Structural Bioinformatics

Assignment 2 - Homology modelling

Question 1

We used the sequence from CASP Target T0882[1], using this sequence we identified homologs using the HHpred tool from the MPI Bioinformatics toolkit[2]. We used the 2LRU_A[3] hit as a template to create a model for our target, T0822. The HHPred tool converted this into PIR format. We had to adjust the beginning and ending ranges for the 2LRU sequence, according to the 2LRU PDB file, for MODELLER.

Next we used the example script from last year's students for MODELLER to create different models for our target, T0882. The scores of these models that were created are given in the table below.

Filename	molpdf	DOPE Score	GA341 Score
T0882.B99990001.pdb	372.89200	-6718.61865	0.97076
T0882.B99990002.pdb	400.99597	-6529.19287	0.57844
T0882.B99990003.pdb	387.10696	-6768.50391	0.68367
T0882.B99990004.pdb	327.69962	-6838.76270	0.66705
T0882.B99990005.pdb	345.09473	-6845.7504	0.74111

Table 1: Output from MODELLER showing 5 different models with corresponding 'molpdf'-, 'DOPE'- and 'GA341'-scores.

[1]: http://predictioncenter.org/casp12/target.cgi?id=27&view=all
[2]: https://toolkit.tuebingen.mpg.de/#/tools/hhpred
[3]: http://www.rcsb.org/pdb/explore/explore.do?structureId=2LRU

Question 2

To assess how modeller handles gaps in the sequence alignment, we visualized both protein models. In figure 1 we can see both protein structures, because our protein alignment already starts with a gap we created a new alignment file with the same protein structure, but removed the gap at the start.
This results in a shift for the whole alignment and for the models created by MODELLER . We can see that the structure itself has not changed, but the position in space has changed for the complete structure. We think that this is effect is largely because the gap is at the start of alignment, which caused this shift in space.

Figure 1: Protein structure of the T0882 target. Blue protein structure has a gap at the start of the sequence. In the white protein structure we removed the gap.

Question 3
Global Distance Test (GDT) is a measure we used to assess the quality of our protein structure predictions. GDT maximizes the percentage of superimposed (or matched) residue pairs under four different threshold in each model, and reports the average of these percentages as the final GDT score.

The GDT score can be defined as:
GDT_TS = (GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8)/4.0
Where GDT_Pn is the estimation of the percent of residues that can fit under distance cutoff <= n.0 (1,2,4,8) Angstroms.

Model:	**GDT_TS SCORE:**
T0882.B99990001	74.658
T0882.B99990002	73.973
T0882.B99990003	73.288
T0882.B99990004	73.973
T0882.B99990005	74.315

Table 2: This table shows two columns, where the first column 'Model' shows the models that were created using MODELLER, and the second column 'GDT_TS Score' shows the GDT_TS score generated by LGA.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3607910/
http://predictioncenter.org/local/lga/lga_description.html

Question 4

In Figure 2 we can clearly see a correctly modeled helix turn helix structure. The white color shows the solution structure according to the pdb file of our target, the red color shows the model build using the 2LRU homolog found using pairwise sequence alignment and the blue color shows the model build using the HHpred tool. They both accurately follow the solution structure, this might be due to the fact that helices are easier to form due to their native structure. Figure 3 shows an example of a poorly modeled region, where the blue model lacks beta-sheets. We think this is due to the fact that the local alignment that was used excluded prior residues that could contain information on how the original bend should be formed. The red model does show beta-sheets, however due to gaps in the initial alignment the angle is slightly off and therefore the beta-sheets do not have a good overlap.

