Structural Bioinformatics
Assignment 3 - Structure prediction

Sanne Abeln, Maurits Dijkstra, Juami van Gils,
Dea Gogishvili, Katharina Waury

March 12, 2023

Introduction

The aim of this practical is to predict the 3D structure of a protein based on its sequence, and to evaluate the similarity of the resulting model to the true structure. For this purpose, you will work with two target sequences from the CASP14 competition. You will choose template structures, perform sequence alignment between the template and target sequence, and build homology models of the target structures. You will then evaluate the performance of this approach. Additionally, you will evaluate the 3D structures predicted by AlphaFold for these sequences and compare them to your models.

Assignment

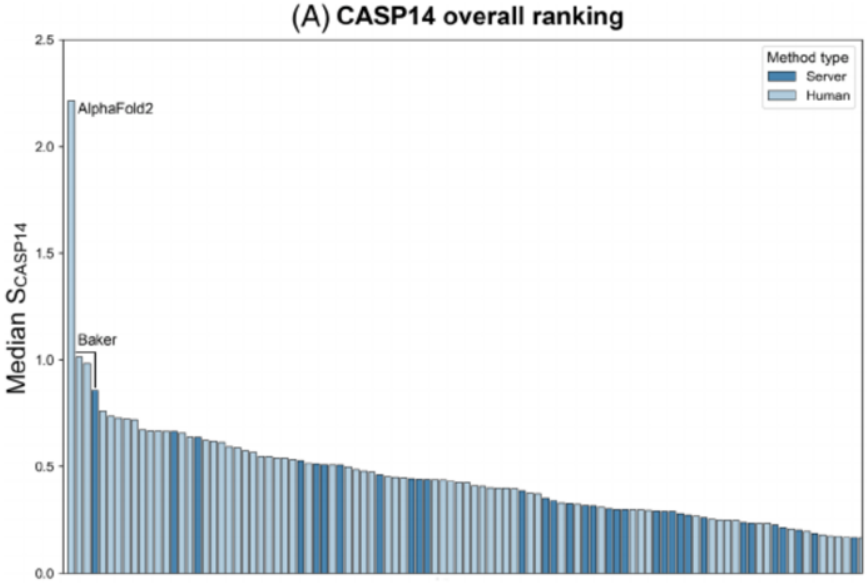
You will do this assignment in groups of two. Please hand in one report per group. You only have to answer the indicated questions. There are some additional questions in the text to help you think about the assignment, but these will not be graded. In the last question, you have to state how each of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

Note that you are also required to hand in the files you generate during this assignment. Specifications can be found in the assignments below. Please also consider the instructions on how to hand in your files on the Canvas assignment pages, as this helps us to grade your work most efficiently.

CASP14 and targets

Critical **A**ssessment of protein **S**tructure **P**rediction (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. After submission, these are compared to the actual experimentally solved 3D structures to determine who came closest to the true structure by the CASP team. During CASP14 in 2020, AlphaFold demonstrated exceptional results and largely outperformed all other groups in both easy and especially in difficult target categories (see Figure 1). AlphaFold is believed to come closest to solving the most difficult biological task of protein structure prediction. In this practical, you will experience the process of protein 3D structure prediction and compare your work to the results of the most famous and hyped program in the field of structural biology. You will also step into the shoes of the CASP14 organizers, and score the performance of both methods.

Figure 1: Results of CASP14. AlphaFold2 was much more accurate than any other competitor. Source: Pereira et al. (2021), `https://doi.org/10.1002/prot.26171`



In this assignment, we have chosen two targets for you from the CASP14 contest (`https://predictioncenter.org/casp14/targetlist.cgi`):

- T1026
- T1064

T1026 is considered an easy prediction target, while T1064 is a difficult target to predict (What does it depend on if a target is considered easy or hard to predict?).

Create two folders named "T1026" and "T1064". Please read once through the entire assignment before beginning your work.

First, we ask you to follow the described workflow below for T1026 and to save all your results into its corresponding folder. Please answer Question 1 - 5 with regards to T1026 and its models. Repeat the analysis for T1064 and save all your results into its corresponding folder. Please answer Questions 6 - 8 with regards to T1064 and its models. Questions 9 - 10 should be answered generally.

