

Projet de Bioinformatique structurale

Introduction

Dans ce TP, nous allons étudier construire un modèle d'une protéine appartenant à la superfamille Major Facilitator Superfamily (MFS). C'est la plus grande famille de transporteurs secondaires. J'ai choisi la séquence Uniprot Q8IY34 **(1)** de l'espèce *Homo sapiens*. La protéine se nomme Solute carrier family 15 member 3, elle est codé par le gène SLC15A3. Cette protéine ne contient pas de structure obtenue par cristallographie. Notre but dans ce TP est de modéliser la structure de la protéine afin d'étudier la conformation inward ou outward **(2)** de la protéine et grâce aux modes normaux étudier la dynamique de la protéine et identifier les endroits qui fluctuent beaucoup dans cette protéine membranaire afin de comprendre le mécanisme de transport. Nous nous intéresserons surtout à la comparaison des outils de prédiction et de construction d'un modèle par homologie. Dans notre cas on considère que 2 séquences sont homologues s'ils ont au moins 20 à 30% d'identité.

Dans un premier temps, je vais présenter l'étude bibliographique réalisé sur la protéine d'intérêt Q8IY34**(1)** et la super famille à laquelle elle appartient. Dans un second temps, je vais exposer la recherche de template à partir de notre séquence d'intérêt à différentes procédures et la construction de modèles à partir de différents logiciels et pipelines (MéMoir, Swissprot, HHPRED).

I – Etude Bibliographique

La séquence d'intérêt a pour identifiant Uniprot Q8IY34 (1). La séquence a été revu manuellement par Swiss-Prot avec un score d'annotation de 4 sur 5. Notre protéine contient 581 acides aminés. Elle se nomme Solute carrier family 15 member 3 et elle est exprimée par le gène SLC15A3. Cette protéine joue un rôle dans les cellules du cerveau

Proton oligopeptide cotransporter. Transports free histidine and certain di- and tripeptides (By similarity). [By similarity](#)

GO - Molecular functionⁱ

- [peptide:proton symporter activity](#) [Source: Reactome](#)

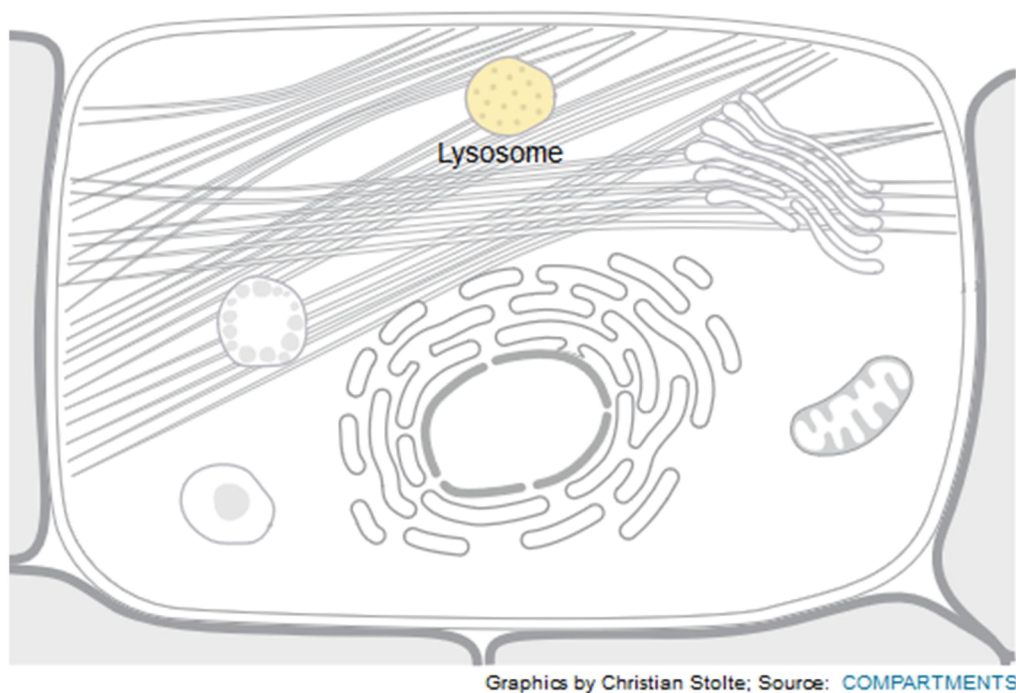
[View the complete GO annotation on QuickGO ...](#)

GO - Biological processⁱ

- [ion transport](#) [Source: Reactome](#)
- [protein transport](#) [Source: UniProtKB-KW](#)

Figure 1 – Fonction de la protéine

La figure 1 ci-dessus montre la fonction de cette protéine. Elle réalise le transport de peptide grâce à un transport symport actif d'ion.



 Manual annotation Automatic computational assertion

Figure 2 – Illustration montrant la localisation de la protéine.

