Report of the structural bioinformatics Project

Realised by

MESSAK Imane
ELHARTY Hager

January 4th, 2021

M2BI Students
Paris Diderot Univerity

1 Introduction

In biology, there are a large number of databases such as the PDB (Protein Data Bank) which contains structural information on biological macromolecules, or the GenBank which contains genomic information. In the case of proteins, the Uniprot database contains approximately 209.10^6 protein sequences. Among these sequences, $0.56.10^6$ are annotated Swissprot (or reviewed) and 209.10^6 are annotated TrEMBL (or unreviewed). The Swissprot annotated sequences are the sequences which have been verified manually, that is to say that their existence in the cell has been verified. The sequences noted TrEMBL are only a simple translation of the genes present in the GenBank database, their existence in the cell has therefore not been verified. It should be noted that we do not necessarily know the biological function and/or structure of all the proteins in Uniprot, whether they are annotated reviewed or unreviewed.

Having this information is an important challenge in bioinformatics when a protein is involved in the development of a disease. Indeed, having precise information on a protein makes it possible to better understand the defective or mutant proteins involved in a disease. The same is true for DNA and RNA molecules. There are experimental methods such as NMR (Nuclear Magnetic Resonance) or X-ray crystallography which makes it possible to obtain the structures of unknown proteins. Having the structure can allow us to understand the biological functions of proteins and by extension understand disease. However, these methods are cumbersome and do not necessarily give an exploitable or hardly exploitable result [1]. Obtaining the biological macromolecule structures can take up to 6 months and is relatively expensive. In addition, obtaining experimentally usable molecules can be relatively difficult. In our case, for example, we want to study a membrane protein, in order to extract them using organic solvents or detergents such as triton X100, SDS is essential [2]. However, these detergents can also alter the target protein.

To have the structure of a protein and in a short time, in silico methods for predicting the 3D structure of a biological molecule are widely favored. Today our project consists of the prediction of the 3D structure of an unreviewed membrane protein whose accession number on uniprot is W0K4U9. Membrane proteins play an important role in molecular transport, signal transduction, energy utilization and other basic physiological processes [3]. They are attached to the cell membrane, either on the extracellular, cytosolic or transmembrane side, mediating signal transduction between cells and the outside world, and performing many important cellular functions [3,4]. In our project, we will focus on a membrane protein encoded by the HALDL113775 gene. On Uniprot, the protein is called Membrane Protein and belongs to the Heliorhodopsin (HeRs) family. These proteins are found in the organism Halobacterium sp. DL1, which is a class of Euryarchaeota, found in water saturated or nearly saturated with salt [5].

2 Materials & Methods

2.1 Preliminary study of the protein

Before launching any modeling of the protein structure, it is essential to learn as much as possible about the target protein. Information can be obtained from scientific studies or from existing databases. The objective is to collect as much information as possible to be able to launch various predictions, for example by imposing certain information, a template or letting the program run autonomously.
Uniprot gives the protein a very general name : Membrane Protein. However, there are two types of membrane proteins : transmembrane and intrinsic membrane [3]. According to the literature, this protein belongs to the HeR family, it is probably a transmembrane protein since the proteins of this family are all transmembrane [5]. Remember that a protein is transmembrane when it crosses the membrane at least once. In contrast, a membrane protein is said to be intrinsic when it is covalently attached to the membrane by a phospholipid or another protein (see Diagram 1). To determine which type of membrane protein it is, we will study the hydrophobicity profile of the protein, and we will use membrane protein prediction methods. When studying the hydrophobicity profile, we tested several scales, and observed that it gave more or less the same result, so we decided to keep the Kyte Doolittle scale proposed first in the list. Due to lack of time, we did not have time to learn about the different scales in order to choose more precisely which one is the most suitable for the construction of the hydrophobicity profile [6].
We were also interested in information such as a thorough search for proteins of the same family and conserved domains. Our protocol is summarized in Table 1 below.

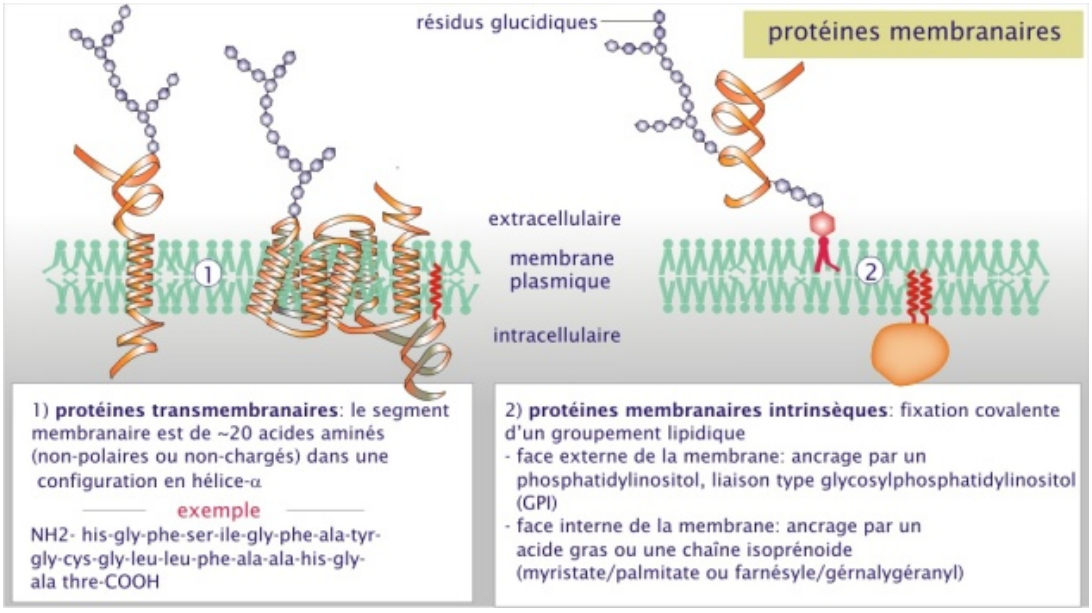


Diagram 1 : Diagram of the two types of membrane proteins existing in the living world

Tools	**P**urpose
Uniprot [7]	General information
Blastp [8]	Finding similar sequences
Pfam [5]	Information on the family of proteins find the Blastp.
T-Coffee [9]	Multiple alignment for the search of conserved domain (s)
Protscale & TMHMM Server[6]	Define the hydrophobicity profile & transmembrane helix predictions

2.2 Generation of models

2.2.1 Prediction of the tertiary structure

There are several methods to predict a 3D structure of a protein, just from the sequence. The general diagram below (cf. Diagram 2), explains which method to choose according to the data available for the prediction of the 3D structure. Indeed, from the FASTA sequence of the target protein, we performed a BLASTp (Basic Local Alignment Search Tool) alignment of the protein against the PDB (Protein Data Bank) structure database in order to find a structurally similar sequence. to the target protein. This alignment is based on the assumption that structure is more conserved than sequence. In other words, if the sequences are sufficiently similar, and therefore homologous, then so are the structures. Note that the homology is binary. That is, two sequences are either homologous or they are not.
To determine if two sequences are homologous, we analyzed three criteria : the percentage of identities, the E-value and the percentage of recovery. The percentage of identities is the number of identical amino acids at the same position divided by the length of the alignment multiplied by 100. The E-value (expected-value) is a parameter indicating whether our alignment is statistically significant [10]. The lower the E-value, the more significant the result. If the percent identity is high at least 50%, and a low E-value then we can assume a homology hypothesis between our target sequence and the PDB sequences.

Obviously the choice of the prediction method depends on the results obtained during the alignment. In our case, we found a template in the PDB with 51.38% identity with the target sequence. However, today there is a debate about what percentage of identity to take to consider a homology. Some scientists suggest that 30% is enough while others explain that at least 60% identity is needed. As we do not know which value considered, we therefore decided to use several programs and to play on one parameter. Indeed, we have launched several predictions with different web servers, imposing or not two templates found after our BLASTp against the PDB. The difference between these tools will lie in the construction mode, in particular for indels, and in the score function used. Each tool will have its own criteria to select and order predictions from best to worst.

