Computational biology project

Rapport

Rédigé par :
Guillaume Tricknot

Haute École en Hainaut

Master en Life Data Technologies

Année académique 2024-2025

Table des matières

Introduction

Ce projet en bioinformatique vise à explorer les structures biomoléculaires via l'analyse, la classification, la prédiction et la modélisation 3D. Il développe des compétences en interprétation scientifique, utilisation d'outils comme PyMOL et LaTeX, et réflexion critique.

1 Partie 1 : Visualisation et analyse de la structure biomoléculaire

1.1 Qualité des structures expérimentales

Les protéines portant les codes PDB **1N0S** et **1T0V** sont des lipocalines modifiées. La différence entre les deux entrées réside dans la méthode expérimentale utilisée pour déterminer leur structure.

1N0S : La structure a été déterminée par diffraction des rayons X, avec une résolution de 2Å. Les valeurs du facteur R sont :

- R-libre : 0,243
- R-travail : 0,193
- R-observé : 0,196

Le facteur R évalue l'adéquation entre le modèle théorique et les données expérimentales. Un R-libre, calculé à partir d'un ensemble distinct de données ($\sim 10\%$), donne une estimation moins biaisée de la qualité du modèle.

1T0V : Cette structure a été déterminée par résonance magnétique nucléaire (RMN). Parmi les 100 conformères calculés, 20 ont été soumis à la Protein Data Bank. Les conformères représentent différentes conformations possibles d'une molécule dues aux rotations autour de certaines liaisons simples.

1.2 Visualisation des structures

En comparant les structures **1N0S** et **1T0V** à l'aide des rapports de validation complets :

- **Diagrammes des propriétés des résidus :** 1N0S montre 72% de résidus sans valeurs aberrantes contre 25% pour 1T0V. Les résidus marqués en rouge (RSRZ $>$ 2) pour 1N0S incluent : `Y59, K65, V172, M173` (chaîne A) et `P35, D120, I150, M173` (chaîne B).
- **Modèle et résidus :** 1N0S omet environ 6% des résidus, souvent aux extrémités. En revanche, 1T0V conserve tous les résidus, bien que certains soient mal définis (RMN).
- **Séquence de résidus consécutifs :** Moins de résidus consécutifs avec des écarts dans 1T0V.
- **Diagramme de Ramachandran :** Aucun résidu aberrant pour 1N0S, mais 8% hors normes pour 1T0V.

Rapports Procheck : Les facteurs G indiquent les propriétés atypiques d'un modèle.

— Pour 1N0S : une seule propriété inhabituelle, moyenne générale : $-0,22$.

— Pour 1T0V : deux propriétés très inhabituelles, moyenne générale : $-0,15$.

Un facteur $G < -0,5$ indique une anomalie, et $G < -1$ signale une anomalie majeure.

Visuel de 1N0S :

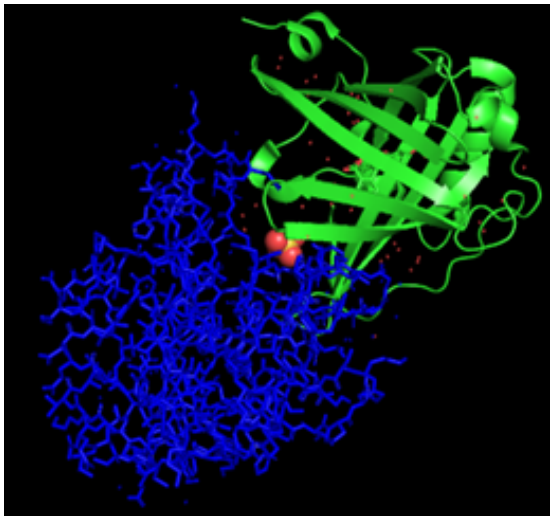


FIGURE 1 – Visuel 1N0S dans PyMOL. En vert la représentation "cartoon", en bleu la représentation "sticks".

Densité électronique de 1N0S : Voici une représentation dans Pymol d'un résidu (M173), nous remarquons des mauvais ajustements électroniques, surtout en augmentant le niveau de densité dans les options Pymol.

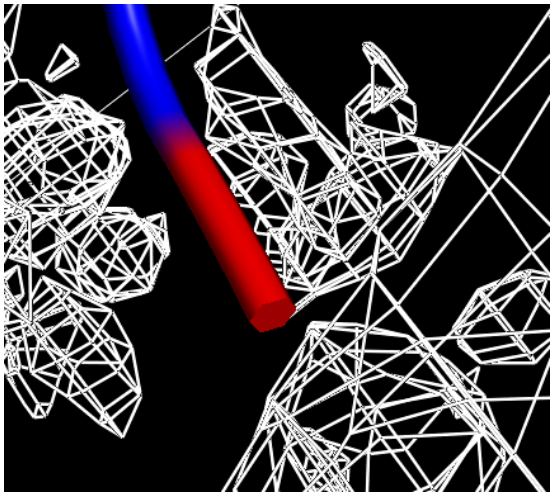


FIGURE 2 – Vue du résidu M173 de la structure 1N0S. En rouge, le résidu, en bleu la chaîne A et en blanc la densité électronique.

Visuel de 1T0V et analyse des conformères :

FIGURE 3 – Visuel 1T0V dans PyMOL.

