eau.

2. Structure secondaire

La structure secondaire correspond aux organisations spatiales régulières de la chaine polypeptidique. Ainsi une protéine peut être décrite par un enchaînement d'éléments de la structure secondaire qui prend des conformations se trouvant nettement favorisées car stabilisées par des liaisons hydrogènes entre les groupements amine (-NH) et carbonyle (-CO) du squelette peptidique. Elle est engendrée par la rotation des atomes de la chaîne peptidique les uns par rapport aux autres au cours de la synthèse de la chaîne in vivo. Les angles possibles et les structures qu'ils engendrent le plus souvent sont représentés sur la table de RAMACHANDRAN.

On reconnaît deux grands types de structure secondaire :

2.1 L'hélice alpha : Lorsque le squelette carboné de la protéine adopte un repliement hélicoïdal périodique, on parle d'hélice alpha. L'hélice α qui fait tourner la chaîne carbonée par rapport à elle-même d'un tour tous les 4 acides aminés environ. Elle est stabilisée par des liaisons hydrogènes entre le carbonyle de la liaison peptidique qui suit l'acide aminé n° 1 avec l'amine de la liaison peptidique qui précède l'acide aminé n° 5, puis de même entre les acides aminés 2 et 6, etc….

- **Tours et boucles**

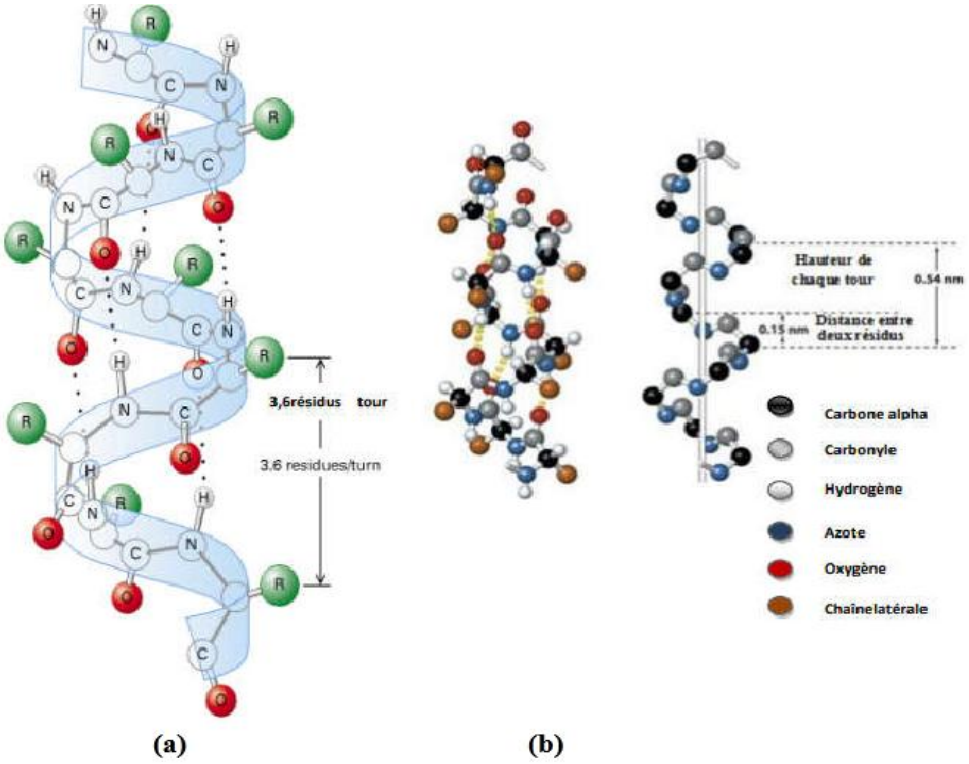

Figure : Représentation de la structure de l'hélice alpha. (a) : structure en ruban ; les radicaux R sont à l'extérieur de la chaîne. (b) : Détails de la structure en boules

2.2 Feuillets plissés β : Contrairement à l'hélice alpha, ce ne sont pas des segments continus d'une unique chaîne polypeptidique, mais des combinaisons de segments différents ne se suivant pas obligatoirement et provenant d'une ou plusieurs chaînes polypeptidiques.

Ces brins β sont arrangés les uns à côté des autres de telle sorte que des liaisons hydrogène puissent se former entre les groupements CO et NH de brins voisins. Les deux brins peuvent être parallèles ou antiparallèles.