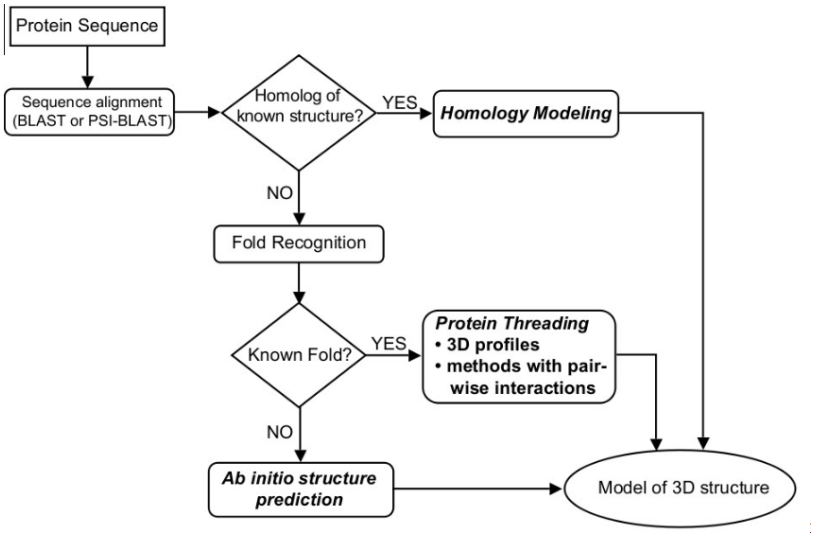


FIGURE 1 – General diagram for the prediction of the 3D structure of proteins

For this, we used *I-TASSER* [12], *ModWeb* [13], **Phyre2** [14] and *RaptorX* [15]. With *I-TASSER*, we launched a prediction without imposing anything [16] on the program and just giving the FASTA sequence of the protein of interest, then a prediction by imposing a first template [17], and finally another prediction in imposing another template [18]. For the *RaptorX* [19], *Phyre2* [20], and *ModWeb* [21] programs we didn't force anything and run the programs just by giving the FASTA sequence.

I-TASSER (Iterative Threading Assembly Refinement) modeling identifies structural models in the PDB by a threading approach. *I-TASSER* generates tens of thousands of conformations (called "decoys"). Then, to select the final models, the *SPICKER* program clusterizes these decoys according to their structural similarities and selects a maximum of five models belonging to the five largest clusters of similar structures. The models are classified by rank (from 1 to 5) according to the size of the cluster, then the program assigns a value of C-score corresponding to a confidence value of the structure. This C-score is calculated from the free energy of the structural model belonging to the cluster [26] (Cf. Figure 1 Annex).

To judge the quality of a model predicted with *RaptorX*, three measures are taken into account together : P-value, the score and the uGDT *(unnormalized Global Distance Test)* which is a measure of similarity between two protein structures with matches of known amino acids. A model with both a good P-value and a uGDT value above 50 will most likely be a high quality model. The program gives 5 predictions ranked from best to worst.

With *Phyre2*, four steps are designed to have our model predicted, first it collects homo-

logous sequences, then scans them, against a database of HMM *(Hidden Chain of Markov)* proteins, with a known structure . The best alignments from this research are selected to build only basic rough models. Then, the program performs loop modeling so that insertions and deletions in these patterns are corrected, and finally it adds amino acid side chains to generate the final pattern. The predicted models are ranked by a raw alignment score based on the number of aligned residues and the quality of the alignment.

The *ModWeb* web server relies on the *MODELLER* program, a computer program for homology modeling to produce models of tertiary structures of proteins. The program is inspired by protein nuclear magnetic resonance (NMR) spectroscopy, called satisfaction of spatial constraints, whereby a set of geometric criteria is used to create a probability density function for the location of each atom in the protein. The method relies on an input sequence alignment between the target protein to be modeled and a protein whose structure has been resolved. With *ModWeb*, the peer template search is done automatically. *ModWeb* selects the templates with the highest identity score and MPQS *(ModPipe quality Score)*.

2.2.2 Comparison and pre-selection of prediction models

At the end of the predictions we get a certain number of 3D models. The objective of the next step is to pre-select a reduced number of models per tool.Each tool gives a different number of results. For example *Phyre2* gives 20 prediction models but only selects 2 models as being the best. However, *I-Tasser* and *RaportX* give 5 models per prediction. So to pre-select the models, we compared the structures within the same tool. We have aligned the structures two by two of each tool on *PyMol* to calculate the RMSD (Root Mean Square Deviation) in order to select two models by tools having the most distant structures while always keeping the first model given as output by the tool prediction. Recall that the RMSD corresponds to a geometric difference value between two structures. The lower the value, the more the two structures are considered identical.

2.2.3 Evaluation, improvement and selection of the final model

After our pre-selection, we evaluated our models with *Process* [22] which is a web server designed to evaluate predicted biological structures, it integrates a variety of already developed, well-known and tested methods to evaluate both the overall specificities of the residues : covalent and geometric quality, non-bonded/packing quality, torsion angle quality, chemical transfer quality and NOE *(nuclear Overhauser effect)* quality. We then refined the structures with *GalaxyRefine2* [23] which is also a web server for the refinement of loops and side chains, which adapts an iterative optimization approach involving various sets of structure movements to improve local and global structures. After that, we re-evaluated our models with *Process*, to see if our predictions were improved. This step consists in evaluating the significance of our models in order to select a single final model. In fact, the more a model is improved, the more it will be taken into consideration.

We then used *PPM server* [24] a web server to provide the spatial positions of the structures of the membrane proteins within the lipid bilayer. This step helped us select our final model, because the model selection was relatively tricky from a structural similarity point of view. Finally, we visualized our protein structures using *PyMol*.

2.3 Study of the conformation over time of our models

In this part, we studied the conformation over time of our predicted model.

2.3.1 Study of normal modes on the basis of an elastic network

After choosing our model, we performed an elastic lattice (ENM) -based normal mode analysis to observe potential movements of the protein such as folding. ENM is a reliable computational method to characterize the flexibility of proteins and, by extension, their dynamics. This analysis was performed using *WEBnma* [26], a web server to facilitate the assessment of protein flexibility within protein families and super-families, giving a good view of structural movement and preservation of flexibility on the different structures. Then we visualized our results on *PyMol* [27], and performed a *porcupine* plot which allowed us to observe the direction of significant movements of our protein.

2.3.2 Coarse Grain simulation with the Martini force field

We carried out a coarse-grained simulation of molecular dynamics of our model which consists of several steps [29]. First, we converted our protein structure into a large grain model. Then we generated an appropriate Martini-like topology, minimization, solvation of the protein in water, equilibration and then we launched molecular dynamics. Our first two steps are performed using the publicly available *martinize.py* script which generates topology and structure files in a format suitable for Gromacs [30].The last steps were also performed with the tools available with Gromacs .

2.3.3 Analysis of the molecular dynamics trajectory

In this final step we used the *PyMol* visualization tool, which is a molecular graphics program designed for the display and analysis of molecular assemblies, in particular biopolymers such as proteins and nucleic acids, which allowed us to visualize and analyze our results of molecular dynamics simulations. We then carried out a more in-depth analysis of our dynamics using the *bio3d* package on R, in order to observe the evolution of RMSD and RMSF over time and.

3 Results and interpretation

3.1 Study of the protein

Based on the information that can be found on *Uniprot*, we find that we are dealing with a small *unreviewed* annotated membrane protein consisting of 268 amino acids. This protein comes from an archaeobacteria **Halobacterium sp.D1** and is encoded by the **HALDL1-13775** gene. From Uniprot we can see that in the String database, They demonstrated interactions of the target protein with other proteins whose accession numbers on Uniprot are as follows : **W0K2E7** encoded by the same gene that our target protein, the **W0K4U9** protein encoded by the **HALDL1-13780** gene, and the **W0K0V0** protein encoded by the **HALDL1-13770** gene. Unfortunately our protein is not mentioned in the literature, no function has yet been associated with it and the same is true for the proteins with which our target protein interacts.

