Protein structure prediction

Introduction:

The aim of this project is to predict the structure of a chosen protein using its sequence. That is possible with the use of tools and algorithms which are mentioned in the “Material and methods” section.

The protein Q9UGQ3 encoded by the SLC2A6 gene, also known as the *Solute carrier family 2, facilitated glucose transporter member 6*, or GLUT6, for short, is a protein from the *Major Facilitator Superfamily* (MFS). This protein superfamily gathers many membrane transporters which allow small molecules to cross the plasma membrane. These proteins are organized into three classes according to their resemblance. Our protein of interest belongs to the third class.

The human protein, GLUT6, previously known as GLUT9, allows the transport of intestinal glucose through plasma membranes. This protein is a symport, which means that both the sugar and an ion go through the plasma membrane in the same direction.

Its expression mainly occurs in human brain, in the spleen and in lymphoid tissues.

Material and methods :

I. MATERIAL:

From the accession number of the protein (Q9UGQ3), the protein sequence has been retrieved on Uniprot, a protein sequence database. GLUT6 sequence is composed of 507 amino acids. The amino acid sequence deduced from its cDNA predicts 12 putative membrane-spanning helices.

According to the literature, GLUT6 sequence contains all the motifs that are characteristic of the family of sugar transporters. In particular motifs corresponding with the PESPR/PETKGR motifs after helices 6 and 12, the GRR motifs in loops 2 and 8, and glutamate and arginine residues in the intracellular loops 4 and 10. Loop 1 in GLUT6 is short and lacks a glycosylation site. A glycosylation site appears to be located in the larger loop 9. A striking characteristic of GLUT6 is the presence of two arginines in the putative helices 7 and 8 at positions where the organic anion transporters harbour basic residues [2].

II. METHODS :

A. Template protein selection:

Since our protein structure is not known, a BLASTp of its sequence has been done on the Protein Data Bank (PDB), a crystallographic database for the three-dimensional structural data of large biological molecules.

We fixed the *E-Threshold* parameter to the default value (10). This parameter represent the number of matches between our sequence (query) and those from the database. We used the default matrix (BLOSUM62). The alignment program used is blastp (version 2.2.29+).

The selection of a template protein has been based on its quality. We estimate that a protein has a good quality when its coverage represent at least 80% of the query sequence. Furthermore, the template protein must have a low e-value. Indeed, the lower is that parameter (if the value tends towards zero), the better is the significance of the match.

This approach allows us to identify the proteins which have a putative similar structure to GLUT6.

B. Search for conserved residues:

A BLASTp alignment of our query protein has been made with the same parameters as above but on the UniprotKB databank. This approach allows us to identify proteins that share a sequence similarity with GLUT6.
Here, our aim is to identify conserved residues. It is not in our interest to select proteins with high identity percentage, but those which diverged a bit from ours (but still having a good coverage of the query sequence).

First of all, the WebLogo3 tool has been used to visualise the conservation of the residues throughout the sequence. To do so, we gave our multiple sequence alignment as an input to the tool. The bigger a letter (amino acid) is, the more it is represented at a given position in the sequence. [6]

We then made multiple alignments of the selected sequences using alignment tools named MUSCLE (version 3.8) and T-COFFEE (PSI/TM-COFFEE). We used the UniRef100 database with T-COFFEE for better results.

Finally, the Conserved Domain tool from NCBI has been used. For a given sequence (here, we gave our query protein), the tool searches through external databases (Pfam, SMART, …) sequences that match the query to identify conserved regions.[7]

C. Secondary structure prediction and transmembrane region detection:

GLUT6 is a transmembrane protein. It is in our interest to determine the transmembrane regions of it as well as its secondary structure organization.
These informations will next allow us to conclude on the quality of the generated structural models.

In addition to the multiple alignment, T-COFFEE returns a prediction of the secondary structure of the protein, which is based on the sequences that has been given for the alignment.

The HCA tool (Hydrophobic Cluster Analysis) creates a plot that allows us to decipher a protein sequence into its secondary structure. This tool clusters hydrophobic residues into globular regions. These regions allow us to identify the underlying secondary structures. [5]

The Bioinformatic Groupe at the University College of London offers a tool named PSIPRED that allows us to predict secondary structure. For a given sequence, many tools were used: PSIPRED version 3.3 for secondary structure prediction, MEMPACK (SVM Prediction of TM Topology and Helix Packing) and MEMSAT3 & MEMSAT-SVM (Membrane Helix Prediction).

D. Protein modelling and model refinement:

The aim of this project is to estimate the three dimensional (3D) structure of a given protein. This is important because the ability of a protein to interact with others depends on its 3D structure. Knowing the latter makes it possible to better understand its mode of action. Within in silico methods to model proteins exist two types of modelling : homology based modelling and *ab initio* modelling, also known as threading.
According to the identity percentage of the chosen templates, one method or both will be used. If the identity percentage is greater than 40%, homology based modelling will be

prefered. Otherwise, if the identity percentage is lower than 30%, *ab initio* modelling will be prefered.