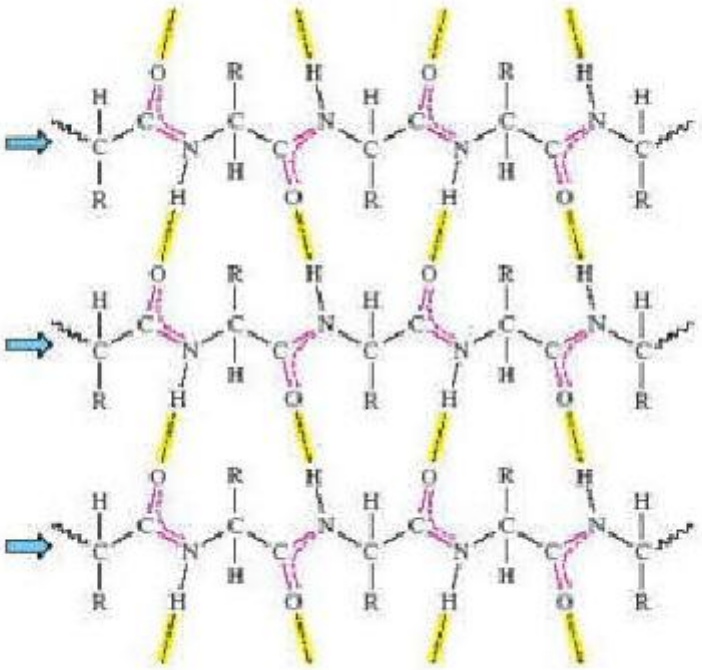

Figure : Représentation du modèle structural de feuillets béta

- **Les tours :** C'est la structure qui connecte deux brins β antiparallèles. Les tours sont généralement courts : 2 à 4 acides aminés en dehors des brins.

- **Les boucles :** Les structures des protéines sont souvent des combinaisons d'hélice et de feuillets reliées par des boucles de longueurs très variables : de 1 à 12 résidus (voire jusqu'à 22) avec le plus fréquemment 1, 3, 4 ou 7 résidus. La comparaison de structures tridimensionnelles montre que les boucles adoptent un nombre limité de conformations.

3. La structure tertiaire

Elle correspond au repliement et à l'assemblage des différents éléments de la structure secondaire. Cette structure correspond en fait à la structure tridimensionnelle (structure 3D) de la protéine. Le repliement et la stabilité des protéines sont guidés par plusieurs forces ; les liaisons hydrogènes, les forces de Van der Waals, les forces électrostatiques et les interactions hydrophobes.

III. Approches bioinformatiques pour la prédiction des structures 2D et 3D des protéines

La séquence primaire des protéines contient toutes les informations qui vont déterminer les structures secondaires et tertiaires des protéines. La fonction d'une protéine dépend en grande partie de sa structure c'est-à-dire de la manière dont se replie la chaînes d'acides aminés

L'identification de la fonction d'une protéine commence par la recherche de similarité de séquence en comparant sa séquence primaire à d'autres séquences (alignements). Si la similarité est significative (> 70%), alors deux postulats peuvent être posés :

- Premièrement, les deux séquences sont homologues, autrement dit, elles sont liées phylogénétiquement.

- Deuxièmement, l'homologie entre deux séquences peut laisser supposer que les protéines ont la même structure voire la même fonction.

La recherche de similarité au sein d'un échantillon de séquences passe par leur alignement multiple. Celui-ci permet de :

- Déterminer si la protéine s'organise en

- Déterminer des motifs conservés

- Retracer son évolution et ses liens phylogénétiques

- Identifier sa fonction

Les programmes d'alignement multiple sont nombreux. Les plus utilisés sont Clustal Oméga, Multalign, Clustal W, Dialign, T-coffee, MAFFT et MUSCLE.

- **Méthodes pour la prédiction de structures secondaires**

La fonction d'une protéine dépend en grande partie de sa structure c'est-à-dire de la manière dont elle se replie au tour d'elle-même. La prédiction de la fonction protéique passe d'abord par les prédictions structurelles 2D (et 3D). Prédire la structure 2D revient à prédire les éléments conformationnels locaux : helices α, feuillets β et coudes.

Les premières méthodes de prédiction de structures secondaires reposaient sur une analyse statistique de la propension des acides aminés à se trouver dans l'un des éléments de structures secondaires. Les méthodes de Chou et Fasman et GOR sont les plus utilisées.

- **Méthodes pour la prédiction des structures tertiaires**

La structure 3D est importante puisqu'elle détermine les propriétés biochimiques et la fonction biologique des protéines. En général, les structures 3D des protéines sont déterminées soit par cristallographie aux rayons-X soit par RMN. Cependant, ces méthodes sont coûteuses et restent un processus long.

Deux types de méthodes sont utilisés pour la prédiction de la structure 3D des protéines : la modélisation comparative et les méthodes dites ab initio.

1. **La modélisation comparative**

- **Modélisation par homologie**

Son principe consiste à aligner la séquence d'une protéine donnée dont la structure est inconnue (cible) avec la séquence d'une ou de plusieurs protéines ayant une structure expérimentale connue (RMN ou Rayons-X) (patrons ou références).

- **Les méthodes de reconnaissance de repliement (Threading)**

Cette méthode est préconisée quand la similarité de séquence entre la cible et les patrons est comprise entre 20 et 30%. Elle consiste à enfiler la séquence de la cible sur une bibliothèque de repliement afin de déterminer les structures qui correspondent le mieux à la séquence sur la base d'un critère énergétique ou de score.

2. **Les méthodes ab initio**

Ces méthodes sont utilisées pour prédire la structure tertiaire de la cible pour des pourcentages d’identité de séquence très faibles à partir de sa séquence primaire en acides aminés, en se basant sur leurs interactions physicochimiques entre les atomes des résidus.

Au sein de cette méthode, on distingue deux types de méthodes dites : ab initio pures, basées uniquement sur des principes physiques, et de novo qui utilisent une batterie d’informations issues de bases de données.