Structural Bioinformatics
Assignment 2 - Homology Modeling

Sanne Abeln, Maurits Dijkstra and Juami van Gils

February 12, 2021

Introduction

The aim of this practical is to predict the 3D structure of a protein based on its sequence. For this purpose, you will choose one of two target sequences, choose a template structure, perform sequence alignment between the template and target sequence, and build a model of the target structure. Additionally, you will test the effect of the sequence alignment between the template and the target on the quality of the final model.

Assignment

You will do this assignment in groups of two. Please hand in one report per group. You only have to answer the indicated questions. There are some additional questions in the text to help you think about the assignment, but these will not be graded. In the last question, you have to state how each of you has contributed to the assignment.

Selecting a target and sequence alignment

CASP is a bi-annual contest to predict a 3D protein structure from sequence alone. Sequences of proteins for which the structure has been solved, but not yet published, are provided and researchers can submit models of what their method predict the structure will look like. These are compared to the actual 3D structures to determine who came closest to the true structure.

In this assignment, you can choose between two targets of the CASP12 contest (`http://predictioncenter.org/casp12/index.cgi`):

- T0868
- T0882

Create a folder named "HHpredModel". This folder should contain all files related to the model you build using a sequence alignment from HHPred.

Use the HHpred server to find a template for your target sequence. Use the default HHpred parameters but changing the "MSA generation method" command into PSI-BLAST=>nr70. You are not allowed to choose any templates that were deposited after May 2016.

(Why would it be unfair to choose a template deposited after May 2016? Which characteristics does a good template have?)

Convert the alignment to PIR format. The PIR format is an alignment format consisting of three parts. The first line specifies the protein code; this is generally the PDB identifier and the PDB chain, separated by an underscore, e.g. "1fwx_A". The second line contains the PDB ID and specifies which part of the PDB structure to use for the model. The remaining lines consist of the sequence and must end with "*". For more information on the PIR format, see `https://salilab.org/modeller/manual/node501.html`. Note that the positions indicated in the PIR file must match the residue numbers in the PDB file.

It is possible to obtain a PIR format alignment directly from HHpred. In order to do this, select the alignment with the protein you would like to use as a template and click "Model using selection". This should give you a largely correct PIR formatted file. Sometimes HHpred does not get the format entirely correct so you may need to make edits, particularly to the amino acid ranges specified on the second line.

Creating the model

Use MODELLER (one of the most popular homology modelling programs, `https://salilab.org/modeller/`) to create models from the alignment with your template. You can use the example script that is available on Canvas; MODELLER itself is already installed on the VU servers. Before trying run your script, activate the structbio environment using the following command: "`conda activate structbio`". Note that there is some randomness in the way

MODELLER fits the models, so it may be preferable to let it build multiple models and choose the best one.

Question 1 [15 points]

Please hand in the following files in a single folder named "HHpredModel":

- PIR alignment file
- MODELLER script (build_model.py)
- Models (DOPE scores)

Specify the target you selected (T0882 or T0868). Explain the strategy you used to select a template. Indicate how you modified the ".pir" alignment and the "build_model.py" script. Choose a suitable model from your MODELLER results and explain why you chose that one.

Question 2 [10 points]

What does MODELLER do with any regions that are gaps in the alignment in your template sequence? You will need to inspect your model in a protein structure viewer such as UCSF Chimera to answer this question. See the "Structural comparison" section below for instructions on how to run Chimera. (If your alignment does not contain any long gaps, you can introduce them at one of the termini, and see what happens.)

Scoring your model

Calculate GDT_TS scores between your model and the solution structure. You can use the LGA program (`http://proteinmodel.org/AS2TS/LGA/lga.html`). You will need to use the parameters

"`-3 -o2 -gdc -lga_m -stral -aa1:begin:end -aa2:begin:end`".

The "`-aa1:begin:end`" and "`-aa2:begin:end`" parameters should give the ranges that you know are corresponding between the reference structure and your model structure. See `http://predictioncenter.org/local/lga/lga_description.html` for a comprehensive description of what all possible parameters mean.

If you are having trouble determining the correct ranges it is also possible to use the "`-4`" option to have LGA determine them for you; you can then get them from its output and subsequently input them into the "`-3`" command above.

Question 3 [5 points]

State your GDT_TS score and discuss the results. What does your GDT score indicate?

For all the models you get from MODELLER, compare the DOPE scores to the GDT_TS scores. Was the DOPE score predictive?
Please hand in the LGA output by adding it to the folder “HHpredModel”.

Question 4 [15 points]

By carefully inspecting the LGA output of your selected model state which regions are accurately modelled and which are poorly modelled. Give a screenshot from a structure viewer showing a poorly and an accurately modelled region. Give a brief explanation, why the specific regions are modelled accurately/poorly.

Structural comparison

Use Chimera to compare the template structure to your model and to compare the solution structure to your model.

You need to load both the reference structure and your model. You can do this in the “File” menu using the “open” command. Note that it’s possible for a PDB file to contain multiple chains. In this case it’s recommended to only keep the chain you are modelling. You can do this with:

1) “Select” → “Chain” → “<letter of the chain you want to keep>”

2) “Select” → “Invert (all models)”

3) “Actions” → “Atoms/Bonds” → “delete”

This will throw away any data not related to the chain your built your homology model for. You can now tell Chimera to superimpose your model on top of the reference structure with “Tools” → “Structure comparison” → “MatchMaker”. Make sure that you have the correct structures selected in both the “Reference structure” and “Structure(s) to match” lists. All other options can be left at their default values. There are a lot of other tools in the “Structure comparison” submenu that may be helpful. This includes the “Match → Align” tool which will let you calculate the RMSDs.

Question 5 [10 points]

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?

Pairwise sequence alignment

Create a new folder called “PairwiseModel”. Create a pairwise alignment between the target and template sequence that is based on sequence alone (so not using a profile based method). You can use a server at the EBI `https:`

`//www.ebi.ac.uk/Tools/psa/`. Create and score a model based on the pairwise alignment using the same steps as before. The amino acids in your alignment need to be present in the ATOM record of the PDB file, otherwise MODELLER can not model the 3D structure of your target. In this case, you need to correctly modify the PIR alignment.

Question 6 [15 points]

Please give the name of the alignment program you used, and any options you chose. Give a short justification for using local or global alignment. Also state the GDT_TS score of your model.

Please hand in the following files in a single folder "PairwiseModel":

- PIR alignment file
- MODELLER script
- Models + DOPE scores
- LGA output (model vs. solution)

Question 7 [10 points]

Now compare your model built from the pairwise sequence alignment with the model built from the HHpred alignment. Clearly state which model you think is most accurate and how you are able to observe this. Give a rationale for your findings.

Paper by Forrest et al. (2006)

Forrest et al. (2006) compare models for transmembrane proteins, using an approach similar to the method you have just used.

Question 8 [10 points]

Describe in a flow diagram all the steps needed to obtain the results given in Figure 1B.

Question 9 [10 points]

Consider Figure 2B. What is most accurately aligned and modelled: the TM regions or the regions outside the TM? Does this match your expectations? Briefly explain how these results may be rationalised.

Contributions

Question 10

Please indicate who you worked with, and how each of you contributed.