# PROJET DE STRUCTURE :
# Construction de modèle tridimensionnel de la protéine O15244

**LA Kévin**
**Master 2 Bioinformatique**
**Université Paris Diderot**

# Sommaire

# Introduction :

Le but de ce projet est de générer un modèle de structure 3D à partir d'une séquence. La protéine a une fonction de transporteur et appartient à la famille des *Major facilitator superfamily* (MFS).

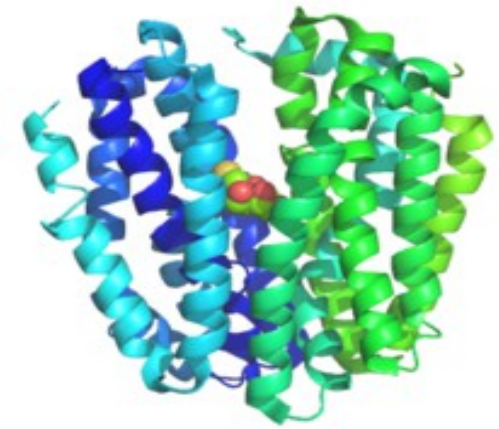


*Figure 1 : Une structure 3D des transporteurs de la famille MFS*

Pour ce projet je vais m'intéresser à la protéine *Solute carrier family 22 member 2* (code uniprot : O15244) appartenant à la famille des MFS. Cette famille est très importante car elle permet entre autre le passage du glucose, essentiel au bon fonctionnement des cellules. Cette protéine est codée par le gène SLC22A2 chez l'homme et est située sur le chromosome 6. Cette protéine est un transporteur de cation organique situé sur la membrane des cellules. Elle est présente dans plusieurs organes dans le foie, les reins, l'intestin car elle joue un rôle essentiels pour l'élimination de nombreux petits cations organiques endogènes. Dans cette famille il y a 12 hélices transmembranaires.
Sur O15244, il y a une modification post-transcriptionnel (glycosylation) sur asparagine en position 72 du côté extracellulaire. La fonction d'une protéine étant directement lié à sa structure, résoudre la structure de O15244 permettra une meilleur compréhension de son fonctionnement.

D'après l'étude d'Afinsen, il a été montré que toute l'information nécessaire au repliement et contenu dans sa seul séquence. De plus, nous savons qu'une structure de protéine est plus conservé que sa séquence. Il est donc théoriquement possible de construire une structure en se basant sur des séquences dont nous connaissons leurs structures. En se basant sur ces principes, deux grandes méthodes ont été développé, la construction de structure par homologie ou threading.

**Modèle d'homologie :** Pour la construction de structure, ce modèle recherche une séquence similaire à notre séquence d'intérêt dont les séquences sont connues. Lorsque des séquences similaires sont trouvées, elles partagent un ancêtre commun et donc la construction d'un modèle est possible.

**Modèle de threading :** Pour la construction de structure, ce modèle se base sur l'idée que la fonction est plus conservée que la séquence. À partir d'un ensemble de séquence, il faut identifier qu'elles sont les bon repliements à travers une fonction de score.

Pour ce projet, dans le but de générer le meilleur modèle possible, j'ai utilisé plusieurs approches basé sur les deux types de modèle cité précédemment.

# Matériels et méthodes :

## *protéine :*

Sur la figure 2 est représentée la place du gène SLC22A2 sur le chromosome 6 (https://www.ncbi.nlm.nih.gov/gene/6582). Ce gène code pour 6 transcrits dont 2 protéines O15244 et Q5T7Q5.

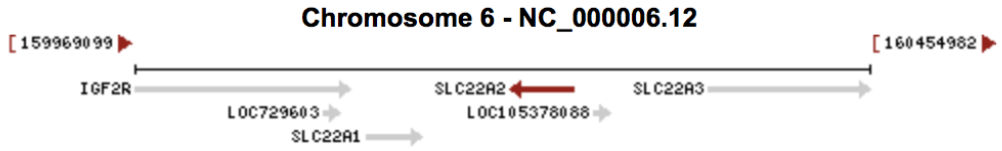



*Figure 2 : Transcrit du gène SLC22A2 situé sur le chromosome 6 chez l'Homme*

Dans un premier temps, pour m'assurer que la structure de O15244 n'a pas été résolu, j'ai entrer le numéros d'accession dans UNIPROT dans lequel aucune structure PDB n'a été trouvé (http://www.uniprot.org/uniprot/O15244#structure). J'ai choisie cette protéine car sa présence a été démontré expérimentalement et a été « reviewed ». De plus, elle joue un rôle très important dans le transport d'ion et le transport de glucose chez l'homme donc elle représente un intérêt certain dans la résolution de sa structure.
La protéine O15244 a une longueur de 555 résidus avec 1 domaine résolu d'après le site pfam (http://pfam.xfam.org/protein/O15244).


*Figure 3 : Domaine identifié par pfam pour la protéine O15244*

Le domaine Sugar_tr (transporteur de sucre) débute du résidu 91 au résidu 529. Comme O15244 est principalement exprimé dans le rein, cette protéine est impliqué dans des maladies comme l'hyperuricémie, l'insuffisance rénale et des liaisons rénale.

