PROJET BIOINFORMATIQUE STRUCTURALE

I. INTRODUCTION

Dans ce projet de Bioinformatique Structurale, on va s'intéresser à une protéine particulière dont la structure tridimensionnelle n'est pas encore réalisée. Le but final de ce projet est ainsi de modéliser cette protéine en 3D, en utilisant différents outils bioinformatiques ainsi que de bases de données. La protéine cible choisie est « Glucose-6-phosphate exchanger (UniProt P57057) ». Cette protéine est codée par le gène *SLC37A1*, et est présente chez l'homme. Elle est localisée au niveau de la membrane du réticulum endoplasmique (RE). Le gène codant pour cette protéine est exprimé dans divers tissus, il présente une expression élevée surtout au niveau du pancréas, fois et intestin grêle.

Cette protéine entre en interaction avec d'autres protéines. La masse de la protéine est environ 57,648 Daltons, elle comprend 533 acides aminés. Elle joue un rôle principal en tant qu'un antiporteur du glucose-6-phosphate et du phosphate inorganique. Elle est capable de transporter le glucose-6-phosphate à l'intérieur du lumen du réticulum endoplasmique, et le phosphate inorganique dans le sens opposé. Elle est impliquée aussi dans le transport des carbohydrates.

Le gène codant pour cette protéine est localisé sur le chromosome 21 (Bartoloni L. *et al,* 2000). Des études sur la détermination de la séquence cDNA et la structure génomique du gène *SLC37A1* pourront être utiles afin de comprendre l'implication de ce gène dans l'apparition des phénotypes monogéniques et polygéniques (Bartoloni L. *et al,* 2000). La triplication du *SLC37A1* en cas de la trisomie 21, pourrait jouer un rôle dans l'apparition du phénotype Syndrome de Down (Bartoloni L. *et al,* 2000).

Les gènes porteurs de solutés (SLC) comprennent plus de 50 familles de gènes codant pour des transporteurs transmembranaire (Vasiliou K. *et al.,* 2009). La famille SLC37 comprend les 4 protéines d'échange sucre-phosphate suivantes : SLC37A1, SLC37A2, SLC37A3 et SLC37A4. Ces protéines ont été initialement regroupées dans la famille SLC37 sur la base d'une homologie de séquence avec les échangeurs organo-phosphate. Le membre le mieux caractérisé est SLC37A4, mieux connu sous le nom de transporteur de glucose-6-phosphate (G6PT) (Chou *et al.,* 2013). Il semble que chez les patients ayant une déficience en glycérol kinase, SLC37A1 ne présente aucune mutation, cela suggère que le glucose-3-phosphate ne semble pas être son substrat primaire (Pan C. *et al,* 2011). L'activité du substrat SLC37A1 n'est pas encore déterminée. D'après les expériences réalisées, SLC37A1 n'est pas capable de se coupler avec la G6Pase-α et ne semble pas impliquée dans « blood glucose homeostasis » (Pan C. *et al,* 2011). SLC37A1 est ''up-regulated'' par les facteurs de croissance épidermique dans les cellules cancéreuses de sein, cela pourrait suggérer une implication de cette protéine dans la biosynthèse des phospholipides (Lacopetta D. *et al,* 2010).

La protéine Glucose-6-phosphate exchanger SLC37A1 appartient à la grande super famille des faciliteurs (Major Facilitator Superfamily MFS), Un des plus grands groupes de transporteurs actifs secondaires conservés de bactéries aux humains. L'analyse

bioinformatique a prédit que la majorité des membre MFS comprennent 12 hélices transmembranaires (TM), certaines contenant plus (Reddy, V.S *et al.,* 2012).

Les extrémités N et C d'un MFS sont généralement situées du côté cytoplasmique de la membrane (Yan N. *et al.,* 2013). Dans toutes les structures MFS, les hélices transmembranaires numéro 1, 4, 7 et 10 sont positionnées au centre du transporteur, contribuant à la majorité des résidus essentiels, à la coordination du substrat et au couplage du co-transport. Chez un grand nombre de membres MFS, deux séquences conservées, DRXXRR, se trouvent aux extrémités cytoplasmiques étroitement jointes des hélices transmembranaires 2 et 3 dans les domaines N terminaux, 8 et 9 dans les domaines C terminaux.

II. METHODE ET DISCUSSION

a) Séquence primaire P57057:

La séquence primaire de la protéine d'intérêt P57057 représentée ci-dessous, peut être téléchargé sur UniProt sous format FASTA, ce dernier sera ainsi très utiles pour les études réalisées dans la suite du travail.

Légende :

- sp : pour UniProtKB/Swiss-Prot
- P57057 : UniqueIdentifier, c'est le numéro d'accès principal de l'entrée UniProt
- G6PT2_HUMAN : EntryName, c'est le nom d'entrée UniProtKB
- Glucose-6-phosphate exchanger SLC37A1 : Nom de la protéine
- OS : nom scientifique de l'organisme, Homo sapiens dans notre cas
- GN : nom de gène, SLC37A1 dans notre cas
- PE : la valeur numérique décrivant l'existence prévue de la protéine P57057
- SV : le numéro de la version de la séquence

>sp|P57057|G6PT2_HUMAN Glucose-6-phosphate exchanger SLC37A1 OS=Homo sapiens GN=SLC37A1 PE=2 SV=2

