Exercise 5: Correcting protein models using statistical potentials and secondary structure prediction.

We propose the following problem: we have obtained a structure model that has a wrongly modeled region. Only this region is wrong, the rest of the structure is correctly modeled. In this tutorial we will learn to identify wrongly modeled regions and some strategies to correct those. To do so we will make use of statistical potentials and programs for secondary structure prediction and identification. It is important to take into account that another way to correct a wrong model is to use a different template.

Theoretical concepts:

Comparative modeling and statistical potentials: see exercise 4.

Secondary structure prediction: it is known that some amino acids are prone to be located in specific types of secondary structures, such as helices or sheets. We can use this knowledge to evaluate the assignation of the secondary structure in a protein model. If we find that some regions of the protein have a secondary structure that differs from the prediction is probable that this region has been wrongly modeled.

Tutorial:

Step 0: Get inside the cluster and load the modules

You have to get inside the cluster and then get inside a computation node. For this tutorial you have to load the modules for modeller, perl, dssp and psipred. To load the modules use the commands “module av” and “module load”.

Step 1: compare the predicted secondary structure with the modeled one using PSIPRED and DSSP

Move to the “secondary_structure” directory in your working directory. Here you have the same two PDB files as before.

Now we will compare the secondary structure of the model with the predicted one. If we find a strong discrepancy between both it will mean that the error that we are looking for involves a wrong modeling of the secondary structure. To do so we will use two programs: PSIPRED and DSSP.

PSIPRED is a program to predict the secondary structure of one protein. It takes as input one protein sequence in fasta format and returns a prediction of the secondary structure for that sequence.

DSSP is a program that extracts the secondary structure from a PDB file. It takes as input a PDB file and returns, for each amino acid, in which type of secondary structure is placed.

Follow the next instructions:

1. Extract the sequence of the model with PDBtoSplitChain:

```
perl ~/Documents/perl_scripts/PDBtoSplitChain.pl -i model.pdb -o model
```

2. Execute PSIPRED with the model sequence (in the citrix, I'm sorry):

```
runpsipred_single  model.fa
```

This will return two files:

- model.ss2 → contains the assignation of the predicted secondary structure for each of the residues in the model sequence.
- model.horiz → contains the scores associated to the predictions done for each residue. The highest the score the most reliable the prediction.

3. Transform the model.ss2 file into a pir alignment containing the model sequence and its secondary structure:

```
perl  ~/Documents/perl-scripts/psipred.pl  model.ss2  >  psipred.pir
```

4. Extract the secondary structure from the model using DSSP:

```
dssp  model.pdb  model.dssp
```

Some of you (mostly MacOS users) will have to use a slightly different command:

```
mkdssp  model.pdb  model.dssp
```

5. Transform the DSSP output into a PIR alignment between the model sequence and its secondary structure.

```
perl  ~/Documents/perl_scripts/aliss.pl  model.dssp  >  dssp.pir
```

6. Concatenate both PIR alignments in a single file:

```
cat  psipred.pir  dssp.pir  >  compare.pir
```

7. Transform the concatenated PIR alignment into a clustalw alignment:

```
perl  ~/Documents/perl_scripts/aconvertMod2.pl  –in  p  –out  c
<compare.pir>compare.clu
```

Now we have an alignment in which the sequence of the model, its secondary structure and the predicted secondary structure are aligned. Therefore, we can look at the regions of the alignment in which the predicted secondary structure differs from the actual one. Secondary structure types are indicated by the following letters:

- Helix: H, G, I
- Strand: B, E
- Loop (or coil): C, T, S, and any other unmentioned letter.

