Protein Modelling

1. Introduction

Protein A6NKX4 is a transmembrane anion transporter in *Homo sapiens.* It is the member number 31 of the putative solute carrier family 22. This protein is coded by the gen SLC22A31, located in the chromosome 16. Although there is any bibliography related to this protein, we aim to make a model of his tertiary structure by homologue modeling.

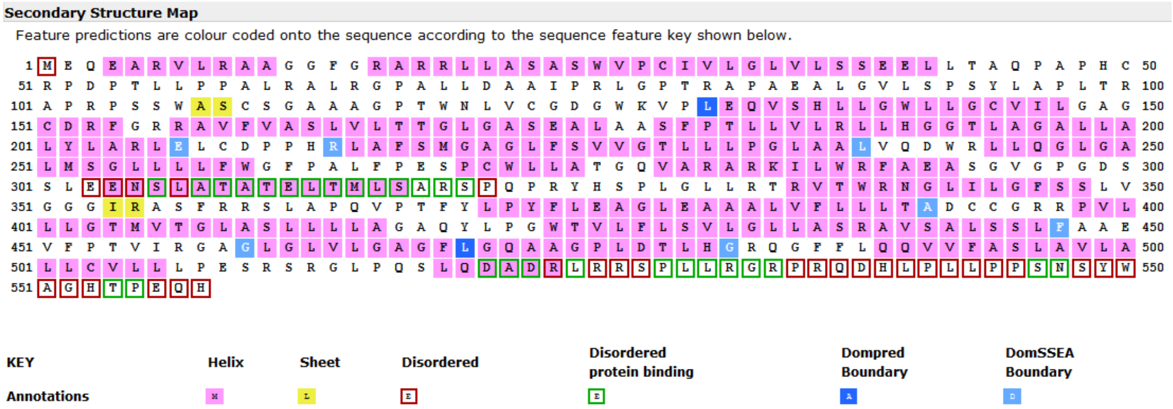
We know that this protein belongs to de major facilitator superfamily (MFS) [1]. This superfamily is composed entirely by membrane transporters that share some common features in their structures. MFS proteins present 12 transmembrane helixes (TM) divided in two domains: The N-termini domain is composed by the 6 first TMs and the C-termini domain by the last 6 TMS. Another common characteristic is that the two extremes of the proteins are always located in cytoplasm.

As A6NKX4 is coded by a SLC2 gen, we suppose that it is a member of GLUT family [2]. GLUT proteins are hexoses or polyols transporters coded by SLC2 genes. In *Homo sapiens* there are 14 GLUT proteins characterized. They all present the structural features described in MFS. There are different subgroups of GLUT proteins that share more specific structural characteristics and conserved residues [3]. Unfortunately, the lack of bibliography related to our protein impede us to continue predicting which possible structure it may have.

2. Secondary Structure

The first step in this study is to predict the secondary structure of A6NKX4. Reading UNIPROT A6NKX4 entry, we can appreciate than the major part of our protein is in the transmembrane region. In addition to this, the two extremes are located in the cytoplasm, important characteristic of the MFS. The sequence is composed by 558 amino acids.

To go into detail about his secondary structure, we are going to use PSIPRED [4]. PSIPRED is a web server that takes a fasta file as input and predicts general 2D features of the protein. As we can see in Figure 1, our protein is composed mainly by helixes and some loops in between. It is also important to notice that termini regions are disordered. In consequence, these regions cannot be predicted in 3D as they will decrease our quality score. We can also find a disordered region between the residues 303 and 324. This fragment may be a high flexibility loop that binds the ligand and changes his conformation to let him pass across the channel. We can see a more detailed scheme of disordered regions in Figure 2.



Secondary Structure Map

Feature predictions are colour coded onto the sequence according to the sequence feature key shown below.

```
  1 M E Q E A R V L R A A G G F G R A R R L L A S A S W V P C I V L G L V L S S E E L L T A Q P A P H C 50
 51 R P D P T L L P P A L R A L R G P A L L D A A I P R L G P T R A P A E A L G V L S P S Y L A P L T R 100
101 A P R P S S W A S C S G A A A G P T W N L V C G D G W K V P L E Q V S H L L G W L L G C V I L G A G 150
151 C D R F G R R A V F V A S L V L T T G L G A S E A L A A S F P T L L V L R L L H G G T L A G A L L A 200
201 L Y L A R L E L C D P P H R L A F S M G A G L F S V V G T L L L P G L A A L V Q D W R L L Q G L G A 250
251 L M S G L L L L F W G F P A L F P E S P C W L L A T G Q V A R A R K I L W R F A E A S G V G P G D S 300
301 S L E E N S L A T A T E L T M L S A R S P Q P R Y H S P L G L L R T R V T W R N G L I L G F S S L V 350
351 G G G I R A S F R R S L A P Q V P T F Y L P Y F L E A G L E A A A L V F L L L T A D C C G R R P V L 400
401 L L G T M V T G L A S L L L L A G A Q Y L P G W T V L F L S V L G L L A S R A V S A L S S L F A A E 450
451 V F P T V I R G A G L G L V L G A G F L G Q A A G P L D T L H G R Q G F F L Q Q V V F A S L A V L A 500
501 L L C V L L L P E S R S R G L P Q S L Q D A D R L R R S P L L R G R P R Q D H L P L L P P S N S Y W 550
551 A G H T P E Q H
```

KEY	Helix	Sheet	Disordered	Disordered protein binding	Dompred Boundary	DomSSEA Boundary
Annotations	M	L	E	E	A	D