1. Homology based modelling:

Homology based modelling use a template protein to modelize the structure of a given query protein. Here are the tools and algorithms used for homology based modelling :

- **MEMOIR:** MEMOIR is a homology modelling algorithm designed for membrane proteins. The given input are the protein to be modelled and the template from which the 3D structure will be modelized. MEMOIR uses four membrane protein software programs to align (iMembrane), annotate (MP-T), generate coordinate (Medeller) and modelize the loops (Completionist). This algorithm returns at least 3 models : one with the highest accuracy, one with the highest coverage and one representing only the transmembrane part of the protein. [8]

- **Swiss-model:** Swiss-model is a structural web tool dedicated for homology modelling of protein 3D structure. As an input, Swiss-model takes the fasta file or the sequence of the query protein to modelize. No template is requested. Swiss-model will then look for templates using a BLAST. A pairwise alignment between the query and the selected template will be done. Then the structure is modelized and its quality is assessed using QMEAN, a statistical potential of mean force. The more the QMEAN tends towards 0, the better is the model. [SM]

- **HHpred and MODELLER:** HHpred is an open-source program for sequence searching and structure prediction. HHpred is based on the pairwise comparison of profile hidden Markov models (HMMs) which greatly simplifies the list of hits. As an input, HHpred accepts a single query sequence or a multiple alignment. If a single query is given, a multiple alignment is done and the result is given to MODELLER. MODELLER is a program used for homology modelling.

2. Threading (fold recognition):

Ab initio based modelling simply use algorithms to compute a 3D structure for a given protein. This method takes longer time to generate models. Here are the tools and algorithms used for *ab initio* modelling:

- **Phyre2:** Phyre2 is a threading algorithm based on homology based methods.
- **MUSTER:** MUSTER is a standard threading algorithm based on dynamic programming and sequence profile-profile alignment. It also combines multiple structural resources to assist the sequence profile alignment.
- **RaptorX:** RaptorX is a protein structure and function prediction is developed by Xu group. For a Given a protein sequence, RaptorX predicts its secondary and tertiary structures as well as contact map, solvent accessibility, disordered regions and binding sites. RaptorX assigns confidence scores to indicate the quality of prediction results.
- **I-TASSER:** (**I**terative **T**hreading **ASSE**mbly **R**efinement) is a bioinformatic method for predicting three-dimensional structure model of protein molecules from amino acid sequences. The model generated using this method will be used as a control.

To refine our models we first compared each model with the informations retrieved in the literature (conserved residues and domains). Then, we performed a refinement of the underlying models using GalaxyRefine. GalaxyRefine performs repeated structure perturbation and subsequent overall structural relaxation by molecular dynamics simulation. For each model given as an input, 5 refined models are returned. Visualisation of them are made using PyMol.

E. Model validation:

To validate a model, the tool Verify3D has been used. Verify3D Determines the compatibility of an atomic model with its own amino acid sequence. The result is then compared to a good model. The web tool ProSA-web (Protein Structure Analysis) has been used as well. ProSA-web is usually used for refinement and validation of experimental protein structures and in structure prediction and modeling. The *z*-score indicates overall model quality and measures the deviation of the total energy of the structure. The lower is the z-score the better is the structural model. Indeed, a lower z-score reflects a native structural structure. []

F. Protein position in membrane prediction:

The PPM server calculates rotational and translational positions of transmembrane and peripheral proteins in membranes using their 3D structure given as an input. The output is a new PDB file containing both the protein and its membrane. [9]

G. Normal modes analysis:

The aim of this analysis is to visualize the movement of the protein inside the membrane. We used the NOMAD-Ref tool to calculate the normal modes of a given model.

III. RESULTS :

A. Template protein selection:

Two templates has been selected : GLUT1 (Uniprot accession number : P11166) and GLUT3 (P11169). It has been demonstrated that GLUT1 is the second relative of GLUT6 with 28.4% identical residues after GLUT8, which has 44.8% identical amino acids [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1221309/pdf/10970791.pdf].
Both of them belong to the MFS superfamily and are associated to glucose transport. Both GLUT1 and GLUT3 are located in plasma membrane.
GLUT1 and GLUT3 share 66% sequence identity and have 80% homology [1]. To allow sugar transport through the membrane, both GLUT1 and GLUT3 have to change conformation.

GLUT1 is encoded by the SLC2A1 gene. Its expression is mainly found in erythrocytes but the protein can be found in various organs in our body.
GLUT1 very broad substrate specificity, can transport a wide range of aldoses including both pentoses and hexoses. Many 3D structures are available in the PDB database, each having a different resolution. We chose the structure corresponding to the 4PYP PDP identifier. This structure is the only one which is not complexed to

other substrate. Its structure has been obtained using X-Ray diffraction method at a resolution of 3.17Å [2]. 4PYP has only one chain, composed of 492 amino acids.