## *Méthodes :*

MODELLER est une méthode d'homologie. Pour identifier les templates, des structures de templates multiples peuvent être alignées avec différents domaines de la cible (ils ont des petits recouvrement avec la cible). Dans ce cas, la modélisation peut construire un modèle basée sur l'homologie de la séquence cible entière. Deuxièmement, les structures de modèles peuvent être alignées avec la même partie de la cible. Dans ce cas, la modélisation est susceptible de construire automatiquement le modèle sur le meilleur template local. Pour la sélection du meilleur modèle, un Z-score de ProsaII mesure la compatibilité entre un template et une structure. La construction du modèle est obtenu en utilisant une méthode de restriction spatiale (les préférences des longueurs de liaisons et des liaisons d'angles sont obtenues par le champ de force CHARMM-22).
https://doi.org/10.1016/S0076-6879(03)74020-8

SWISS-MODEL est une méthode d'homologie couramment utilisée pour générer un modèle 3D lorsque les séquences ne sont pas résolues expérimentalement. Simple d'utilisation, SWISS-MODEL est un serveur avec une interface très simple s'utilisation. Pour générer un modèle, il suffit

simplement d'entrer la séquence de la protéine d'intérêt. Une fois le calcul lancé, SWISS-MODEL identifie les templates, sélectionne le meilleur template, construit le modèle et estime la qualité du modèle.
Pour l'identification des templates, l'algorithme de SWISS-MODEL utilise blast et HHblits. La sélection est faite par les propriétés d'alignement des templates avec la séquence cible (identité, recouvrement). Une fois la sélection du template fait, MODELLER construit le modèle. Enfin, pour évaluer les modèles il y a un indicateur de qualité global, le QMEAN. Cette indicateur est basé sur une fonction utilisant plusieurs potentiel de force comme la distance interatomique, l'accessibilité au solvant et les angles de torsions. Un bon modèle est caractérisé par un score élevé. https://doi.org/10.1093/nar/gku340

HHPred est une méthode de détection de protéine homologue. Dans un premier temps, à l'aide de multiple itération de psi-blast, Hhpred recherche des séquences homologues, puis les alignent. Dans un deuxième temps, un profile de chaîne de markov caché est généré depuis l'alignement multiple. Ce profile contient entre autre une description de la structure secondaire et des statistiques d'insertion et délétion. Puis pour la construction du modèle, une fois les templates sélectionnés, la construction du modèle est fait par MODELLER.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160169/

Phyre2 est une méthode de construction de modèle suivant le figure 4.

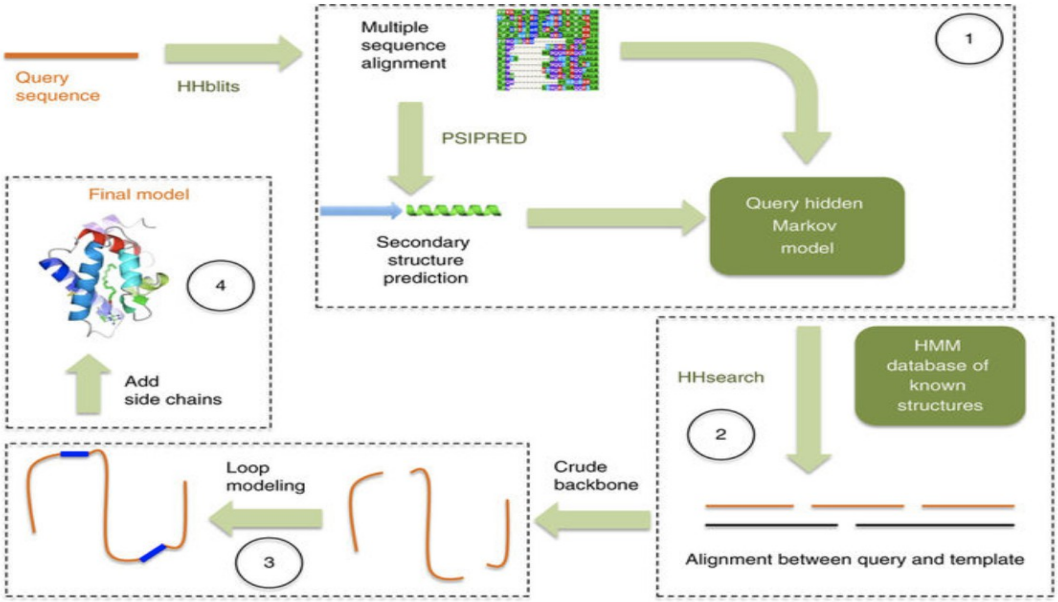



*Figure 4 : Pipeline de l'utilisation de phyre2 pour la construction de modèle 3D*

https://www.nature.com/articles/nprot.2015.053

i-TASSER est une méthode de threading permettant de structure protéique en se basant sur la séquence en acide aminé de la protéine. Le pipeline est décrit dans la figure 5. Dans un premier temps LOMETS identifie des structures homologues. La structure est ensuite construite par assemblage de fragment identifié par LOMETS et enfin après raffinement, la structure ayant la plus faible énergie est sélectionnée.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4871818/