MARLPAGIRFIISFSRDQWYRAFIFILTFLLYASFHLSRKPISIVKGELHKYCTAWDEADVRFSSQNRKSGSA
APHQLPDNETDCGWAPFDKNNYQQLLGALDYSFLCAYAVGMYLSGIIGERLPIRYYLTFGMLASGAFTAL
FGLGYFYNIHSFGFYVVTQVINGLVQTTGWPSVVTCLGNWFGKGRRGLIMGVWNSHTSVGNILGSLIAG
YWVSTCWGLSFVVPGAIVAAMGIVCFLFLIEHPNDVRCSSTLVTHSKGYENGTNRLRLQKQILKSEKNKPL
DPEMQCLLLSDGKGSIHPNHVVILPGDGGSGTAAISFTGALKIPGVIEFSLCLLFAKLVSYTFLFWLPLYITNV
DHLDAKKAGELSTLFDVGGIFGGILAGVISDRLEKRASTCGLMLLLAAPTLYIFSTVSKMGLEATIAMLLLSG
ALVSGPYTLITTAVSADLGTHKSLKGNAHALSTVTAIIDGTGSVGAALGPLLAGLLSPSGWSNVFYMLMFA
DACALLFLIRLIHKELSCPGSATGDQVPFKEQ

b) Informations du modèle 1PW4:

Le modèle adapté dans la bibliographie pour notre protéine d'intérêt (P57057) est la protéine 1PW4 (code Protein Data Bank PDB). 1PW4 est une protéine membranaire, c'est une structure d'un Crystal du transporteur glycérol-3-phosphate, appartenant à MFS. Le gène

codant pour cette protéine est *glpT*, elle est présente chez l'*Escherichia coli*. D'après la bibliographie, 1PW4 joue aussi le rôle d'un antiporteur de phosphate inorganique. Cette protéine contient 452 acides aminés, et 12 hélices transmembranaires. D'après l'analyse de sa structure 2D présente sur UniProt, les hélices transmembranaires sont connectées via des coudes (petite boucle de quelques acides aminés) et absence de grande boucle dans la structure. La méthode X-Ray admet une résolution de 3,3 A° (figure 2). L'identité de la séquence entre la protéine cible (P57057) et la structure du modèle (1PW4) est généralement considérée comme un premier indicateur de la précision attendue d'un modèle, comme l'ont confirmé diverses études. Ce modèle est basé sur l'alignement des séquences target-template, dont l'identité de la séquence est égale à 21% (figure 3).

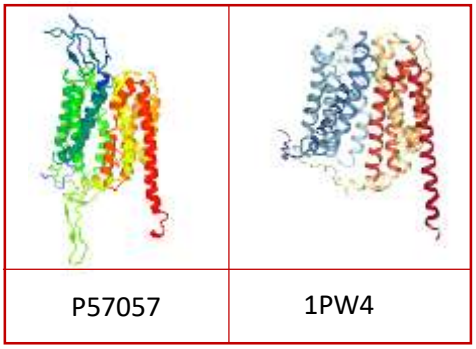


Figure 2: Comparaison de la structure du target (protéine cible P57057) et celle du template (1PW4).

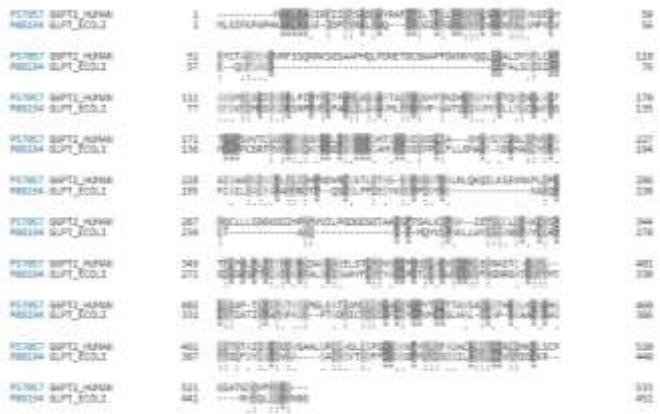


Figure 3: Résultat d'alignement des séquences de la structure target (protéine cible P57057) avec celle du template (1PW4).

Légende :

- ***** signifient que le résidu d'acide aminé en cette position a était conservé
- **.** signifient que sur 4 acides aminés, 2 sont conservés et les 2 autres se diffèrent des 4 acides aminés
- **:** signifient que sur 4 acides aminés, 2 sont conservés et les 2 autres sont semblables entre eux seulement

L'alignement ci-dessus est réalisé sur UniProt, qui utilise CLUSTALO comme programme d'alignement, le pourcentage d'identité dans la structure des deux protéines est 17,446. Il y a présence de 170 positions similaires et 97 positions identiques entre les séquences primaires des deux protéines.

c) Prédiction de la structure secondaire du P57057:

Le serveur PSIPRED, est une méthode de prédiction de structure secondaire simple et précise, qui effectue une analyse sur la sortie obtenue à partir de PSI-BLAST (Position Specific Iterated- BLAST). Après avoir téléchargé le fichier FASTA de la protéine d'intérêt (P57057), ce dernier servira en tant qu'un input pour PSIPRED. Le résultat figure ci-dessous.

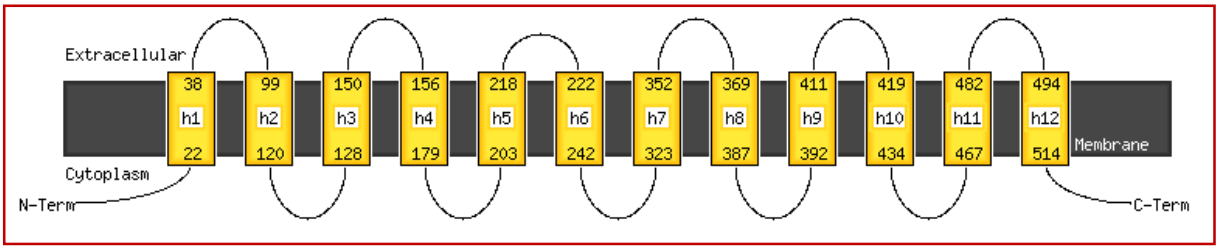


Figure 4: Résultat de prédiction de la structure de la protéine d'intérêt P57057 d'après PSIPRED.