GLUT3 is encoded by the SLC2A3. It is mainly located in human brain but can be found in other organs.
We chose the outward-occluded conformation of this template. Its structure has been obtained using X-Ray diffraction method at a resolution of 1.5Å [3]. 4ZW9 is the PDB code of this structure. It has only one chain, composed of 518 amino acids.

B. Search for conserved residues and secondary structure prediction:

The multiple alignment on the UniprotKB database allowed us to retrieve 10 MFS proteins belonging to different organisms. These sequences' identity percentage range from 70 to 98%.
First, using multiple alignment tool such as MUSCLE and T-COFFE, we observed that even if the algorithm behind these tools differs, the same residues are conserved throughout the whole sequence. Here is a WebLogo3 representation of the conservation of the residues.

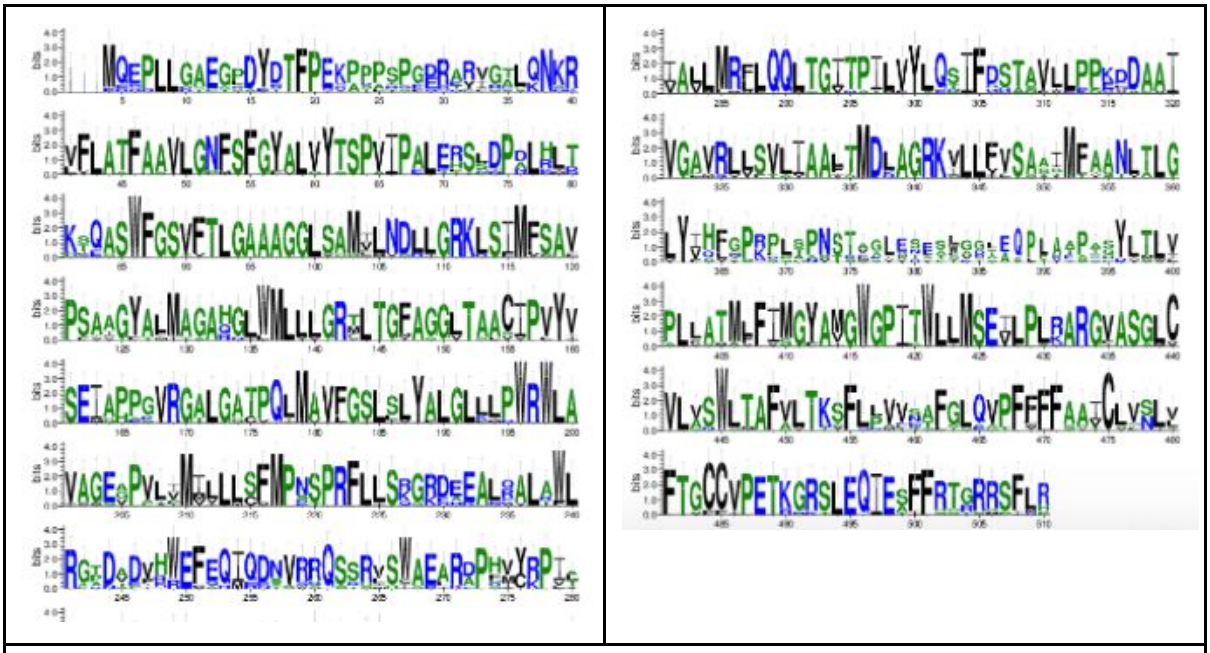


Figure 1: Residues conservation throughout the whole sequence. WebLogo plot.
Color legend: Blue = Hydrophilic amino acids; Green = Neutral; Black = Hydrophobic

A dileucine motif can be seen in the N-terminal region at position 8 and 9. The PETKGR motif in C-terminal is in fact a PETKG in our sequence between position 484 and 489.

In addition to the multiple alignment, T-COFFEE returns a prediction of the secondary structure of the protein.