Figure 1. A6NKX4 secondary structure map

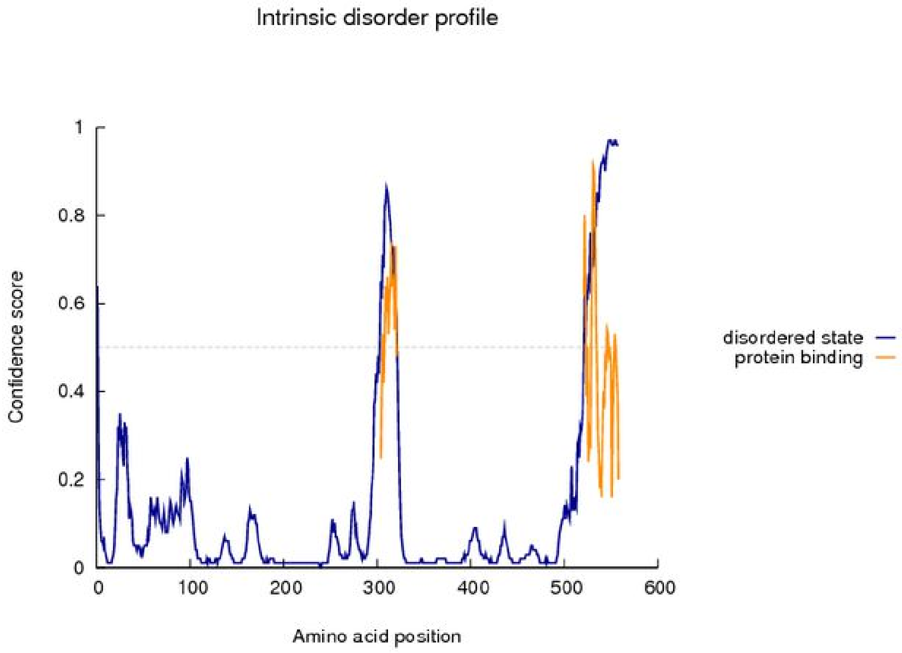


Figure 2. Disordered regions prediction map

In Figure 3 we have topology scheme where it is much simpler to imagine how our protein is folded. A6NKX4 is composed by 12 TM and his extremes are located in the cytoplasm. As these two features are in agreement with MFS structural characteristics we can prove the reliability of our results.

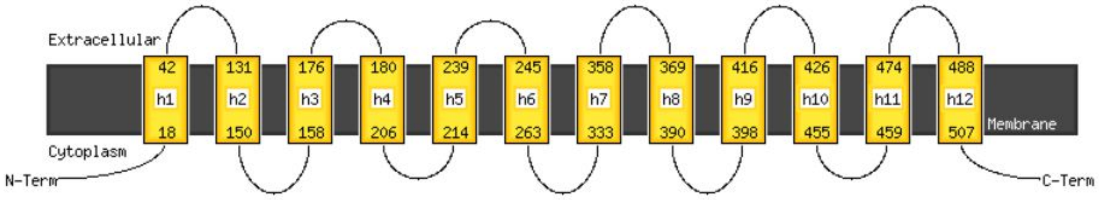
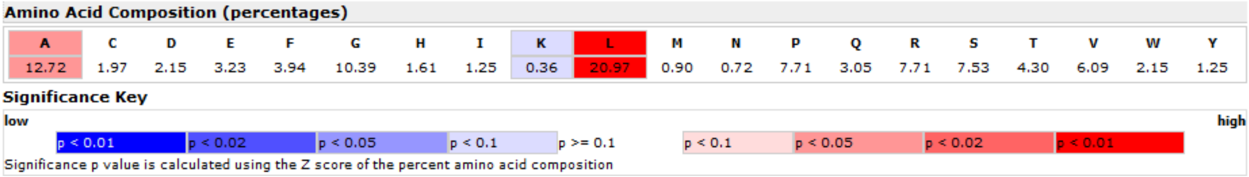


Figure 3. Predicted transmembrane topology

It is also very interesting to study the amino acid composition of our protein. As we can see in Figure 4, we have an overrepresentation of Alanine and Leucine, two non-polar residues. This is a key point in the understanding of the membrane protein structure: A6KNX4 is mainly located in the transmembrane region so it will be surrounded by lipids, a non-polar environment. In contrast, the frequency of apparition of Lysine, a polar residue, is very low. We expect to find these polar residues in the protein central channel that will interact with the substrate.



Amino Acid Composition (percentages)

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
12.72	1.97	2.15	3.23	3.94	10.39	1.61	1.25	0.36	20.97	0.90	0.72	7.71	3.05	7.71	7.53	4.30	6.09	2.15	1.25

Significance Key

low									high
	p < 0.01	p < 0.02	p < 0.05	p < 0.1	p >= 0.1	p < 0.1	p < 0.05	p < 0.02	p < 0.01

Significance p value is calculated using the Z score of the percent amino acid composition

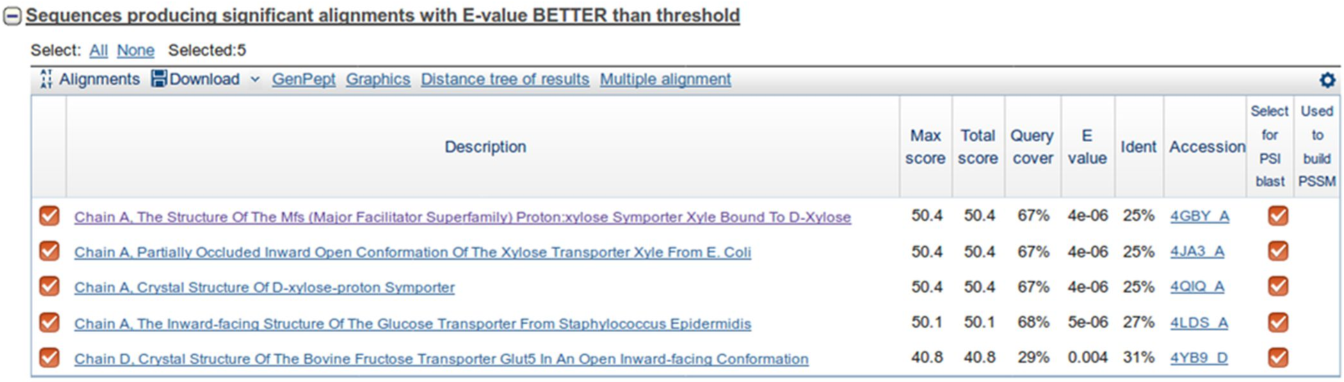
Figure 4. A6NKX4 amino acid composition.