L'interprétation de la structure prédite de la protéine cible montre la présence de 12 hélices transmembranaires réparties tout au long de la séquence (représentées par les rectangles jaune et numérotées du h1 à h12). Il y a aussi présence d'une grande boucle allant du résidu **39 à 98**. Une deuxième boucle encore plus grande que la précédente allant du résidu **243 à 322**.

Afin de confirmer le résultat de prédiction obtenu par PSIPRED, une autre étude a été mené sur le serveur TMHMM (figure 5) est un serveur qui permet de prédire les hélices transmembranaires dans la protéine. Il prédit la topologie de la protéine membranaire en se basant sur le modèle de Markov caché. La figure ci-dessous confirme ce qui a été trouvé dans la littérature, présence de 12 hélices transmembranaire dans la protéine cible. Les 12 hélices transmembranaires semble être distribuées selon sur les résidus suivant, 20-37, 99-121, 128-150, 156-185, 192-214, 219-241, 328-350, 365-387, 394-413, 423-445, 457-479, 494-513 (représenté par les rectangles rouges). La présence de deux boucles, la première est large et située dans le côté extra membranaire allant du résidu **38 à 98**. La deuxième boucle est encore plus large et située dans le côté intra-membranaire allant du résidu **242 à 327**. D'après la comparaison entre les deux méthodes de prédiction, il paraît que la différence est négligeable, plus ou moins 3 acides aminés, donc la structure prédite dans les deux outils semble correspondre à la structure réelle de la protéine cible P57057.

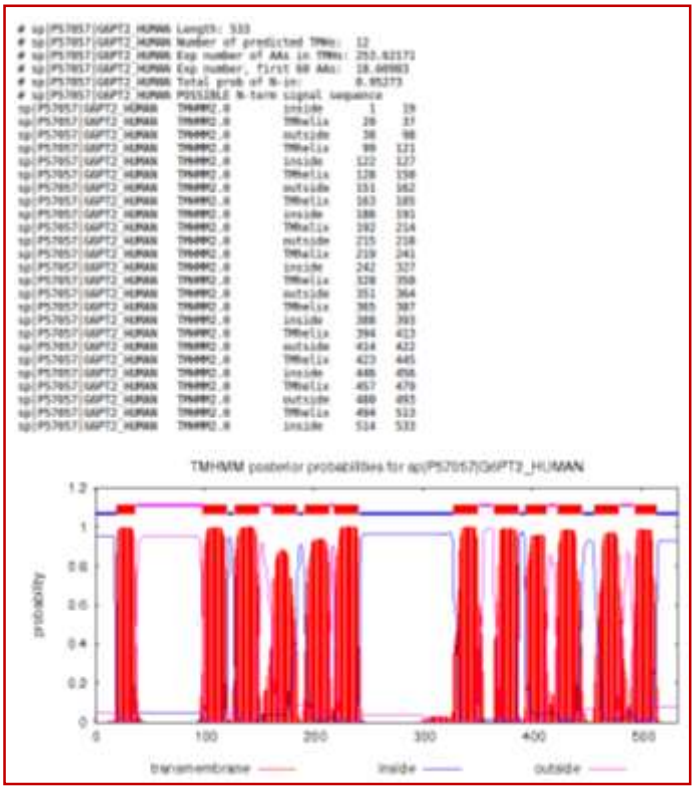


Figure 5: Résultat de prédiction de la structure de la protéine d'intérêt P57057 d'après TMHMM.

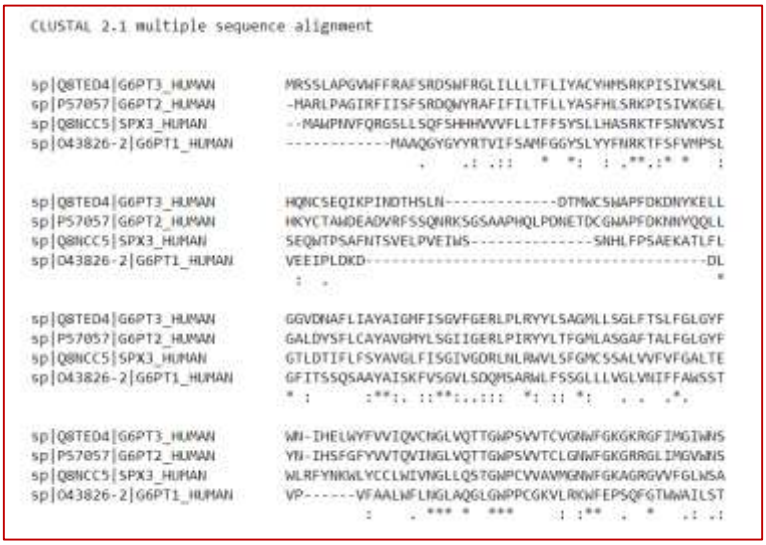
Le but ainsi est d'essayer d'améliorer le modèle pour notre cible de façon qui correspond le plus proche possible à sa structure ainsi prédit et également d'arriver à obtenir un modèle dont la résolution est meilleur que le modèle adapté (résolution inférieur à 3,3 A°).

d) Etude des domaines conservés dans la famille SLC37:

Au premier temps, on s'est intéressé à étudier la conservation des résidus d'acides aminés dans les séquences peptidiques des 4 protéines appartenant à la famille SLC37. Après avoir récupérer les formats FASTA de chacune des protéines sur UniProt, un alignement de leur séquence peptidique est réalisé. Le résultat figure ci-dessous et est obtenu par ClustalW, c'est un outil d'alignement multiple pour aligner des séquences protéiques. (http://www.genome.jp/tools-bin/clustalw).