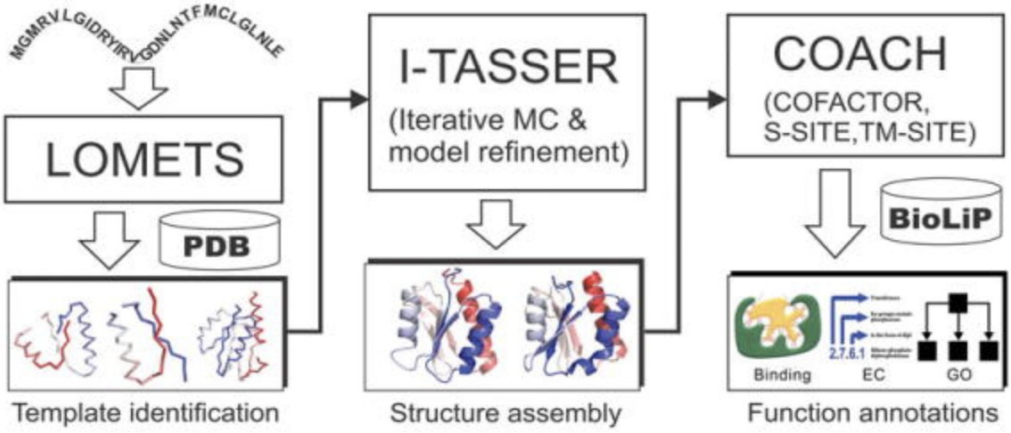



*Figure 5 : Pipeline de l'utilisation de i-TASSER pour la construction de modèle 3D*

T-Coffee et plus précisément TM-Coffee est une méthode d'alignement plus spécialisée dans les protéines transmembranaires. En effet, elle recherche les séquences homologues dans les bases de données spécifiques aux protéines transmembranaires. De plus, celle méthode permet de déterminer les régions transmembranaires à l'aide de l'algorithme HMMTOP.
https://www.ncbi.nlm.nih.gov/pubmed/27106060

Le serveur de PSIPRED est une méthode de prédiction de structure secondaire à partir d'une séquence protéique mis en input. De plus, se serveur nous permet en parallèle d'utiliser d'autre méthode permettant de prédire les hélices transmembranaires et les domaines de la protéine.
https://www.ncbi.nlm.nih.gov/pubmed/10869041

RaptorX est une méthode de threading permettant la construction de modèle 3D avec une méthode de prédiction de structure secondaire se basant sur la séquence.
https://www.ncbi.nlm.nih.gov/pubmed/24573471

# Résultats

## *Régions conservées*

Ces régions ont été identifié sur des séquences similaires à O15244. C'est à dire que j'ai choisie une séquences faisant partie de la famille des MSF mais pouvant appartenir à différents organisme. Pour ce faire, j'ai effectué un BLAST avec la séquence de O15244 sur la base de donnée UNIPROT et j'ai sélectionné les séquences avec un taux d'identité allant de 90 à 50 pourcent chez plusieurs espèce. À partir des séquences sélectionnées, j'ai récupéré 16 séquences au format fasta, puis j'ai lancé l'alignement de ces séquences sur t-coffee pour des protéines transmembranaires. Les résultats sont décrits sur la figure 6. Sur cette figure, les résidus sont globalement bien conservés notamment ceux situés dans la membrane. Il est donc possible de construire un modèle de O15244 car malgré la distance génétique, les séquences sont bien conservées. De plus, l'algorithme HMMTOP a prédit 12 hélices transmembranaires. Cette prédiction est en accord avec le nombre

d'hélice transmembranaire cité dans l'article de Robert Augustin and Eric Mayoux (http://dx.doi.org/10.5772/58325). Dans cette article, les auteurs ont trouvé 12 ou 14 hélices selon la famille de transporteur de glucose. De plus, on peut noter que la prédiction de structure secondaire (hélice) est la même pour toute les séquences.