La figure 2 ci-dessus montre la localisation de la protéine dans la cellule. Nous pouvons remarquer qu'elle joue un rôle au niveau de la membrane du lysosome (3) qui possède un pH acide. Ce paramètre sera à prendre en compte pour la réalisation du modèle final de la protéine. Si nos protéines supports sont localisées dans la membrane plasmique qui a un pH neutre.

Figure 3 - Tableau montrant les modifications post traductionnelles de la protéine.

La Figure 3 montre les modifications post traductionnelles, nous remarquons que la protéine contient 6 résidus asparagines (N) qui subissent une glycosylation. Une glycosylation de la protéine transmembranaire permet la création de glycoprotéine. Vous pouvez observer sur la figure 3 ci-dessus la description de ces résidus.

Figure 4 - Tableau montrant les protéines interagissant avec la protéine cible.

La base de données BioGrid que notre protéine interagit avec 8 autres protéines (4) dont 3 dans la membrane du lysosome qui sont impliqués dans l'apparition de syndrome. Le tableau de la figure 4 montre le nom des protéines qui interagissent avec notre protéine cible.

Figure 5 - Illustration montrant la position des 12 hélices transmembranaires selon l'annotation.

Ensuite notre protéine cible est une protéine transmembranaires qui contient 12 hélices transmembranaire selon la fiche Uniprot ci-dessus sur la figure 5. Pour valider ce résultat j'ai effectué des recherches contre des logiciels de prédiction à partir de la séquence protéique.

Vous pouvez trouver sur la figure 6 la comparaison des prédictions obtenus à partir de différent outils (HCA(6), MEMPACK méthode contenu dans PSIPRED(5))

Figure 6-A Résultat obtenu avec HCA(6)
Figure 6-B Résultat obtenu avec MEMSAT3

Le modèle finale doit ainsi posséder 12 hélices transmembranaires afin de le considérer comme cohérent. Nous pouvons voir que toutes les méthodes de prédictions ont le même résultat. La protéine contient 12 hélices transmembranaires.

Figure 7- Tableau montrant certains résidus caractéristiques de la superfamille.
La figure 7 montre les sous séquences qui sont conservées dans la famille. Il y a notamment la séquence Poly_Leu LLLLGL en position 165 à 170 et Poly-Val VVVV en position 180-183

Figure 8 - Illustration montrant la prédiction de PSIPRED(5)
La figure 8 ci montre que les partis Nter et Cter sont non structurés. Leurs structures secondaires sont donc des boucles qui sont désordonnés en plus. Pour la génération des modèles, nous allons enlever la partie Nter (position 1 à 32) et Cter dans notre fichier Fasta (position 566-581) afin d'obtenir un meilleur modèle.

Ensuite, nous pouvons remarquer que cette séquence possède déjà un modèle 3D obtenu à partir de la structure PDB support 4OH3 et la chaîne A. La figure montre le

modèle, il possède un score d'identité de 27% et qui couvre la région 15 à 568 qui correspond à un RMSD de 5 à 6 Å selon la fiche Uniprot.

Identification des résidus conservés

Avec un Blastp, nous allons tout d'abord rechercher les séquences qui ont une structure 3D résolue expérimentalement. J'ai réalisé plusieurs recherches avec différents paramètres. J'ai changé le paramètre de la matrice de score utilisée, j'ai tout d'abord essayé avec une matrice de score construite à partir des séquences qui ont 62% d'identité puis avec des séquences qui se ressemblent peu (45% d'identité) et finalement des séquences qui se ressemblent (BLOSUM80 avec des séquences qui ont 80% d'identité). Vous pouvez trouver ci-dessous les paramètres utilisés.

Figure 9 – Illustration montrant les paramètres de Blastp utilisés

Ensuite, j'ai sélectionné 6 séquences, sur la figure ci-dessous vous pouvez trouver les caractéristiques des séquences.

Code Uniprot	Identité (%)	Espèce	Revision
Q9JP7	21.5	Mus Musculus	oui
Q63424	22.3	Rattus Norvegicus	oui
Q0585	30.5	<i>Arabidopsis thaliana</i> (Mouse-ear cress)	oui
Q8EKT7	30	<i>Shewanella oneidensis</i> (strain MR-1)	non
A0A0M3KKZ1	31.7	<i>Yersinia enterocolitica</i>	non

Figure 10 Tableau montrant les séquences sélectionnées

Alignement des séquences

Nous avons effectué les alignement avec différentes méthodes ClustalW, PSI/Coffee-TM et TM-align

Figure 11 - Alignement obtenu à partir de Clustalw

La figure 11 ci-dessus montre un extrait d'alignement de la séquence cible Q8IY34 contre les séquences trouvées manuellement. Nous voyons bien que des résidus sont conservés (représentées par l'asterix) dans l'alignement.

Figure 12 - Alignement du modèle support (séquence QO5085)
La figure 12 montre que la séquence de la structure support QO5085 et de l'autre support Q8EKT7 contre les autres séquences.