The hydrophobicity study confirms that the protein is indeed of the transmembrane type and more precisely that the protein crosses the cell membrane several times. in fact, the hydrophobicity profile below shows an alternation of hydrophilic and hydrophobic peaks. The amino acids that make up the hydrophobic part (score$>$ 0) should correspond to the transmembrane part while the amino acids that make up the hydrophilic parts (score $<$0) should correspond to the extracellular and intracytosolic parts. The protein is expected to cross the membrane at least 6 times.

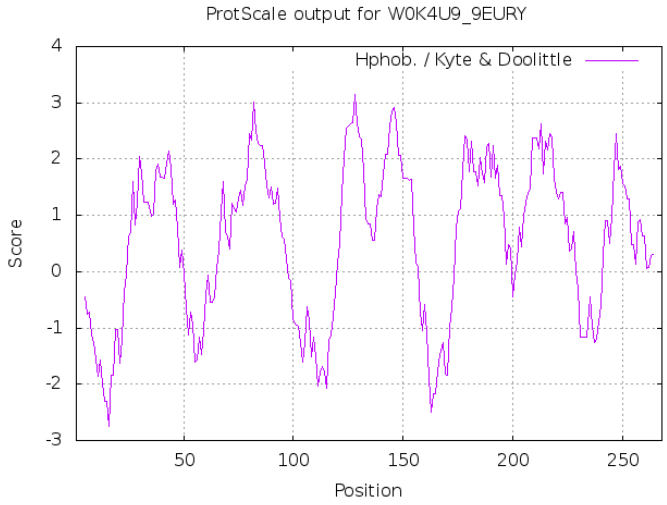


FIGURE 2 – General diagram for the prediction of the 3D structure of proteins

With the TMHMM server [6], we observe that there are indeed 6 transmembrane segments and all alpha helices (see Figure 3 bis below).

```
# tr|W0K4U9|W0K4U9_9EURY Length: 268
# tr|W0K4U9|W0K4U9_9EURY Number of predicted TMHs:  6
# tr|W0K4U9|W0K4U9_9EURY Exp number of AAs in TMHs: 145.89759
# tr|W0K4U9|W0K4U9_9EURY Exp number, first 60 AAs:  22.74203
# tr|W0K4U9|W0K4U9_9EURY Total prob of N-in:        0.99625
# tr|W0K4U9|W0K4U9_9EURY POSSIBLE N-term signal sequence
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        inside       1    20
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        TMhelix     21    43
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        outside     44    75
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        TMhelix     76    98
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        inside      99   118
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        TMhelix    119   136
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        outside    137   139
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        TMhelix    140   157
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        inside     158   169
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        TMhelix    170   192
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        outside    193   206
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        TMhelix    207   229
tr|W0K4U9|W0K4U9_9EURY  TMHMM2.0        inside     230   268
```

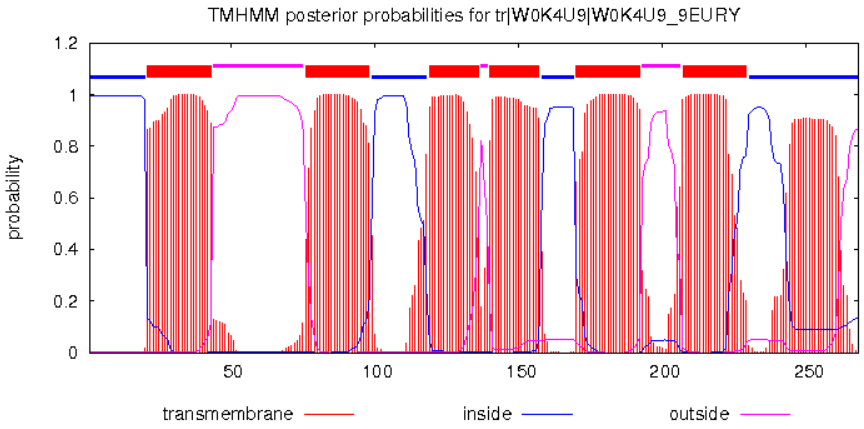
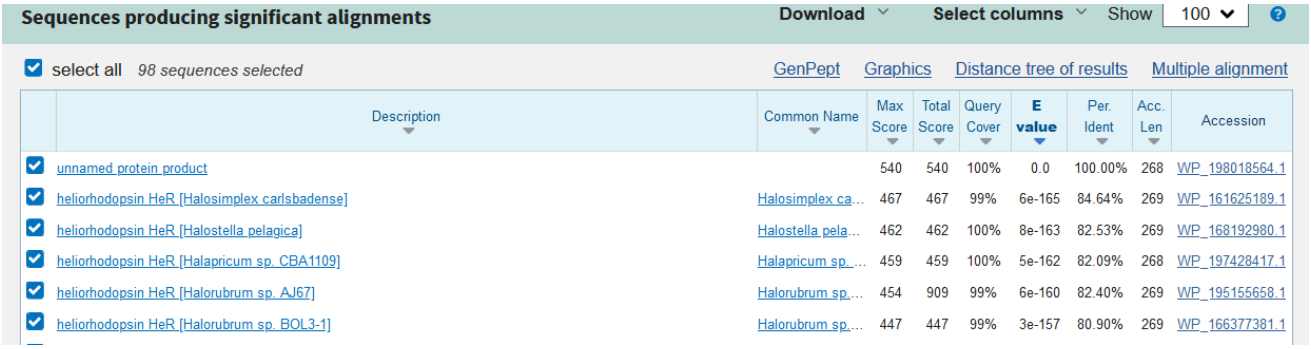


FIGURE 3 – Prediction profile of transmembrane helices of the W0K4U9 protein with ProtScale

Afterwards, we were interested in the conserved domains of the family to which our target protein belongs. For this we started by searching in the libraries by sequence similarity. We therefore launched a BlastP against the *nr (non-redundant protein sequences)* database. We obtained 98 protein sequences (cf. FIG. 4) having similarities with our target protein. We then started a *T-Coffee* multiple alignment to look for the presence of conserved regions in these sequences. From the results obtained during our Blastp, we selected a set of sequences (about thirty) to achieve a multiple alignment with the *ClustalW* program. We constructed a sample of closely related sequences (the first 30 sequences) by including the protein of interest to bring out conserved areas during the evolution. In this alignment [9], our target sequence carries the identifier **WP-198018564.1** (first sequence) and we could see that the conserved area starts from the 10th amino acid approximately until the end of the sequence. From this alignment, we can deduct that almost the entire sequence would be a conserved domain. In other words, the whole transmembrane part.



Sequences producing significant alignments Download Select columns Show 100

select all *98 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
unnamed protein product		540	540	100%	0.0	100.00%	268	WP_198018564.1
heliorhodopsin HeR [Halosimplex carlsbadense]	Halosimplex ca...	467	467	99%	6e-165	84.64%	269	WP_161625189.1
heliorhodopsin HeR [Halostella pelagica]	Halostella pela...	462	462	100%	8e-163	82.53%	269	WP_168192980.1
heliorhodopsin HeR [Halapricum sp. CBA1109]	Halapricum sp. ...	459	459	100%	5e-162	82.09%	268	WP_197428417.1
heliorhodopsin HeR [Halorubrum sp. AJ67]	Halorubrum sp....	454	909	99%	6e-160	82.40%	269	WP_195155658.1
heliorhodopsin HeR [Halorubrum sp. BOL3-1]	Halorubrum sp....	447	447	99%	3e-157	80.90%	269	WP_166377381.1

FIGURE 4 – Blasp results against nr databases.

