Exam 1

In this exercise we are working with a target protein for which we have its sequence but we don't know its structure or function. This is its sequence:

```
>target
MASSTQGQVITCKAAVAWEANKPMTIEDVQVAPPQAGEVRVKILFTALCHTDHYTWSGKD
PEGLFPCILGHEAAGIVESVGEGVTEVQPGDHVIPCYQAECRECKFCKSGKTNLCGKVRA
ATGVGVMMNDRKSRFSINGKPIYHFMGTSTFSQYTVVHDVSVAKINPQAPLDKVCLLGCG
VSTGLGAVWNTAKVEAGSIVAIFGLGTVGLAVAEGAKSAGASRIIGIDIDSKKFDVAKNF
GVTEFVNPKDHDKPIQQVIVDLTDGGVDYSFECIGNVSVMRSALECCHKGWGTSVIVGVA
ASGQEISTRPFQLVTGRVWKGTAFGGFKSRSQVPWLVEKYLNKEIKVDEYVTHSMNLTDI
NKAFDLLHEGGCLRCVLATDK
```

1) What is the function of this protein?

```
blastp -query target_exam_1.fa -db /shared/databases/blastdat/uniprot_sprot.fasta -out target_uniprot_exam_1.out
```

We find the following hit: A2XAZ3
We go to uniprot and find the function: Alcohol dehydrogenase

2) Get a HMM in the PFAM database that fits your target sequence and name it 2.hmm.

```
hmmscan /shared/databases/pfam-3/Pfam-A.hmm target_exam_1.fa > exam_1_HMM_PFAM.out
```

Here we can see the different models. We pick the first one (ADH_N) and fetch it:

```
hmmfetch /shared/databases/pfam-3/Pfam-A.hmm ADH_N > 2.hmm
```

3) Use the HMM you just obtained to retrieve 4 sequences from the uniprot database. Then, align these 4 sequences and your target using the same HMM you used to get the sequences. Name the alignment 3.aln.

```
hmmsearch 2.hmm /shared/databases/blastdat/uniprot_sprot.fasta > exam_1_seq_uniprot.out
```

A2XAZ3 (target)
P42328
P12311
Q9UYX0
O58389

Retrieve the sequences and upload to the cluster.
Make the alignment:

```
hmmalign 2.hmm seq_retrieved_exam_1.fa > MSA_exam_1.sto
```

Change format:

```
perl /shared/PERL/aconvertMod2.pl -in h -out c < MSA_exam_1.sto > 3.aln
```

4) Discuss briefly your alignment. Does it show sequence conservation or not? Why some regions have capital letters and other regions have lowercase letters?

There is some conservation.
Capital letters are the ones that align with the model.
Lowercase letters are the ones that do NOT align with the model (we should not consider them).

5) Obtain 4 templates for your target. Indicate their corresponding PDB IDs. Should you use the whole PDB entry or just a part of it?

Use UniProt to create an accurate PSSM (unbiased and non-redundant DB):

```
psiblast -query target_exam_1.fa -num_iterations 5 -out_pssm exam_1.pssm -out trash.out -db
/shared/databases/blastdat/uniprot_sprot.fasta
```

exam_1.pssm is the PSSM created from the hits
trash.out is the hits (not templates because they are not structures) obtained from UniProt

Use the PSSM to search for templates in the PDB:

```
psiblast -db /shared/databases/blastdat/pdb_seq -in_pssm exam_1.pssm -out templates.out
```

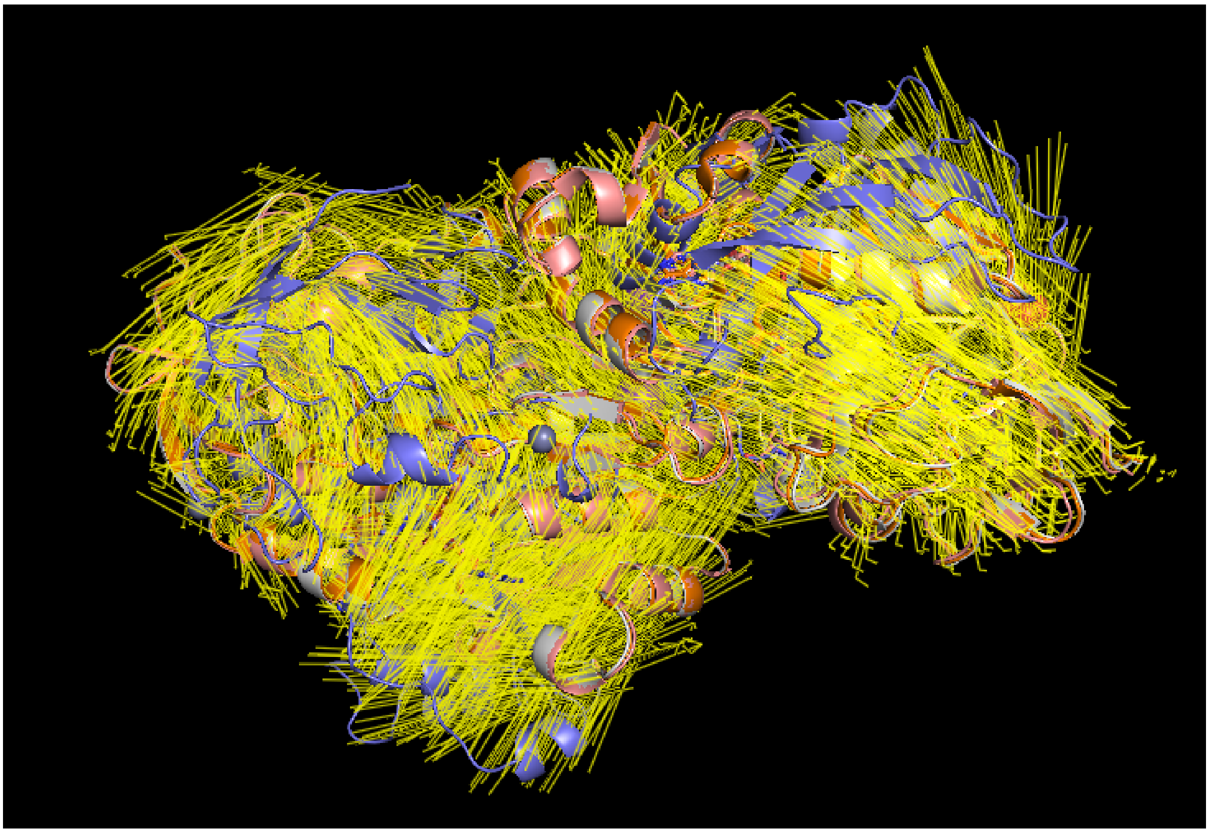
The obtained templates (we should not copy the chain of the PDB ID):
1mc5
2fze
1teh
1mp0