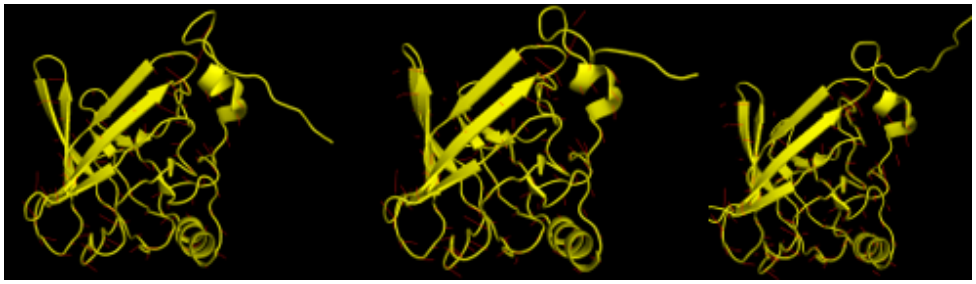


FIGURE 4 – Visuel des conformers 1/5/15 de 1T0V dans PyMOL.

Les 20 conformères de la structure 1T0V présentent tous des structures secondaires identiques. Cependant, la structure tertiaire varie légèrement d'un conformère à l'autre, avec des différences dans l'étirement et l'orientation des segments. Ces variations mineures dans la conformation tertiaire illustrent la flexibilité de la structure dans différentes conditions expérimentales, tout en conservant la même organisation de base.

1.3 Complexe enzyme-ligand

La structure 1DG5 de la dihydrofolate réductase de Mycobacterium tuberculosis (oxydoréductase) permet d'étudier l'interaction entre l'enzyme et le triméthoprime. Ce ligand forme des liaisons hydrogène directes avec trois acides aminés : deux isoleucines et un acide aspartique. Par ailleurs, 45 autres résidus stabilisent indirectement cette interaction, sans liaison directe.

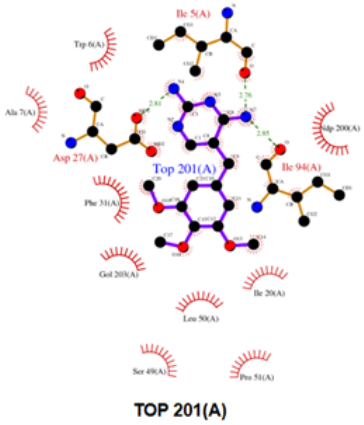


FIGURE 5 – Interactions de 1DG5 avec le triméthoprime

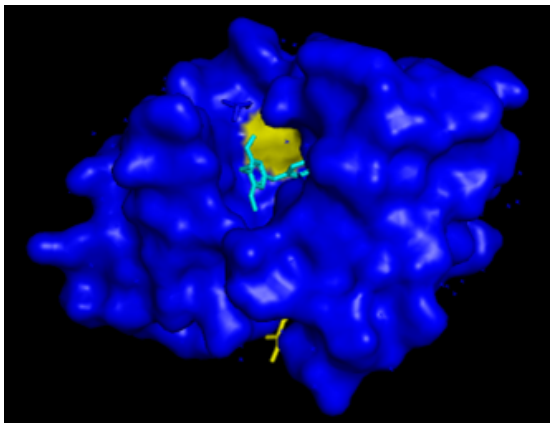


FIGURE 6 – Interaction entre la dihydrofolate réductase et le triméthoprime (en turquoise). En jaune sont les acides aminés responsables de la liaison

L'interaction est clairement visible : le triméthoprime s'insère dans une cavité de la protéine et s'y fixe via l'un de ses cycles.

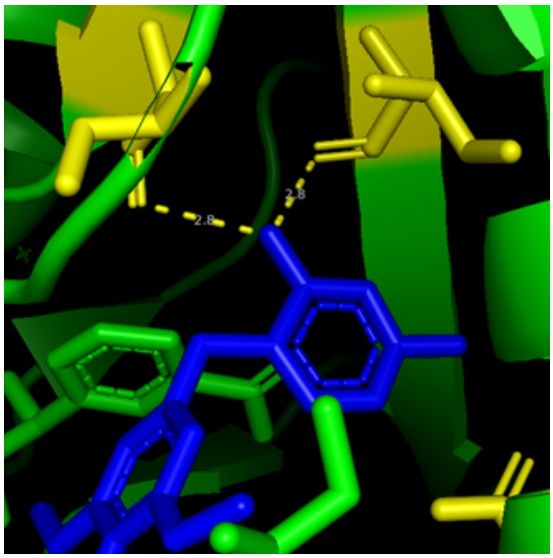


FIGURE 7 – Distances entre la dihydrofolate réductase et le triméthoprime (en bleu) et les acides aminés responsables de la liaison (jaune)

En utilisant l'outil Measurement de PyMOL, nous avons mesuré les distances ont été mesurée et coincides avec la figure 5.