3.2 Prediction of the tertiary structure

The Blastp alignment against PDB to find homologue sequences to our target sequence provided us with 4 sequences potentially similar to our target sequence (see Figure 5) and whose structures are known. However among these 4 results, only the first two sequences seem to be templates for the prediction of the tertiary structure. These proteins, whose PDB identifiers are 6IS6 and 7CLJ, belong to the *HeRs* family and a percentage identity of 51.38% and 50.99% respectively is observed. We observe a recovery percentage of 94%, and a good E-value of the order 10^{-85}. From the values of these parameters we can assume a homology between these sequences and our target sequence.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Crystal structure of Thermoplasmatales archaeon heliorhodopsin [Thermoplasmatales archaeon SG8-52-1]	Thermoplasmatales arch...	258	258	94%	3e-86	51.38%	259	6IS6_A
Crystal structure of Thermoplasmatales archaeon heliorhodopsin E108D mutant [Thermoplasmatales archaeo...	Thermoplasmatales arch...	257	257	94%	1e-85	50.99%	259	7CLJ_A
Crystal structure of bacterial heliorhodopsin 48C12 [Actinobacteria bacterium]	Actinobacteria bacterium	170	170	91%	1e-51	41.70%	252	6UH3_A
Crystal structure of the 48C12 heliorhodopsin in the violet form at pH 8.8 [Actinobacteria bacterium]	Actinobacteria bacterium	170	170	91%	2e-51	41.70%	264	6SU3_A

FIGURE 5 – Blasp results against PDB

After launching all our 3D structure predictions, we finally obtained 26 models : 15 with the *I-Tasser* program (5 by predictions), 2 with *Phyre2*, 5 with and 4 with ModWeb. As described previously, for some tools we have imposed templates, for others not. For each tool we classified the models by structural similarity by aligning the structures together to calculate the RMSD, then select the two most distant models. *Phyre2* gives us 20 predictions however this tool selects only two final models with sufficient identity percentage to consider the predictions as sufficiently exploitable [20] so we decided to keep only the first two models (see Figure 6).

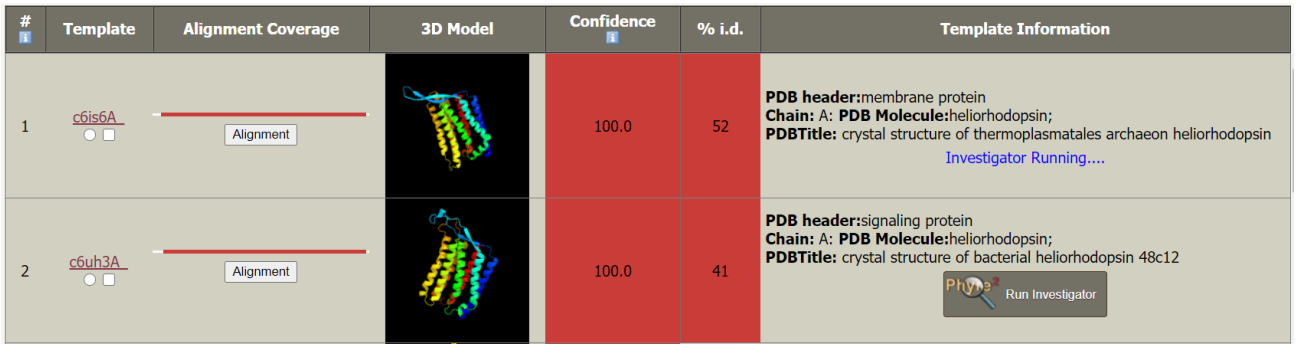


FIGURE 6 – The 2 results kept with Phyre2

As said previously with ModWeb, we launched a Fast mode and a Slow (or intensive) mode and we obtained 2 results per mode. In both modes, we get a prediction of the 3D structure with the 6IS6 template, however, we decided to take the results from the slow mode, because the study was more thorough and the percentage of identity found with the slow mode (52%) is closer than that of Fast mode (53%) compared to the results of our Baslp against PDB (51.38%). In the end with ModWeb we selected three prediction models (see Figure 7).

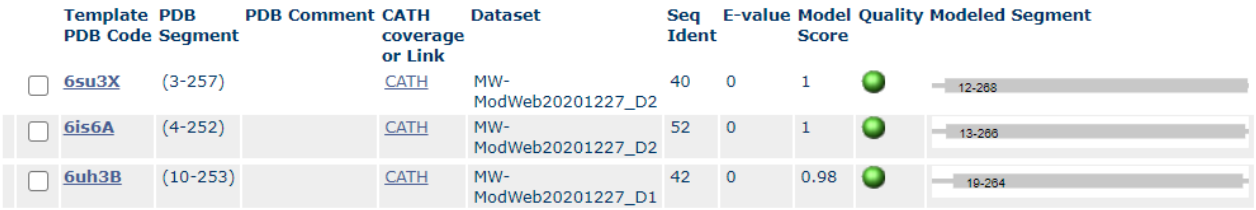


FIGURE 7 – The 3 results kept with ModWeb

For RaptorX, we observe (see Table 2) that the RMSDs between structures are very close. We therefore decided to keep only the first model, classified as the best model by RaptorX (see Figure 8).

1&2 = 0.046	1&3 = 0.052	1&4 = 0.049	1&5 = 0.045
2&3 = 0.055	2&4 =0.044	2&5 =0.042	
3&4 = 0.044	3&5 = 0.040	4&5 = 0.033	

Table 2 : RMSD calculation table between each structure for RaptorX

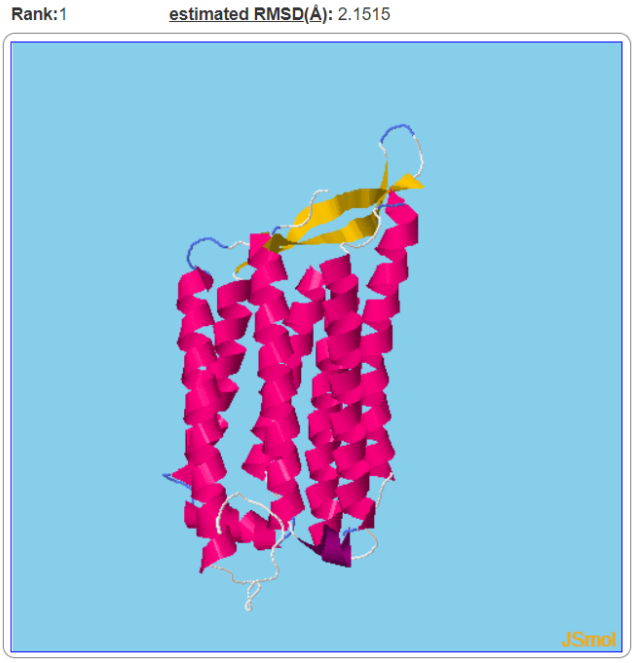


FIGURE 8 – First model of RaptorX

For the results with *I-Tasser*, we decided to keep by modeling the 2 most distant models structurally. As *I-Tasser* classifies the structures from 1 to 5, we kept for each modeling the model classified number 1 and aligned it with the 4 other structures to calculate the RMSD and see with which structure it is the most distant. Then we looked with which model, is this structure the closest (lowest RMSD) and looked at its ranking in *I-Tasser*. We have therefore classified each structure by modeling so as to keep the two models having the most distant structures and having the best *I-Tasser* classification.

I-Tasser, gives us 5 results by prediction classified according to the parameters explained previously including the C-score.

For example, for modeling without imposing a template on *I-Tasser*, we looked with which prediction model number 1 is the furthest away (cf. Table 3). We can see that this is model 5 with an RMSD between these two structures of 0.484 Å. Then we looked with which model, model 5 is the closest, we observed that it was model 4 with an RMSD between the two structures of 0.378 Å. Since model 4 is classified before 5 and having the highest C-score in this model (see Figure 9), we therefore decided to keep models 1 and 4.