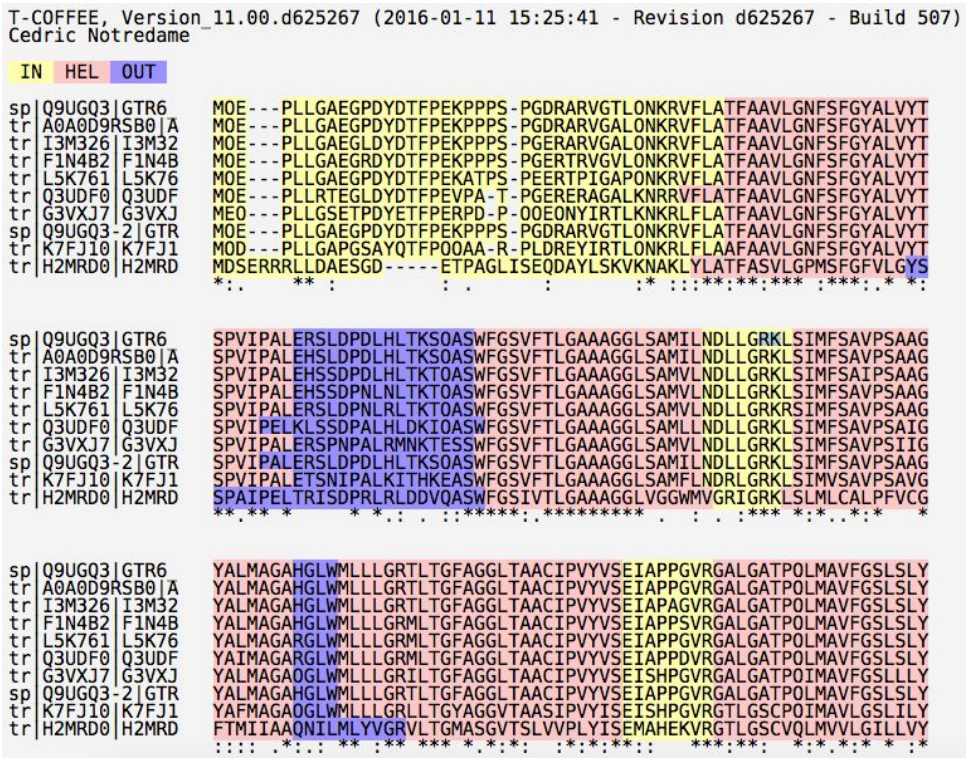


Figure 2: Insight of the multiple alignment and secondary structure prediction of Q9UGQ3. T-COFFEE tool. Legend : Yellow : Loops inside the membrane; Blue : Loops outside the membrane; Red : Transmembrane helix. See Supplementary Figure 1.

A total of 12 transmembrane helices has been predicted using T-COFFEE, 8 transmembrane loops and 7 outside of it.

In respectively intramembrane loop 2 and loop 8, the GRK motif and the GR motif are conserved for the 10 proteins selected. Two arginines can be found respectively in intramembrane loop 4 and 8. In intramembrane loop 5, the PNSPR motif is conserved. In helix 6, two glutamines (Q) are conserved. In helix 10, the GWGPITW motif is conserved.

These motifs are characteristic of the family of sugar transporters. And correspond to the figure below (Figure 3), which is a prediction of GLU6 (formerly GLUT9) sequence signatures.

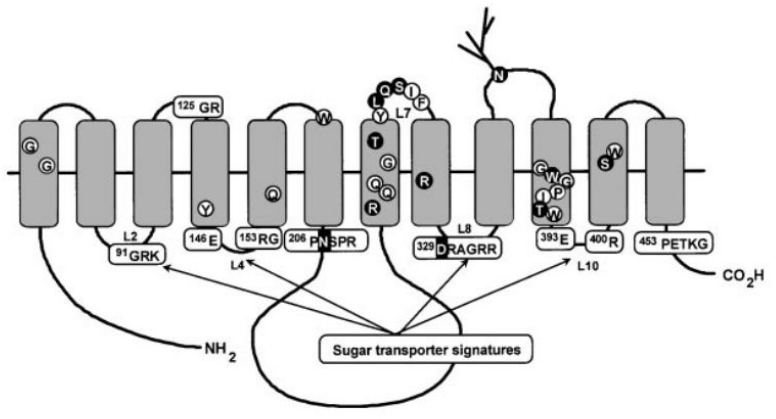


Figure 3: Putative membrane topology of GLUT6 and sequence signatures. [0]

The same results as above have been found using UCL's tool PSIPRED and MEMSAT.

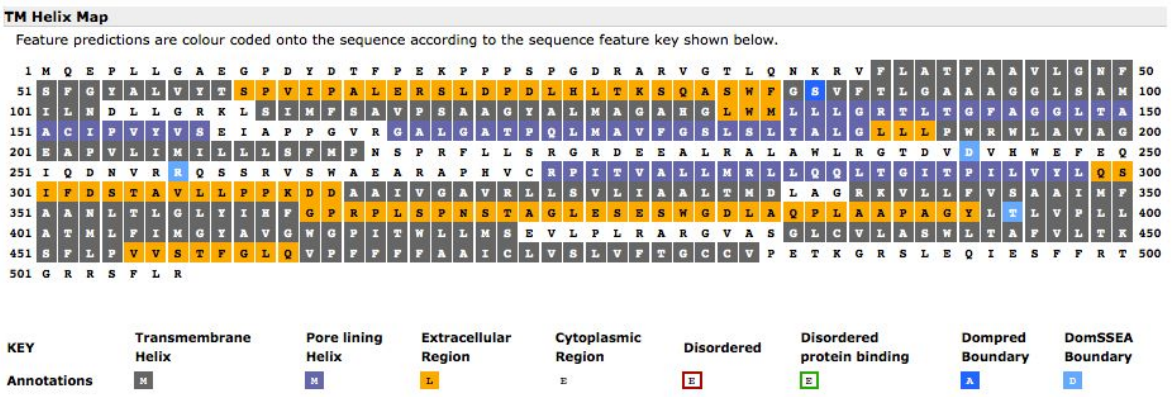


Figure 4.A: Transmembrane helix MEMSAT map.