Le code de la protéine dans UniProt correspondant au nom de la protéine figure dans le tableau ci-dessous.

Membre de la famille SLC37	Identifiant de la protéine (UniProt)
SLC37A1	P57057
SLC37A2	Q8TED4
SLC37A3	Q8NCC5
SLC37A4	O43826



D'après l'alignement des 4 séquences, il paraît que le pourcentage de résidus conservés est très négligeable.

e) Etude des domaines conservés de P57057 avec d'autres protéines:

Dans un deuxième temps, on va s'intéresser à identifier la conservation de domaines dans la protéine d'intérêt en comparant son alignement avec d'autres protéines dont le pourcentage d'identité est compris entre 40 et 65 afin de tirer les régions conservées. Le résultat obtenu par BLAST correspond à 520 protéines, dans lequel des protéines présentent des similarités allant de 47 jusqu'à 99 % avec P57057.

Les critères pour sélectionner les protéines sont les suivants :

1. Faut qu'elles appartiennent à la même famille de protéine
2. Faut que le taux de similarité soit faible (entre 40% et 65%), car les pourcentages très élevés ou très faible biaisent le résultat
3. Faut sélectionner au moins 6 protéines pour que le résultat soit significatif à interpréter

Les protéines sélectionnées figurent dans le tableau ci-dessous.

Nom de la protéine (UniProt code)
P57057 (G6PT2_HUMAN)
A0A1W2W5D5 5A0A1W2W5D5_CIOIN)
E0W2V1 (E0W2V1_PEDHC)
R0KWX2
A0A1S3NW06
Q5U3T2
A0A1A8PC42
W5U6I5

Le résultat de l'alignement des 8 protéines réalisé sur UniProt est présenté dans la figure ci-dessous.

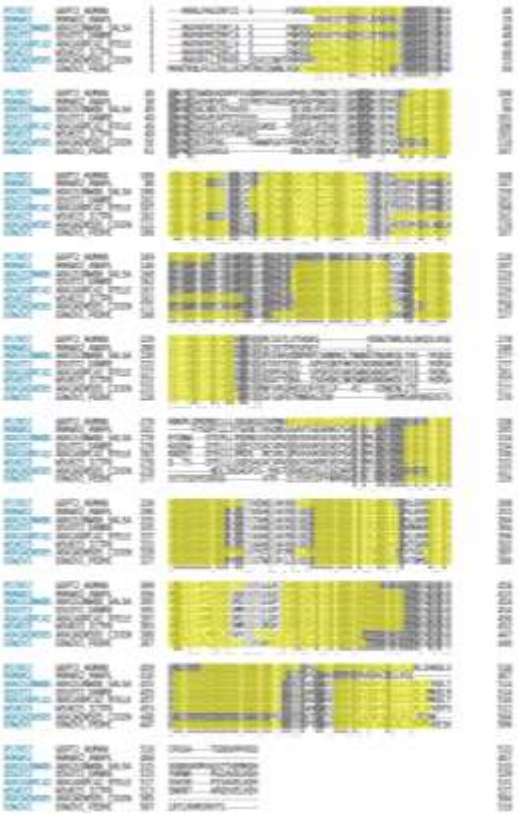


Figure 5: Résultat d'alignement sur UniProt (Align).

Légende :

- Couleur jaune représente les domaines transmembranaires
- Couleur grise représente les résidus d'acides aminés similaires

D'après l'analyse du résultat de l'alignement, il semble qu'un grand pourcentage d'acides aminés conservé est localisé au niveau de domaine transmembranaire. D'où ce domaine semble jouer un rôle essentiel pour le fonctionnement de la protéine. Le tableau ci-dessous présente l'interprétation des domaines similaires entres les séquences primaires des 4 protéines de la famille SLC37A, et leur correspondance d'après la prédiction de la structure secondaire de la protéine P57057 déjà vue.

Résidus d'acides aminés conservés	Leur correspondance dans la structure 2D
34 – 53	Partie boucle extra-membranaire, N-terminal
169 – 228	Boucle, hélice N°5, boucle
324 – 354	Hélice transmembranaire N°7
370 – 379	Hélice transmembranaire N°8
434 – 451	Boucle intra membranaire
459 – 480	Hélice transmembranaire N°11

f) Recherche d'un bon template par la modélisation d'homologie :

La modélisation comparative de protéines repose sur l'identification d'une ou plusieurs structures protéiques connues susceptibles de ressembler à la structure de la séquence d'acides aminés recherchée et sur la production d'un alignement qui mappe des résidus dans la séquence d'acides aminés recherchée à des résidus dans la séquence d'acides aminés modèle. Deux outils seront ainsi utilisés BLAST et Swiss-model.

i. Méthode BLASTp (3D) :

L'étude suivante sert à trouver des protéines présentant une structure 3D dans la base de données UniPort qui peuvent servir comme des templates pour notre protéine d'intérêt. Un BLASTp est ainsi réalisé, en tenant en compte les paramètres suivants :

1. Structures 3D des protéines
2. BLOSUM 80 (Les autres options de BLOSUM ont déjà été traité, il semble qu'avec BLOSUM 80, le nombre de protéines sorti est 8. Par contre un nombre de protéine inférieur à 8 est obtenu avec les autres BLOSUM)

Le résultat du BLASTp (structures 3D) figure ci-dessous.