1&2 = 0.409	1&3 = 0.393	1&4 = 0.353	1&5 = 0.484
2&3 = 0.337	2&4 = 0.331	2&5 = 0.416	
3&4 = 0.269	3&5 = 0.421	4&5 = 0.378	

Table 3 : Calculation table of the RMSD between each structure for I-Tasser without template

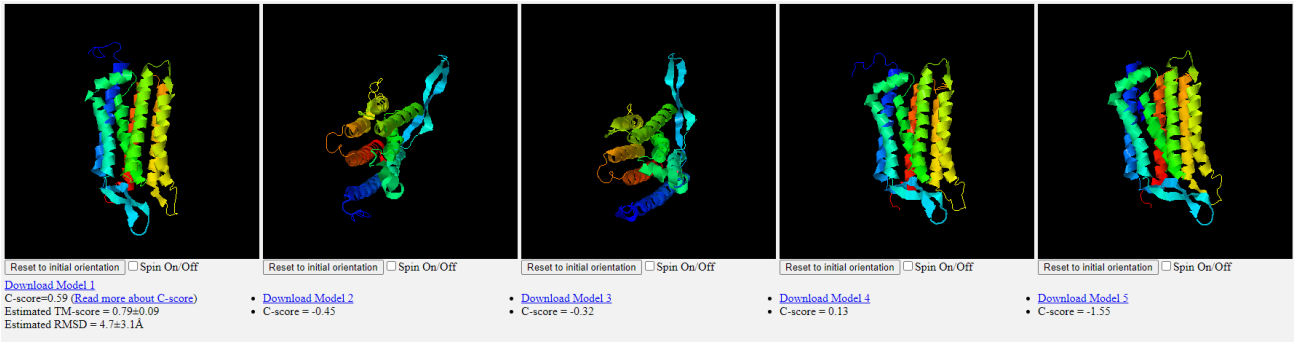


FIGURE 9 – the 5 models predicted by I-TASSER without template

For the modeling with *I-Tasser* by imposing the 6IS6 template, we decided to keep models 1 and 3 (Cf. Table 1 and Figure 1 in the Annex). Finally for the modeling with *I-Tasser* by imposing the 7CLJA template, we decided to keep models 1 and 4 (Cf. Table 2 and Figure 2 in the Annex). The preselected models are summarized in the following tables (see Table 4). After selecting our models, we evaluated, refined, and then re-evaluated them to select the final model.

	Phyre2	ModWeb	RaptorX	I-Tasser
Number of models	2	3	1	6
Template	6is6 6uh3	6is6 6su3x 6uh3	-	6is6 3abw 6is6 *2 7clja*2

Table 4 : Summary of our pre-selection of prediction models.

3.2.1 Evaluation, improvement, and selection of the final model

In this part, we have evaluated the models selected in the previous step with Process. Then we refined areas like curls and side chains with *GalaxyRefine2*. We re-evaluated them, to see if there was improvement or not in the structures. All the results of the models evaluated before and after refinement are presented in the Annex from Table 4 to Table 12.

For *Phyre2* we could see that the quality values after refinement are almost identical between the two models, however the quality scores of model 1 predicted from template 6IS6 have decreased. While for model 2 predicted from template 6UH3A after refinement have

been improved. So *Phyre2* overestimated the prediction quality of its first model. (See Table 3 in the report and Table 3 in the Annex). We therefore kept model 2 of *Phyre2*. Regarding the results of *ModWeb*, we see that the model predicted from the 6IS6 template is the model with the best score value before and after refinement. (See Table 4 in the report and 4.5 in the Annex). For the results with *RaptorX*, we observed that the score values after and before refinement are the same whatever the model, so we decided to keep model 1 because the program ranks the results from best to worst. (See Table 5 in the report and 6 in the Annex). Finally for *I-TASSER*, we started by comparing the models without template, then those with template. It is observed that in all cases, model 1 is the best improved after refinement (Cf Table 6 in the report and 7 to 11 in the Annex).

We then compared the models between tools, and we found that 3 predictions have the same score values after refinement. This is *Phyre2* model 2 predicted from template 6UH3, model predicted from template 6IS6 from *ModWeb* and *I-TASSER* model 1 predicted by imposing template 6IS6. Since two models are predicted from the same template, we aligned them to calculate the RMSD, we got an RMSD of 0.45 Å. Given the relatively small RMSD value, we decided to choose the model predicted from *ModWeb*'s 6IS6 template because the tool uses Modeller and does comparative modeling, while *Phyre2* uses a Threading approach.

	Modèle initial	Modèle 1 raffiner
Overall Quality	4.5	5.5
Covalent Bond Quality	6.5	7.5
Non-Covalent/Packing Quality	3.5	4.5
Torsion Angle Quality	5.5	6.5

Tableau 3 : Tableau de comparaison du modèle 2 6uh3 en utilisant Phyre2

	Modèle initial	Modèle 1 raffiner
Overall Quality	4.5	5.5
Covalent Bond Quality	6.5	7.5
Non-Covalent/Packing Quality	3.5	4.5
Torsion Angle Quality	5.5	6.5

Tableau 4 : Tableau de comparaison du modèle 6is6A avec 52%id en utilisant ModWeb

	Modèle initial	Modèle 1 raffiner
Overall Quality	3.5	4.5
Covalent Bond Quality	3.5	7.5
Non-Covalent/Packing Quality	2.5	3.5
Torsion Angle Quality	4.5	6.5

Tableau 5: Tableau de comparaison du modèle 1 en utilisant RaptorX

	Modèle initial	Modèle 1 raffiner
Overall Quality	3.5	5.5
Covalent Bond Quality	6.5	7.5
Non-Covalent/Packing Quality	3.5	4.5
Torsion Angle Quality	1.5	6.5

Tableau 6: Tableau de comparaison du modèle 1 en utilisant *I-TASSER* avec template 6is6

We then proceeded to use the *OPM* program [28] to visualize the orientation of proteins within the lipid bilayer. Comparing the model of *ModWeb* and that of *Phyre2*, we noticed that it has a similar orientation, so we made an alignment of its two structures within the membrane. We observed that the topology of the two 3D models within the cell membrane is identical. A beta sheet on the extracellular side and 6 transmembrane alpha helices (see Figure 10). This confirms the results obtained with THMM, just by having provided the fasta sequence to the web server.

In the end, we decided to keep the *ModWeb* prediction because it uses the 6IS6 template which, according to the blast, has a higher percentage of identity, compared to the 6UH3 template used by *Phyre2* (see figure 5). In addition, comparative modeling generally gives better results than by threading or ab initio.

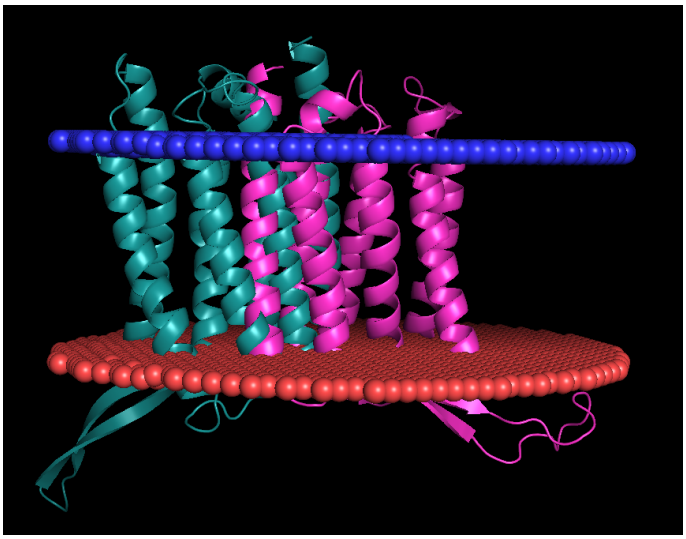


FIGURE 10 – Visualization of ModWeb and Phyre2 results on OPM using Pymol

After choosing our final model, we decided to compare it with that of *SWISSModel* already available on *Uniprot* which was predicted with the 7CLJ template. The RMSD between these two predicted structures is 0.609 Å. We deduce that our structure and that of *Swiss-Model* are not different even if we do not use the same template. Remember that the 7CLJ template is 50.99% identical to our protein of interest.