Sequence alignment

Use the HHpred server (`https://toolkit.tuebingen.mpg.de/tools/hhpred`) with the default HHpred parameters but change the "MSA generation method" command to PSI-BLAST=>nr70. Find a suitable template for your target sequence. You are not allowed to choose any templates that were released after May 18, 2020 (Why would it be unfair to choose a template released after this date? 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 homology 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 to 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.

Scoring the models

Calculate GDT_TS scores between your homology 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://proteinmodel.org/AS2TS/LGA/lga_format.html` for a comprehensive description of what all possible parameters mean. In case you need to specify a gap in one of the structures, you can find which additional parameters to use here.

If you are having trouble determining the correct ranges, try to run LGA with all parameters listed above except for the range specifications. Looking at the output will give you a further idea of how to specifiy the ranges correctly. When selecting an input for the LGA, pay attention that the chain you are indicating corresponds to the HHpred results.

The CASP website provides the created models of all participants. You can find 5 AlphaFold models for each target in the CASP14 data archive, in the folders containing the models as originally submitted by the CASP competitors: `https://predictioncenter.org/download_area/CASP14/predictions/regular/` Use the first predicted protein structure of AlphaFold for this target and calculate the GDT_TS scores between the AlphaFold model and the solution structure in the same way as described above.

Structural comparison

Use Chimera to compare the solution structure to your models.

You need to load both the solution structure and your homology and AlphaFold models. 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/Structure

analysis" → "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 1 [15 points]

Please hand in the following files, with "T1026" included in each file/folder name.

- PIR alignment file
- MODELLER script (build_model.py)
- The homology models you created for this target, in a zipped folder

Please paste a screenshot of your pir alignment file in your report as well. Explain the strategy you used to select a template. Indicate how you modified the ".pir" alignment (in case necessary) 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]

State the GDT_TS scores of your homology model and the AlphaFold model compared to the solution structure and discuss the results.Explain what GDT_TS scores indicate. Clearly state which model you think is most accurate and how you are able to observe this. Give a rationale for your findings.
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 outputs for the homology and AlphaFold models,specifying "T1026"in the filenames. '.

Question 3 [15 points]

By carefully inspecting the LGA outputs state which regions are accurately modelled and which are poorly modelled. Do the MODELLER and AlphaFold models show the same regions as poorly modelled or do you notice differences? 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.

Question 4 [10 points] (max 150 words, please include your word count)

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 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.)

Question 5 [10 points] (max 200 words, please include your word count)

Do you observe any flanking stretch of residues at the terminus of the AlphaFold model? If so, why is this not included in the solved protein structure?

Now proceed with T1064, a target classified as 'hard' by the CASP14 organizers.

Question 6 [5 points]

Please hand in the following files, including "T1064" in all filenames.

- PIR alignment file
- MODELLER script (build_model.py)
- THe homology models you created

Please also paste a screenshot of your PIR alignment file in your report. Explain the strategy you used to select a template. Indicate how you modified the ".pir" alignment (in case necessary) and the "build_model.py" script. Choose a suitable model from your MODELLER results and explain why you chose that one.

Question 7 [10 points]

Both the MODELLER and the AlphaFold model perform less well on the hard target compared to the easy target. Discuss for each method what explains the drop in accuracy.
Please hand in your LGA output for the Modeller and AlphaFold models.

Question 8 [10 points]

For this target (T1064) use all five structures predicted by AlphaFold during CASP14. State the GDT_TS scores of all five AlphaFold models compared to the solution structure and discuss the results. Do you consider the AlphaFold predictions for this target stable? Discuss how your findings would influence your interpretation of an AlphaFold structure model in a real-life application (i.e. when you do not have access to the solution structure to compare)?

Question 9 [10 points]

How is AlphaFold able to outperform Modeller for the hard target, while for the easier target, they perform similarly? Include in your discussion the distances between the predicted structures, template structure(s), and solution structure.

Contributions

Question 10 [5 points]

Please state your contributions to this project per person using the taxonomy outlined in `https://journals.plos.org/ploscompbiol/s/authorship#loc-author-contributions`