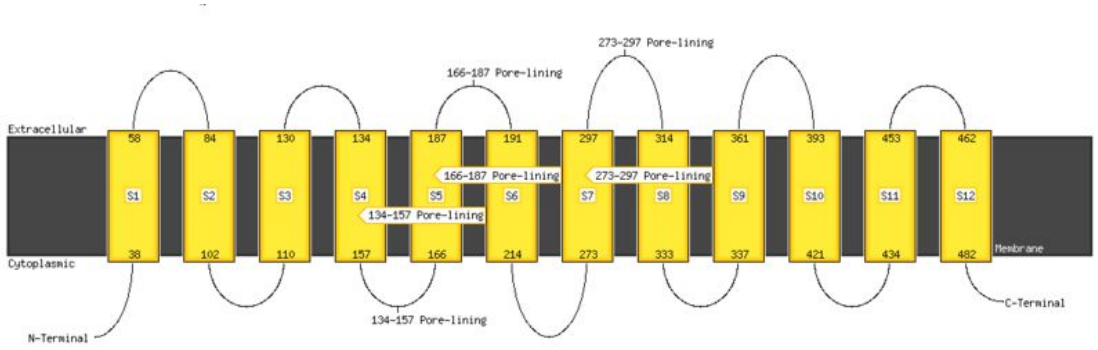


Figure 4.B: MEMSAT-SVM Cartoon : Representation of the proteins secondary structure.

In figure 4.A, 12 transmembrane helices have been predicted. Three of them are lining the ligand pocket (pore). This figure brings additional precision to the secondary structure of GLUT6. Figure 4.B shows the previous informations and completes the membrane topology predicted in figure 2.

To compare these results, we used the HCA tool to predict the secondary structure of our protein.

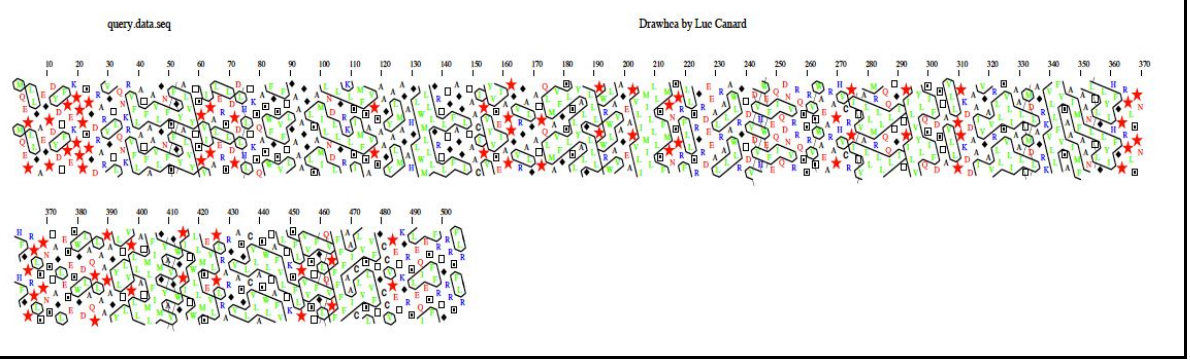


Figure 5: HCA plot of the Q9UGQ3 protein.

Hydrophobic clusters are represented in green. These clusters are representative of the transmembrane regions of the protein that are either alpha helixes of beta sheets. In our case, most of these regions are alpha helixes.

Finally, using the Conserved Domain tool from NCBI, we identified that the same regions are conserved. The figure below allows us to conclude that these regions are conserved not only for the 10 proteins selected but also for every proteins in the Sugar Transporter Superfamily (which accession number is pfam00083).

Figure 6: Conserved domains alignment plot, using NCBI Conserved Domains tool. The alignment has been done against the Sugar Transporter Superfamily. E-value = 4.04e-77. The alignment is biologically significant.

C. Protein modelling, refinement and validation:

Aligning our query protein against the whole PDB allowed us to identify two templates that share respectively 28.7% (GLUT1) and 26.7% (GLUT3) sequence identity with our query protein (GLUT6). Since these percentages are lower than 30% it is prefered to use *ab inito* modelling. Here we chose to use both ab initio and homology methods to modelize the structure of our protein.

Using every methods mentioned in the “Material and methods” section, a total of 16 models were generated.