3. Tertiary Structure

3.1 Model construction

Now that we have a general overview of A6NKX4 secondary structure, we aim to construct a 3D model by homology. The first step is searching of homologous proteins that can be use as template. Using a PSI-BLAST [5] against PDB we are going to see which sequences with a known structure are similar to our query.

The results of the PSI-BLAST are shown in Figure 5. We notice that our protein has not a high similarity with any of the proteins found. The ident percentage is between 25 and 31 % and the coverage is lower than 70%. On the other hand, the 5 proteins found are transmembrane transporters member of the MFS, so they will share common features with our protein of interest.



Sequences producing significant alignments with E-value BETTER than threshold

Select: All None Selected:5

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
☑	Chain A, The Structure Of The Mfs (Major Facilitator Superfamily) Proton:xylose Symporter Xyle Bound To D-Xylose	50.4	50.4	67%	4e-06	25%	4GBY_A	☑	
☑	Chain A, Partially Occluded Inward Open Conformation Of The Xylose Transporter Xyle From E. Coli	50.4	50.4	67%	4e-06	25%	4JA3_A	☑	
☑	Chain A, Crystal Structure Of D-xylose-proton Symporter	50.4	50.4	67%	4e-06	25%	4QIQ_A	☑	
☑	Chain A, The Inward-facing Structure Of The Glucose Transporter From Staphylococcus Epidermidis	50.1	50.1	68%	5e-06	27%	4LDS_A	☑	
☑	Chain D, Crystal Structure Of The Bovine Fructose Transporter Glut5 In An Open Inward-facing Conformation	40.8	40.8	29%	0.004	31%	4YB9_D	☑	

Figure 5. Psi-Blast results

The first server used to create our homology model is MEMOIR [6], who is specific for membrane proteins. As the sequence identity is very low, it is impossible to guarantee an accurate model. For models of < 50% sequence identity, the query sequence and template structure diverged a long time ago. In consequence, we are going to use another server called HHPred [7], which is more appropriated for remote protein homology detection.

HHPred realizes a search of remote homologous proteins in PDB. As we can see in Figure 6, the sequences found are not exactly the same that we found with PSI-BLAST, there are only 3 in common. The first 125 residues of the protein and the last 6 ones are not cover by any of the templates. This is another reason why we cannot make an homology model of the protein’s extremes.

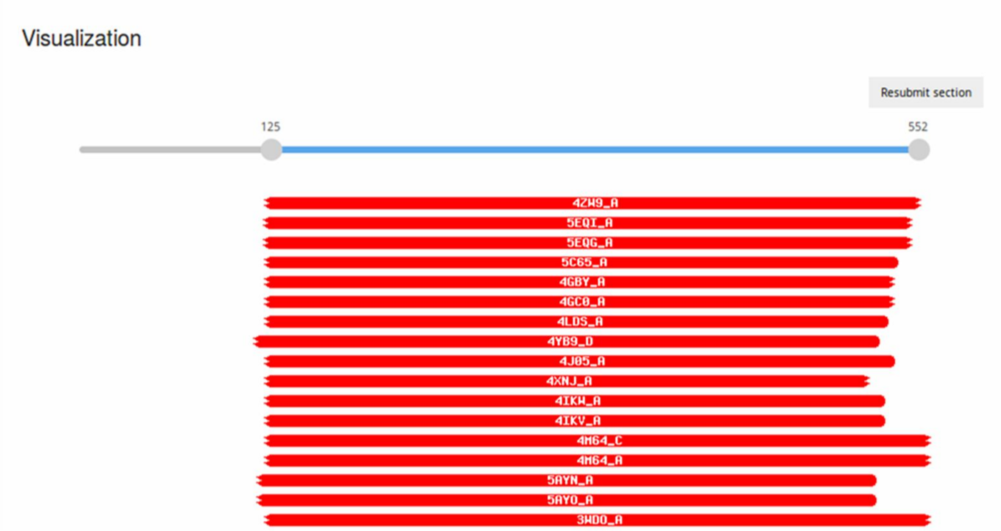
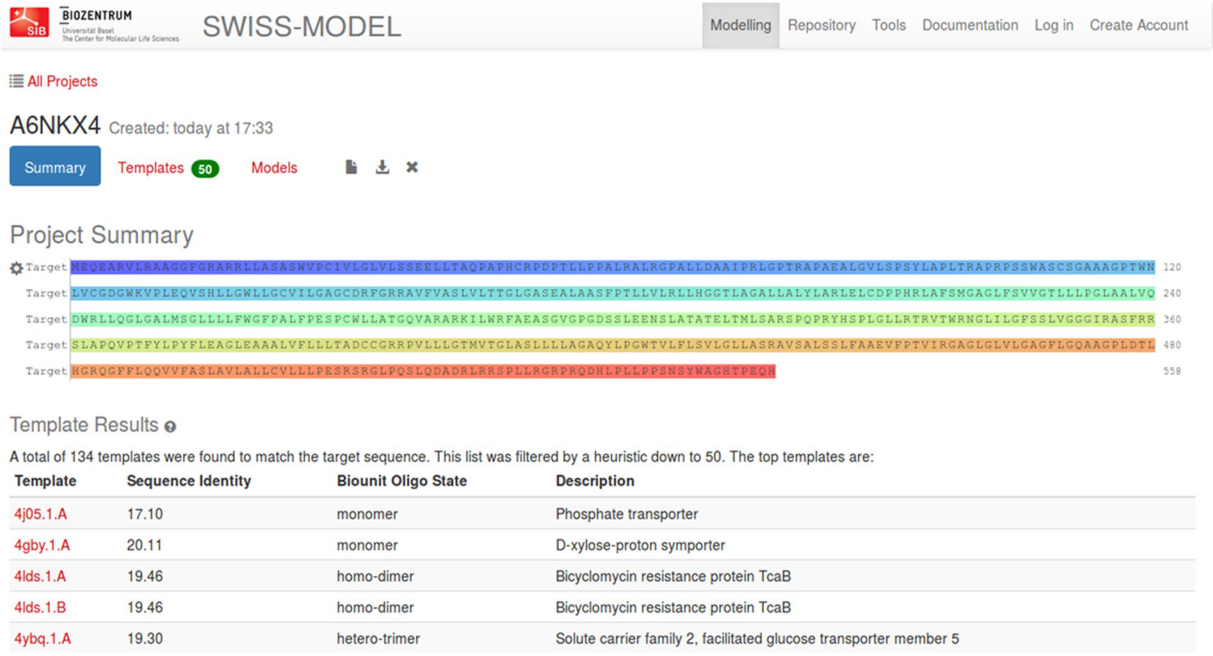