1.4 Compacité du noyau protéique

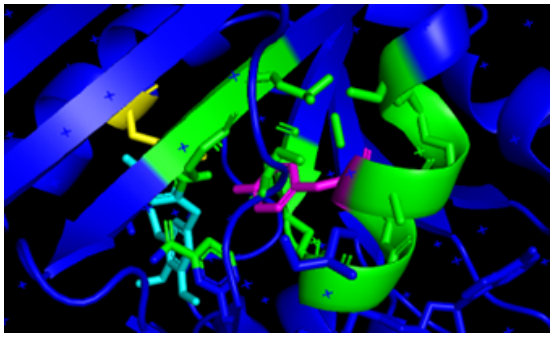


FIGURE 8 – En rose l'acide aminé numéro 100 et en vert les résidus dans un rayon de 5Å

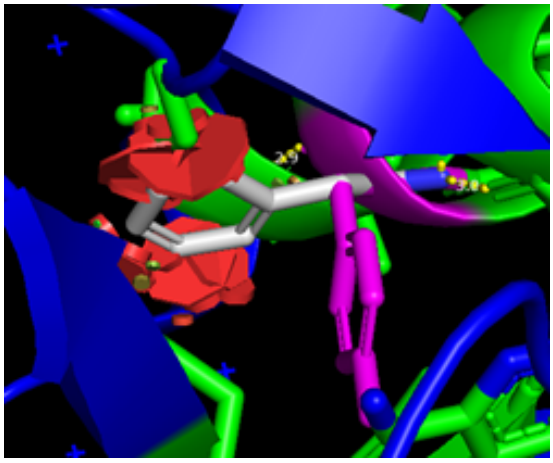


FIGURE 9 – Résultat de la mutation de l'acide aminé 100 en phénylalanine dans la structure de la protéine 1DG5,la mutation est représentée par les bâtonnets gris au centre de l'image

Les disques rouges autour de la chaîne latérale de la phénylalanine indiquent des collisions stériques, ce qui signifie que l'ajout de ce résidu volumineux entraîne des conflits d'espace avec les résidus environnants. Ces collisions suggèrent que cette mutation pourrait causer une contrainte structurelle, perturbant potentiellement la conformation locale de la protéine.

2 Partie 2 : Classification et alignement de la structure des protéines

2.1 Recherche de domaines protéiques et de leur structure

La protéine étudiée, identifiée par le code Uniprot **Q12923**, a une longueur de séquence de 2485 acides aminés et contient plusieurs domaines fonctionnels, dont les domaines **FERM** et **PDZ**. La structure de la protéine avec le code PDB **3PDZ** correspond spécifiquement au domaine **PDZ2**. On trouve 10 régions désordonnées dans cette protéine, dont les annotations ont été réalisées automatiquement, sans mesure expérimentale. Par ailleurs, il n'existe pas de structure expérimentale complète de cette protéine ; seules des structures partielles obtenues par cristallographie aux rayons X ou par RMN sont disponibles. Cependant, une prédiction de la structure complète est accessible via **AlphaFold**.

2.2 Base de données de classification

Le système **CATH** organise les protéines en quatre niveaux :
— **Classe** : basée sur les structures secondaires ;
— **Architecture** : basée sur la forme 3D globale ;
— **Topologie** : l'agencement interne des structures secondaires ;
— **Superfamille homologue** : regroupe les protéines évolutivement proches.

L'annotation de famille protéique se base sur les similitudes de séquence, de structure et de fonction, ainsi que sur les domaines et motifs fonctionnels.

Pour les domaines **3PDZ**, **1Z86** et **1RGW**, la classification CATH les identifie tous comme des domaines **PDZ**, avec une structure principalement bêta et une architecture "rouleau".

L'annotation de famille les place aussi dans la même catégorie, car ils partagent une fonction de liaison protéine-protéine. Ce domaine est impliqué dans la reconnaissance et la liaison de motifs spécifiques de peptides situés à l'extrémité C-terminale des protéines. Ce domaine est essentiel dans les processus de signalisation cellulaire.[1]

2.3 Superposition de structures

En observant l'alignement ClustalOmega, on remarque que certaines positions sont bien conservées entre les séquences des protéines **1RGW**, **1Z86**, et **3PDZ**. Les résidus aux positions montrent des similarités avec des étoiles (*), indiquant des acides aminés identiques. En revanche, d'autres positions présentent plus de variabilité, marquées par : ou . pour des similitudes biochimiques sans conservation complète.

Certaines régions clés sont conservées, suggérant une fonction commune ou des caractéristiques structurales partagées. Cependant, l'alignement montre aussi des différences importantes entre les séquences, indiquant qu'elles ne sont pas hautement conservées sur toute leur longueur.

```
CLUSTAL O(1.2.4) multiple sequence alignment


1RGW_1|Chain      ------MAYSVTLTGPGPWGFRLQGGKDFNM---PLTISRITPGSKAA-QSQLSQGDLVV  50
1Z86_1|Chain      -----RRRVTVRKADAGGLGISIKGGRENKM---PILISKIFKGLAADQTEALFVGDAIL  52
3PDZ_1|Chain      PKPGDIFEVEL-AKNDNSLGISVTGGVNTSVRHGGIYVKAVIPQGAAESDGRIHKGDRVL  59
                            :   . .  *: : ** : .:    : :. :     *     :  ** ::

1RGW_1|Chain      AIDGVNTDTMTHLEAQNKIKSASYNLSLTLQKSKR-- 85
1Z86_1|Chain      SVNGEDLSSATHDEAVQALKKTGKEVVLEVKYMKE-- 87
3PDZ_1|Chain      AVNGVSLEGATHKQAVETLRNTGQVVHLLLEKGQSPT 96
                  :::* . .  ** :* : ::.:.  : * ::  :
```