First, every models which do not start with a loop (from position 1 to 37) were not retained. Indeed, the N-terminal region contains a dileucine motif (LL, at position 8 and 9). It is believed that this motif leads to cell surface expression of the protein. [1]
In the same way, we did not retained every models which did not possess the PETKG motif as well as most conserved residues mentioned before.
After this first step, only 4 models are selected as well as one control model generated by I-TASSER for each template (GLUT1, 4PYP and GLUT3, 4ZW9).

The following table present the four models that are kept after this first step of refinement.

Model	Template	Z-score Before refinement*	Z-score After refinement*
RaptorX	-	-6.17	-6.71
MUSTER	GLUT1	Error within the PDB, no score given	
MEMOIR	GLUT1	-5.72	-6.7
	GLUT3	-5.85	-6.4
I-TASSER (control)	GLUT1	-7.18	
	GLUT3	-7.18	

* Before/After Refinement using GalaxyRefine.

Unfortunately, no Z-score can be obtained for the MUSTER model.
According to those results the best models are both the MEMOIR models with GLUT1 and GLUT3 and RaptorX's.

D. <u>Protein position in membrane prediction:</u>

For the best model, the membrane position has been predicted using PPM-server. This information will be use for the future dynamic modelisation.

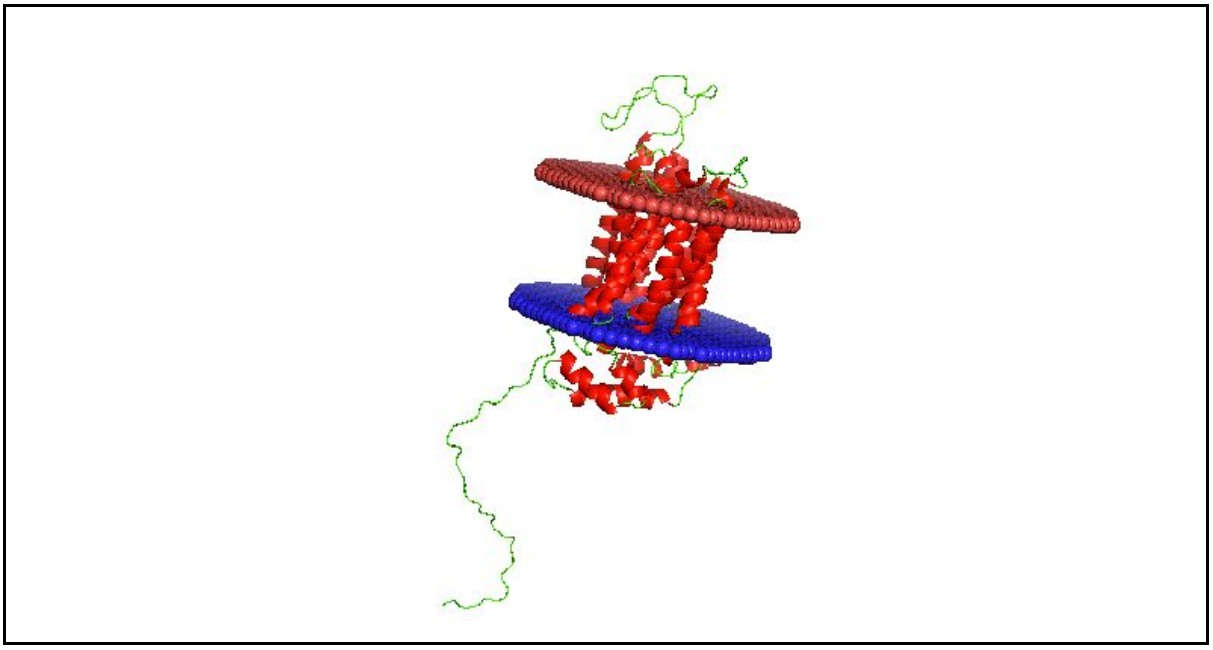


Figure : Membrane position of the model generated by RaptorX.

As we can see in this figure, the 12 transmembrane are contained between the membranes. The C-terminal and N-terminal region are both in the same compartment.

E. Normal modes analysis:

This analysis allows us to view the movement of the protein within the membrane. The following description concerns the best model generated (RaptorX).

The first 7 normal modes returned by NOMAD-ref are named the trivial normal modes, they describes rotation and translation movement. For example, the first mode only represent translation of the transmembrane helices from right to left. In the 14th mode, we can view the changing conformation of the protein. The first states of the protein seems to be either inward or outward-open. Then the helices lining the pore are gathering to close the pore. This is the occluded state of the protein.