Not all predictions made by PSIPRED are equally reliable. We should identify the regions in which the predictions are more reliable. This can be done looking at the model.horiz file obtained previously. We will only consider reliable regions those with a score of 9. Then, we will search for these reliable regions in the alignment containing the PSIPRED and the DSSP information. If we find any region that has a PSIPRED score of 9 and in which PSIPRED and DSSP differ, that region will be surely wrongly modeled. In the next image you can see the model.horiz file. The highlighted region has been determined as a helix with a confidence of 9.

```
# PSIPRED HFORMAT (PSIPRED V3.4)

Conf: 987066854121421003566776534643413343247897438999434888999775
Pred: CCCCEEEEECCCCCCCCCCCCCCCCEEECCCCHHHCCCCCCCEEEEEEECCCCCCCCCC
  AA: MNGEIRLIPYVTNEQIMDVNELPEGIKVIKAPEMWAKGVKGKNIKVAVLDTGCDTSHPDL
              10        20        30        40        50        60

Conf: 453206776557999963234457887651465476518998622016766578998504
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEECCCCCCCCCCCCCCCCEEEEEECC
  AA: KNQIIGGKNFTDDDGGKEDAISDYNGHGTHVAGTIAANDSNGGIAGVAPEASLLIVKVLG
              70        80        90       100       110       120

Conf: 888882126898130000453122898437999993679999997675567999853888
Pred: CCCCCCCCEEEECCCCCCEEECCCEEEECCCCCCCHHHHHHHHHHHHHCCEEEEEECCCC
  AA: GENGSGQYEWIINGINYAVEQKVDIISMSLGGPSDVPELKEAVKNAVKNGVLVVCAAGNE
             130       140       150       160       170       180

Conf: 898655311555422202666411457656433104782444543785301358875455
Pred: CCCCCCCCCCCCCCCCCCCEEEECCHHHHHHHHHCCCCCCEEEEECCCCCCCCCCCCCCCC
  AA: GDGDERTEELSYPAAYNEVIAVGSVSVARELSEFSNANKEIDLVAPGENILSTLPNKKYG
             190       200       210       220       230       240

Conf: 534544455441226787752346789743046589999998643633563304686799
Pred: CCCCCCCCCCCCCCCHHHHHHCHHHHHHHHHHCCHHHHHHHHHHHCCHHHHHHCCCCCEEE
  AA: KLTGTSMAAPHVSGALALIKSYEEESFQRKLSESEVFAQLIRRTLPLDIAKTLAGNGFLY
             250       260       270       280       290       300

Conf: 8274478998752045579
Pred: ECCCHHHHHHHHCCCCCCC
  AA: LTAPDELAEKAEQSHLLTL
             310
```

In the next image we see the alignment between the PSIPRED and the DSSP results. The highlighted region in the model.horiz file is highlighted too in this alignment. We can see that the highlighted region contains a loop while it has been predicted as a helix with a confidence of 9. This means that there is an error there and that we should correct it.

```
CLUSTAL W(1.60) multiple sequence alignment

model.fa.ss2Seq  MNGEIRLIPYVTNEQIMDVNELPEGIKVIKAPEMWAKGVKGKNIKVAVLDTGCDTSHPDL
model.fa.ss2SS   CCCCEEEEECCCCCCCCCCCCCCCCEEECCCCHHHCCCCCCCEEEEEEECCCCCCCCCC
model.dsspSeq    MNGEIRLIPYVTNEQIMDVNELPEGIKVIKAPEMWAKGVKGKNIKVAVLDTGCDTSHPDL
model.dsspSS     ----S------S-------B---HHHHHTTHHHHHHHT---TT-EEEEEES---TT-TTS

model.fa.ss2Seq  KNQIIGGKNFTDDDGGKEDAISDYNGHGTHVAGTIAANDSNGGIAGVAPEASLLIVKVLG
model.fa.ss2SS   CCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEECCCCCCCCCCCCCCEEEEEECC
model.dsspSeq    KNQIIGGKNFTDDDGGKEDAISDYNGHGTHVAGTIAANDSNGGIAGVAPEASLLIVKVLG
model.dsspSS     -TTT-EEEE-TTSSS----TT--SSSSHHHHHHHHH--SSSSB---SSTTSEEEEEE-S-

model.fa.ss2Seq  GENGSGQYEWIINGINYAVEQKVDIISMSLGGPSDVPELKEAVKNAVKNGVLVVCAAGNE
model.fa.ss2SS   CCCCCCCCEEEECCCCCCEEECCCEEEECCCCCCCHHHHHHHHHHHHHCCEEEEECCCC
model.dsspSeq    GENGSGQYEWIINGINYAVEQKVDIISMSLGGPSDVPELKEAVKNAVKNGVLVVCAAGNE
model.dsspSS     TTTS---HHHHHHHHHHHHHHT-SEEEE---BSS--HHHHHHHHHHHHHT-EEEEE--S-

model.fa.ss2Seq  GDGDERTEELSYPAAYNEVIAVGSVSVARELSEFSNANKEIDLVAPGENILSTLPNKKYG
model.fa.ss2SS   CCCCCCCCCCCCCCCCCCEEEECCHHHHHHHHHCCCCCCEEEEECCCCCCCCCCCCCCC
model.dsspSeq    GDGDERTEELSYPAAYNEVIAVGSVSVARELSEFSNANKEIDLVAPGENILSTLPNKKYG
model.dsspSS     ----TT-----BTTTSTTSEEEEEE-TTS-B-TT---STT--EEEE-SSEEEEETTTEEE

model.fa.ss2Seq  KLTGTSMAAPHVSGALALIKSYEEESFQRKLSESEVFAQLIRRTLPLDIAKTLAGNGFLY
model.fa.ss2SS   CCCCCCCCCCCCCCHHHHHHCHHHHHHHHHCCHHHHHHHHHHHHCCHHHHHCCCCCEEE
model.dsspSeq    KLTGTSMAAPHVSGALALIKSYEEESFQRKLSESEVFAQLIRRTLPLDIAKTLAGNGFLY
model.dsspSS     EE-SHHHHHHHHHHHHHHHHH-SS--S---SS-SS-SHHHH-SS-HHHHS---S-SSSS

model.fa.ss2Seq  LTAPDELAEKAEQSHLLTL
model.fa.ss2SS   ECCCHHHHHHHHCCCCCCC
model.dsspSeq    LTAPDELAEKAEQSHLLTL
model.dsspSS     TT-B--HHHHHS-------
```