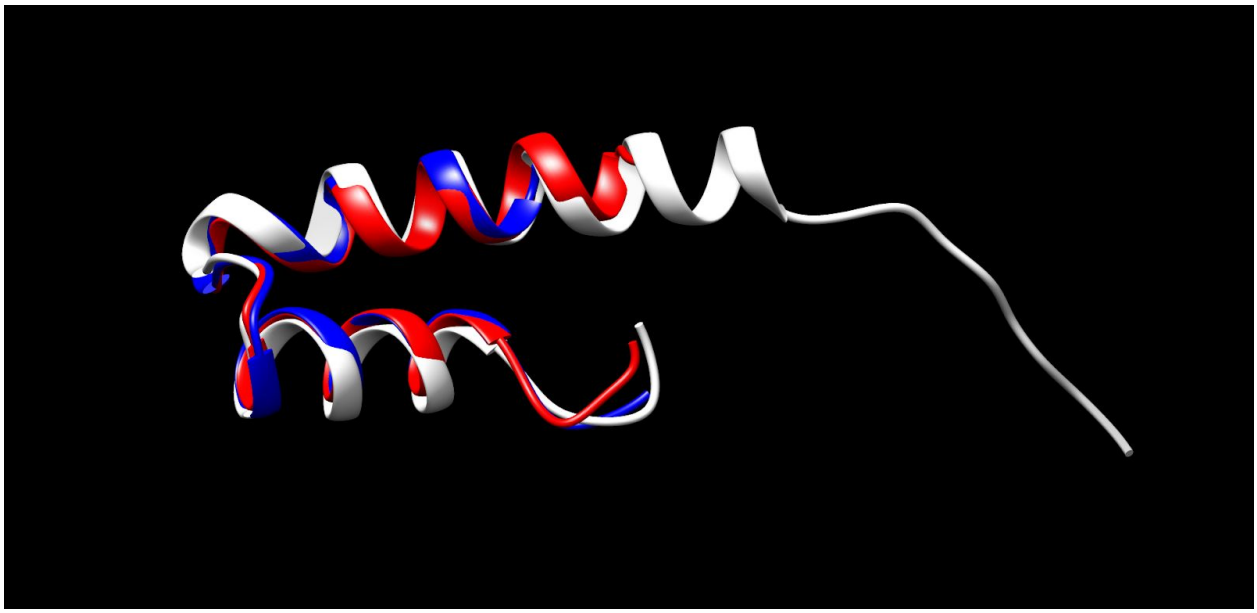


Figure 2: In white: solution structure (5G3Q); In blue: HHPred Model 1; In red: Pairwise Sequence alignment Model 5.

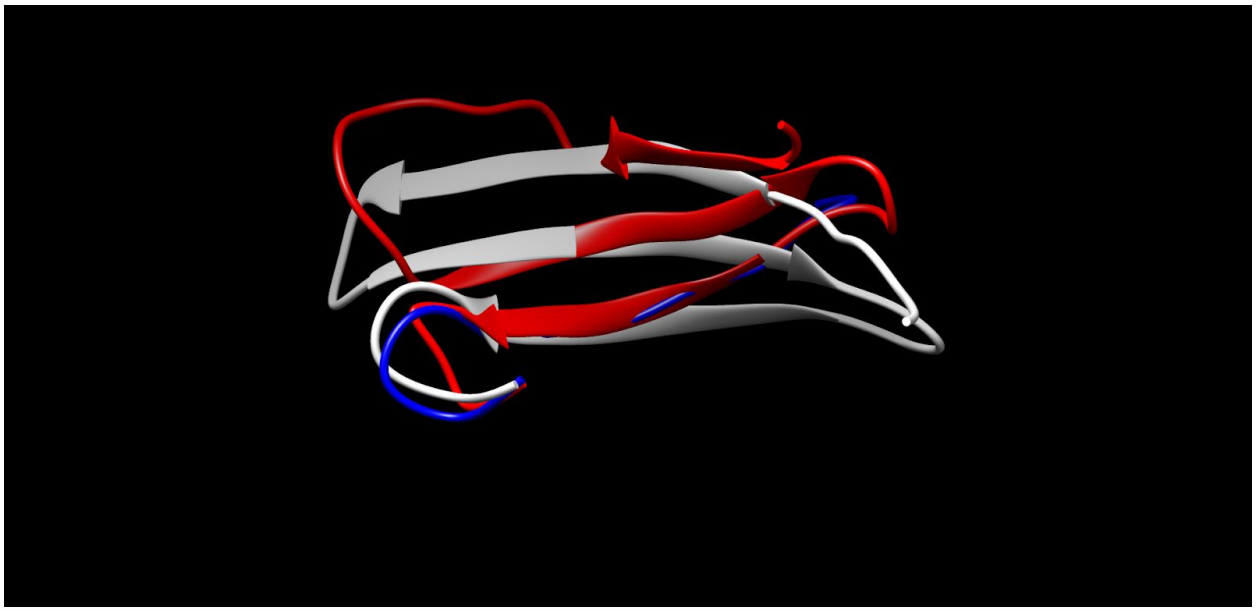


Figure 3: In white: Solution structure (5G3Q); In blue: HHPred Model 1; In red: Pairwise Sequence alignment Model 5.