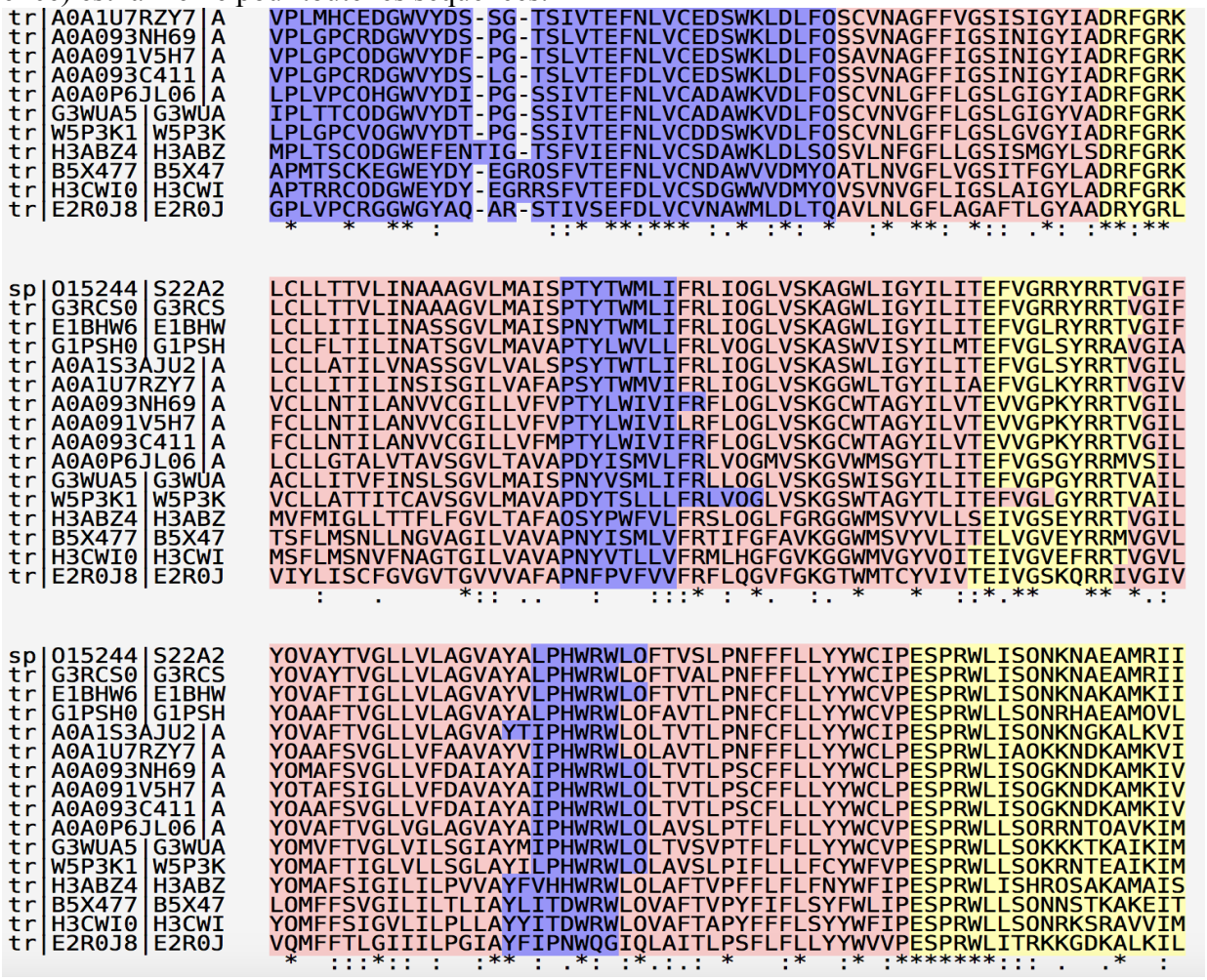
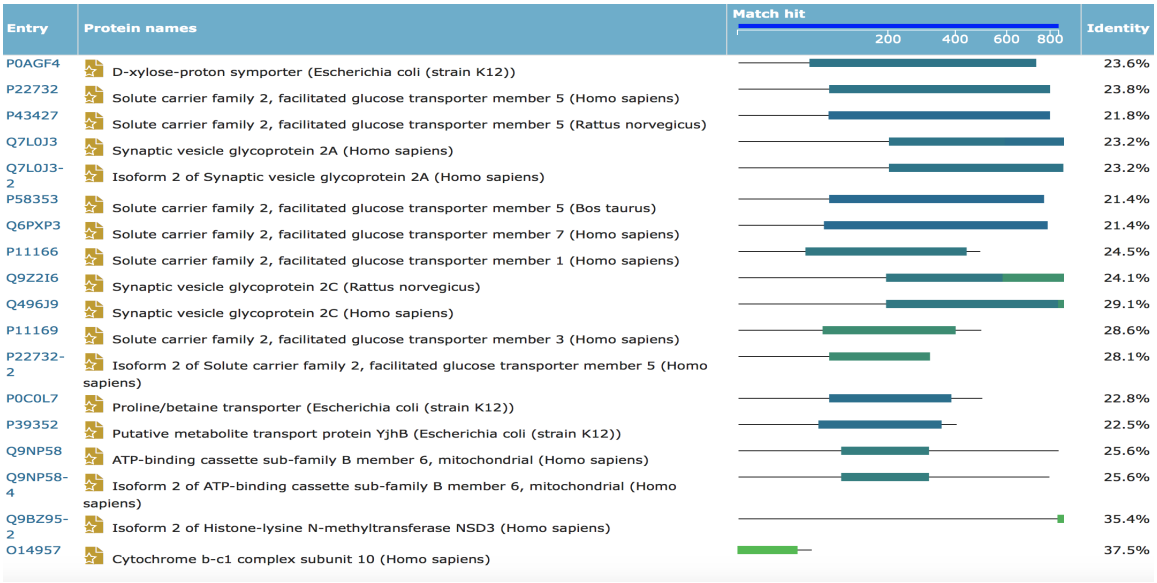



*Figure 6 : Résultats de l'alignement de t-coffee. En jaune sont représentés les résidus situés dans la membrane, en rose les résidus situés sur une hélice et en bleu, les résidus situés à l'extérieure de la membrane*

## *Identification des séquences supports*

Les séquences supports ont été identifié en effectuant un BLAST sur la PDB par rapport à la séquence de O15244. Le résultat du BLAST sur la base de donnée de la PDB a donné 21 structures dont 18 reviewed.