6) Superimpose the 4 templates you obtained in the previous question using pymol. Provide the RMSD value for all the superimpositions and save an image of the superimposition as 6.png.

```
super 2fze, 1mc5, object=aln_exam_1, cutoff=5.0
RMSD =    0.569 (5498 to 5498 atoms)
```

```
super 1teh, 1mc5, object=aln_exam_1, cutoff=5.0
RMSD =   23.491 (5087 to 5087 atoms)
```

```
super 1mp0, 1mc5, object=aln_exam_1, cutoff=5.0
RMSD =    0.641 (5506 to 5506 atoms)
```



7) Generate an structure-based alignment from the superimposition you created in the previous exercise. Name this alignment 7.aln.

We have already done it.
We just need to save it:

```
save 7.aln, aln_exam_1
```

8) Generate a HMM using as input the structure-based aligment you obtained in the previous question and name it 8.hmm.

Transform the format of the alignment, from clustalw to stockholm:

```
perl /shared/PERL/aconvertMod2.pl -in c -out f <7.aln>7.fa

perl shared/PERL/fasta2sto.pl 7.fa > 7.sto

hmmbuild 8.hmm 7.sto
```

9) Make an structural model of the target sequence and name it 9.pdb.

We are going to obtain models for our target protein using a program called modeller. To execute MODELLER you need three files:

- **Target file:** contains the target sequence.

  target_exam_1.fa

- **Alignment file:** contains the alignment between the target and the template/s in PIR format.

  We need to make an alignment with one of the templates obtained in question 5. For example, template “1mc5”.
  We obtain its fasta:

  ```
  perl /shared/PERL/PDBtoSplitChain-pl -i 1mc5.pdb -o 1mc5
  ```

  When doing this, I obtain 2 FASTA files (one for each chain).
  With which chain should I do the alignment?

  ```
  cat target_exam_1.fa > quest_9.fa
  cat 1mc5A.fa >> quest_9.fa
  ```

  Now we make the alignment using the model obtained in question 8.

  ```
  hmmalign 8.hmm quest_9.fa > align.sto
  perl /shared/PERL/aconvertMod2.pl -in h -out c <align.sto>align.aln
  ```

  Change the format to PIR:

  ```
  perl /shared/PERL/aconvertMod2.pl -in c -out p <align.aln>align.pir
  ```

- **Script file (modeling.py)**

Once we have edited the file modeling.py:

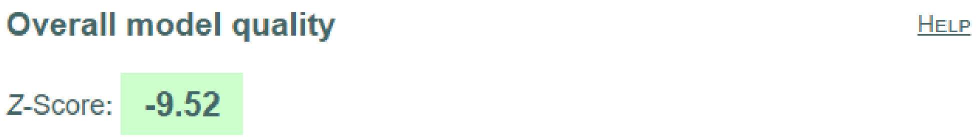
```
module load modeller/10.2
source activate modeller
mod10.2 modeling.py
```

Now we change the name of the model:

```
mv target.B99990001.pdb 9.pdb
```

10) Analyze with prosa the model you generated in the previous question. To perform a reliable analysis, should you compare the analysis of your model with another structure? What structure is this?

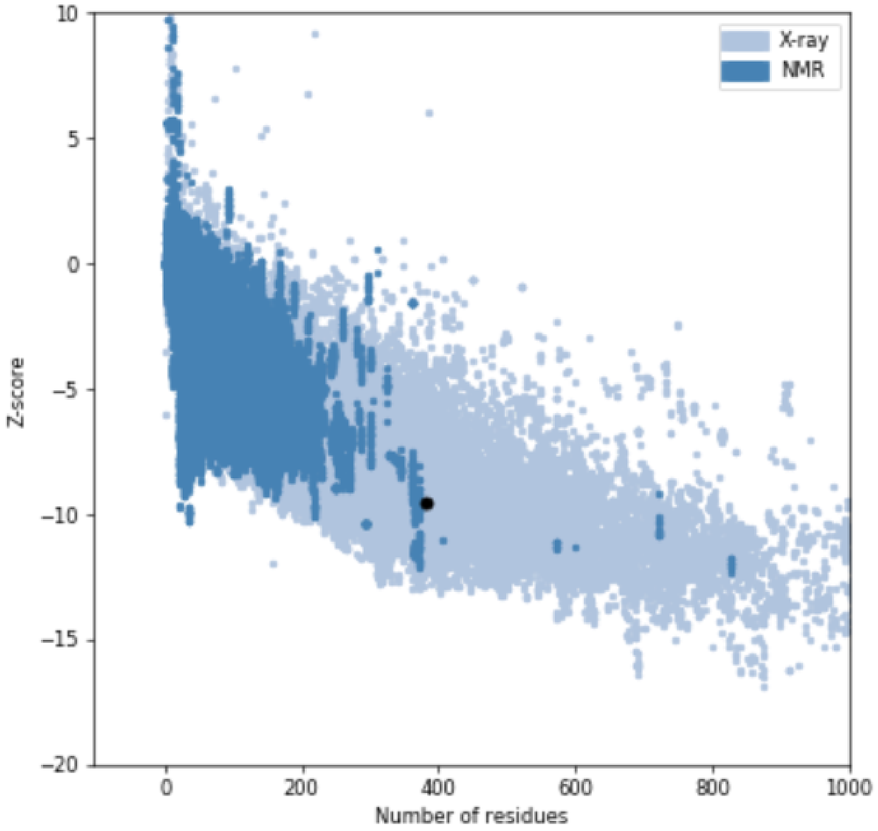
To analyze the model generated, we can check the Z-score and the following plot comparing how good our structure is in comparison with the structures in the PDB.