Figure 6: BLASTp avec structures 3D et BLOSUM 80.

Le pourcentage d'identité des protéines obtenu par rapport à la protéine cible varie entre 28,1% et 44,0%. Les 8 protéines sont représentées dans le tableau ci-dessous.

Code de la protéine (UniProt)
P08194
Q51330
Q96T58
A1IIX4
A0A0M3KKZ1
Q63TA4
P03819

P9WGY3

Chacune de ces protéines a été analysées en se référant à la fois de la fiche PDB correspondante et également de résultat UniProt afin de sélectionner la plus adaptée à notre cible.

Interprétation des résultats de BLASTp (3D) :

La protéine Glycerol-3-phosphate transporter dont le code UniProt est P01894, présente un pourcentage d'identité est égale à 28,1 par rapport à la protéine cible (P57057). Cette protéine n'est autre que le template 1PW4 déjà vu précédemment. D'après la structure PDB, 1PW4 a une structure allant du résidu 3 d'acides aminés, jusqu'au résidu 448 d'acides aminés.

La protéine Oxalate:formate antiporter dont le code UniProt est Q51330, présente un pourcentage d'identité est égale à 29,2 par rapport à la protéine cible (P57057). D'après la structure PDB, 1ZC7 est un modèle, ainsi cette protéine sera rejetée pour la suite de l'étude.

La protéine Msx2-interacting protein dont le code UniProt est Q96T58, présente un pourcentage d'identité est égale à 31,6 par rapport à la protéine cible (P57057). D'après la structure PDB, 4P6Q a une structure allant du résidu 335 d'acides aminés, jusqu'au résidu 620 d'acides aminés.

La protéine Maleylacetate reductase dont le code UniProt est A1IIX4, présente un pourcentage d'identité est égale à 33,3 par rapport à la protéine cible (P57057). D'après la structure PDB, 3W5S a une structure allant du résidu 1 d'acides aminés, jusqu'au résidu 351 d'acides aminés.

La protéine Di-/tripeptide transporter dont le code UniProt est A0A0M3KKZ1, présente un pourcentage d'identité est égale à 44,0 par rapport à la protéine cible (P57057). D'après la structure PDB, 4W6V a une structure allant du résidu 1 d'acides aminés, jusqu'au résidu 511 d'acides aminés.

La protéine Putative membrane protein dont le code UniProt est Q63TA4, présente un pourcentage d'identité est égale à 43,9 par rapport à la protéine cible (P57057). D'après la structure PDB, 4USX a une structure allant du résidu 657 d'acides aminés, jusqu'au résidu 992 d'acides aminés.

La protéine Glutathione-regulated potassium-efflux system protein KefC dont le code UniProt est P03819, présente un pourcentage d'identité est égale à 38,6 par rapport à la protéine cible (P57057). D'après la structure PDB, 3L9X a une structure allant du résidu 401 d'acides aminés, jusqu'au résidu 620 d'acides aminés.

La protéine Putative 4-hydroxy-4-methyl-2-oxoglutarate aldolase dont le code UniProt est P9WGY3, présente un pourcentage d'identité est égale à 39,0 par rapport à la protéine cible (P57057). D'après la structure PDB, 1NXJ a une structure allant du résidu 2 d'acides aminés, jusqu'au résidu 157 d'acides aminés.

Le problème des protéines (Q63TA4, P03819, P9WGY3 et 1NXJ) est le taux de recouvrement qui ne prend pas en compte la totalité de séquence de la protéine cible (taux de recouvrement inférieur à 533). Ces protéines ne seront pas ainsi sélectionnées pour la suite de l'étude. Les

protéines ainsi restantes sont 3W5S et 4W6V. Dans l'étude qui suit, on va comparer les deux templates déduits avec le modèle 1PW4.

Test des templates choisis par BLASTp (3D) :

Le serveur Memoir, membrane protein modelling pipeline, est un algorithme de modélisation d'homologie conçu pour les protéines membranaires. Les entrées sont la séquence qui doit être modélisée (Fichier FASTA de la protéine cible P57057), et la structure 3D d'une matrice de protéine membranaire (Fichier PDB de la protéine template). Memoir intègre ces logiciels de protéines membranaires :

- iMembrane (pour l'annotation de la membrane)
- MP-T (pour l'alignement),
- Medeller (pour générer des coordinations)
- Completionist (modélisation de boucle).

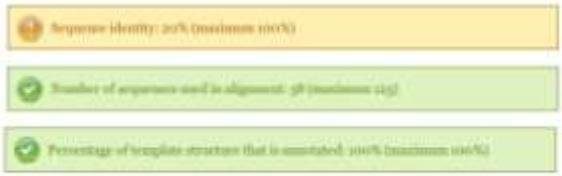


Figure 6: Résultat MEMOIR fait sur 1PW4.

Une analyse sur Memoir a été faite sur le modèle 1PW4, afin de comparer ensuite le résultat avec les deux autres modèles. Le résultat confirme que 1PW4 peut être un modèle fiable pour la protéine P57057, car le pourcentage d'identité de séquence est 20 (dans le seuil d'acceptation), le nombre de séquence utilisé dans l'alignement est 58, supérieur à 50 et à la fin la structure du 1PW4 est 100% annoté (seuil d'acceptation à partir de 95%) (Figure 6).