| Entry | Protein names | Match hit (200 400 600 800) | Identity |
|---|---|---|---|
| P0AGF4 | D-xylose-proton symporter (Escherichia coli (strain K12)) | | 23.6% |
| P22732 | Solute carrier family 2, facilitated glucose transporter member 5 (Homo sapiens) | | 23.8% |
| P43427 | Solute carrier family 2, facilitated glucose transporter member 5 (Rattus norvegicus) | | 21.8% |
| Q7L0J3 | Synaptic vesicle glycoprotein 2A (Homo sapiens) | | 23.2% |
| Q7L0J3-2 | Isoform 2 of Synaptic vesicle glycoprotein 2A (Homo sapiens) | | 23.2% |
| P58353 | Solute carrier family 2, facilitated glucose transporter member 5 (Bos taurus) | | 21.4% |
| Q6PXP3 | Solute carrier family 2, facilitated glucose transporter member 7 (Homo sapiens) | | 21.4% |
| P11166 | Solute carrier family 2, facilitated glucose transporter member 1 (Homo sapiens) | | 24.5% |
| Q9Z2I6 | Synaptic vesicle glycoprotein 2C (Rattus norvegicus) | | 24.1% |
| Q496J9 | Synaptic vesicle glycoprotein 2C (Homo sapiens) | | 29.1% |
| P11169 | Solute carrier family 2, facilitated glucose transporter member 3 (Homo sapiens) | | 28.6% |
| P22732-2 | Isoform 2 of Solute carrier family 2, facilitated glucose transporter member 5 (Homo sapiens) | | 28.1% |
| P0C0L7 | Proline/betaine transporter (Escherichia coli (strain K12)) | | 22.8% |
| P39352 | Putative metabolite transport protein YjhB (Escherichia coli (strain K12)) | | 22.5% |
| Q9NP58 | ATP-binding cassette sub-family B member 6, mitochondrial (Homo sapiens) | | 25.6% |
| Q9NP58-4 | Isoform 2 of ATP-binding cassette sub-family B member 6, mitochondrial (Homo sapiens) | | 25.6% |
| Q9BZ95-2 | Isoform 2 of Histone-lysine N-methyltransferase NSD3 (Homo sapiens) | | 35.4% |
| O14957 | Cytochrome b-c1 complex subunit 10 (Homo sapiens) | | 37.5% |

*Figure 7 : Résultats du BLAST de 015244 sur la PDB*

Sur la Figure 7, il est important de noté qu'aucune structure résolue couvrent la totalité de notre séquence. De plus, le taux d'identité des structures résolues varie de 21 à 37 pourcent. Pour les quelques séquences appartenant à la famille des MSF, 2 ont été résolu par un modèle. Pour les autres structures, elles ont été obtenu de avec la méthode X-ray (résolution moyenne 3,2Å). Il sera donc très difficile de déterminer une structure tridimensionnel de O15244 à cause de ces raisons.

## *Prédiction de structure secondaire par PSI-Pred*

À l'aide de PSI-Pred, j'ai regardé la prédiction de structure secondaire ainsi que la prédiction de domaine pour pouvoir par la suite les comparer avec d'autre méthode. PSI-Pred me prédit 12 hélices transmembranaires ainsi qu'une fonction de transporteur.

## *Construction des modèles*

Dans un premier temps j'ai construit des modèles en utilisant le serveur de SWISS-MODEL. Pour cela, après l'alignement fait par le serveur, j'ai sélectionné les templates ayant un taux d'identité supérieurs ou égale à 20 pourcent. La mailleur structure basé sur le template 4GBY est représentée sur la figure 8.

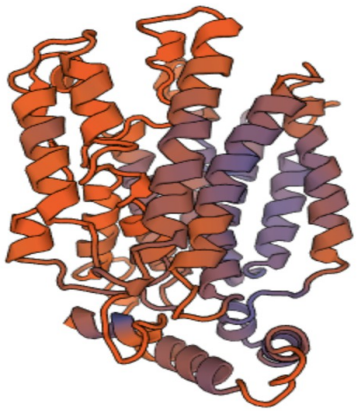


*Figure 8 : Représentation du mailleur modèle de SWISS-MODEL*

Le template sélectionné ne couvre pas la zone N-terminal de la protéine et a un taux d'identité de 26,86 pourcent.
Sur cette figure, on peut voir qu'il y a en majorité une haute estimation de la qualité (zone rouge) par rapport au zone de basse qualité (zone bleu). Les scores donnés par SWISS-MODEL pour ce modèle sont un GMQE de 0.43 et un QMEAN de -5.55. Plus la valeur du GMQE est élevé plus il y a une correspondance entre le template et la séquence d'intérêt. Le QMEAN représente le score des propriétés géométriques. Quand ce score est inférieur à 0.6, la qualité est considérée comme mauvaise. Ce modèle est le meilleur trouvé par SWISS-MODEL mais la structures trouvé n'est pas bonne (GMQE et QMEAN trop faible). Pour vérifier cette hypothèse, j'ai vérifié la qualité du modèle avec Verify3D et Prosa. Le résultat de Verify3D a montré que 54.59 pourcent des résidus ont un score supérieur ou égal à 0.2. Pour confronté le résultat de Verify3D, j'ai également lancé le teste sur Prosa. Prosa renvoi un score global de -3.96 (limite bon pour une résolution à X-ray) et des valeurs énergétiques le long de la séquence plutôt positives (indique que le modèle a beaucoup d'erreur).

Par la suite, j'ai utilisé Phyre2 pour la prédiction de structure secondaire et pour la construction d'un autre modèle. L'algorithme de prédiction des structures secondaire montre 12 hélices transmembranaires comme le montre la la figure 9 et prédit une fonction de transporteur de Glucose. On peut donc commencé les vérifications de la qualité du modèle.

