# Assignment 2 Homology modeling

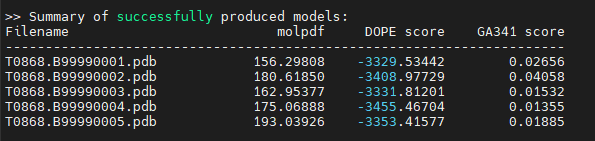
Lianne van Ruijven  
Anna Dekker

# Creating the model

### Question 1

The protein T0868 is selected as target. The sequence of the target is loaded into HHpredModel. The protein with the highest probability and deposited before may 2016 is chosen as a template. This is the protein 3ZIH\_A. In the pir file of the alignment, the name of the target protein is added. Furthermore, the name of the template is also adjusted by making the capital letters small. The beginning and ending residues are checked and assessed as correct.

The ‘build\_model.py’ script is edited by filling in the variables. The pir file is assigned to the variable ‘alnfile’, the template protein is assigned to the variable ‘knowns’ and the target protein is assigned to the variable ‘sequence’. By running the script in Linux, Modeller is used to create models from our alignment. The DOPE scores of the 5 created models are returned and shown in figure 1. A lower DOPE indicates a model with more accuracy and efficiency. Therefore, T0868.B99990004.pdb is chosen as the most suitable model. This model has the lowest DOPE score (-3455.46704).

  
**Figure 1.** *The models created by Modeller.*

### Question 2

The alignment of the target sequence with the template sequence is 160 aa long. The alignment has a gap (114 aa long) at the begin termini. Therefore, the part where the sequences overlap is 46 aa long. Chimera shows a target protein of 46 aa long. This means that the gap is removed from the model.

This answer is checked by adding a gap at the end termini of the proteins. The pir file of the alignment is modified so that the beginning and ending residues are not determined. Again does Chimera show a target protein of 46 aa long. The added gap is also removed from the model.

# Scoring you model

### Question 3

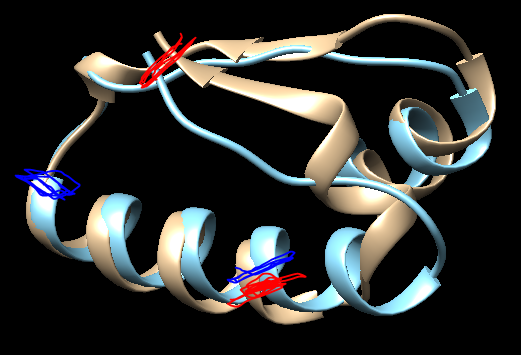
Our GDT\_TS score is 76.63 %. Our GDT\_TS score indicates that on average, 76.63 % of the alpha carbon atoms on the model and the solution, lay within a certain cut-off distance of each other. These cut-off distances are 1, 2, 4, and 8 angstroms. Compared to other models (see figure 6.5 from book Introduction to Protein structural bioinformatics), a model with a GDT\_TS score of 76.63% can be noted as a well predicted model. This leads to that our GDT\_TS score also indicates that the alignment used for the modelling is of good quality.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Range solution | Range model | DOPE | GDT\_TS |
| 1 | 315\_A:270\_A | 01:46 | -332,95 | 73,37 |
| 2 | 315\_A:270\_A | 01:46 | -340,90 | 72,28 |
| 3 | 315\_A:270\_A | 01:46 | -333,18 | 74,46 |
| 4 | 315\_A:270\_A | 01:46 | -345,55 | 76,63 |
| 5 | 315\_A:270\_A | 01:46 | -335,34 | 74,46 |

In the table, the greener the cell, the higher the similarity of between the solution and the model. The dope score was not completely predictive. The DOPE score and GDT\_TS score rank the models in different order from most to least similar to the solution. So when doing a real simulation (where the solution structure is not available to calculate the GDT\_TS score), one should keep in mind that the DOPE score is not completely predictive for the actual similarity between the model and the solution.

### Question 4

This answer is based on the DISTANCE scores in the GDT\_TS\_4.txt file. Atoms 22 till 33 have a low DISTANCE score and therefore are modelled good. This range is shown between the two blue lines. Atoms 34 till 46 have a high DISTANCE score, which is shown between the two red lines.