L'étude Memoir a été réalisé pour les templates 3W5S et 4W6V. Les résultats montrent une identité de séquence de 12% pour 4W6V, inférieur au seuil 20% d'où le modèle choisi est moins fiable que 1PW4. Par contre l'analyse a été rejetée par Mémoir pour le second template 3W5S à cause de la grande différence de la longueur de séquence entre P57057 (533 acides aminés) et 3W5S (350 acides aminés).

ii. Méthode SWISS-MODEL :

Le serveur SWISS-MODEL, permet la modélisation d'homologie de structure protéique entièrement automatisé. Dans l'option « SEQUENCE », il recherche des templates à partir de la séquence primaire d'acides aminés de la protéine cible (P57057). Le résultat obtenu est ci-dessous (Figure 7).

Figure 7: Résultat Swiss-model.

Le pourcentage d'identité le plus élevée dans les templates obtenus est 20,40% correspondant à la protéine 1PW4 (le modèle déjà vu). Tous les autre templates présentent un pourcentage de similarité inférieur à 15% (Figure 7).

D'après les résultats obtenus par BLAST et Swiss-model, il paraît que trouver un modèle par la méthode d'homologie n'est pas représentative. Donc une autre méthode sera alors adaptée qui pourra fournir des résultats significatifs.

g) Recherche d'un bon template par Protein Threading :

Le filetage protéique (protein threading), est une méthode de modélisation des protéines qui est utilisée pour modéliser les protéines qui ont le même pli que les protéines de structures connues, et n'ont pas de protéines homologues de structure connue comme par la méthode de modélisation par homologie de la prédiction de structure. Threading fonctionne en utilisant la connaissance statistique de la relation entre les structures déposées dans le PDB et la séquence de la protéine que l'on veut modéliser.

i. Méthode MUSTER :

L'outil utilisé pour faire le threading est « MUSTER », c'est un nouvel algorithme de threading MUSTER (Multi-Source ThreadER) qui combine différentes informations de séquences et de structures. Il prend en compte les paramètres les profils de séquence, la prédiction des structures secondaires, les profils de structure dépendant de la profondeur, l'accessibilité au solvant, les angles dièdres du squelette et matrice de notation hydrophobe (Wu S. *et al.,* 2009). Le serveur nécessite la séquence de la protéine cible (P57057) sous format FASTA. Le résultat est présenté ci-dessous (Figure 8).

Rank	Template	Align_length	Coverage	Zscore	Seq_id	Type	Target-template alignments	3-D models from threading alignments	Full-length models by MODELLER
1	1pw4A	420	0.787	14.879	0.205	Good	alignment_1	threading_1	model_1
2	3wdoA	405	0.759	9.356	0.111	Good	alignment_2	threading_2	model_2
3	4zowA	385	0.722	9.239	0.143	Good	alignment_3	threading_3	model_3
4	4ikvA	473	0.887	9.139	0.127	Good	alignment_4	threading_4	model_4
5	4iu8A	391	0.733	9.064	0.130	Good	alignment_5	threading_5	model_5
6	4w6vA	471	0.883	8.745	0.121	Good	alignment_6	threading_6	model_6
7	4j05A	401	0.752	8.667	0.120	Good	alignment_7	threading_7	model_7
8	4apsA	436	0.818	8.063	0.115	Good	alignment_8	threading_8	model_8
9	4q65A	432	0.810	7.824	0.113	Good	alignment_9	threading_9	model_9
10	4lepA	434	0.814	7.700	0.092	Good	alignment_10	threading_10	model_10

Figure 8: Résultat MUSTER.

MUSTER nous sort 10 résultats de protéines pourraient servir un template pour la protéine cible (Figure 8). Le but est toujours d'essayer de trouver un modèle mieux que celui adaptée dans la bibliographie (1PW4). 1PW4 présente une longueur d'alignement est égale à 420, d'où l'intérêt de trouver un autre modèle présentant une valeur supérieure et qu'elle soit proche de 533 (longueur de la séquence de notre protéine cible). Ainsi, 3WDO, 4ZOW, 4IU8 et 4J05 seront rejetées car ils ont les longueurs respectivement 405, 385, 391 et 401. Il reste alors 4IKV, 4W6V, 4APS, 4Q65 et 4LEP à tester. Plus le taux de recouvrement est grand plus le modèle approche de la fidélité (le seuil dans ce cas est 0,787 qui correspond à celui du 1PW4). 4IKV et 4W6V présentent le plus important recouvrement (0,887 et 471 respectivement), elles seront ainsi retenues. Ces dernières présentent aussi des valeurs de Zscore et pourcentages d'identité de séquence importants. Donc les protéines 4IKV et 4W6V seront ainsi sélectionnées comme des modèles candidats. D'après PDB, ces protéines sont transmembranaires et jouent le rôle de transporteurs. 4W6V a une résolution de 3,02 A° et contient 14 hélices transmembranaires dans sa structure secondaire. D'un autre côté, 4IKV a une résolution de 1,9 A° et contient 12 hélices transmembranaires. Etant donné que le modèle 1PW4 a une résolution de 3,3 A° et que la prédiction de la structure de la protéine cible P57057 montre la présente de 12 hélices transmembranaire, le modèle numéro 1 ainsi choisi est celui de la protéine 4IKV et le modèle numéro sera alors 4W6V. Chacun des deux modèles présentent un avantage lui permettant d'être le plus adapté. Le domaine MFS de la protéine 4W6V s'étend du résidu d'acide aminé position 19 jusqu'à l'acide aminé position 506 (selon PDB). Le domaine MFS de la protéine 4IKV s'étend du résidu d'acide aminé position 1 jusqu'à l'acide aminé position 207 (selon PDB). La protéine cible P57057 semble d'après InterProt (trouvé sur UniProt), qu'elle présente le domaine MFS le long de sa structure primaire (allant de l'acide aminé 27 à 517). Malgré la présence de 14 hélices transmembranaires chez 4W6V, cela ne diminue pas la probabilité qu'il soit un bon modèle, la prédiction de la structure 2D de P57057 montre la présence de deux grandes boucles (résidus) cela peut être issue d'un calcul non précis des méthodes de prédictions utilisées, car d'après l'analyse de la structure 2D de P57057, il semble possible d'avoir deux hélices transmembranaires au milieu des deux grandes boucles. Chacun des deux modèles est probablement le bon modèle de la protéine P57057.