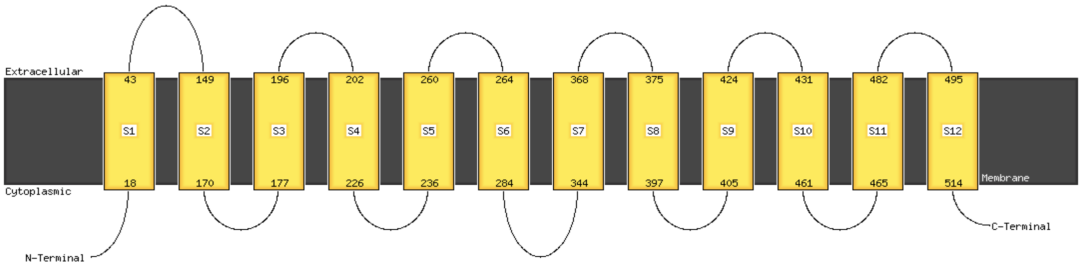



*Figure 9 : Prédiction de structure secondaire par Phyre2*

Pour la construction du modèle, Phyre2 c'est basé sur le template 4YBQ couvrant 78 pourcent de la séquence d'intérêt. La représentation du modèle est donné sur la figure 10.

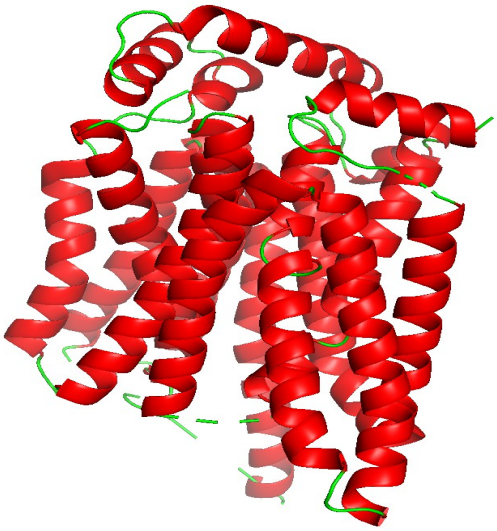


*Figure 10 : Modèle construit par Phyre2*

Tout comme le modèle de SWISS-MODEL, j'ai verifié la qualité du modèle construit par Phyre2 avec Verify3D et Prosa. Verify3D montre qu'il y a seulement 52.07 pourcent des résidus avec un

score supérieur ou égal à 0.2. Pour Prosa, le Z-score est de -1.46 (au dessus de la moyenne des résolutions des structures de X-ray). De plus, la qualité local est mauvaise car sur la courbe donné par Prosa, il y a un seul pique en négative. Donc globalement, la qualité de se modèle est mauvais que se soit en local ou global.

J'ai essayé de construire un modèle à partir de RaptorX. La construction du modèle a été fait sur la template 4GBY recouvrant 78 pourcent de la séquence d'intérêt. La structure est donnée sur la figure 11. Sur la figure, on peut voir que les boucles sont très mal placé. Pour l'étude de la structure, 51.17 pourcent des résidus ont un score supérieur ou égal à 0.2 donné par Verify3D. Le Z-score de Prosa est de -5.37. Ce qui nous donne une résolution de structure comparable au autre structure résolu par la méthode de X-ray. De plus, il y a une qualité au niveau local car sur la distribution des énergie, on trouve en moyenne plus souvent des énergies négative que positive (figure 12).