# Structural comparison

### Question 5

The solution structure, 5j4a, is found with HHpredModel. The model, made with Modeller, is compared with 5j4a by using Chimera. This comparison shows a RMSD of 1.840 Å and a Q-score of 0.048. When only the template (3ZIH\_A) is compared to 5j4a, the RMSD is 2.286 Å and the Q-score is 0.229.

The RMSD is the average distance between atoms of aligned residues. Therefore, a structure with a lower RMSD is closer to the solution structure. The RMSD of the model is lower than the RMSD of the template protein. Therefore, the model is closer to the solution than when only the template is used. This is also confirmed by the Q-score, where a higher value indicates similarity. Therefore, it can be concluded that Modeller improves the prediction.

The RMSD is very sensitive for outliers and therefore not a good measure for quality. Therefore, the GDT\_TS scores should also be considered. However, a GDT-TS score for template and the solution structure can not be determined because there is no similarity in sequence.

# Pairwise sequence alignment

### Question 6

The pairwise alignment between the target (T0868) and the template (3zih) sequences is created with the Water EMBOSS pairwise sequence alignment (<https://www.ebi.ac.uk/Tools/psa/emboss_water/>). This is a local alignment method. We used all default settings but set the output to ‘pearson/fasta’.

We choose for a local alignment method because this enhances the probability to get matches between the sequences. We expected that not a large set of the target and template sequences would be similar. Therefore, we think it is better to look at smaller ranges of sequences to match. Instead of using global alignment and taking the risk of not get any matching ranges between the two sequences.

After running MODELLER with the newly made alignment, model 1 was selected as best model with a DOPE score of -1860.33423. Only this model was compared to the solution (5j4a) and a GDT\_TS of 33,721% was calculated.

### Question 7

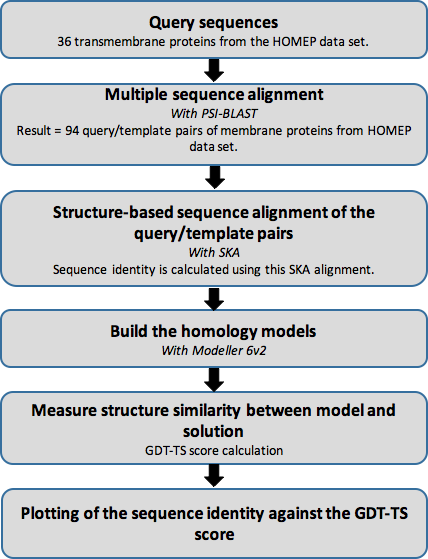
Model with HHpred gives a GDT\_TS of 76,63 %

Model with PSA gives a GDT\_TS of 33,721 %

HHpred is based on the pairwise comparison of profile hidden Markov models (HMMs). Whereas most conventional sequence search methods search sequence databases such as UniProt or the NR, HHpred searches alignment databases, like Pfam or SMART. Also, profile to profile comparison contains more information about other proteins. This all corresponds with our GDT\_TS scores, which state that HHpred creates a better alignment to create a more accurate model than creating an alignment with PSA.

# Paper by Forrest et al. (2006)

### Question 8



### Question 9

The transmembrane regions in the proteins are more accurately aligned and modelled, because the AL0 and the GDT\_TS score are higher than for the whole protein. The whole protein consists of the transmembrane regions and extramembrane regions. This result is expected because transmembrane regions are often conserved. This means that creating a high quality alignment for the TM regions is easier than creating a high quality alignment for the whole protein. And a high quality alignment leads to a higher quality of the model. Transmembrane proteins are often receptors or channels. The extramembrane regions are then binding sites for ligand or downstream proteins. These regions are therefore highly variable in sequence and length. This diversity makes the extramembrane regions harder to model. Furthermore, the extramembrane regions are often loops. Loops are hard to model.

# Contributions

### Question 10

|  |  |
| --- | --- |
| Question | Contributed |
| 1 | Lianne, Anna |
| 2 | Anna |
| 3 | Lianne |
| 4 | Lianne |
| 5 | Anna |
| 6 | Lianne |
| 7 | Lianne |
| 8 | Anna |
| 9 | Anna |

All answers are explained to each other and discussed