FIGURE 10 – Résultat de l'alignement multiple avec Clustal Omega

```
EMBOSS_001         1 PKPGDIFEVELAKND-NSLGISVTGGVNTSVRHGGIYVKAVIPQGAAESD     49
                       ...|.:.|.| ..||||:.||....:   .|.:..:....||:..
EMBOSS_002         1 -----RRRVTVRKADAGGLGISIKGGRENKM---PILISKIFKGLAADQT     42

EMBOSS_001        50 GRIHKGDRVLAVNGVSLEGATHKQAVETLRNTGQVVHLLLEKGQSPT     96
                     ..:..||.:|:|||..|..|||.:||:.|:.||:.|.|.::..:.
EMBOSS_002        43 EALFVGDAILSVNGEDLSSATHDEAVQALKKTGKEVVLEVKYMKE--     87
```

FIGURE 11 – Alignement global des séquences entre 3PDZ et 1Z86 avec Needle

L'alignement global entre les séquences **3PDZ** et **1Z86** donne une longueur de 97 résidus avec une identité de 29,9% et une similarité de 46,4%. Il y a 11,3% de gaps.

structures superposées :

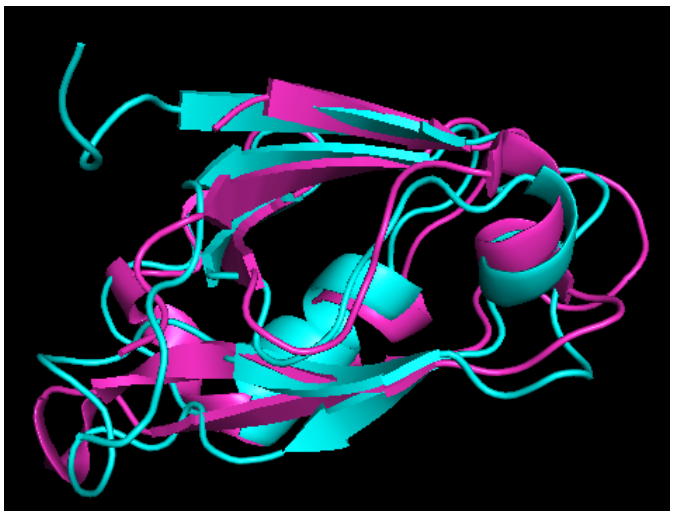


FIGURE 12 – Structures superposées 3PDZ et 1Z86, en rose 3PDZ, en bleu 1Z86

Pour l'alignement des structures **3PDZ** et **1Z86** dans **PDBeFold**, nous obtenons un *Z-score* de 7,63 et un *RMSD* de 1.83Å. Le *Z-score* mesure la pertinence de la correspondance entre les structures : plus il est élevé, plus l'alignement est significatif. Le *RMSD*, quant à lui, évalue la distance moyenne entre les atomes alignés, un faible *RMSD* indiquant une bonne superposition.

En comparant avec l'alignement par **Needle**, on constate que **PDBeFold** introduit des gaps pour optimiser l'alignement sur des zones structurales clés, tandis que **Needle** pénalise les gaps, conduisant à un alignement sans interruptions mais potentiellement moins

précis dans les régions structurales importantes.

L'alignement global des séquences de **1FCF** et **3PDZ** réalisé avec **Needle** commence au résidu 141 en tenant compte du décalage, et se termine au résidu 240. Les scores d'identité (6%) et de similarité (10%) sont faibles, ce qui est attendu, car l'alignement global compare les protéines dans leur intégralité, bien que seul le domaine **PDZ** soit pertinent ici.

Un alignement local, effectué avec **Matcher**, permet de se concentrer spécifiquement sur la région du domaine **PDZ**, ce qui améliore les scores d'identité (32%) et de similarité (53%). L'alignement local commence au résidu 179 de **1FCF** et correspond à la séquence **3PDZ** à partir du résidu 35.

L'alignement local s'avère être un choix plus judicieux pour comparer spécifiquement les domaines **PDZ**, car il fournit une meilleure qualité de correspondance que l'approche globale.

```
EMBOSS_001         1 ------------PKPGDIFEVELAKNDNSLGISVTGGVNTSVRHGG----     34
                                 |....:....||........||| ||...:.:.|
EMBOSS_002        51 DAIRKLLAVLDDPFTRFLEPSRLAALRRGTAGSVT-GVGLEITYDGGSGK     99

EMBOSS_001        35 -IYVKAVIPQGAAESDGRIHKGDRVLAVNGVSLEGATHKQAVETLR-NTG     82
                     :.|....|.|.||..| ...||.::.|:|.:::|.:.....:.|: ...
EMBOSS_002       100 DVVVLTPAPGGPAEKAG-ARAGDVIVTVDGTAVKGLSLYDVSDLLQGEAD    148

EMBOSS_001        83 QVVHLLLEKGQSPT-----------------------------------     96
                     ..|.::|....:|:
EMBOSS_002       149 SQVEVVLHAPGAPSNTRTLQLTRQKVTINPVTFTTCSNVAAAALPPGAAK    198
```

FIGURE 13 – Alignement global de 3PDZ et 1FCF réalisé par Needle

```
EMBOSS_001        35 IYVKAVIPQGAAESDGRIHKGDRVLAVNGVSLEG     68
                     :.|....|.|.||..|. ..||.::.|:|.:::|
EMBOSS_002       101 VVVLTPAPGGPAEKAGA-RAGDVIVTVDGTAVKG    133
```

FIGURE 14 – Alignement global de 3PDZ et 1FCF réalisé par Matcher

3 Partie 3 : Prédiction de la structure secondaire

3.1 Analyse des programmes de prédiction de la structure secondaire

Dans cette partie, l'objectif est d'explorer les méthodes de prédiction de structures secondaires de protéines à l'aide de différents outils bioinformatiques. Voici une explication des approches utilisées par trois programmes de prédiction : **GOR IV**, **HNN**, et **Sympred**.