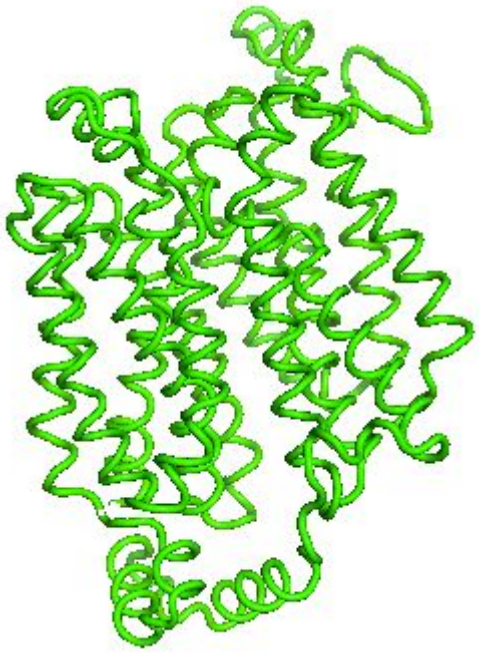


Figure : Either inward or outward-open

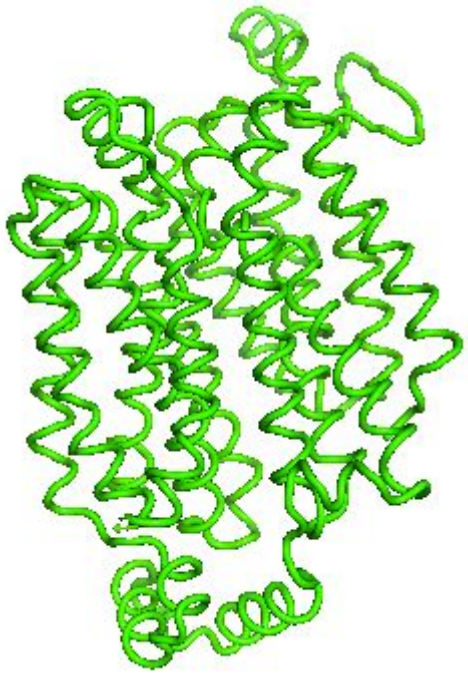


Figure : The pore lining helices seem closer.

IV. Conclusion :

The aim of the project was to predict the structure of the protein Q9UGQ3. Using several tools and algorithm we identified 3 final models that might represent our protein.

V. Références :

[1] Activity and genomic organization of human glucose transporter 9 (GLUT9), a novel member of the family of sugar-transport facilitators predominantly expressed in brain and leucocytes.
Holger DOEGE, 2000. Biochemical Journal, 3.

[2] Crystal structure of the human glucose transporter GLUT1.
Deng D et al. *Nature* **510** 121-5 (2014)

[3] Crystal structure of human GLUT3 bound to D-glucose in the outward-occluded conformation at 1.5 angstrom
Deng D et al. *Nature* (2015)

[4] Molecular basis of ligand recognition and transport by glucose transporters
Deng D et al. *Nature* (2015)

[5] Hydrophobic cluster analysis and secondary structure predictions revealed that major and minor structural subunits of K88-related adhesins of *Escherichia coli* share a common overall fold and differ structurally from other fimbrial subunits
Marie-ClaireMéchin et al. *FEBS Letter* (1995)

[6] WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, (2004)
[7] CDD: a conserved domain database for interactive domain family analysis
Aron Marchler-Bauer et al. Nucleic Acids Res. 2007 Jan; 35(Database issue): D237–D240.