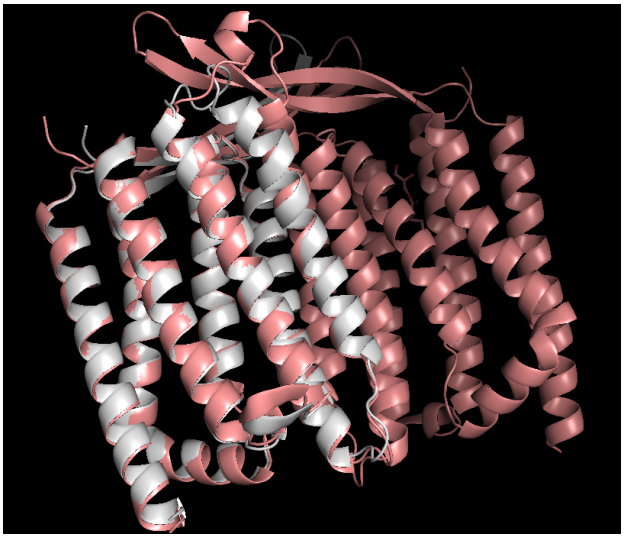


FIGURE 11 – Alignment visualization of the SwissModel (pink) and ModWeb (white) structures

Next, we wanted to observe the structures within the lipid bilayer with *OPM*. We obtained an RMSD of 1.60 Å. We were finally able to conclude from this alignment that our two structures are close.

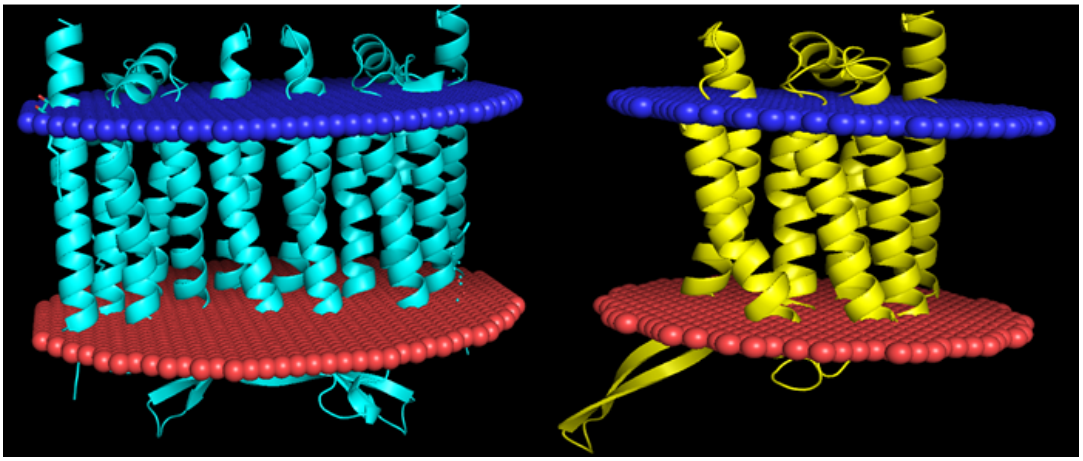


FIGURE 12 – Visualization of the lipid layers of the two structures SwissModel (blue) and ModWeb (yellow)

3.3 Study of the conformation over time of our models

3.3.1 Study of normal modes on the basis of an elastic network

In this part, we launched *WEBnma* on our final model. By using *PyMol* we were able to observe 6 normal modes describing movements of our protein (see Figure 11). There is movement throughout the protein, however the lateral expansion, which corresponds to the beta sheet, is the part of the protein with the most intense movement. We also notice similar movements between modes. Two major movements have been observed : the first corresponds to a sort of left-to-right sweep found in modes 7,8 and 9 (see Figure 11). The second corresponds more to a folding of the sheet on itself found in modes 11 and 12 (see Figure 12).

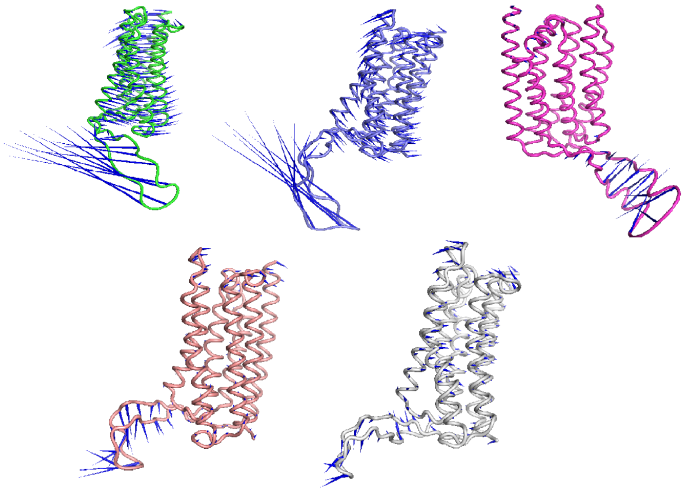


FIGURE 13 – Visualization of the 2 different movements of our structure on the modes

3.3.2 Coarse Grain simulation with the Martini force field and trajectory analysis

The visualization of the trajectory of the molecular dynamics of our solvated protein, allowed us to observe that our model changes conformation over time. Indeed, Figure 12 below shows us the evolution of RMSD over time. It is observed that the value of RMSD increases and does not reach a state of equilibrium. We therefore visualized our protein on *Pymol* and observed that we find a movement, observed during the study of normal modes. Indeed, we notice the folding of the beta sheet on the protein (cf. Figure 13) which is done more intensely than what is observed during the study of the normal modes. We then looked at the fluctuations of the residuals by calculating the RMSF (Root Mean Square Fluctuations). We generally observe a fluctuation of all the residuals (cf. Figure 14). However, a larger peak is observed between residues 45 to 75 roughly corresponds to the beta sheet (cf. Figure 15). We conclude that the only characteristic movement of the protein is that of the beta sheet.