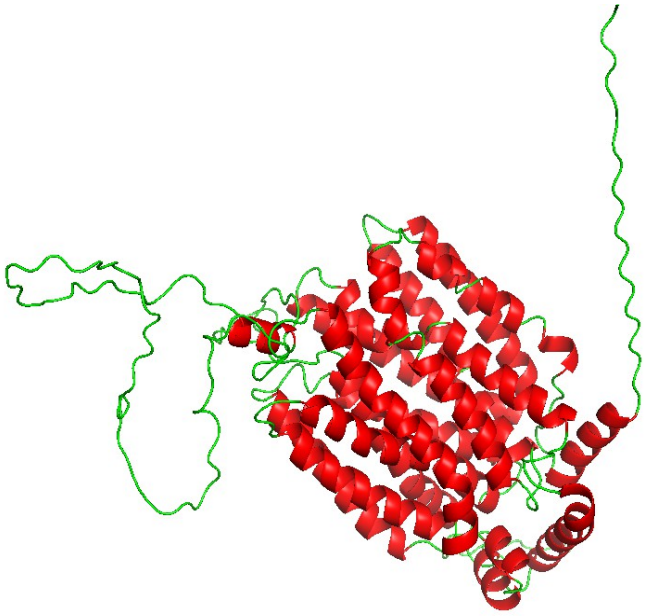


*Figure 11: Modèle construit par RaptorX*

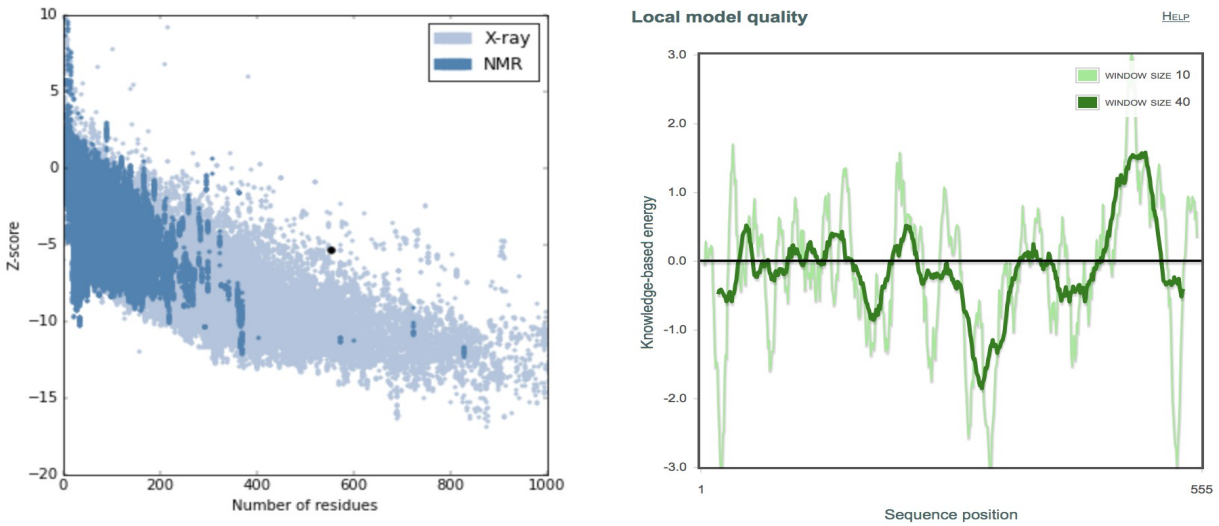



*Figure 12 : Contrôle de la qualité de Prosa du modèle construit par RaptorX*

Pour le modèle de MODELLER, j'ai tout d'abord fait un HHPred sur la séquence de la protéine O15244 puis à partir de l'alignement trouvé, j'ai choisie 9 templates appartenant à la famille des MFS mais ne recouvrant pas la totalité de la séquence (78% de recouvrement). La structure obtenu par MODELLER est donné sur la figure 13. Tout comme le modèle de RaptorX, on peut remarquer qu'une extrémité (boucle) de la protéine n'est pas bien modéliser. Lors du contrôle qualité sur Verify3D, on trouve que 60.09 pourcent des résidus ont un score supérieurs ou égal à 0.2. De plus, le contrôle de Prosa montre une similarité par rapport au modèle de RaptorX (figure 12).

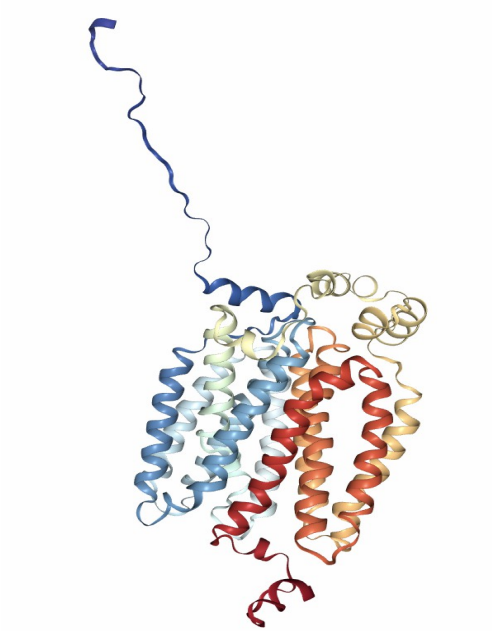


*Figure 13 : Meilleur structure obtenu par MODELLER*

Par la suite, j'ai construis un modèle à l'aide de i-TASSER. Le modèle de i-TASSER est donné sur la figure 14. Les scores associés à ce modèle donnés par i-TASSER sont un C-score (variable de -2 à 5) valant -1.40, une estimation du TM-score de 0.54 et du RMSD de 10.9. Pour le C-score, plus cette valeur est haute plus le modèle est considéré comme bon. Un TM-score se rapprochant de 1 et un RMSD le plus proche de 0 signifie une très bonne qualité de la structure.
Ici, on a un C-score et un TM-score moyens mais une valeur de RMSD très mauvaise. Cependant parmi tous les modèles générés, il semble que c'est la structure la plus probable.

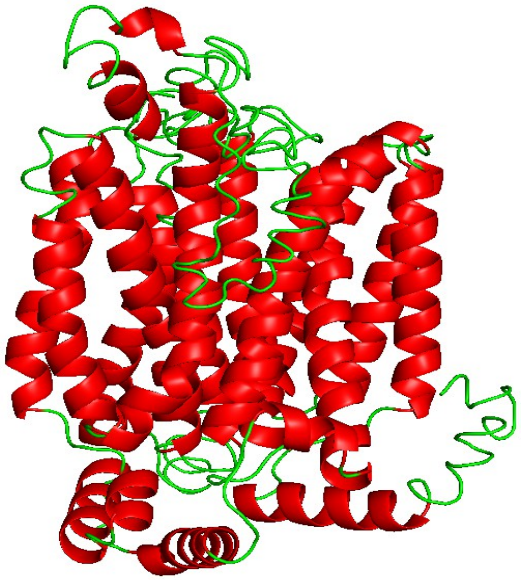


*Figure 14 : Modèle construit par i-TASSER*

Pour le vérifier, j'ai regardé les scores des programme Verify3D et Prosa. Le score que donne Verify3D est de 66.49 pourcent. Pour Prosa (résultat décrit sur la figure 15), la valeur du Z-score est de -1.71 qui est très largement en dehors des résolution classique de X-ray (point noire). De plus, sur la courbe d'énergie local, cette structure a 4 piques avec une énergie négative.