Overall model quality

HELP

Z-Score: -9.52



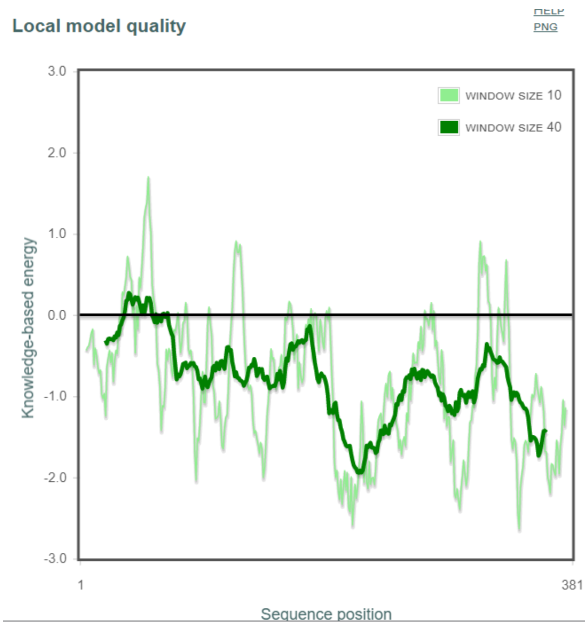
It has a small z-score, which a priori is good. Meaning that our protein model is likely to have similar structural properties to the ones in the reference set and, thus, that our model is accurate and reliable.

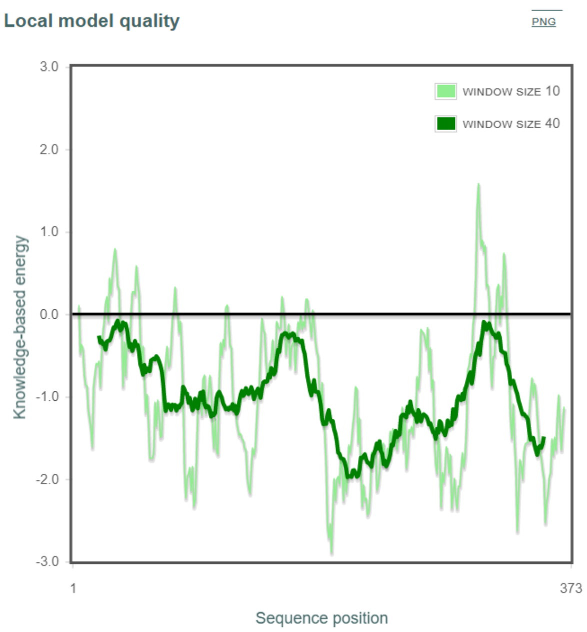
It is important to keep in mind that statistical potentials are relative measurements. This means that we cannot apply a score threshold to discriminate good from bad models. They have been intended to compare models between them or with experimentally determined structures.

A good strategy to assess whether a protein has been correctly modeled is to compare the statistical potentials energies between the model and its template. We expect the template to have a good energetic profile, since it is an experimentally determined structure. Therefore, we can use the template as a reference. If the model has higher energies than the template, it is very likely that it has been wrongly modeled.

Our template is 1mc5A.pdb
We load it to PROSA and we obtain a Z-score of -10.49.
Thus, our model is pretty good since it has a similar score.

We can now look at the energetic profile of the model (left) and the template (right). This profile shows the statistical potentials scores across the amino acid sequence.





As we can see, our model does not have peaks of high energy, meaning that all regions are well modeled.
Note that the lowest the energy, the most stable is the structure.

Exam 2

In this exercise we are working with a target protein for which we have its sequence but we don't know its structure or function. This is its sequence:

```
>target
MFRKSVRPLVGAALVMAAAVAAGPTPASAAESEFYVNPDMASAQWVRDNPQDSRADVIA
NRVADVPQGTWFTSYNPNEVQGQVDALVGSAESAGKTPIMVVYNIPNRDCSNHSGGGAP
DHSSYRDWVDQVAAGLNGRPATIVVEPDVLSLMDSCMDQSQQNHVMDSIAYAGKTLMAG
SSQARVYFDAGHSGWHSPGEIASRLNGADIANSAHGIATNTSNYNWTDDEVSYTRQIIDAT
GHSGLRAVVDTSRNGNGPQGSEWCDPEGRAIGTESTTNTGSSHVDAFLWVKLPGEADG
CAAGAGEFVPQLAYDMAVAADPEPDPEPSPDPEPTPDPEPTPEPGEGCEAAYSVANEWS
DGFQAEVTVTAGADLDGWEVAIDFPDGQGIEQAWNAEVSGSGGAYTASDVSHNGSLSAG
ESTGFGFTGTHSGANGEPELTCSAA
```

1) What is the function of this protein?

```
blastp -query target_exam_2.fa -db /shared/databases/blastdat/uniprot_sprot.fasta -out target_uniprot_exam_2.out
```

We find the following hit: P26222
We go to uniprot and find the function: Endoglucanase E-2

2) What is the fold of this protein?

Use UniProt to create an accurate PSSM (unbiased and non-redundant DB):

```
psiblast -query target_exam_2.fa -num_iterations 5 -out_pssm exam_2.pssm -out trash.out -db /shared/databases/blastdat/uniprot_sprot.fasta
```

exam_2.pssm is the PSSM created from the hits
trash.out is the hits (not templates because they are not structures) obtained from UniProt

Use the PSSM to search for templates in the PDB:

```
psiblast -db /shared/databases/blastdat/pdb_seq -in_pssm exam_2.pssm -out templates_2.out
```

The obtained templates (we should not copy the chain of the PDB ID):
2bod

Go to SCOP DB and look for the fold:
7 stranded beta/alpha barrel

3) Obtain a HMM for this protein in the PFAM database. Name this HMM 3.hmm.

```
hmmscan /shared/databases/pfam-3/Pfam-A.hmm target_exam_2.fa > exam_2_HMM_PFAM.out
```

Here we can see the different models. We pick the first one (Glyco_hydro_6) and fetch it:

```
hmmfetch /shared/databases/pfam-3/Pfam-A.hmm Glyco_hydro_6 > 3.hmm
```

4) Use the HMM you just obtained to retrieve 4 sequences from the uniprot database. Then, align these 4 sequences and your target using the same HMM you used to get the sequences.
Name the alignment 4.aln.

```
hmmsearch 3.hmm /shared/databases/blastdat/uniprot_sprot.fasta > exam_2_seq_uniprot.out
```

P26222 (target)
B0XWL3
Q4WFK4
A1DJQ7
P49075

Retrieve the sequences and upload to the cluster.
Make the alignment:

```
hmmalign 3.hmm seq_retrieved_exam_2.fa > MSA_exam_2.sto
```