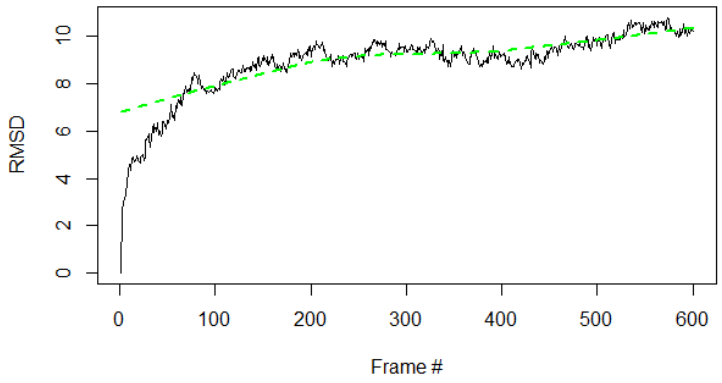


FIGURE 14 – Evolution of the RMSD over time of our model

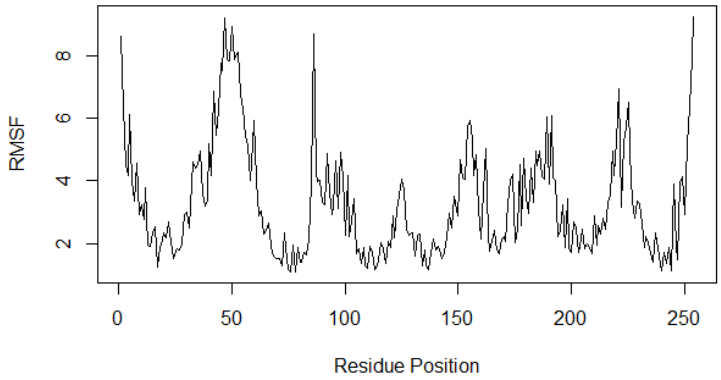


FIGURE 15 – Evolution of the RMSF over time of our model

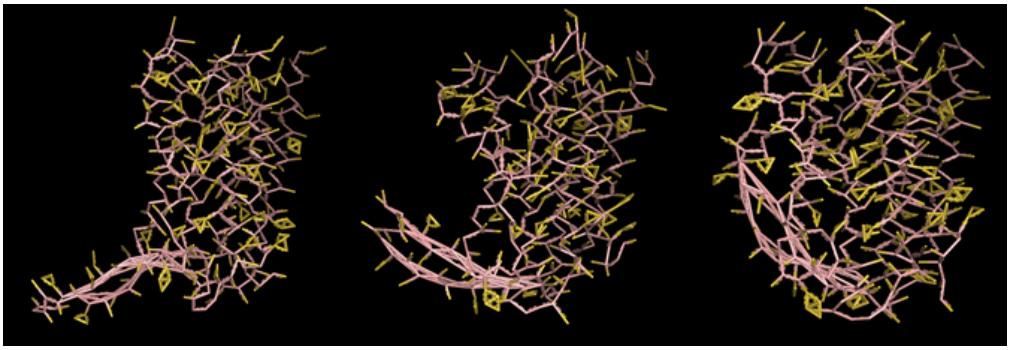


FIGURE 16 – Different states of our protein by applying molecular dynamics

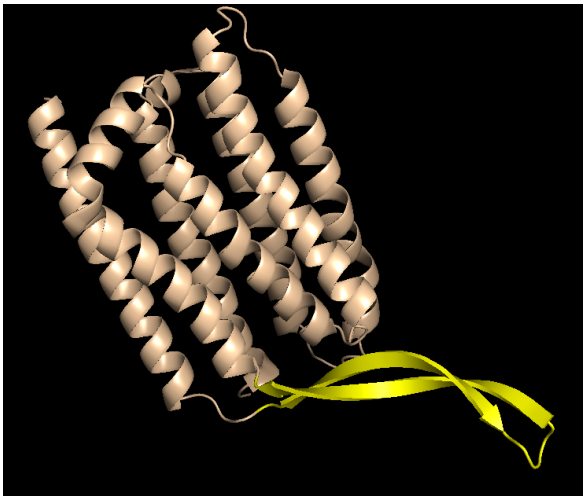


FIGURE 17 – Structure of our protein with the beta sheet colored yellow

4 Conclusion & Discussion

To conclude, to model our protein, we used different 3D structure prediction tools using different approaches (comparative modeling, *ab-initio*, threading, template imposed or not. . .). Homology modeling methods usually give better results. In our case, it was quite difficult to choose a final model for the study of movements over time. Indeed, almost all of our models evaluated after refinement have been improved. A first selection was made keeping the models with the best evaluation scores after refinement and then the final model was selected according to the prediction approach and the template used. The ideal would have been to study several good models predicted by different approaches : threading, *ab-initio*. We therefore succeeded in obtaining a structure prediction thanks to *ModWeb*. This prediction was made by comparative modeling using the 6IS6 template having the best identity rate with our sequence of interest. Predicting the 3D structure allowed us to observe the folded structure of the 6 alpha helices and the beta sheet that make up the transmembrane protein. At the end of this first major step of our project, we can say that our modeling of the W0K4U9 protein was a success. In fact, the comparison of the predicted structure with *SwissModel* and deposited on *Uniprot*, allowed us to observe a similar structure, and a similar orientation within the lipid bilayer.

The study of conformational changes using normal modes, allowed us to identify different movements such as the folding of the beta sheet and its sweeping from left to right. However, the molecular dynamics study did not allow us to visualize this sweeping movement. Our coarse grain simulation with a Martini force field was relatively short (600ps). Indeed, it does not allow us to reach a state of equilibrium of the structure as shown in figure 12 where the values of RMSD continue to increase over time. For this, the ideal would have been to increase the simulation time so that the dynamic stops when the structure of the protein no longer changes over time. A longer simulation would surely allow us to detect sweeping motion or even unobserved motion either using normal modes or in our first simulation. In addition, we recall that the dynamic was made with the protein solvated in water. As we are sure that our protein is membrane and more precisely transmembrane, the objective now would be to study the movements of the protein in its biological environment, that is to say in the lipid bilayer.

5 Annex

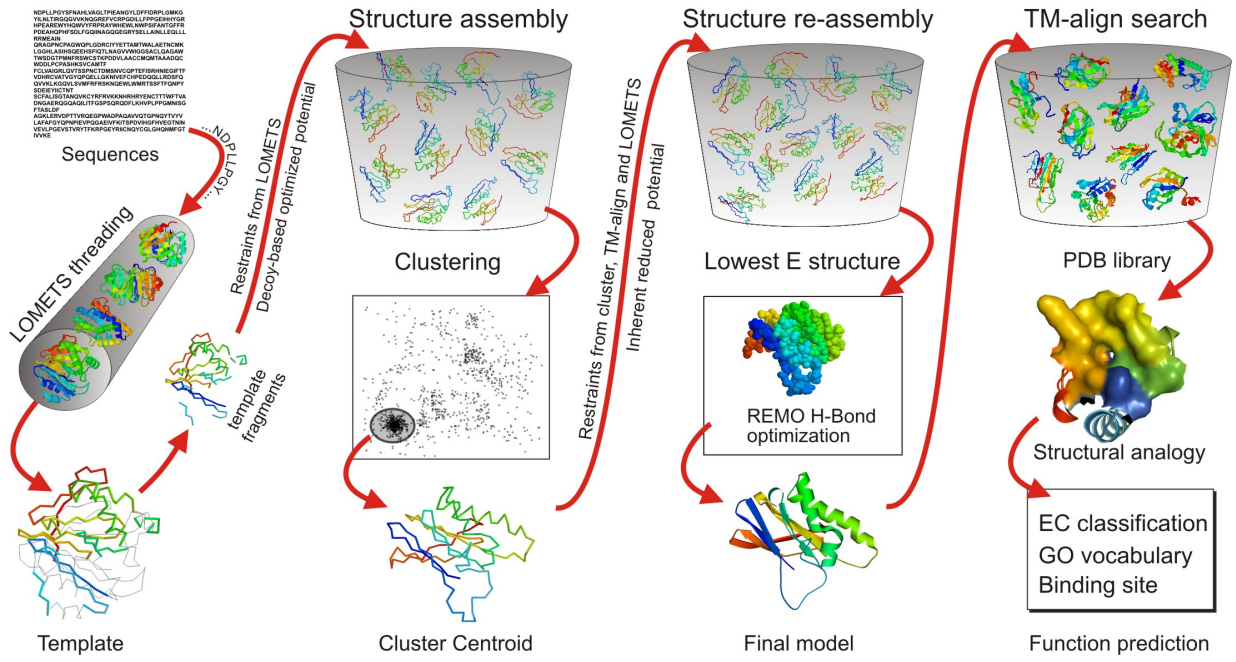


Figure 1 : RaptorX prediction method

1&2 = 0.259	1&3 = 0.230	1&4 = 0.259	1&5 = 0.243
2&3 = 0.223	2&4 = 0.273	2&5 = 0.282	
3&4 = 0.235	3&5 = 0.262	4&5 = 0.262	

Table 1 : Table of the calculation of the RMSD between each structure for I-Tasser with the 6IS6 template