Figure 6. Remote homologous found by HHPred

Once we have found the remote homologous, we are going to select the first 10 templates and the server is going to align them. This multiple alignment is going to be used as template for the homologous modelling.

We want now to use another homologous modelling server to compare the different models created. The second server is called SWISS-MODEL. It also finds the homologous proteins, but It creates a different model for every template found. For our query, we get 5 templates and the server give us a predicted structure for each template. As we can see in Figure 7, 3 of the proteins found are also found in HHPred prediction and in the previous PSI-BLAST.



A6NKX4 Created: today at 17:33

Project Summary

```
Target MEQEARVLRAAGGFGRARRLLASASWVPCIVLGLVLSSEELLTAQPAPHCRPDPTLLPPALRALRGPALLDAAIPRLGPTRAPAEALGVLSPSYLAPLTRAPRPSSWASCSGAAAGPTWN 120
Target LVCGDGWKVPLEQVSHLLGWLLGCVILGAGCDRFGRRAVFVASLVLTTGLGASEALAASFPTLLVLRLLHGGTLAGALLALYLARLELCDPPHRLAFSMGAGLFSVVGTLLLPGLAALVQ 240
Target DWRLLQGLGALMSGLLLLFWGFPALFPESPCWLLATGQVARARKILWRFAEASGVGPGDSSLEENSLATATELTMLSARSPQPRYHSPLGLLRTRVTWRNGLILGFSSLVGGGIRASFRR 360
Target SLAPQVPTFYLPYFLEAGLEAAALVFLLLTADCCGRRPVLLLGTMVTGLASLLLLAGAQYLPGWTVLFLSVLGLLASRAVSALSSLFAAEVFPTVIRGAGLGLVLGAGFLGQAAGPLDTL 480
Target HGRQGFFLQQVVFASLAVLALLCVLLLPESRSRGLPQSLQDADRLRRSPLLRGRPRQDHLPLLPPSNSYWAGHTPEQH 558
```

Template Results

A total of 134 templates were found to match the target sequence. This list was filtered by a heuristic down to 50. The top templates are:

Template	Sequence Identity	Biounit Oligo State	Description
4j05.1.A	17.10	monomer	Phosphate transporter
4gby.1.A	20.11	monomer	D-xylose-proton symporter
4lds.1.A	19.46	homo-dimer	Bicyclomycin resistance protein TcaB
4lds.1.B	19.46	homo-dimer	Bicyclomycin resistance protein TcaB
4ybq.1.A	19.30	hetero-trimer	Solute carrier family 2, facilitated glucose transporter member 5

Figure 7. SWISS-MODEL templates.

We obtain 5 different models from SWISS-MODEL. They all have a similar structure and the extremes have not been modelized. They are automatically evaluated by a QMean score. None of them have a valid score but we know that this score is not appropriated for membrane proteins. In consequence, we are going to validate our 6 final models with other servers.

3.2 Model evaluation

We are going to use ProQ to do a first evaluation of our models. The results are shown in Table 1. To validate our model, the LGscore should be greater than 1.5 and the MaxSub greater than 0.1. A we can see, only 3 of our 6 models have a correct score (highlighted on yellow). They are all models created by SWISS-MODEL.

	LGscore	MaxSub
Modeller	2.016	0.089
SWISS-MODEL01	2.223	0.181
SWISS-MODEL02	2.344	0.158
SWISS-MODEL03	1.714	0.136
SWISS-MODEL04	1.787	0.134
SWISS-MODEL05	2.302	0.188

Table 1. Evaluation score by ProQ.

The score of the three validated models are very similar. To decide which is the best one we are going to study their Ramachandran plot (Tab. 2). These plots are done with a web server called RAMPAGE.

	Favoured (%)	Allowed (%)	Outlier (%)
SWISS-MODEL01	87.9	8.5	3.6
SWISS-MODEL02	92.2	4.8	3
SWISS-MODEL05	95.1	3.4	1.6

Table 2. Ramachandran plot results

In the Ramachandran plots we obtain in every case some residues in the outlier region. The best of our models is the SWISS-MODEL number 5 so this is the one that we are going to choose. We are going to study where are this outlier residues located in the structure to try to refine it.

3.3 Model refinement

In Figure 7 we can observe our chosen model. We have silhouetted in red the residues that outlie in the Ramachandran plot. They are all located in loops. We can appreciate that two of them are found in de center of the protein, in a loop colored in green. This loop may correspond to the disordered region found in Figure 2, so it may be important for the function of the protein. There are another two loops colored in green that are also in the transmembrane region. As the three colored loops have outlier residues, we are going to refine them using GalaxyWeb server.