Change format:

```
perl /shared/PERL/aconvertMod2.pl -in h -out c < MSA_exam_2.sto > 4.aln
```

5) Obtain 4 templates for your target. Indicate their corresponding PDB IDs. Should you use the whole PDB entry or just a part of it?
Just take the PDB ID (without the chain) from the file templates_2.out:
2bod
1tml
2bof
2boe

6) Superimpose the 4 templates you obtained in the previous question using pymol. Provide the RMSD value for all the superimpositions and save an image of the superimposition as 6.png.
Download the structures, load them to the cluster, separate chains and obtain the correct chain.

```
super 1tml, 2bod, object=aln_exam_2, cutoff=5.0
```

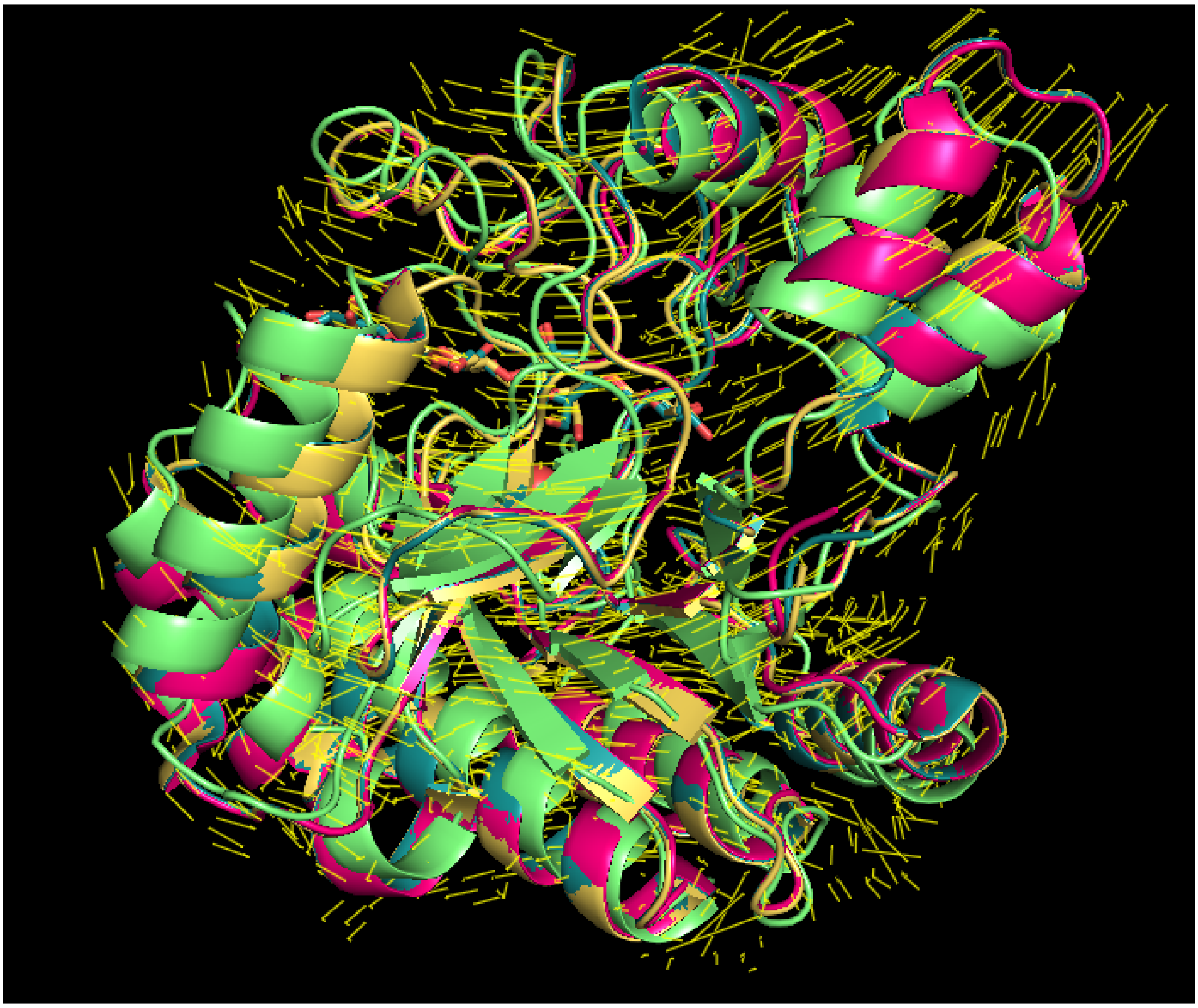
RMSD = 2.413 (2086 to 2086 atoms)

```
super 2bof, 2bod, object=aln_exam_2, cutoff=5.0
```

RMSD = 0.189 (2021 to 2021 atoms)

```
super 2boe, 2bod, object=aln_exam_2, cutoff=5.0
```

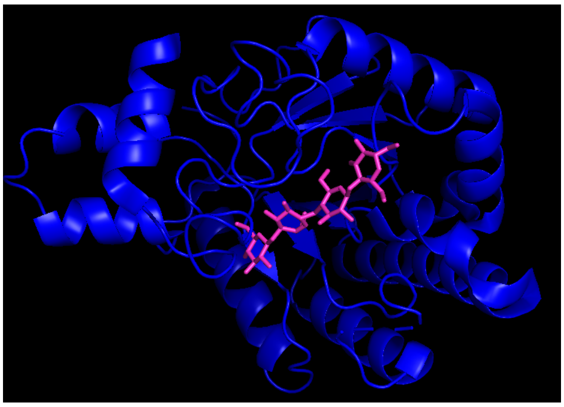
RMSD = 0.342 (2015 to 2015 atoms)



7) Your target is an enzyme. Find a template that contains a substrate molecule in its active site.

Provide an image of the template you chose where the substrate can be seen and name the image 7.png.

By looking at the templates from “templates_2.out”, we can select the first template, 2bod. If we go to the pdb, we can see that it is forming a complex with its substrate.