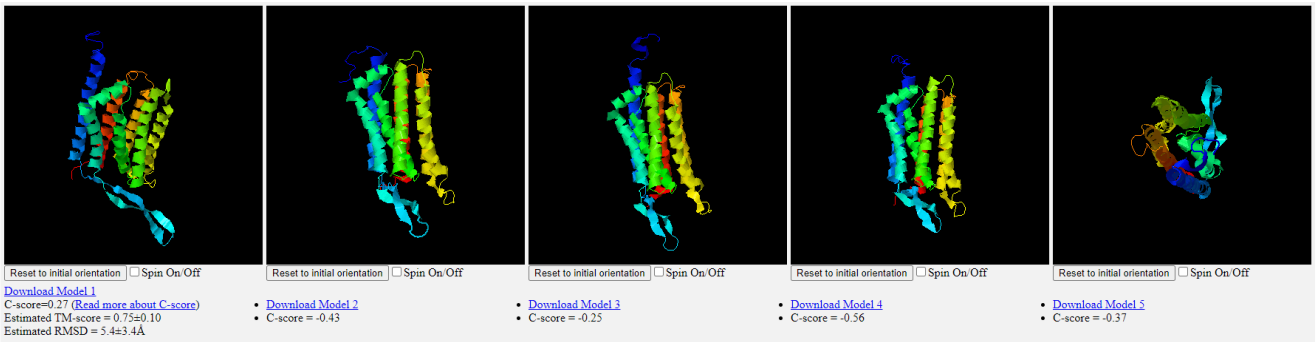


FIGURE 18 – the 5 models predicted by I-TASSER by imposing the 6IS6 template

1&2 = 0..387	1&3 = 0.458	1&4 = 0.388	1&5 = 0.296
2&3 = 0.552	2&4 = 0.380	2&5 = 0.446	
3&4 = 0.361	3&5 = 0.421	4&5 = 0.421	

Table 2 : Calculation table of the RMSD between each structure for I-Tasser with the 7CLJA template

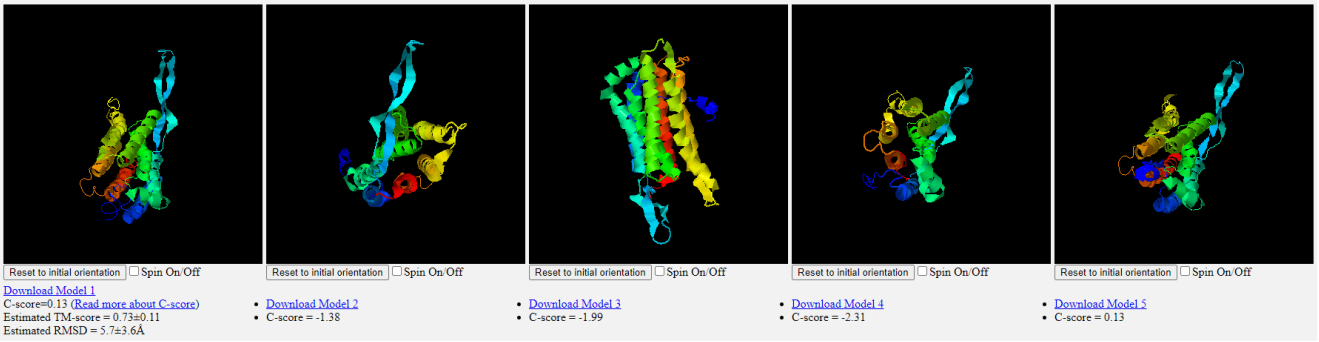


FIGURE 19 – the 5 models predicted by I-TASSER by imposing the 7CLJA template

Table 3 : 6is6 model comparison table using Phre2

	Modèle initial	Modèle 1 raffiner
Overall Quality	4.5	5.5
Covalent Bond Quality	7.5	6.5
Non-Covalent/Packing Quality	3.5	4.5
Torsion Angle Quality	8.5	6.5

Table 4 : 6su3X model comparison table using ModWeb

	Modèle initial	Modèle 1 raffiner
Overall Quality	4.5	4.5
Covalent Bond Quality	6.5	7.5
Non-Covalent/Packing Quality	3.5	3.5
Torsion Angle Quality	4.5	6.5

Table 5 : 6uh3B model comparison table using ModWeb

	Modèle initial	Modèle 1 raffiner
Overall Quality	4.5	4.5
Covalent Bond Quality	7.5	7.5
Non-Covalent/Packing Quality	3.5	3.5
Torsion Angle Quality	3.5	6.5

Table 6 : Model 3 Comparison Table Using RaptorX

	Modèle initial	Modèle 1 raffiner
Overall Quality	3.5	4.5
Covalent Bond Quality	3.5	7.5
Non-Covalent/Packing Quality	2.5	3.5
Torsion Angle Quality	4.5	6.5

Table 7 : Model 4 comparison table using I-TASSER without imposing template

	Modèle initial	Modèle 1 raffiner
Overall Quality	2.5	5.5
Covalent Bond Quality	6.5	6.5
Non-Covalent/Packing Quality	3.5	4.5
Torsion Angle Quality	1.5	6.5

Table 8 : Model 3 comparison table using I-TASSER with template 6IS6

	Modèle initial	Modèle 1 raffiner
Overall Quality	3.5	4.5
Covalent Bond Quality	7.5	7.5
Non-Covalent/Packing Quality	3.5	3.5
Torsion Angle Quality	1.5	6.5

Table 9 : Comparison table of model 4 using I-TASSER with template 7CLJA

	Modèle initial	Modèle 1 raffiner
Overall Quality	3.5	3.5
Covalent Bond Quality	7.5	7.5
Non-Covalent/Packing Quality	3.5	3.5
Torsion Angle Quality	1.5	6.5

Table 10 : Model 1 comparison table using I-TASSER with template 7CLJA

	Modèle initial	Modèle 1 raffiner
Overall Quality	2.5	4.5
Covalent Bond Quality	6.5	7.5
Non-Covalent/Packing Quality	3.5	3.5
Torsion Angle Quality	1.5	6.5

Table 11 : Model 1 comparison table using I-TASSER with template 6IS6

6 References

[1] https://fr.wikipedia.org/wiki/Mod%C3%A9lisation_de_prot%C3%A9ines_par_homologie

[2]https://fr.wikipedia.org/wiki/Prot%C3%A9ine_membranaire

[3] https://fr.wikipedia.org/wiki/Prot%C3%A9ine_membranaire

[4]http://ressources.unisciel.fr/biocell/chap1/co/module_Chap1_9.html

[5] https://pfam.xfam.org/family/Heliorhodopsin

[6]https://web.expasy.org/protscale/ & http://www.cbs.dtu.dk/services/TMHMM/

[7] https://www.uniprot.org/uniprot/W0K4U9

[8] https://blast.ncbi.nlm.nih.gov/Blast.cgi

[9]http://tcoffee.crg.cat/apps/tcoffee/result?rid=ba4707ff

[10] http://pedagogix-tagc.univ-mrs.fr/courses/bioinfo_intro/BI4U2/concepts.html

[11]https://www.nature.com/articles/nprot.2012.085

[12] https://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html (documentation)

[13]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC168950/

[14]https://www.nature.com/articles/nprot.2015.053

[15] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3226909/

[16] https://zhanglab.ccmb.med.umich.edu/I-TASSER/output/S589836/ (I-TASSER results without template)

[17] https://zhanglab.ccmb.med.umich.edu/I-TASSER/output/S589905/ (I-TASSER results with template 1)

[18] https://zhanglab.ccmb.med.umich.edu/I-TASSER/output/S589906/?fbclid=IwAR1kvl91Fuyk TASSER results with template 2)

[19] http://raptorx.uchicago.edu/ContactMap/myjobs/57552762_603517/ (Raptor X results)

[20] http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/6a2f9176f1f54891/summary.html (Phyre2 results)

[21]https://modbase.compbio.ucsf.edu/modbase-cgi/model_details.cgi?searchmode=default&displaymode=modpage&seq_id=1cc6c495d8479a1a59c8212b9ff8be08MSSHNSPI&model_id=c8e9768a83a02a04f1d968ce50831ca1&queryfile=1609175809_9043 (Mode slow & intensive/fast)

[22] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896095/

[23]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692086/

[24]https://www.dsimb.inserm.fr/dsimb_tools/OREMPRO/en/index.php

[25] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0427-6

[26]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1904280/

[27] https://pymol.org/2/

[28] https://opm.phar.umich.edu/

[29] http://www.cgmartini.nl/index.php/force-field-parameters

[30] http://www.gromacs.org/About_Gromacs