Question 5
Discuss the added value of MODELLER: is the model created by MODELLER closer to the solution structure than the template is to the solution structure?

The protein structure from the template is more similar to the solution structure than the models created by MODELLER. This is not very strange, since the template we choose, 2LRU, is a subunit of our target and therefore is a correct representation of the structure. The models that we build using MODELLER do approach the actual structural conformations, however due to many available options in MODELLER and a basic understanding on how to use this program this could have caused biases in the model prediction.

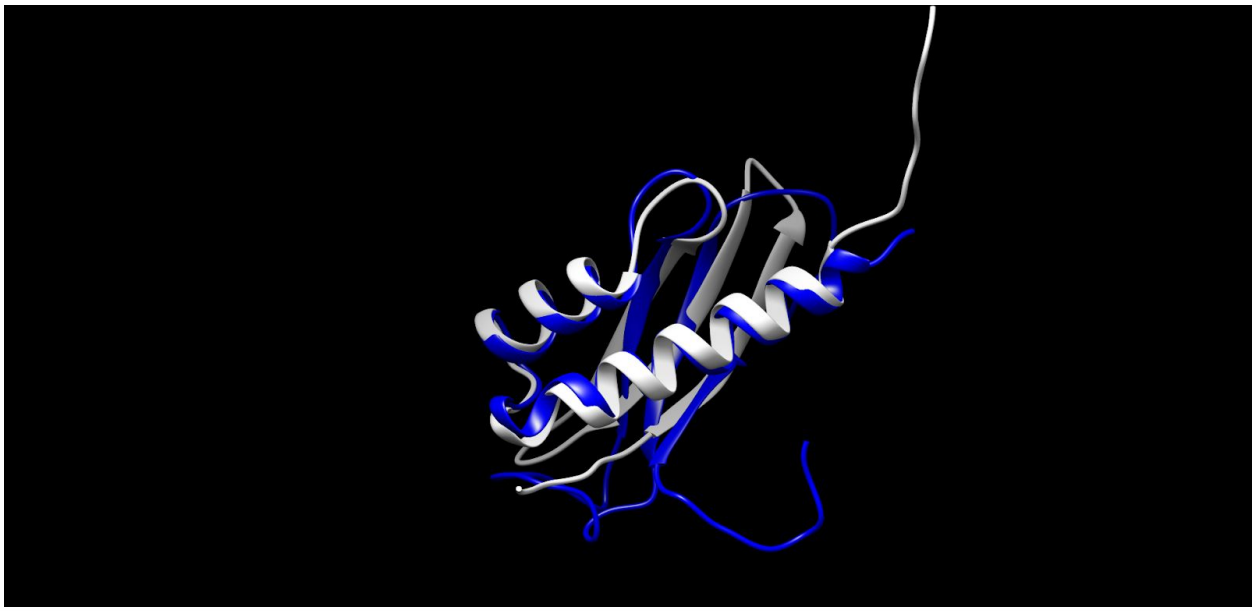


Figure 4: In white: solution structure 5G3Q; In blue: template structure (2LRU).

Question 6

We used the sequence from CASP Target T0882, using a local psi-blast [4] on the CASP sequence searching in the PDB database, we found the same homolog 2LRU_A. The alignment was converted into a PIR format by hand. We had to update the residue ranges to correctly align the sequences.

Next we used the example script from last year's students for MODELLER to create different models for our target T0882. We verified using the solution structure of T0822 (5G3Q.pdb), using LGA we found the GDT_TS scores. All these models that were created are given in table 3.