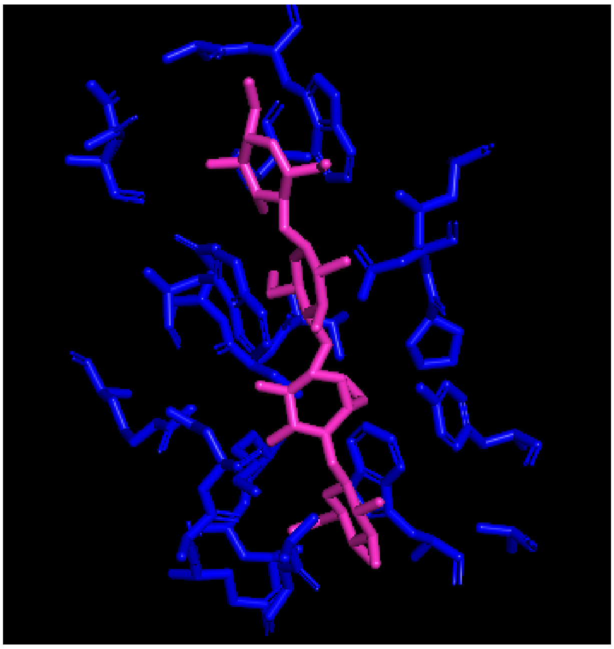
8) Analyze the template you selected in the previous question: identify 3 aminoacids that are making contacts with the substrate molecule.
Provide 3 images (one for each aminoacid) where one of the aminoacids can be seen interacting with the substrate.
Name the images active1.png, active2.png and active3.png.

Select the substrate and rename it to "subst".

Execute:

show sticks, byres all within 5 of subst

Go to object 2bod and hide "cartoon".

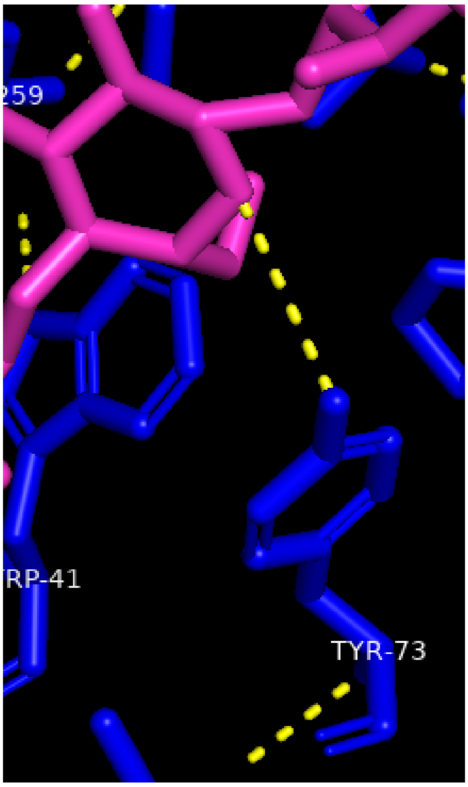


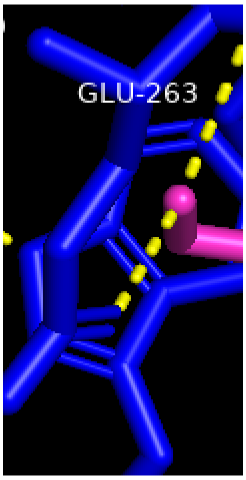
Click the residues individually and rename the selection to "act_site".

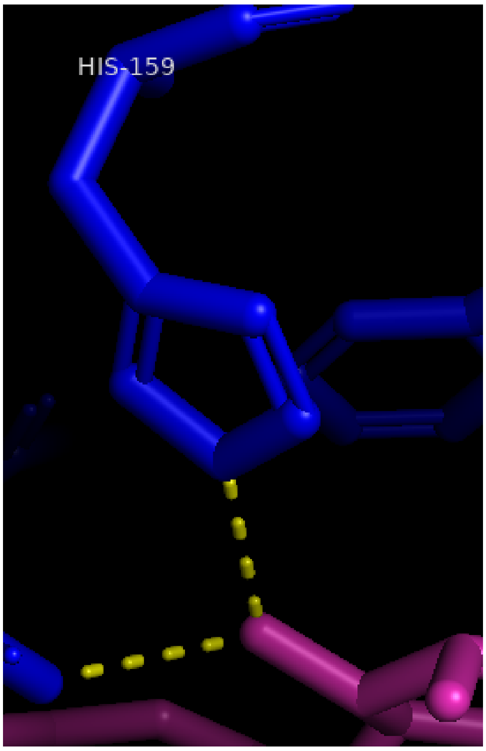
Now, let's look at the interactions made with the substrate:

- Go to the object "act_site" → A → Find polar contacts to any atoms

Now we can see that TYR-73, GLU-263 and HIS-159 interact with the substrate







https://www.youtube.com/watch?v=mBIMI82JRfl&list=PLUMhYZpMLtal_Z7to3by2ATHP-cl4ma5X&index=2

9) Make an structural model of the target sequence and name it 9.pdb.

We are going to obtain models for our target protein using a program called modeller. To execute MODELLER you need three files:

- **Target file:** contains the target sequence.

  target_exam_2.fa

- **Alignment file:** contains the alignment between the target and the template/s in PIR format.

  We need to make an alignment with one of the templates obtained in question 2. For example, template "2bod".
  We obtain its fasta:

  ```
  perl /shared/PERL/PDBtoSplitChain.pl -i 2bod.pdb -o 2bod
  ```

  When doing this, I obtain FASTA file

  ```
  cat target_exam_2.fa > quest_9.fa
  cat 2bodX.fa >> quest_9.fa
  ```

  Now we make the alignment using the model obtained in question 3.

  ```
  hmmalign 3.hmm quest_9.fa > align.sto
  perl /shared/PERL/aconvertMod2.pl -in h -out c <align.sto>align.aln
  ```

  Change the format to PIR:

  ```
  perl /shared/PERL/aconvertMod2.pl -in c -out p <align.aln>align.pir
  ```

- **Script file (modeling.py)**

Once we have edited the file modeling.py:

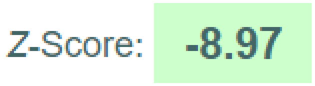
```
module load modeller/10.2
source activate modeller
mod10.2 modeling.py
```

Now we change the name of the model:

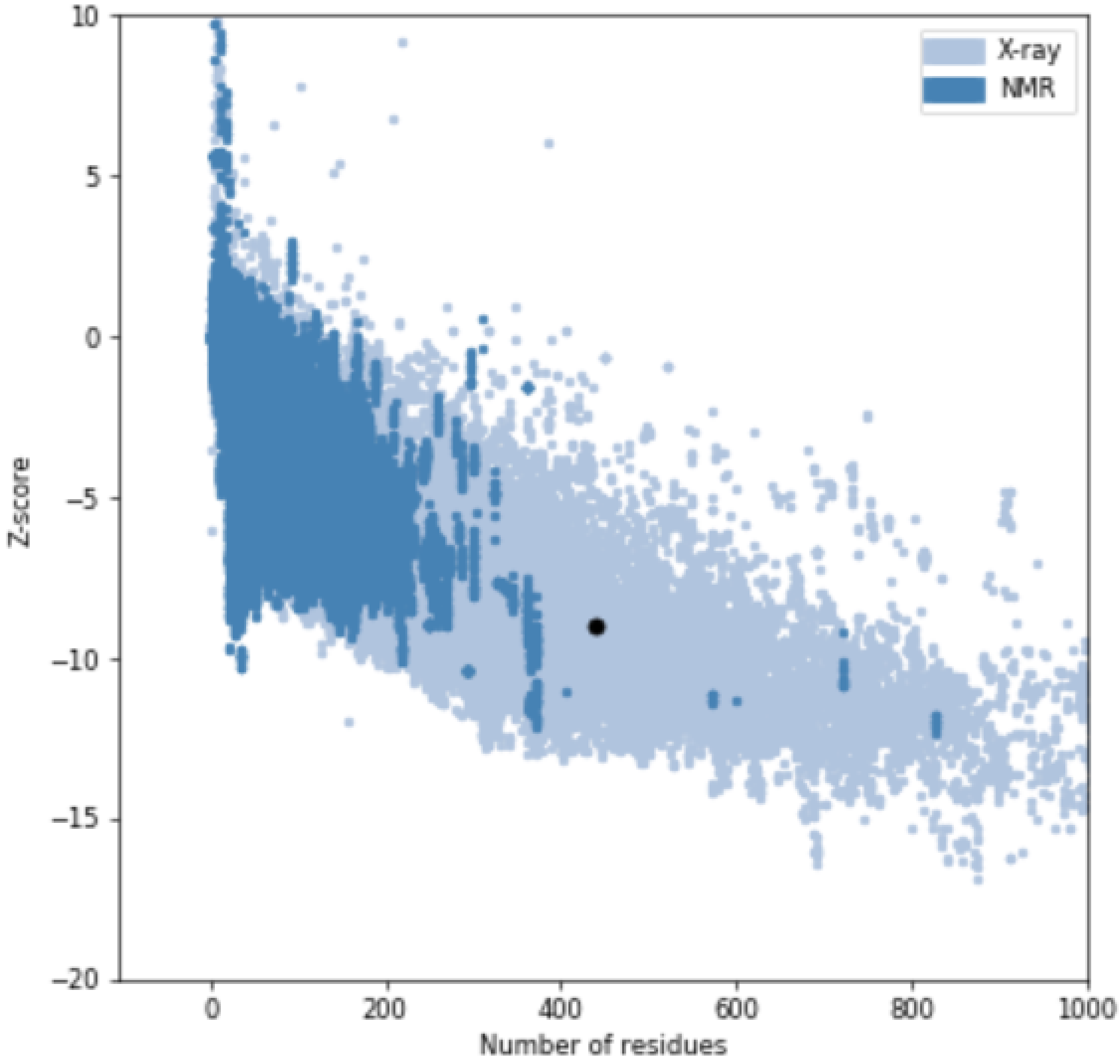
```
mv target.B99990001.pdb 9.pdb
```

10) Analyze with prosa the model you generated in the previous question. To perform a reliable analysis, should you compare the analysis of your model with another structure? What structure is this?

To analyze the model generated, we can check the Z-score and the following plot comparing how good our structure is in comparison with the structures in the PDB.



Z-Score: -8.97



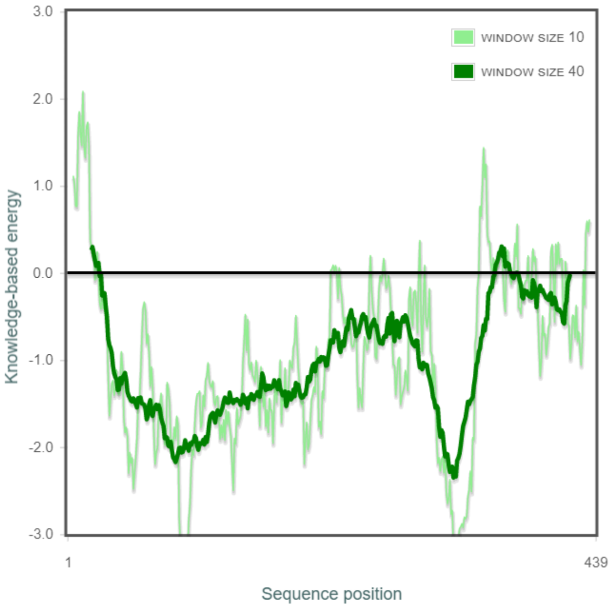
It has a small z-score, which a priori is good. Meaning that our protein model is likely to have similar structural properties to the ones in the reference set and, thus, that our model is accurate and reliable.

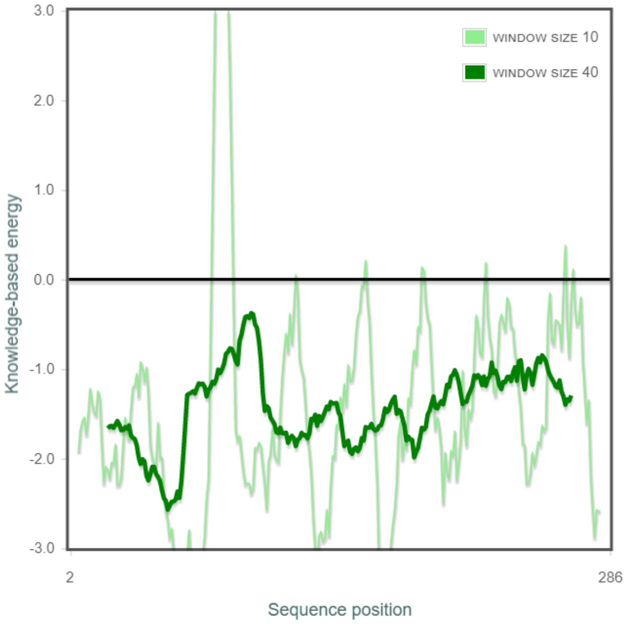
It is important to keep in mind that statistical potentials are relative measurements. This means that we cannot apply a score threshold to discriminate good from bad models. They have been intended to compare models between them or with experimentally determined structures.