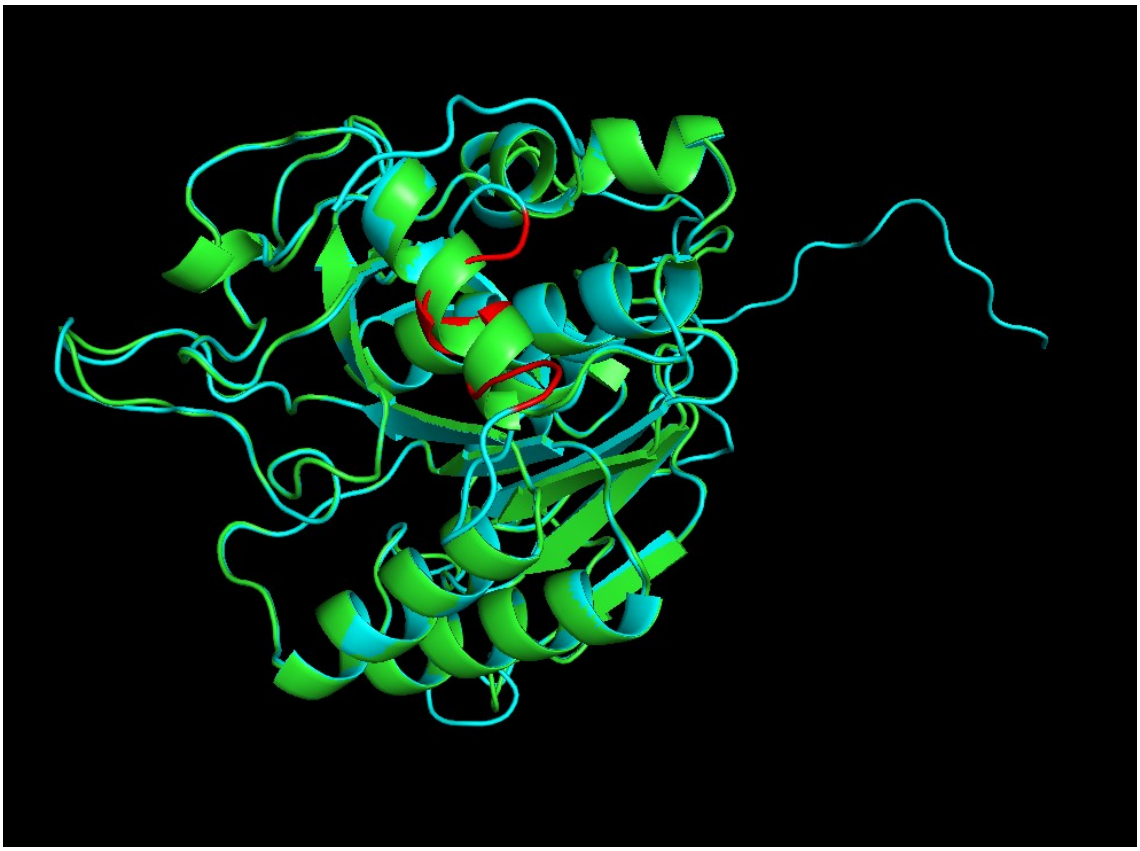
Step 2: inspect the model using pymol

So far, we have detected an error in our model by both PROSA and PSIPRED and DSSP comparison. Now, we will inspect the model using pymol.