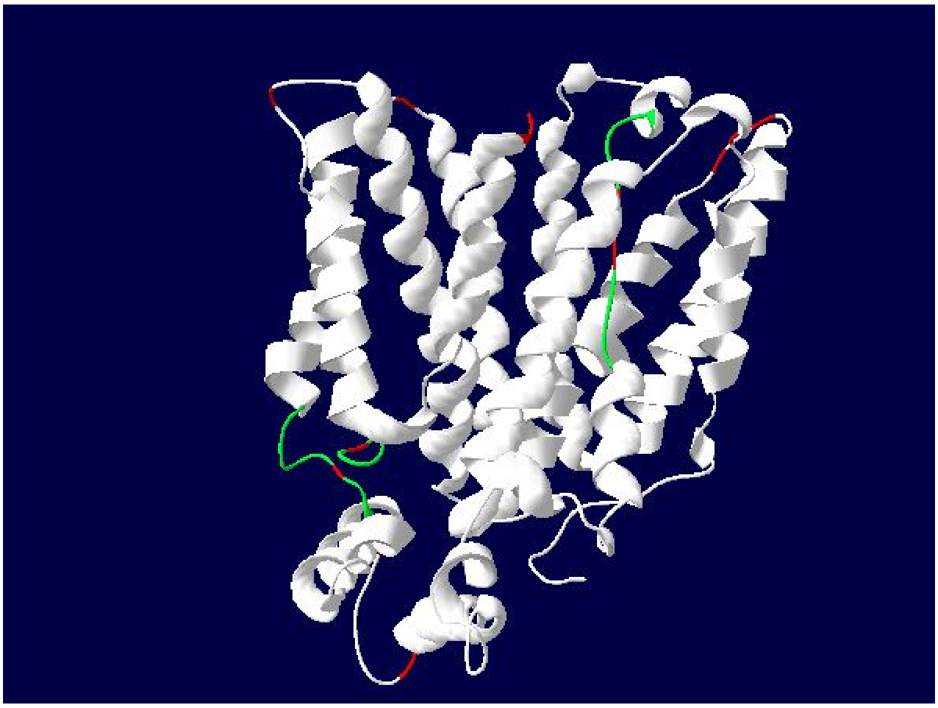


Figure 7. Chosen model for A6NKX4 protein. In red: outlier residues in the Ramachandran plot. In green: three loops that will be refined.

GalaxyWeb refines our loops and gives as result 5 new models. These models are evaluated in Table 3. It is surprising that the 5 refined models obtain lower ProQ scores that the initial one. In the Ramachandran plot, they present a higher number of outlier residues. This is maybe because the loops that we try to refine are located in very flexible disordered regions that change their conformations in order to let the ligand pass across the membrane. In consequence, the model that we are going to use in this study is the initial one and not the refined ones.

	LGscore	MaxSub	Favoured (%)	Allowed (%)	Outlier (%)
SWISS-MODEL05	2.302	0.188	95.1	3.4	1.6
SM05-Galaxy01	1.797	0.082	94.6	3.6	1.8
SM05-Galaxy02	1.800	0.087	94.6	3.6	1.8
SM05-Galaxy03	1.786	0.082	94.0	3.1	2.8
SM05-Galaxy04	1.794	0.08	94.6	3.6	1.8
SM05-Galaxy05	1.813	0.082	94.0	4.1	1.8

Table 3. Evaluation by ProQ and Ramachandran Plot of the 5 GalaxyWeb refined models.

4. Discussion

The extremes of the protein are eliminated to create the model. The structure model starts at residue 130 and ends at residue 517 (Fig. 8). In consequence we cannot observe the final regions in our model and we cannot prove that they are found in the cytoplasm. This is also the reason why we cannot find the first helix of our protein, so the model only shows 11 TMs.

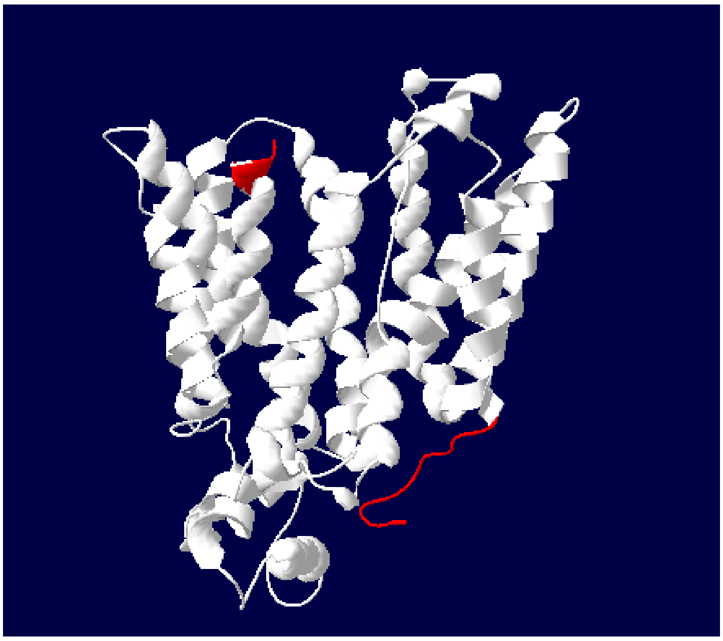
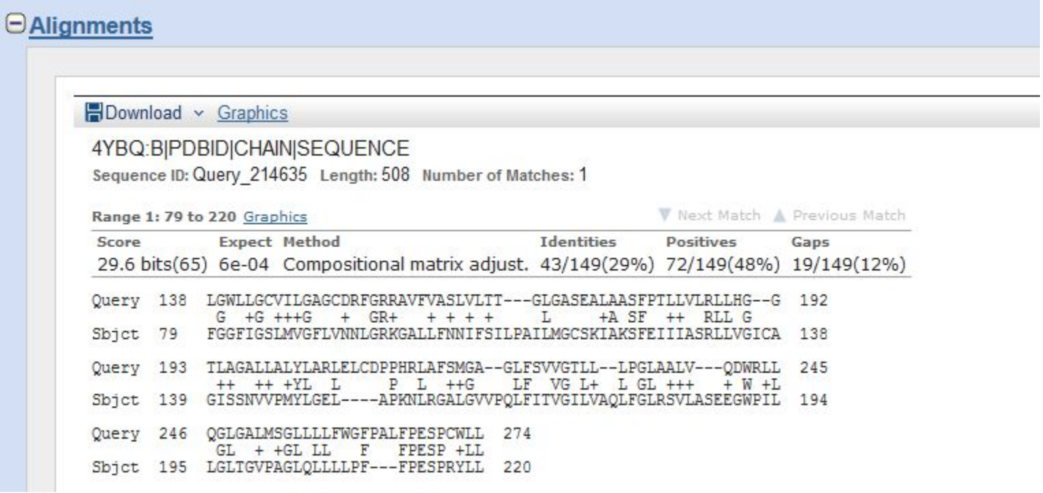