A good strategy to assess whether a protein has been correctly modeled is to compare the statistical potentials energies between the model and its template. We expect the template to have a good energetic profile, since it is an experimentally determined structure. Therefore, we can use the template as a reference. If the model has higher energies than the template, it is very likely that it has been wrongly modeled.

Our template is 2bod.pdb
We load it to PROSA and we obtain a Z-score of -10.01.
Thus, our model is pretty good since it has a similar score.

We can now look at the energetic profile of the model (left) and the template (right). This profile shows the statistical potentials scores across the amino acid sequence.





As we can see, our model does not have peaks of high energy, meaning that all regions are well modeled.
Note that the lowest the energy, the most stable is the structure.

Exam 3

In this exercise we are working with a target protein for which we have its pdb but we don't know its sequence or function. The structure of this protein is stored in the file unknown.pdb.

1) Get the sequence corresponding to this structure and save it in a fasta file named 1.fa.

```
perl /shared/PERL/PDBtoSplitChain.pl -i unknown.pdb -o 1
```

2) What is the fold of this protein?

Use UniProt to create an accurate PSSM (unbiased and non-redundant DB):

```
psiblast -query 1.fa -num_iterations 5 -out_pssm exam_3.pssm -out trash.out -db
/shared/databases/blastdat/uniprot_sprot.fasta
```

exam_3.pssm is the PSSM created from the hits
trash.out is the hits (not templates because they are not structures) obtained from UniProt

Use the PSSM to search for templates in the PDB:

```
psiblast -db /shared/databases/blastdat/pdb_seq -in_pssm exam_3.pssm -out templates_3.out
```

The obtained templates (we should not copy the chain of the PDB ID):
2dly

Go to SCOP DB and look for the fold:
SH2-like

3) Obtain a HMM for this protein in the PFAM database. Name this HMM exam_3.hmm.

```
hmmscan /shared/databases/pfam-3/Pfam-A.hmm 1.fa > exam_3_HMM_PFAM.out
```

Here we can see the different models. We pick the first one (Glyco_hydro_6) and fetch it:

```
hmmfetch /shared/databases/pfam-3/Pfam-A.hmm SH2 > exam_3.hmm
```

4) Use the HMM you just obtained to retrieve 4 sequences from the uniprot database. Then, align these 4 sequences and your target using the same HMM you used to get the sequences. Name the alignment exam_3.aln.

```
hmmsearch exam_3.hmm /shared/databases/blastdat/uniprot_sprot.fasta >
exam_3_seq_uniprot.out
```

Target
P29349
P41499
P35235
Q06124

Retrieve the sequences and upload to the cluster.
Make the alignment:

```
hmmalign exam_3.hmm seq_retrieved_exam_3.fa > MSA_exam_3.sto
```

Change format:

```
perl /shared/PERL/aconvertMod2.pl -in h -out c < MSA_exam_3.sto > exam_3.aln
```

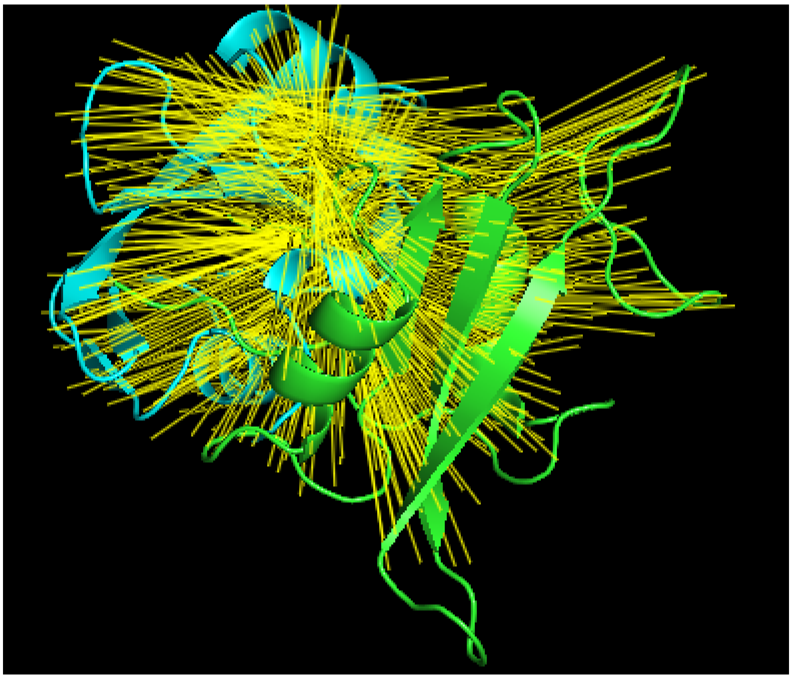
5) Find an homologous protein to our protein of interest with available structure. Superimpose both structures and provide the RMSD and an image of the superimposition named 5.png.

We can just look at the file “templates_3.out” and choose a template:
3hck

We go to the PDB and obtain the file in PDB format.
Go to PyMol and superimpose both structures (unknown.pdb and 3hck.pdb).

```
super unknown, 3hck, object=aln_exam_3, cutoff=5.0
```

RMSD =   10.094 (538 to 538 atoms)



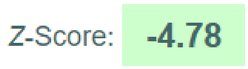
6) Generate an structure-based alignment from the superimposition you created in the previous exercise. Name this alignment 6.aln.