ii. Méthode HHPred :

Afin de confirmer le choix de notre deux templates, un second outil de protein threading est utilisé, HHPred est un serveur rapide et populaire de protein threading, largement utilisé et basée sur la comparaison par paires de modèles de Markov cachés. En effet, la présence de la protéine modèle dans différents outils de prédiction augmente la probabilité que le choix de ce template est vrai. Les protéines 4IKV et 4W6V sont présentes dans les résultats obtenus par HHPred, elles présentent respectivement une valeur

d'espérance E-value est égale à $1,6x10^{-28}$ et $4,4x10^{-25}$. Plus la valeur E-value est petite, plus est hautement improbable que le score d'alignement obtenu soit le fait du hasard, d'après la littérature cette valeur est comprise entre 10^{-10} et 10^{-200}.

Dans les études suivantes, on va s'intéresser à affiner les boucles et également les chaînes latérales des deux modèles sélectionnés 4IKV et 4W6V.

h) Affinement des boucles :

La structure d'une protéine peut être prédite précisément à partir de sa séquence par modélisation basée sur un modèle lorsque l'identité de séquence est suffisamment élevée (>30%). Cependant, malgré une identité de séquences élevée, la structure de la chaîne latérale peut être moins précise que la structure du squelette, même à une identité de séquences faible, les structures prédites peuvent avoir des erreurs significatives dans les structures de la chaîne latérale et du squelette (Heo L. *et al.,* 2013).

i. Méthode GalaxyRefine :

Le serveur GalaxyRefine, est une méthode de raffinement qui reconstruit d'abord les chaînes latérales et effectue l'emballage à nouveau de ces chaînes par simulation dynamique moléculaire. Cette méthode affine donc les modèles générés par les serveurs de prédiction de structure protéiques (Heo L. *et al.,* 2013).

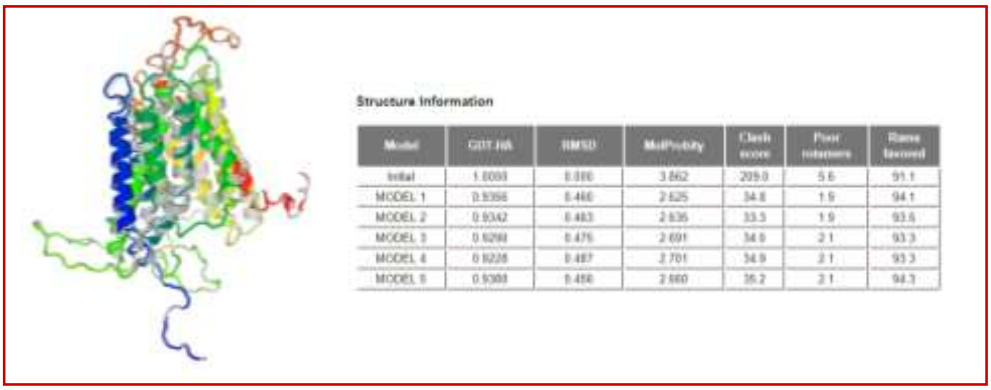


Figure 9: Résultat GalaxyRefine pour 4W6V.

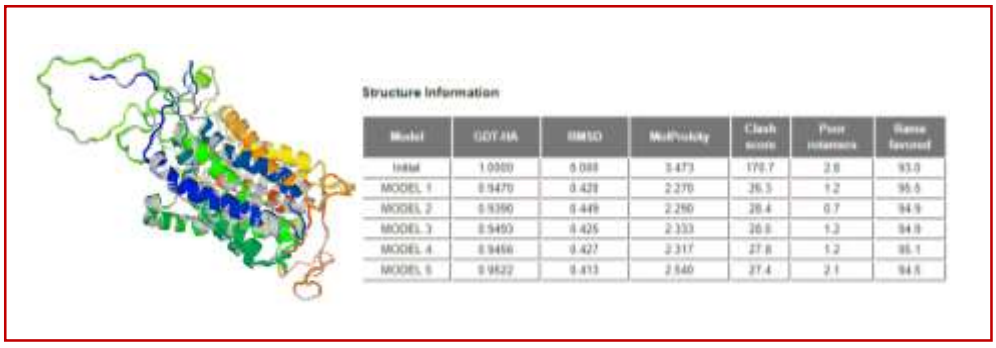


Figure 10: Résultat GalaxyRefine pour 4IKV.

Le serveur GalaxyRefine nécessite les fichiers PDB des modèles sortant des outils de prédiction, dans notre cas Muster et HHPred. GalaxyRefine reconstruit premièrement les chaînes latérales en plaçant les rotamères les plus probables. Le modèle avec les chaînes latérales construites est ensuite affiné par deux méthodes de relaxations douces et agressives.