GOR IV :
Cet outil applique une méthode statistique basée sur l'analyse des fréquences de résidus dans des structures secondaires spécifiques (hélices, feuillets, etc.). Il prend en compte les probabilités conditionnelles des résidus dans une fenêtre de 17 acides aminés pour prédire leur rôle structurel, en tenant compte des interactions locales.

HNN :
HNN utilise des réseaux de neurones artificiels pour modéliser les motifs complexes présents dans les séquences protéiques. Ces réseaux sont entraînés sur des données expérimentales pour apprendre à associer les séquences aux structures secondaires correspondantes (hélices, feuillets, ou boucles).

Sympred :
Sympred adopte une approche consensuelle, combinant les résultats de plusieurs méthodes de prédiction (comme GOR IV et Chou-Fasman). En intégrant ces résultats, il cherche à améliorer la précision globale de la prédiction.

Ces outils offrent des perspectives complémentaires pour mieux comprendre la structure secondaire des protéines, chaque méthode ayant ses propres forces et limitations.

3.2 Comparaison des performances des outils de prédiction de la structure secondaire.

Pour l'analyse de la structure secondaire de la protéine **1E2F**, il est demandé de comparer les assignations fournies par **PDBSum** et le fichier PDB, car ces deux sources utilisent des méthodologies distinctes qui peuvent générer des différences notables.

3.2.1 Hélices (HELIX)

Les limites des hélices diffèrent légèrement. Par exemple, la première hélice est indiquée dans PDBSum comme allant de `LYS 19` à `ALA 32` (14 résidus), tandis que dans le fichier PDB, elle s'étend de `GLY 18` à `ALA 33` (16 résidus). Ces variations reflètent des critères géométriques distincts, tels que la longueur minimale ou l'angle des liaisons hydrogène.

3.2.2 Feuillets (SHEET)

Dans PDBSum, 7 feuillets sont décrits, contre 5 dans le fichier PDB. Par exemple, PDBSum inclut un feuillet de `GLY 208` à `GLU 209`, absent du fichier PDB. Ces différences

proviennent des seuils utilisés pour identifier les feuillets β, comme la longueur minimale ou la distance entre les chaînes principales.

3.2.3 Origine des différences

PDBSum utilise des outils comme `DSSP`, qui reposent sur des algorithmes standards pour déterminer les structures secondaires. Le fichier PDB, quant à lui, peut inclure des annotations manuelles ou basées sur des données expérimentales, ce qui conduit à des interprétations plus souples.

Ces divergences mettent en évidence l'importance des méthodologies utilisées et l'intérêt de combiner plusieurs approches pour une compréhension complète de la structure secondaire des protéines.

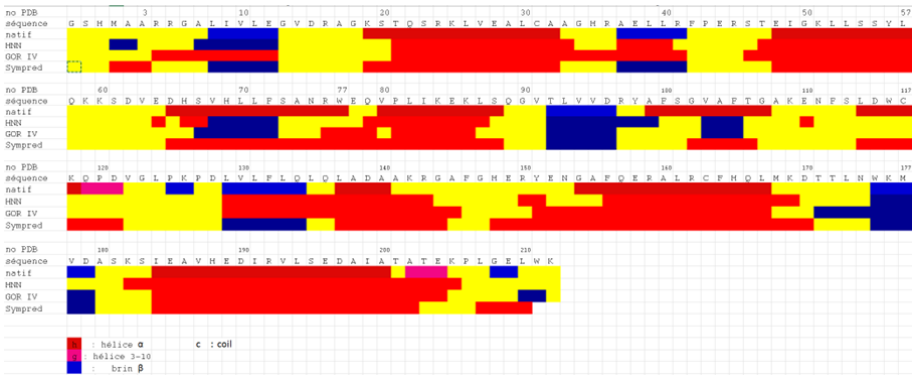


FIGURE 15 – Résultats de HNN, GOR IV et Sympred pour prédire la structure de 1E2F par rapport à la structure PDB

Les résultats montrent une cohérence globale entre les différentes méthodes. Quelques différences mineures apparaissent entre les prédictions. **Sympred** semble être tout de même un peu plus robuste que les autres outils peut-être grâce à son approche basée sur le consensus.

3.3 Analyse de deux séquences inconnues ayant des propriétés particulières

Les deux protéines inconnues ont été identifiées via `BLAST` :
— **Unknown 1** : Protéine prion humaine (`PRNP`, *Homo sapiens*).
— **Unknown 2** : Précurseur de l'amyloïde-β A4 (*Galemys pyrenaicus*).

Les structures secondaires ont été prédites avec **GOR IV**, **HNN**, et **Sympred**, puis comparées aux structures expérimentales.

3.3.1 Protéine prion

— **GOR IV** : Moins précis, ajoutant des fragments de brins inexistants et coupant la dernière hélice.
— **HNN** : Performances correctes mais moins détaillées que **Sympred**.
— **Sympred** : Le plus proche de la structure native, mais manque une petite région en brin au début et introduit deux erreurs (une hélice et un brin fictifs).

3.3.2 Protéine amyloïde-β

— **GOR IV** : Identifie une hélice absente dans la structure native et échoue à distinguer certaines régions.
— **HNN** & **Sympred** : Moins performants, confondant hélices et brins dans la région terminale.