Open the model and the template with pymol and superimpose the two structures using the super command as we saw in practical 3:

```
super 1meeA, model_, object=aln
```

Use the sequence display to visualize the structural alignment and identify the wrongly modeled region. You can find it by looking at the regions that didn't aligned properly between the template and the model. Remember that this happens because pymol is very sensitive to structural differences between superimposed structures. If you want to change this, remember that you can use the cutoff command as we saw on practical 3. In the next image you can see the wrongly modeled region highlighted in red. This region should be completely helical:



We see that the model has a loop in the middle of one helix and that the template doesn't. Furthermore, PSIPRED was also identifying a loop that should be modeled as a helix. Finally, loops happening in the middle of helices are unlikely in nature, while they are one of the most common artifacts in protein modeling. This is the modeling error we were looking for, now we are going to fix it.

Step 3: correct the model by modifying the input alignment of MODELLER

We can fix this error by modifying the input PIR alignment between the target and the template. On the regions of the alignment where the target is aligned with a gap of the template, MODELLER introduces a loop. If this gap happens in the middle of a helix or a strand, a loop is placed in there. Therefore, a good strategy to fix the model is to move the gap outside of the helix, to a region where the template also has a loop. It is recommendable to change the alignment as less as possible, to maintain the sequence similarity correspondence established by the alignment program.

In the following image you have the alignment that we have used so far. The highlighted region is the one that is wrongly modeled. You see how we are having two gaps in that region:

```
CLUSTAL 2.1 multiple sequence alignment


P11018          MNGEIRLIPYVTNEQIMDVNELPEGIKVIKAPEMWAKGVKGKNIKVAVLDTGCDTSHPDL
1meeA           ------------------AQSVPYGISQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDL
                                  .:.:* **. **** : ::* .*.*:****:*:* *:*****

P11018          KNQIIGGKNFTDDDGGKEDAISDYNGHGTHVAGTIAANDSNGGIAGVAPEASLLIVKVLG
1meeA           N--VRGGASFVP---SETNPYQDGSSHGTHVAGTIAALNNSIGVLGVAPSASLYAVKVLD
                :  : ** .*.    .: :. .* ..*********** :.. *: ****.***  ****.

P11018          GENGSGQYEWIINGINYAVEQKVDIISMSLGGPSDVPELKEAVKNAVKNGVLVVCAAGNE
1meeA           S-TGSGQYSWIINGIEWAISNNMDVINMSLGGPTGSTALKTVVDKAVSSGIVVAAAAGNE
                . .*****.******::*:.:::*:*.******:. . ** .*.:**..*::*..*****

P11018          GDGDERTEELSYPAAYNEVIAVGSVSVARELSEFSNANKEIDLVAPGENILSTLPNKKYG
1meeA           GSSGS-TSTVGYPAKYPSTIAVGAVNSANQRASFSSAGSELDVMAPGVSIQSTLPGGTYG
                *.... *. :.*** * ..****:*. *.: :.**.*..*:*::*** .* ****. .**

P11018          KLTGTSMAAPHVSGALALIKSYEEESFQRKLSESEVFAQLIRRTLPLDIAKTLAGNGFLY
1meeA           AYNGTSMATPHVAGAAALILSKHPTWTN---------AQVRDR---LESTATYLGSSFYY
                  .*****:***:** *** * .    :         **:  *   *: : *  *..* *

P11018          LTAPDELAEKAEQSHLLTL
1meeA           GKGLINVQAAAQ-------
                 ..  ::   *:
```