Filename	molpdf	DOPE Score	GA341 Score	GDT_TS
T0882.B99990001.pdb	191.50830	-3290.98682	0.74157	83.140
T0882.B99990002.pdb	166.94431	-3451.71045	0.67513	83.721
T0882.B99990003.pdb	143.39203	-3438.75415	0.65193	86.628
T0882.B99990004.pdb	208.66315	-3269.26025	0.60963	86.628
T0882.B99990005.pdb	189.50381	-3387.12280	0.63645	87.209

Table 3: Output from MODELLER showing 5 different models with corresponding 'molpdf'-, 'DOPE'- and 'GA341'-scores.

[4]: https://blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_388325532

Question 7

The model that was built using the HHpred tool from the MPI Bioinformatics Toolkit is in our opinion more accurate than the model that we build using pairwise sequence alignment. Our reasoning for this is that due to the local alignment the sequence from the pairwise sequence alignment only had 43 residues. This is shown in Figure 5 (Red). The beta-sheet is not fully formed in the model derived from the pairwise sequence alignment, due to this the model that was build using HHpred is more accurate.

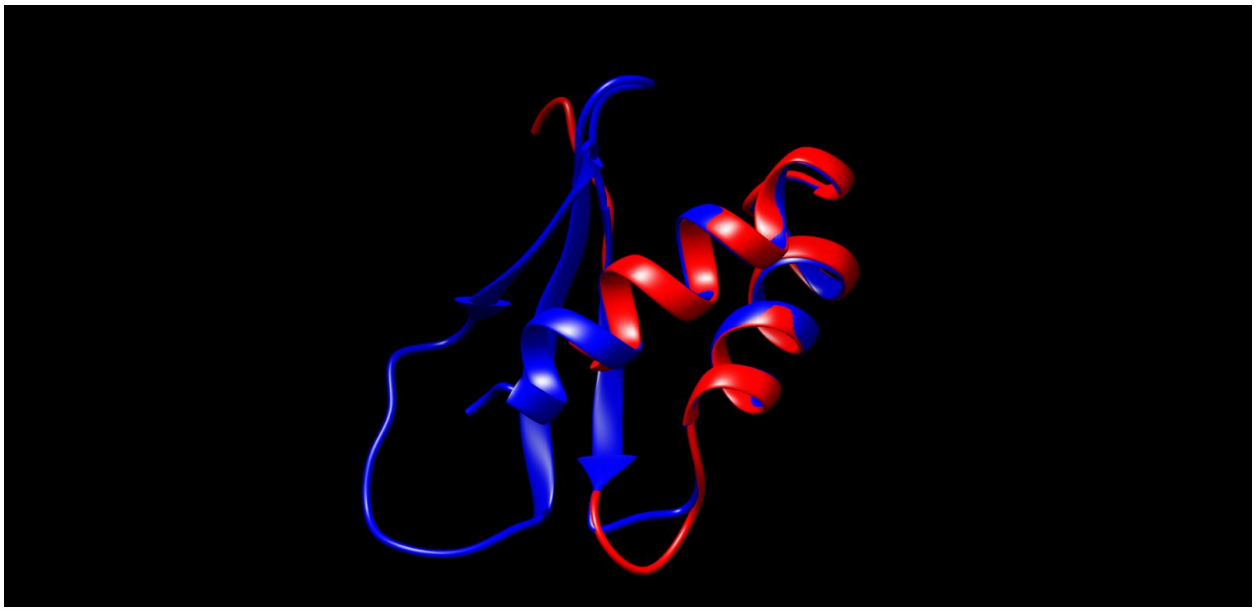
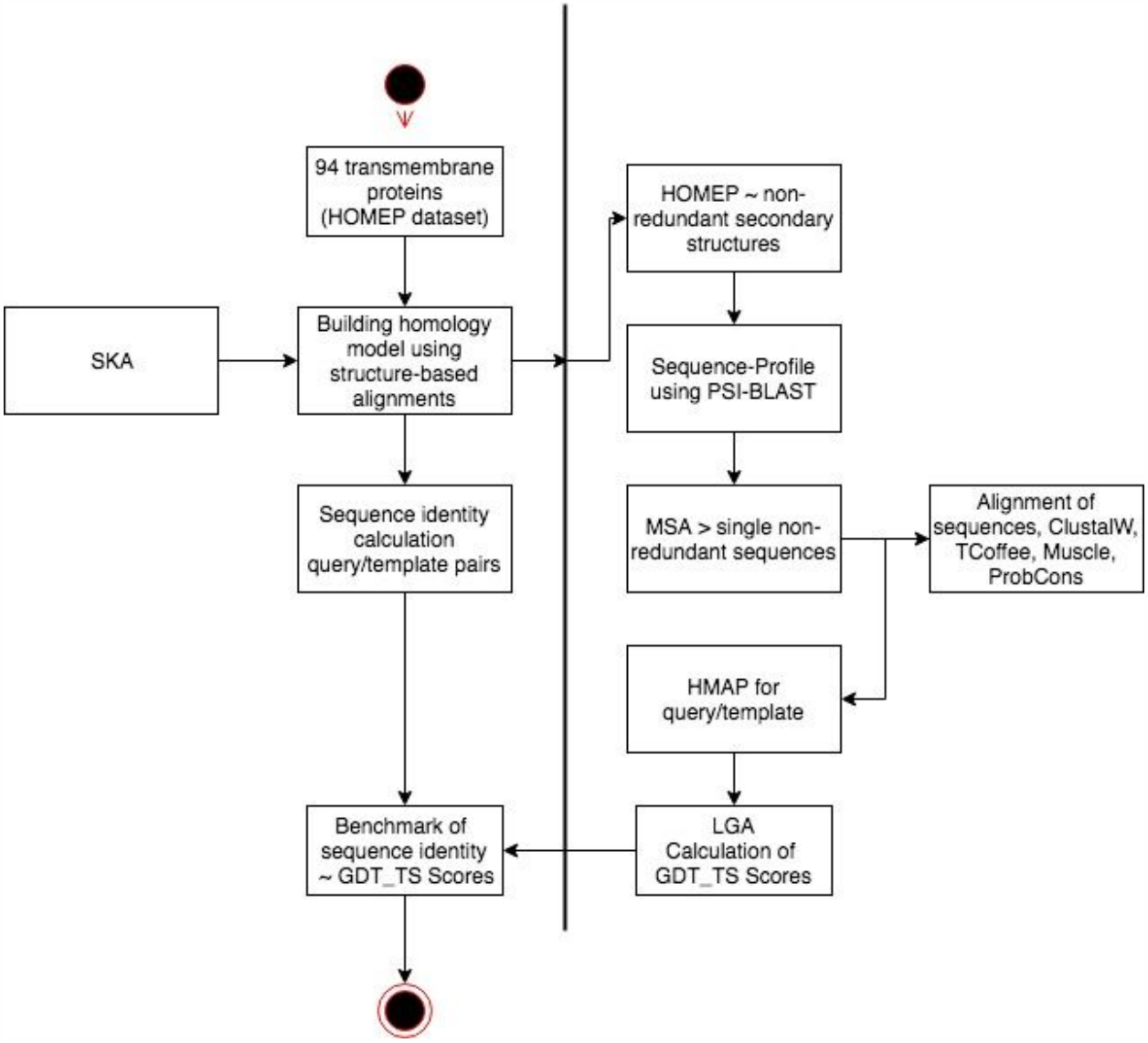


Figure 5: Blue is protein structure modelled with the HHpred method. Red is the protein structure we created with sequence alignment using psi-blast.

Question 8


94 transmembrane proteins (HOMEP dataset)
SKA
Building homology model using structure-based alignments
Sequence identity calculation query/template pairs
Benchmark of sequence identity ~ GDT_TS Scores
HOMEP ~ non-redundant secondary structures
Sequence-Profile using PSI-BLAST
MSA > single non-redundant sequences
Alignment of sequences, ClustalW, TCoffee, Muscle, ProbCons
HMAP for query/template
LGA Calculation of GDT_TS Scores

Question 9

The models for the transmembrane regions are more accurate, this is probably because the transmembrane domains are better conserved than regions outside the transmembrane domain. Other regions, such as the termini and loops between secondary structures are less preserved than the transmembrane domain. Due to this fact it is harder to model these ‘lesser preserved’ regions because the variability is much higher. Determining x-ray crystallography models for soluble proteins may be easier than trying the current method for transmembrane proteins, because transmembrane bound domains are found in and around the cell membrane.

Question 10

This assignment was done by Nick Keur and Erik Schutte. The workload was distributed evenly, we both worked our way through the questions together.