Sympred est le plus fiable pour ces cas, mais les prédictions restent limitées. Les écarts avec les structures natives soulignent l'importance de méthodes complémentaires pour des séquences très divergentes.

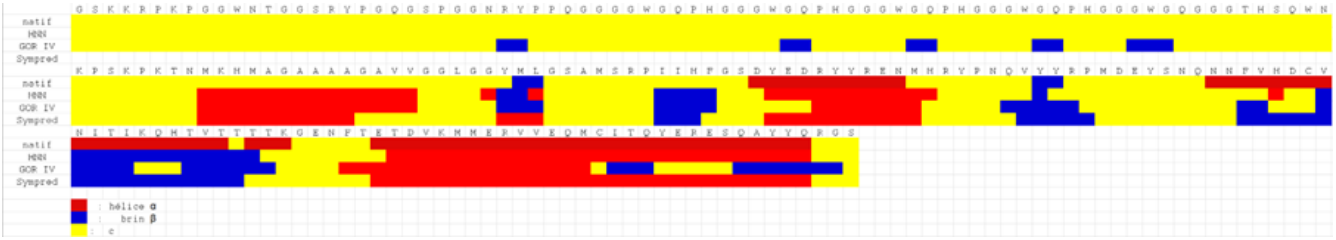


FIGURE 16 – Résultats de GOR IV, HNN et Sympred sur la première séquence inconnue

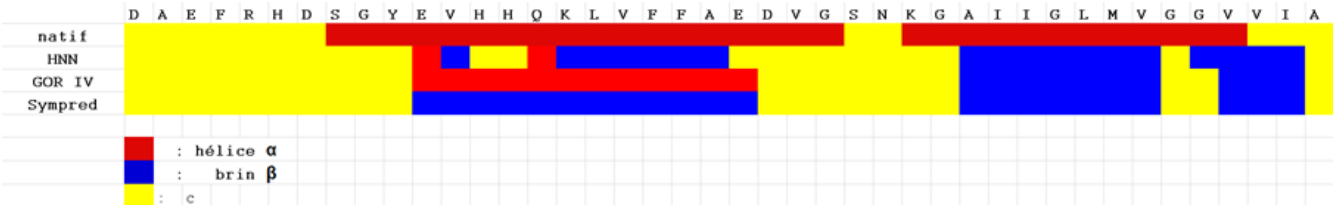


FIGURE 17 – Résultats de GOR IV, HNN et Sympred sur la deuxième séquence inconnue

4 Partie 4 : Prédiction de la structure tridimensionnelle

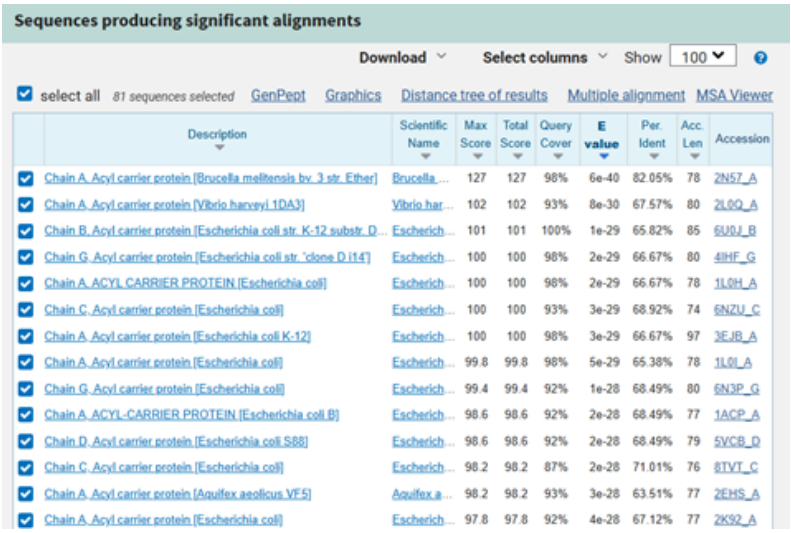
Dans cette partie, l'objectif est de prédire et d'évaluer la structure tridimensionnelle des protéines à l'aide de méthodes de modélisation comparative et de reconnaissance de repliement. Pour cela, on utilise des outils comme **QMEAN** et **Procheck**, qui permettent d'évaluer la qualité des modèles créés.

Le **QMEAN score** est une métrique combinant des potentiels statistiques globaux et locaux pour évaluer la fiabilité d'un modèle structurel. Ce score compare le modèle à des structures expérimentales de la base PDB. Un *Z-score QMEAN* proche de 0 indique une qualité comparable à des structures natives, tandis qu'un score inférieur à -4 reflète des incohérences significatives. Des valeurs élevées traduisent une meilleure correspondance entre le modèle et une structure biologiquement plausible.[2]

4.1 Modélisation comparative : approche manuelle

Pour identifier un modèle de structure, un `BLAST` a été effectué en utilisant la séquence de l'*ACP* de *Rhodospirillum centenum* (UniProt : `B6IN76`) contre la base de données PDB. Les résultats montrent plusieurs hits avec des identités de séquence et des couvertures variables. Nous avons pas mal de modèles ayant une bonne couverture (>98 %).

Parmi les résultats `BLAST`, le modèle choisi présente une forte identité de séquence et une couverture de requête proche de 100 %. Il s'agit de **2N57_A**.