In the next image we see the modified alignment in which we have placed the loops outside of the highlighted region (that is a helix):

```
CLUSTAL 2.1 multiple sequence alignment


P11018          MNGEIRLIPYVTNEQIMDVNELPEGIKVIKAPEMWAKGVKGKNIKVAVLDTGCDTSHPDL
1meeA           ------------------AQSVPYGISQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDL
                                  .:.:* **. **** : ::* .*.*:****:*:* *:*****

P11018          KNQIIGGKNFTDDDGGKEDAISDYNGHGTHVAGTIAANDSNGGIAGVAPEASLLIVKVLG
1meeA           N--VRGGASFVP---SETNPYQDGSSHGTHVAGTIAALNNSIGVLGVAPSASLYAVKVLD
                :  : ** .*.    .: :. .* ..*********** :.. *: ****.***  ****.

P11018          GENGSGQYEWIINGINYAVEQKVDIISMSLGGPSDVPELKEAVKNAVKNGVLVVCAAGNE
1meeA           S-TGSGQYSWIINGIEWAISNNMDVINMSLGGPTGSTALKTVVDKAVSSGIVVAAAAGNE
                . .*****.******::*:.:::*:*.******:. . ** .*.:**..*::*..*****

P11018          GDGDERTEELSYPAAYNEVIAVGSVSVARELSEFSNANKEIDLVAPGENILSTLPNKKYG
1meeA           GSSGS-TSTVGYPAKYPSTIAVGAVNSANQRASFSSAGSELDVMAPGVSIQSTLPGGTYG
                *.... *. :.*** * ..****:*. *.: :.**.*..*:*::*** .* ****. .**

P11018          KLTGTSMAAPHVSGALALIKSYEEESFQRKLSESEVFAQLIRRTLPLDIAKTLAGNGFLY
1meeA           AYNGTSMATPHVAGAAALILSKHPT---------WTNAQVRDRLESTATY---LGSSFYY
                  .*****:***:** *** * .             **:  *   *: : *  *..* *

P11018          LTAPDELAEKAEQSHLLTL
1meeA           GKGLINVQAAAQ-------
                 ..  ::   *:
```

After modifying the alignment in clustal format we should transform it to pir format in order to use it as modeller input:

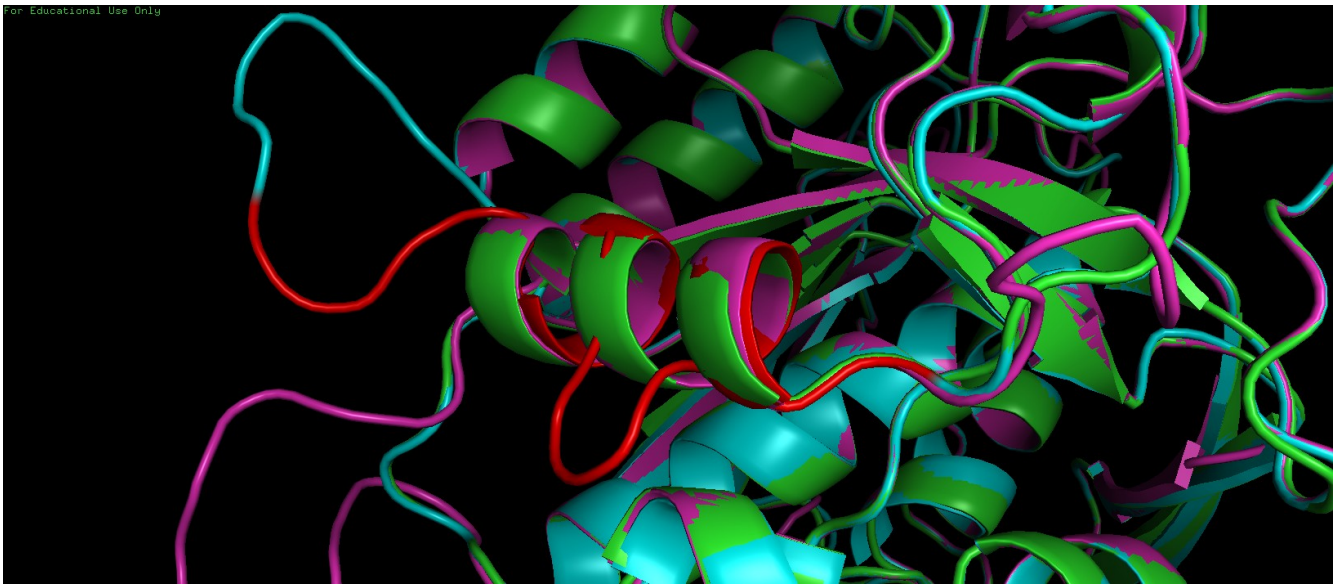
```
perl ~/Documents/perl_scripts/aconvertMod2.pl -in c -out p

<target_template.aln>target_template.pir
```