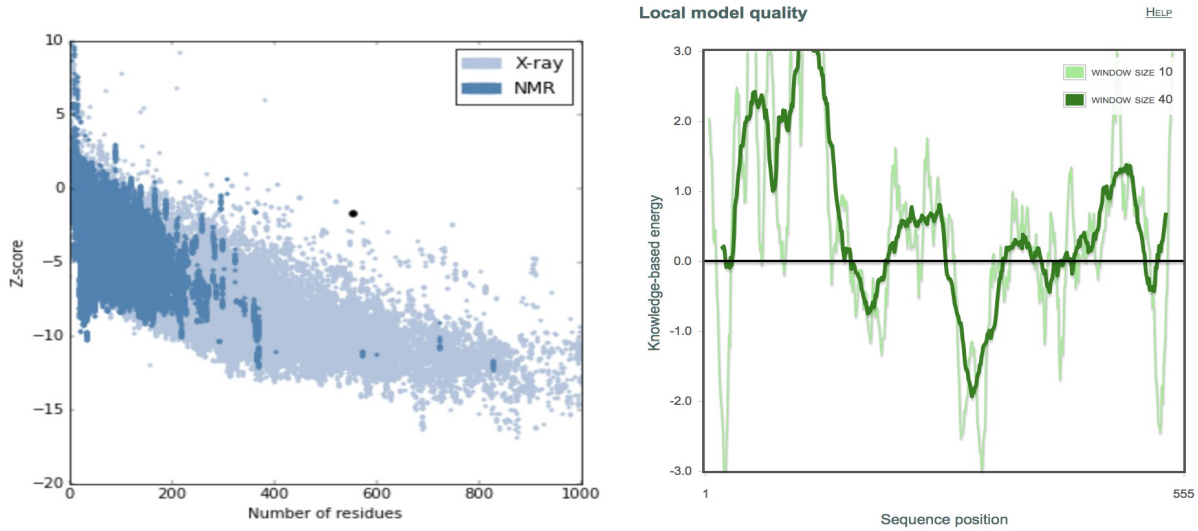



*Figure 15 : Contrôle de la qualité de Prosa du modèle construit par i-TASSER*

Les résultats de Verify3D et de Prosa ne sont pas excellent mais donne tout de même des valeurs de score les plus correcte pour les modèles de MODELLER et i-TASSER.

Partant de cette hypothèse, j'ai donc effectué un raffinement des boucles en utilisant GALAXY-refine. Cette algorithme permet entre autre de limiter les clashs stériques.
Pour le modèle de i-TASSER, le contrôle de la qualité de Prosa reste le même mais le score de Verify3D est passé à 75.14 pourcent des résidus ont un score supérieurs ou égal à 0.2. Pour le modèle de MODELLER, après le raffinement, il y a aucune amélioration lorsque que je teste la qualité dans Verify3D et Prosa.

# Conclusion

La structure de la protéine O15244 faisant partie de la famille des MFS impliquée entre autre dans le transport de glucose n'est à ce jour pas encore été résolu. Dans le but de construire une structure tridimensionnel de O15244, j'ai donc testé plusieurs programme permettant la construction de modèle par la méthode d'homologie ou de threading. Tous les programmes testé ont permis de construire un modèle de la protéine O15244 mais seulement 3 ont construits un modèle acceptable (qualité moyenne).
Cependant la qualité des modèles construit n'a pas permis d'étudier les modes normaux car pour les modèle possédant la longue boucle, seule celle-ci bouge. Pour le modèle de i-TASSER, il y a que des mouvements ne d'écrivant pas les mouvements d'ouvertures et de fermetures d'un transporteur.

# Références :

ProSA-web : interactive web service for the recognition of errors in three-dimensional structures of proteins, Markus Wiederstein and Manfred J. Sippl
http://www.scfbio-iitd.res.in/software/newProsav/Pdf/ProSA.pdf

MODELLER : Modeller: Generation and Refinement of Homology-Based Protein Structure Models, AndrásFiser AndrejŠali
https://doi.org/10.1016/S0076-6879(03)74020-8

SWISS-MODEL : modelling protein tertiary and quaternary structure using evolutionary information, Marco Biasini et al
https://doi.org/10.1093/nar/gku340

Phyre2 : The Phyre2 web portal for protein modeling, prediction and analysis, Lawrence A Kelley
https://www.nature.com/articles/nprot.2015.053

Hhpred : The HHpred interactive server for protein homology detection and structure, Johannes Söding
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160169/

i-TASSER : Protein Structure and Function Prediction Using I-TASSER, Jianyi Yang
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4871818/

RaptorX : RaptorX server: a resource for template-based protein structure modeling.
Källberg M
https://www.ncbi.nlm.nih.gov/pubmed/24573471

Psi-Pred : The PSIPRED protein structure prediction server. McGuffin LJ1, Bryson K, Jones DT.
https://www.ncbi.nlm.nih.gov/pubmed/10869041

TM-Coffee : PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. Floden EW
https://www.ncbi.nlm.nih.gov/pubmed/27106060