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Chain A, Acyl carrier protein [Brucella melitensis bv. 3 str. Ether]	Brucella ...	127	127	98%	6e-40	82.05%	78	2N57_A
Chain A, Acyl carrier protein [Vibrio harveyi 1DA3]	Vibrio har...	102	102	93%	8e-30	67.57%	80	2L0Q_A
Chain B, Acyl carrier protein [Escherichia coli str. K-12 substr. D...	Escherich...	101	101	100%	1e-29	65.82%	85	6U0J_B
Chain G, Acyl carrier protein [Escherichia coli str. 'clone D i14']	Escherich...	100	100	98%	2e-29	66.67%	80	4IHF_G
Chain A, ACYL CARRIER PROTEIN [Escherichia coli]	Escherich...	100	100	98%	2e-29	66.67%	78	1L0H_A
Chain C, Acyl carrier protein [Escherichia coli]	Escherich...	100	100	93%	3e-29	68.92%	74	6NZU_C
Chain A, Acyl carrier protein [Escherichia coli K-12]	Escherich...	100	100	98%	3e-29	66.67%	97	3EJB_A
Chain A, Acyl carrier protein [Escherichia coli]	Escherich...	99.8	99.8	98%	5e-29	65.38%	78	1L0I_A
Chain G, Acyl carrier protein [Escherichia coli]	Escherich...	99.4	99.4	92%	1e-28	68.49%	80	6N3P_G
Chain A, ACYL-CARRIER PROTEIN [Escherichia coli B]	Escherich...	98.6	98.6	92%	2e-28	68.49%	77	1ACP_A
Chain D, Acyl carrier protein [Escherichia coli S88]	Escherich...	98.6	98.6	92%	2e-28	68.49%	79	5VCB_D
Chain C, Acyl carrier protein [Escherichia coli]	Escherich...	98.2	98.2	87%	2e-28	71.01%	76	8TVT_C
Chain A, Acyl carrier protein [Aquifex aeolicus VF5]	Aquifex a...	98.2	98.2	93%	3e-28	63.51%	77	2EHS_A
Chain A, Acyl carrier protein [Escherichia coli]	Escherich...	97.8	97.8	92%	4e-28	67.12%	77	2K92_A

FIGURE 18 – Séquences produisant des alignements significatifs

Un alignement global entre la séquence de l'ACP (`B6IN76`) et celle du modèle sélectionné a été réalisé à l'aide de l'outil `EMBOSS Needle`. L'alignement a une longueur totale de 79 résidus, avec un taux d'identité élevé de 81,0 % et une similarité de 89,9 %. Seulement un gap est présent (1,3 %), indiquant une très bonne correspondance entre les deux séquences. Le score global obtenu est de 320,0, confirmant une forte conservation entre la séquence cible et celle du modèle.

Ces résultats suggèrent que le modèle est bien adapté pour la modélisation comparative, grâce à la forte similitude des séquences et à la continuité presque complète dans l'alignement.

```
#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 79
# Identity:      64/79 (81.0%)
# Similarity:    71/79 (89.9%)
# Gaps:           1/79 ( 1.3%)
# Score: 320.0
#
#
#=======================================

>EMBOSS_001 ..
MSDTAERVKKIVIEHLGVEESKVTESASFIDDLGADSLDTVELVMAFEEE
FGIEIPDDAAEKILTVKDAIDFINQKTAA
>EMBOSS_001 ..
MSDTAERVKKIVVEHLGVDADKVTEGASFIDDLGADSLDTVELVMAFEEE
FGVEIPDDAAETILTVGDAVKFIDKASA-
```

FIGURE 19 – Résultats de l'alignement global entre la séquence de l'ACP (B6IN76) et celle du modèle sélectionné

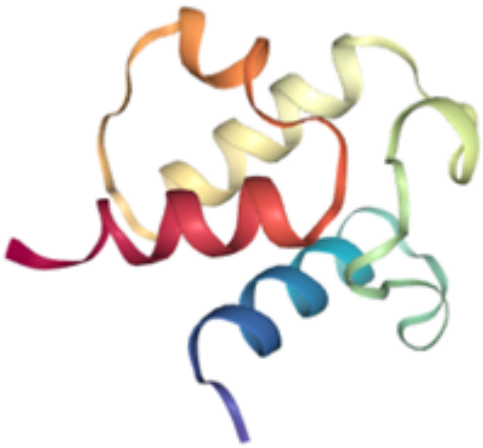


FIGURE 20 – Modélisation comparative de la structure des protéines à l'aide de MODELLER

Un alignement des séquences de l'ACP et du modèle sélectionné a été soumis à `Modeller` après conversion au format `PIR`. Le serveur a généré un modèle tridimensionnel de la protéine ACP, basé sur l'alignement fourni.

FIGURE 21 – Résultats avec Q mean, approche manuelle

Le modèle généré par `Modeller` a été évalué en utilisant les outils **QMEAN**. Le *Z-score* global reflète la qualité globale du modèle en le comparant à des structures expérimentales similaires. En utilisant **QMEAN** nous obtenons une valeur de 0,25, la version **QMEANDisCo** nous donne une valeur de 0,78 ± 0,10.

4.2 Modélisation comparative : approche semi-automatique

Dans cette approche semi-automatique, la séquence de l'ACP a été soumise au serveur `HHPred`, qui utilise des alignements de profils HMM pour identifier des modèles potentiels. Par rapport à `BLAST`, le premier modèle qui est pertinent de sélectionner est **5ISW_A**, c'est le hit avec la plus haute probabilité (96,9 %).