Figure 8. The N and C termini regions shown in red do not correspond to the real protein extremes.

To be able to discuss our model, we have to take a look into the template structure. Our template has the 4YBQ PDB code. Surprisingly, this template was not found by PSI-BLAST or HHPred. It is a fructose transporter found in Rattus norvegicus. It is a member of the GLUT5 family and of MFS. It is known that GLUT5 in rats that share ~ 81% sequence identity to human [8], so it is a good template for our protein.

The first thing we are going to do is to check the alignment between our protein and the template to see which part of our sequence is covered by 4YBQ. This alignment will be done by BLAST (Fig. 9).



4YBQ:B|PDBID|CHAIN|SEQUENCE

Sequence ID: Query_214635 Length: 508 Number of Matches: 1

Range 1: 79 to 220 Graphics

Score	Expect	Method	Identities	Positives	Gaps
29.6 bits(65)	6e-04	Compositional matrix adjust.	43/149(29%)	72/149(48%)	19/149(12%)

```
Query  138  LGWLLGCVILGAGCDRFGRRAVFVASLVLTT---GLGASEALAASFPTLLVLRLLHG--G  192
             G  +G +++G   +  GR+   + + + +     L     +A SF  ++  RLL G
Sbjct  79   FGGFIGSLMVGFLVNNLGRKGALLFNNIFSILPAILMGCSKIAKSFEIIIASRLLVGICA  138

Query  193  TLAGALLALYLARLELCDPPHRLAFSMGA--GLFSVVGTLL--LPGLAALV---QDWRLL  245
             ++  ++ +YL  L     P  L  ++G    LF  VG L+  L GL +++   + W +L
Sbjct  139  GISSNVVPMYLGEL----APKNLRGALGVVPQLFITVGILVAQLFGLRSVLASEEGWPIL  194

Query  246  QGLGALMSGLLLLFWGFPALFPESPCWLL  274
             GL  + +GL LL   F   FPESP +LL
Sbjct  195  LGLTGVPAGLQLLLLPF---FPESPRYLL  220
```

Figure 9. Alignment between the query and the template

In Figure 9 we can observe that the region with a high similarity starts at the residue 138 of our query and ends at the residue 274. In Figure 10, we can see in red that this region corresponds to 5 TH of our protein. That suggest that the covered region is the N-termini domain of A6NKX4. In Figure 11, we show the high similarity region in the template protein. It corresponds exactly to the same helixes, the major part of the 4YBQ N-termini domain. This domain is shared by all MFS members, so we suppose that the first part of our structure is well modelized.

