Algorithms in Structural Bioinformatics
Homology Modeling Practical

TME-4

Exercice - 1 MODELLER

You will use the programm *MODELLER* (`http://toolkit.tuebingen.mpg.de/modeller`) to model the structure of the mystery protein which sequence is written in the file *protInconnue.fasta*.

1- Perform a *BLAST* search (`http://www.ncbi.nlm.nih.gov/blast/`) with the mystery protein's sequence as the query. The objective is to find proteins with known structures sharing high sequence identity with the query protein.

a- Which database do you have to choose?
b- What are the first hits and what percentage of identity do they share with the query protein?
c- You should have identified, among others, the proteins with PDB codes 1IX5 and 2K8I. Download the corresponding coordinate PDB files and sequence fasta files from the PDB website (`http://www.pdb.org`)

2- Use the program *Stretcher* (from the Mobyle portal `http://mobyle.pasteur.fr/`) to perform pairwise sequence alignments between the mystery protein and 1IX5 and 2K8I respectively. Choose fasta format for the outputs and save them.

3- Go to the *MODELLER* website and copy-paste one of the alignment: the first sequence should correspond to the mystery protein and the second sequence to one of the proteins with known structures. Indicate only the PDB code of the protein in the title line. Enter the code "MODELIRANJE" in the field "Please insert your MODELLER-key", and submit your request. Perform the two model buildings in two separate windows.

a- Download the coordinate PDB files corresponding to the predictions.
b- The three last tabs report the evaluation of the quality of the models. Compare the results from *Anolea* for the two predictions.
c- Submit the two downloaded PDB files to another quality evaluation method: *ProsaWeb* (`https://prosa.services.came.sbg.ac.at/prosa.php`). Analyse all these results.

Exercice - 2 Copenhagen Models and HHpred

You will make a homology model of the H7N7 hemagglutinin (HA) or neuraminidase (NA).

1- Download a (recent) H7N7 HA or NA protein sequence you would like to work with from the influenza sequence database (`http://flu.lanl.gov/`).

a- Go to the "Influenza Virus Resource" on the front page.
b- Select "Database" from the top menu bar.

c- Select a H7 or N7 sequence from influenza A. Make sure you get a full-length sequence.
d- Make a note of the strain (e.g. A/Bar-headed Goose/Qinghai/12/05).

2- Submit the sequence to the *PSIPRED* secondary structure prediction server (`http://bioinf.cs.ucl.ac.uk/psipred/`). You will need the secondary structure prediction later on and the prediction can take up to 20 minutes to run. To run this job, you will need to supply an email address to which the results will be sent.

3- You will try two different methods of creating a homology model: *CPHmodels* and *HHpred*.

CPHmodels (`http://www.cbs.dtu.dk/services/CPHmodels/`) was created by Ole Lund *et al.* from CBS. It is very simple to use and very fast. It compares well with other servers in producing accurate models for high homology problems, but with regions of low homology, it is quite conservative and only models the part it is most confident about. *CPHmodels* will create a model based on the template it finds as the best, and it is not possible to force it to use another template.

HHpred (`http://toolkit.tuebingen.mpg.de/hhpred`) is one of the servers which performed best at the latest CASP experiment. It allows to modify the sequence alignment by hand, if necessary. It is also possible to select which PDB entry you want to use as a template for your structure. This gives you the possibility for deselecting "bad" structures as templates. To use *HHpred* you will need the license key for the modeling program *MODELLER*. The license key is "MODELIRANJE".

a- Create two homology models of your protein (one with *CPHmodels* and one with *HHpred*) using the HA or NA sequence downloaded from the influenza sequence database.
b- If you are modeling HA, you may discover that you have a problem with the length of the model. Try to compare your model with the template it was built on in *PyMOL*. What is going on? How will you solve this problem?
c- In *HHpred*, try to play around with the different possibilities for selecting different templates. Note that it is also possible to base your model upon multiple templates. What advantage does that bring?
d- Inspect the two models visually in *PyMOL*. Try to superimpose them with the templates they were built on and also with each other. To align two structures, you can use the following command in *PyMOL*:

```
align structure1, structure2
```

Another possibility is to use the *CE* server (`http://cl.sdsc.edu/`) which enables to compare structures from the PDB or uploaded by the user. It produces an RMSD value and a PDB file with the superimposed structures. To inspect the results from *CE* with *Pymol*, select "show all states" under the "movie" menu.
e- Were the two models built on different templates? If so, which of the templates do you consider to be of the best quality?
f- Do the models cover the same part of the input sequence? If the models differ, are the differences in important or in not so important parts of the structure?
g- Try to check the model quality using the *ProQ* web server (`http://www.sbc.su.se/~bjornw/ProQ/ProQ.html`). *ProQ* works best if you also supply a secondary structure prediction. Simply copy the line with the secondary structure prediction corresponding to the sequence covered by your model into the field in the *ProQ* server. Be aware that the two models may not cover exactly the same part of the sequence, so you may need slightly different regions of the secondary structure prediction for the two models. Which of the two models scores best in *ProQ*?