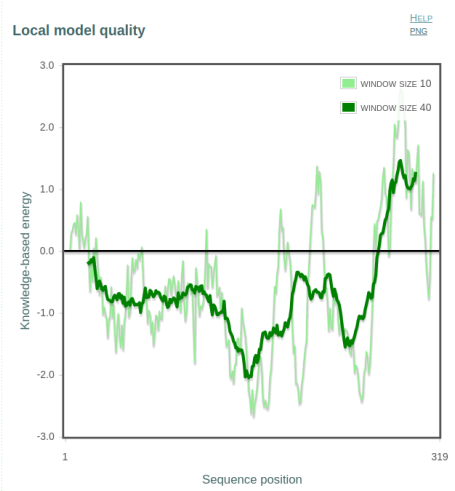
Now, we can execute MODELLER and obtain a new and corrected model:

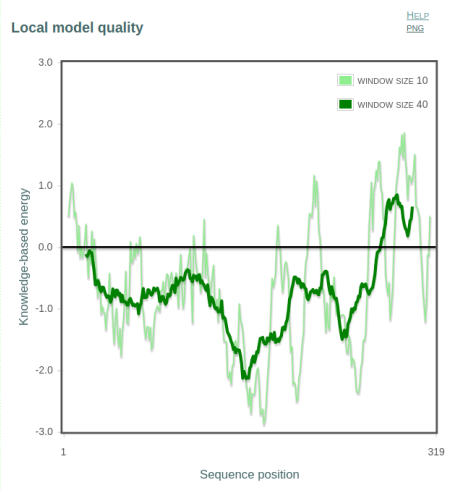
```
mod10.5 modeling.py
```

Rename the corrected model and open it with pymol. In the next image we can see the superimposition of the template (green), our first model (blue) and the corrected model (purple). The corrected helix of the model is highlighted in red. You can see how now the template and the model fit well in the corrected region and how the loops surrounding the helix now are longer.



We can also check that the quality of the new model using prosaweb. Go to prosaweb and open three windows to compare your template, your initial model and your corrected model. You can see that the correction has improved a little bit the energetic profile of the model, however there is still room for improvement.





Step 4: Choose other templates. Use HMMs.

To do this part of the practice you have to load the module for the hmmer package.

Change to the hmm directory. We have explored one way to correct protein models. Another way to improve a model is to choose a different template or using different tools to build the alignment. Since we created the previous model using only blast and clustalw, we could try to obtain a new model but using HMMs instead.

Our strategy will be finding which HMM of PFAM can fit the best our target sequence (with hmmscan). Then, we will use this HMM to search for templates in the PDB (with hmmsearch) and to align the target with the template (with hmmalign). Finally, we can compare the obtained model with the template or with the other models that we have generated so far.

Find the best HMM from PFAM for our target:

```
hmmscan ~/Documents/databases/Pfam-A.hmm P11018.fa
> hmmscan.out
```

Fetch this HMM from PFAM. Check that instead of “hmm_name” you should put the name of the HMM identified by hmmscan.

```
hmmfetch ~/Documents/databases/Pfam-A.hmm “hmm_name”
> peptidase.hmm
```

Search for templates in the PDB:

```
hmmsearch peptidase.hmm ~/Documents/databases/pdb_seq
> peptidase_pdb.out
```

Get the template PDB and sequence:

```
perl ~/Documents/perl_scripts/PDBtoSplitChain.pl -i 1sbh.pdb -o 1sbh
```

Align the sequences of the target and the template:

```
cat P11018.fa 1sbhA.fa > target_template.fa
```

```
hmmalign peptidase.hmm target_template.fa > target_template.sto
```

```
perl ~/Documents/perl_scripts/aconvertMod2.pl -in h -out p
<target_template.sto>target_template.pir
```

Set modeller input files and run modeller:

mod10.5 modeling.py

Check the model with prosa and with pymol.

Questions from the tutorial:

1) Compare the corrected protein model with the one obtained by HMMs. Do they have the same energy profiles? Which is the best model?

2) Do you think that the model obtained with HMMs can be corrected? Is so, how would you do that? Try to carry out your plan to correct this model and check if the quality of the model has improved.

3) In the target-template alignment obtained by HMMs the N and C terminal of the target are not aligned with the template. How is this affecting this to the model? Should we make any change in the model because of this lack of alignment between target and template?