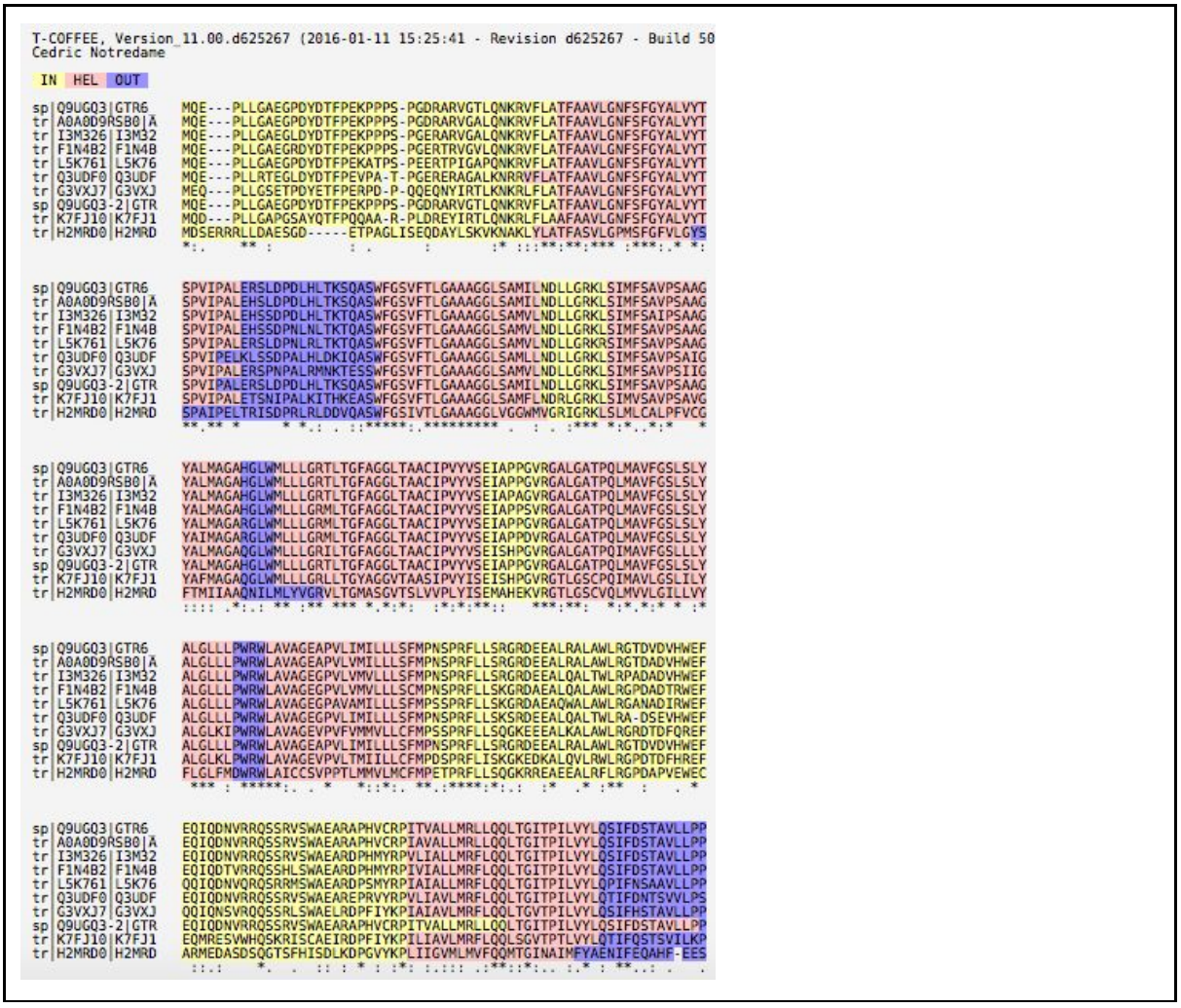
[8] Memoir: template-based structure prediction for membrane proteins.
Ebejer JP et al. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W379-83. doi: 10.1093/nar/gkt331. Epub 2013 May 2.

[9] Lomize M.A., Pogozheva I,D, Joo H., Mosberg H.I., Lomize A.L. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res., 2012, 40(Database issue):D370-6

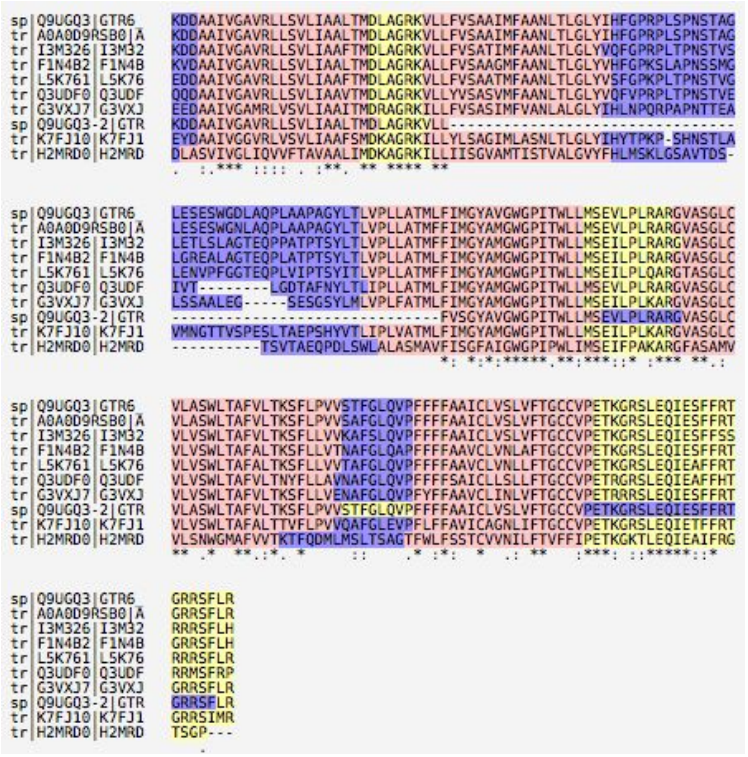
[10] SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information *Nucleic Acids Research* 2014 (1 July 2014) 42 (W1): W252-W258

Protein structure prediction

Supplementary figures:



Protein structure prediction



Supplementary Figure 1: Whole alignment and secondary structure prediction of Q9UGQ3. Using T-COFFEE PSI/TM-COFFEE tool.