Figure 13 – Extrait de l'alignement obtenu avec PSI/TM-Coffee

La figure 13 ci-dessus montre que les segments transmembranaires ont beaucoup de résidus dans les segments transmembranaires représentés en rouge et qu'ils sont globalement de même taille dans la superfamille.

II- Génération des modèles

A) Recherche de Template

Dans un premier temps nous avons effectué une recherche d'une structure support et construire notre modèle avec une méthode comparative par homologie. Pour cela nous avons utilisé plusieurs logiciel, un blastp en utilisant le serveur web Swiss-Model et le serveur HHPred.

A)1- Blast-Méthode manuelle

Pour choisir les différentes structures supports afin de construire le modèle par homologie, j'ai sélectionné à partir des fiches Uniprot des séquences sélectionnées après le Blastp. Pour cela je me suis basé sur différents critères. Est-ce que le fichier PDB recouvre une très grande partie de notre séquence cible. Est-ce que la résolution de la structure obtenue est raisonnable?

Enfin suite à une analyse bibliographique de l'article de fiche Uniprot j'ai regardé quelle était la conformation de la protéine transmembranaire

Uniprot	PDB	Résolution (Å)	Recouvrement PDB	Conformation
Q05085 (Revu manuelment)	AOH3	3.25	9-576	inward
	5A2N	3.7	21-573	outward
Q8EKT7 (Non revu)	2XUT	3.0	13-500	
	4UVM	3.6	8-512	

Figure 14 - Tableau montrant les caractéristiques des fichiers PDB sélectionnées.

Ensuite, j'ai construit les modèles à partir du serveur MEMOIR ([MEMbrane prOtein modeller](#)) (7) qui comprend un pipeline qui use Medeller. Medeller(8)

est une amélioration de Modeller(9) pour mieux prendre en compte les spécificités des protéines transmembranaires.

A)2- Swiss-Model

Dans le but de comparer les différentes méthodes de recherches de Template, j'ai utilisé le serveur Swiss-Model (8). Il utilise des chaînes de Markov cachées (HMM) afin de trouver des structures supports qui ont une similitude avec la séquence cible rentrée en paramètre. Vous pouvez trouver ci-dessous les résultats de la recherche et voir que dans le top 10, les structures supports trouvées manuellement à partir de la procédure BLAST sont retrouvées. Ce serveur donne une réponse intéressante, en effet, il indique qu'il est plus judiciable d'utiliser la chaîne B du fichier PDB 4OH3 que la chaîne A. Ceci pourrait peut améliorer le modèle déjà obtenu précédemment dans la fiche Uniprot de la séquence cible. Dans la prochaine étape de Génération des modèles je vais construire des modèles à partir de ce serveur qui utilise MODELLER.

A)3- HHPRED

Le serveur HHPRED (10) est le dernier outils qui sera utiliser pour chercher des structures support. Il utilise aussi des chaînes de Markov cachées (HMM). A partir de serveur nous retrouvons globalement les mêmes structures support obtenus précédemment. Nous lancerons aussi la génération de modèles à partir

Figure 15 – Illustration montrant les paramètres utilisées pour HHPRED

B)Construction des modèles

Figure 16 - Modèle obtenu avec la structure support 4OH3 et la chaîne B

C) Raffinement des modèles

Pour raffiner les modèles nous avons utilisé le serveur Galaxy Refine(11) qui permet le raffinement des boucles et d'améliorer le modèle obtenu.

Figure 17 - Modèle obtenu à partir de la structure support 4OH3 et la chaîne B.
La figure 17 ci-dessus montre que le modèle contient bien 12 hélices transmembranaires

Figure 18 - Modèle final obtenu avec la structure support 4OH3 et la chaîne B.
La figure 18 a été obtenue avec le serveur PPM montre bien que les 12 segments transmembranaires sont localisés dans la membrane.

D)Évaluation des modèles

Ensuite, Afin d'évaluer les modèles j'ai utilisé les outils Prosa et Verify3D ainsi que le serveur ppm qui permet de déterminer la localisation de la des segments transmembranaires dans un fichier pdb.

Figure 19 - Extrait du graphique obtenu avec verify3D

La figure ci-dessus montre que le score obtenu est 58,3% supérieur à 0,2 alors qu'il devrait être au moins de 65%.

Conclusion

En conclusion, à partir de l'analyse Bibliographique et la recherche de plusieurs protéines supports plusieurs serveurs. J'ai construit différents modèles à partir de différents serveurs. Suite à ces comparaisons entre les serveurs nous pouvons supposer que le serveur MEMOIR est plus adapté dans notre cas. En effet, ce pipeline utilise MEDELLER pour construire les modèles alors que Swiss-Model et HHPRED bien qu'ils utilisent des chaînes de Markov cachées pour la recherche des structures supports, il construit les modèles à partir de MODELLER. Cependant, Swiss-Model et HHPRED permettent de construire des modèles plus rapidement. Les perspectives de ce projet seront d'identifier

Bibliographie