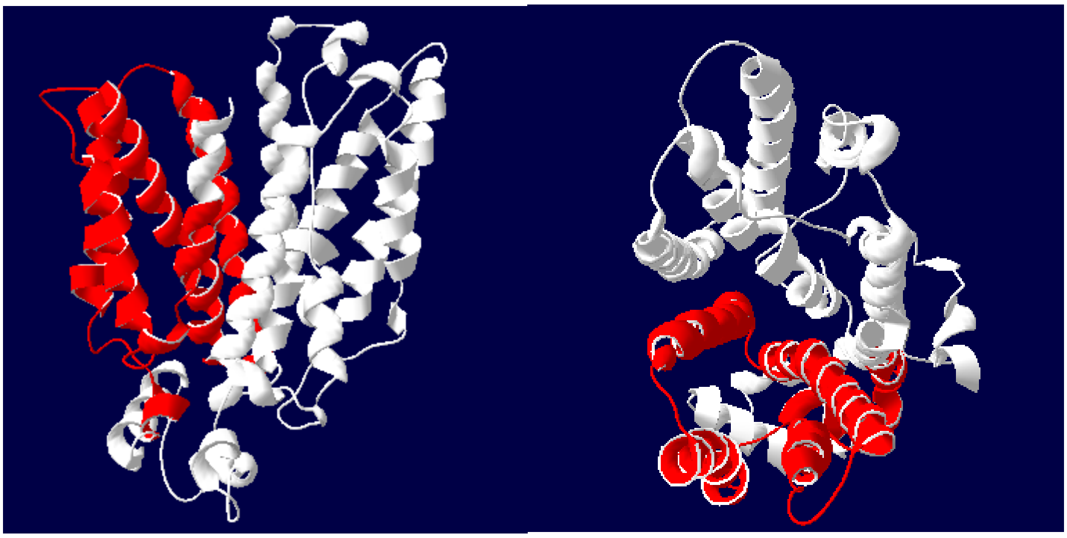


Figure 10. 3D model of A6NKX4. In red: covered region by the template

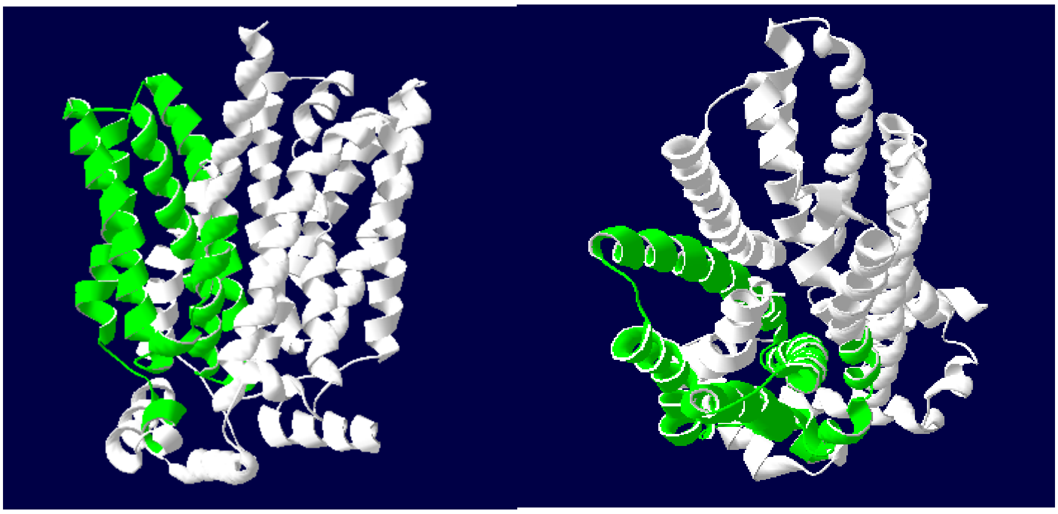


Figure 11. 3D structure of 4YBQ. In green: region that covers the query.

The GLUT5 structure shows the typical MFS fold, plus five additional helices on the intracellular side [8]. In our model, we can also find 5 small helixes that are not located in the transmembrane region (Fig. 12). This suggest that our protein matches the specific characteristics of GLUT5.

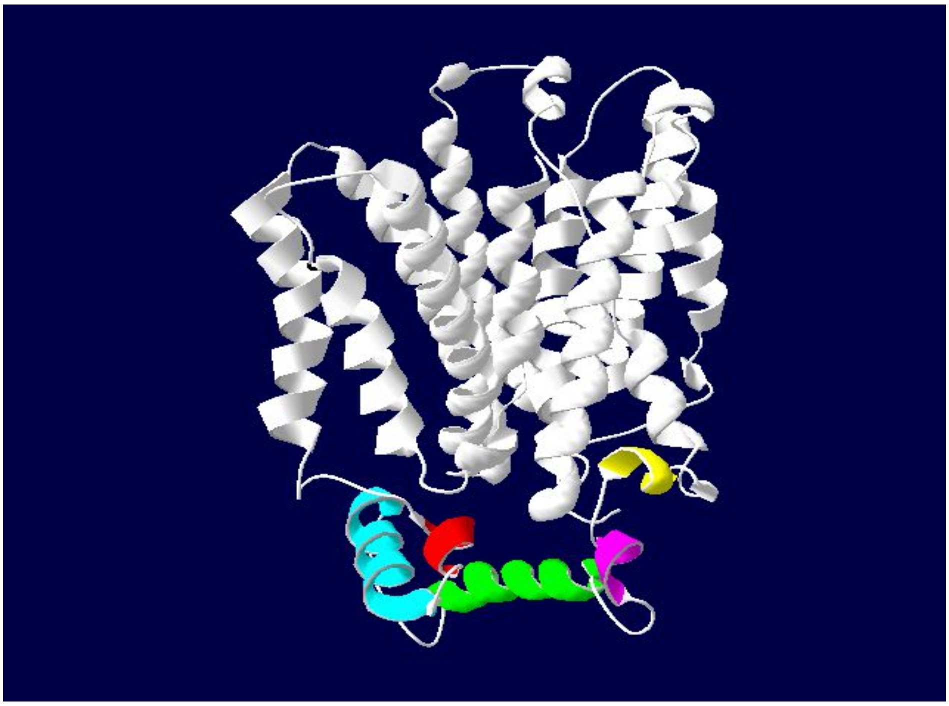


Figure 12. 3D model of A6NKX4. Colored: five small helix that are not located in transmembrane region.

To check if these helixes are in the cytoplasmic region, we are going to use two web servers that predict the position of the protein in the membrane (Fig 13 and 14). The first server used is called PPMserver (Fig. 13) and it makes the difference between the cytoplasmic region (in blue) and the extracellular region (in red). The second server is called OREMPRO (Fig. 14) and it does not specify with part is cytoplasmic and which part is extracellular. Anyway, the two servers predict a similar position for the membrane. The five-small helix of interest (Fig. 12) are located in the cytoplasmic region for the first server and outside the transmembrane region for the second server. This validates the reliability of our model and the use of 4YBQ as template.

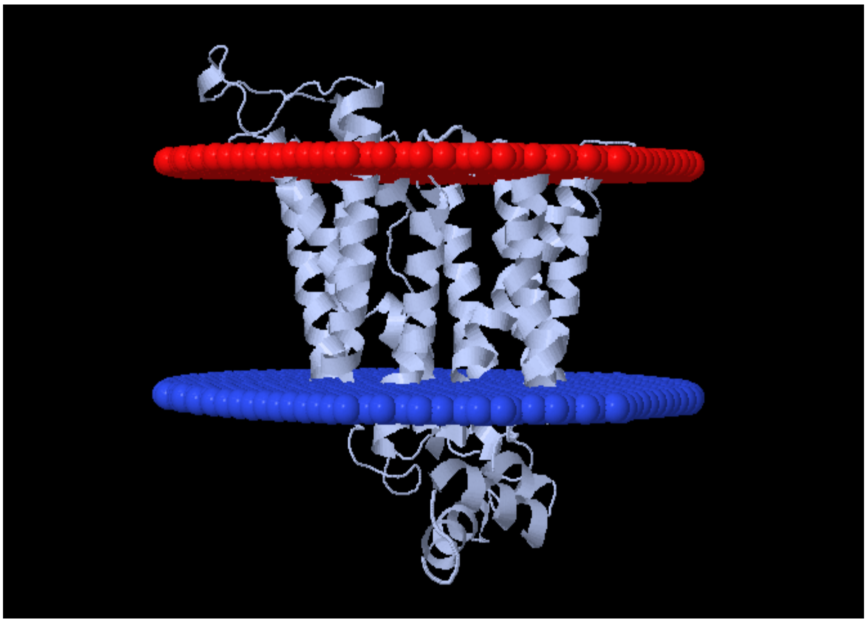


Figure 13. PPMserver membrane position prediction. In blue: cytoplasmic region. In red: extracellular region.