FIGURE 22 – Résultats avec Q mean, approche semi-automatique

Une fois le modèle généré par `Modeller`, il a été évalué à l'aide de **QMEAN** et de la version **DisCo**. Les scores obtenus sont :

— **QMEAN global** : -0.71
— **QMEANDisCo** : 0.58 ± 0,10

Ces valeurs indiquent une qualité modérée pour le modèle. Ces scores reflètent une structure globalement plausible, mais avec des régions pouvant être moins fiables. Cette approche semi-automatique s'avère efficace pour identifier rapidement un template pertinent et générer un modèle structuré, bien qu'elle propose parfois des templates moins spécifiques que ceux identifiés par `BLAST`.

4.3 Modélisation comparative : approche automatique

Le modèle généré avec `SwissModel` utilise le template **2N57.1.A**, une structure NMR d'acyl carrier protein de *Brucella melitensis*, avec une identité de séquence de 81,82 % et une couverture complète. Les scores obtenus montrent une bonne qualité :

— **GMQE** : 0.86
— **QMEANDisCo Global** : 0.80 ± 0.11

Ces résultats indiquent un modèle fiable, bien adapté grâce à un template pertinent et des évaluations structurelles élevées.

4.4 Reconnaissance des repliements

La reconnaissance de repliement a été effectuée avec `GenThreader`, un outil qui identifie un template optimal pour la modélisation. À cette étape, le template sélectionné est **2N57**, déjà utilisé dans une modélisation précédente. Par conséquent, aucune nouvelle modélisation n'a été réalisée pour éviter de générer un modèle redondant.

4.5 Modèle AlphaFold

Le modèle *AlphaFold* de l'ACP a été téléchargé depuis la page `UniProt` (section Structure) et évalué à l'aide de **QMEAN**. Les scores obtenus sont :

— **QMEANDisCo Global** : 0.74 ± 0.10, indiquant une qualité modérée pour le modèle.
— **QMEAN** : 1.66, reflétant des incohérences possibles dans certaines régions de la structure.

4.6 Comparaison des modèles

Les structures des quatre modèles ont été comparées en termes de *RMSD* (Root Mean Square Deviation), mesurant la différence structurelle entre chaque paire de modèles. Voici les résultats obtenus :

— **alphaFold avec automatic** : RMSD de 1.720Å
— **alphaFold avec manual** : RMSD de 1.896Å
— **alphaFold avec semi-auto** : RMSD de 2.047Å
— **automatic avec manual** : RMSD de 1.969Å

— **automatic avec semi-auto** : RMSD de 4.568Å
— **manual avec semi-auto** : RMSD de 2.479Å

Ces résultats montrent que le modèle **automatic** est le plus similaire à *alphaFold* avec une RMSD de seulement 1.720Å, ce qui indique une structure prédite relativement précise par rapport à la structure expérimentale. En revanche, la plus grande différence structurelle est observée entre **automatic** et **semi-auto** avec une RMSD de 4.568Å, suggérant que le modèle semi-automatique présente des divergences plus marquées. Lors de l'utilisation du `super` de `PyMOL` pour superposer les structures, des régions de différences importantes peuvent être visualisées, en particulier dans les modèles qui montrent un RMSD plus élevé. Ce processus a permis d'identifier des zones où les prédictions diffèrent significativement de la structure expérimentale, ce qui peut être crucial pour affiner les méthodes de modélisation.

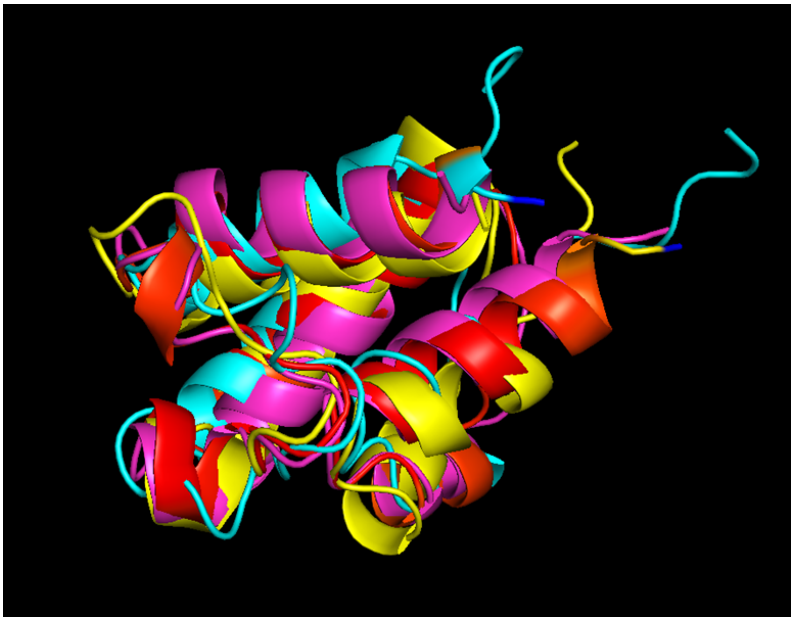


FIGURE 23 – Superposition de tous les modèles

Références

[1] RCSB PDB Help. Cath : What is cath ? https://www.rcsb.org/docs/search-and-browse/browse-options/cath#:~:text=CATH%20is%20a%20free%2C%20publicly,evolutionary%20relationships%20of%20protein%20domains. Consulté le 7 Novembre 2024.

[2] SWISS-MODEL. Qmean : Introduction. https://swissmodel.expasy.org/qmean/help. Consulté le 26 Novembre 2024.