Le modèle d'énergie la plus basse des 32 modèles générés par la relaxation douce est renvoyé en tant que « MODEL 1 », et les 4 modèles générés par la relaxation agressive sont renvoyés en « MODEL 2-5 ». Les résultats ainsi obtenus sont présentés ci-dessus (Figure 9 et 10).

Légende :

- GDT-HA : test de la distance globale est une mesure de la similarité entre deux structures protéiques, couramment utilisée pour comparer les résultats de la prédiction de la structure protéique à la structure déterminé expérimentalement.
- RMSD : erreur quadratique moyenne (Root-Mean-Square Deviation), mesure la différence entre les valeurs prédites par un modèle et les celles observées.
- MolProbity : score de validation de structure qui fournit une évaluation de la qualité des modèles protéiques (Chen V. *et al.,* 2009).
- Clash score : chevauchement atomique résultant de l'énergie Van der Waals supérieur à 0,3 Kcal/mol à l'exception que si les atomes sont liés ou forment de liaison disulfures ou liaisons hydrogènes, etc (Ramachandran S. *et al,* 2011).
- Rotamers : Il définit les angles de torsion autorisées à varier. Jusqu'à 50 rotamers, chaque chaîne latérale flexible peut en être définit. C'est l'énergie des angles de torsion dans les chaînes latérales.
- Rama favored : C'est l'énergie des angles de torsion dans le squelette protéique.

Après analyse des résultats obtenus par GalaxyRefine pour les 5 modèles appartenant à 4W6V, il paraît que les valeurs de scores varient très légèrement. Ainsi MODEL 1 a été sélectionné pour la suite du traitement (d'où son PDB résultant du serveur GalaxyRefine a été téléchargé)(Figure 9). Il paraît également que la différence entre les valeurs de scores est très négligeable pour les 5 modèles appartenant à 4IKV, ainsi MODEL 1 a été aussi sélectionné pour la suite du traitement (Figure 10). Une interprétation sur le modèle le mieux adaptée est difficile dans ce cas car les valeurs de scores entre les résultats obtenus pour 4W6V et 4IKV sont proches.

i) Affinement des chaînes latérales :

Un affinement supplémentaire tenant en compte les chaînes latérales pourrait été effectué grâce aux divers outils tels que RASP, SidePro, etc. Par contre une erreur produise lors d'affinement a empêché de récupérer ainsi les résultats. Les résultats des modèles (sous format PDB) sortant du GalaxyRefine seront donc utilisés.

j) Evaluation des modèles :

Afin d'évaluer nos modèles, une étude comparative de l'alignement du modèle avec la protéine initiale correspondant est capitale, afin de suivre l'évolution de la structure du modèle créer par rapport à sa structure initiale trouvée dans le PDB. TM-align, un algorithme qui permet la détermination du meilleur alignement entre les paires de protéines. Il existe une corrélation significative entre l'exactitude de la structure prédite et la similarité structurelle

du modèle avec les autres protéines du PDB. Cette corrélation pourrait être utilisée pour aider à la sélection de modèles prédits. Les résultats obtenus par TM-align sont représentés ci-dessous (Figure 11 et 12) (Zhang Y. *et al.,* 2005).

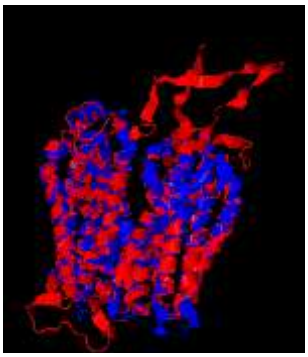


Figure 11: Résultat TM-align pour le modèle 4IKV.

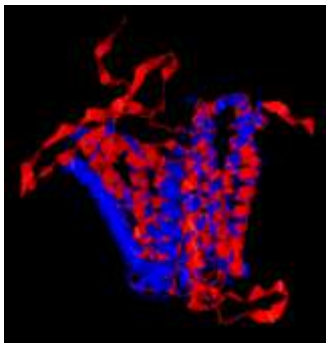


Figure 12: Résultat TM-align pour le modèle 4W6V.

Le tableau ci-dessous compare les valeurs des scores obtenus par TM-align des deux modèles 4IKV et 4W6V.

	TM-score	RMSD
4IKV	0.78928	2.25
4W6V	0.79698	1.96

TM-align attribue un score qui s'appelle TM-score, est défini qu'un bon pliement est obtenu pour TM-score comprise entre 0,5 et 1. La valeur moyenne des TM-score est négligeablement différente entre les deux modèles, par contre un RMSD paraît plus faible pour le modèle 4W6V par rapport à 4IKV. Vu que les deux modèles présentent des RMSD faible, cela confirme encore la fidélité des modèles choisis.

La dernière étape afin d'évaluer la fidélité des deux modèles utilise des outils d'évaluation de modèles tel est l'exemple de SaliLab Model Evaluation Server Submission (ModEval), ce dernier donne un RMSD score prédit du modèle évalué. RMSD est égale à 6 A° en évaluant le modèle 4IKV, par contre une valeur beaucoup plus élevée 17 A° a été obtenu en évaluant le second modèle 4W6V. Ce qui pourrait poser une hypothèse que le modèle 4IKV est mieux adapté que le second modèle. Une recherche sur d'autres outils d'évaluation de modèle prédit serait intéressante, car le temps de calcul pris par ModEval est environ quelques secondes, donc un calcul rapide comme ce dernier pourrait biaiser le résultat et donner de fausses hypothèse.