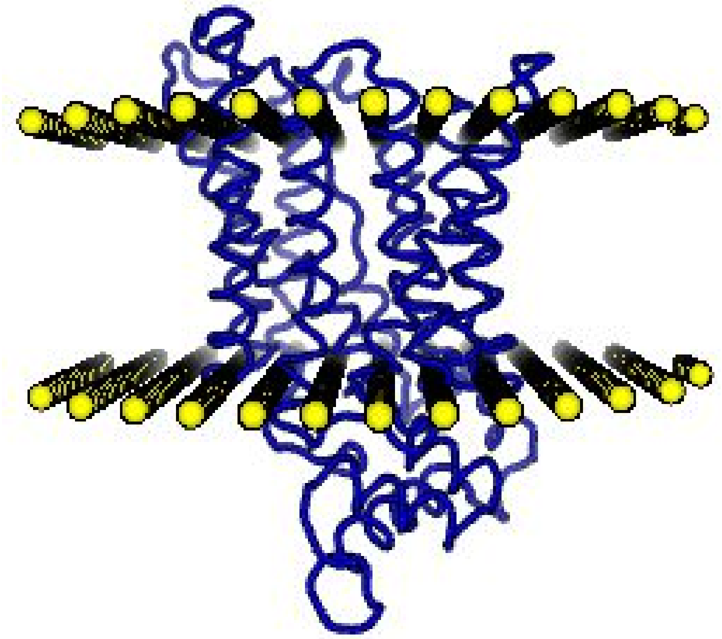


Figure 14. OREMPRO membrane position prediction

Finally, we are going to study the polarity of our residues. As we can see in Figure 15, the majority of our residues are non-polar ones. This was already predicted by PSI-PRED in Figure 4. Non-polar residues are located in the exterior side of the transmembrane helixes. These residues are surrounded by lipids, so the environment is non-polar. If we look at the protein central cavity, we can find some polar residues, the ones that will interact with the ligand. These final results support the reliability of our model.

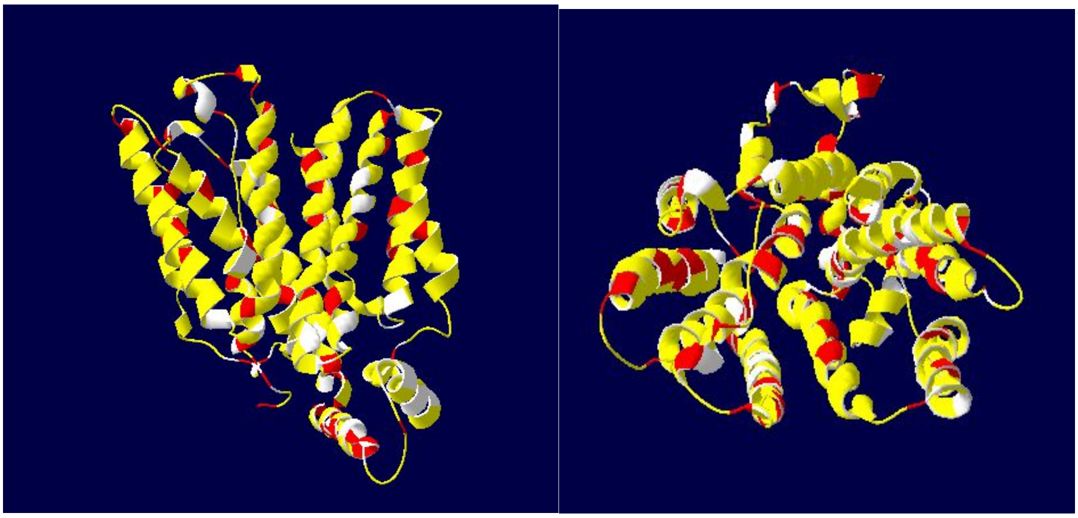


Figure 15. Modelized structure. In yellow: non-polar residues. In red: polar residues

5. Conclusions

We aim to create a model for the structure of the membrane protein A6NKX4. Although we have no information about this protein, we achieved to create a 3D model using SWISS-MODEL server. This model has been validated by ProQ score and It has an acceptable Ramachandran plot. Comparing the structure to the protein template, 4YBQ, we have seen that our protein shares common characteristics with GLUT5 family. It also presents the principal features of MFS. The polar residues distribution helps to validate the reliability of the model.

6. Bibliography

1. Yan, N. (2013) Structural advances for the major facilitator superfamily (MFS) transporters. Trends Biochem Sci 38(3):151-9.

2. Mueckler M., Thorens B. (2013) The SLC2 (GLUT) family of membrane transporters 34(2013): 121-138

3. Agustin R., Mayoux E. (2014) Mamalian Sugar Transporters: Chapter 1.

4. McGuffin LJ[1], Bryson K, Jones DT (2000). The PSIPRED protein structure prediction server. 16(4):404-5.

5. Bhagwat M., Aravind L. (2007). Chapter 10. PSI-BLAST Tutorial. Comparative Genomics: Volumes 1 and 2

6. Ebejer JP. Et al (2013). Memoir: template-based structure prediction for membrane proteins. 41(Web Server issue): W379–W383.

7. Söding G. et al (2005). The HHpred interactive server for protein homology detection and structure prediction. 33(Web Server issue): W244–W248.

8. Nomura N. et al. (2015). Structure and mechanism of the mammalian fructose transporter GLUT5. 526(7573):397-401