We just need to save the alignment:

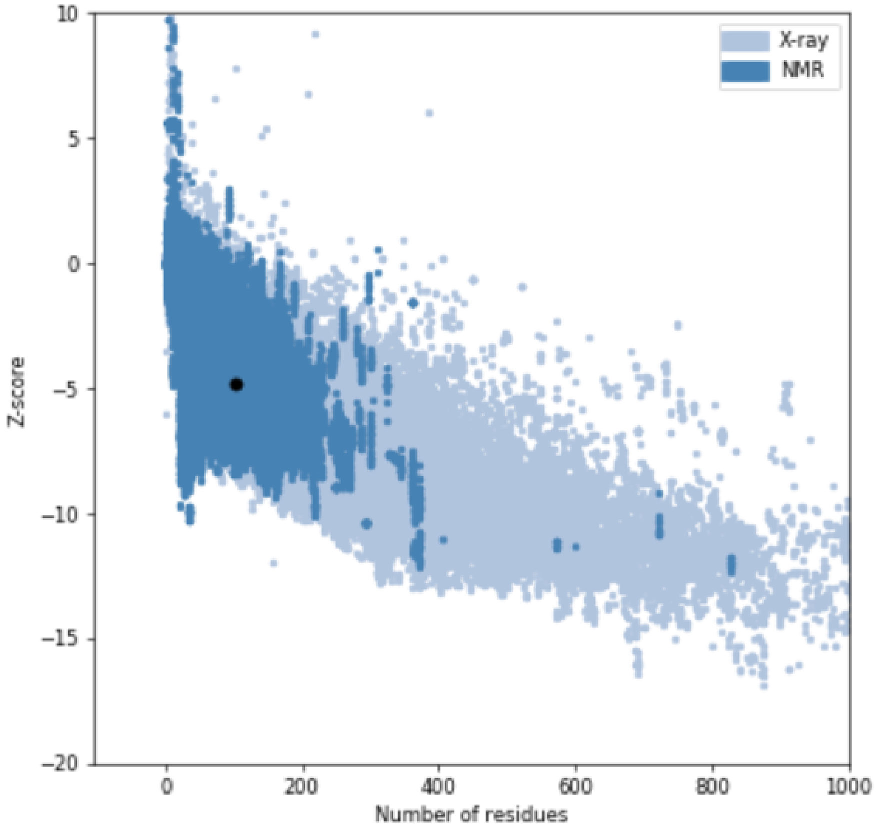
```
save 6.aln, aln_exam_3
```

7) Analyze with prosa the structure of your protein of interest.
To perform a reliable analysis, should you compare the analysis of your model with another structure? What structure could be this?

To analyze the structure of the protein of interest, we can check the Z-score and the following plot comparing how good our structure is in comparison with the structures in the PDB.



Z-Score: **-4.78**



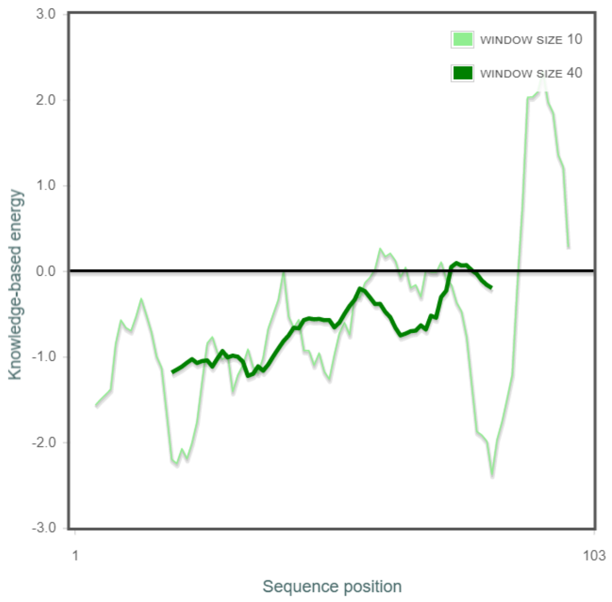
It has a small z-score, which a priori is good. Meaning that our protein of interest is likely to have similar structural properties to the ones in the reference set and, thus, that our protein of interest is correct.

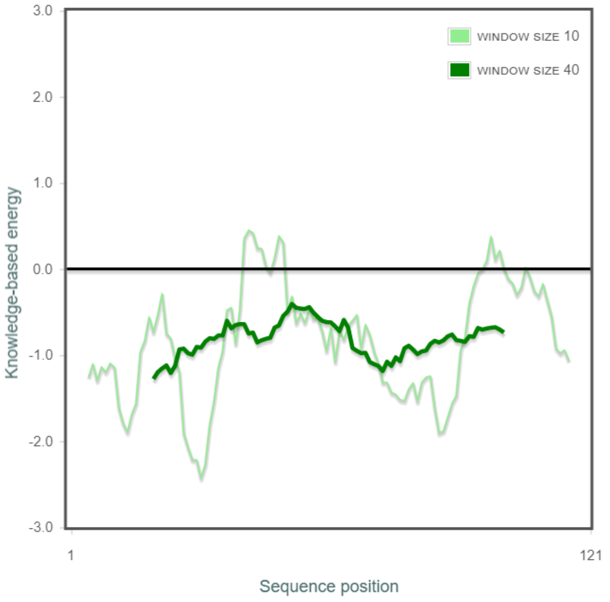
It is important to keep in mind that statistical potentials are relative measurements. This means that we cannot apply a score threshold to discriminate good from bad protein structures.

A good strategy to assess whether a protein has been correctly modeled is to compare the statistical potentials energies between the model and its template. We expect the template to have a good energetic profile, since it is an experimentally determined structure. Therefore, we can use the template as a reference. If the model has higher energies than the template, it is very likely that it has been wrongly modeled.

Our template is 2dly.pdb
We load it to PROSA and we obtain a Z-score of -5.23.
Thus, our model is pretty good since it has a similar score.

We can now look at the energetic profile of the model (left) and the template (right). This profile shows the statistical potentials scores across the amino acid sequence.





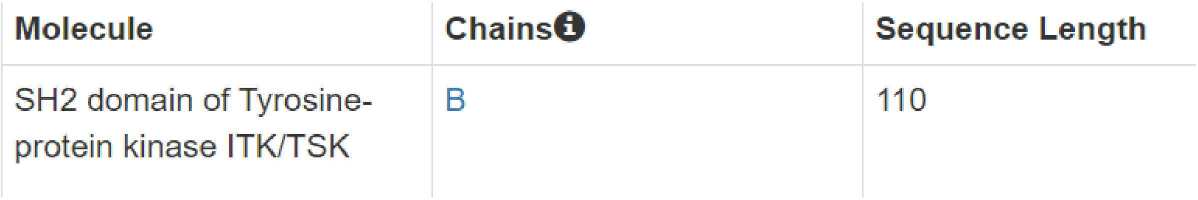
As we can see, our model does not have peaks of high energy, meaning that all regions are well modeled. Note that the lowest the energy, the most stable is the structure.

8) PDB entry 2k79 contains an interaction between two proteins.
One of these two proteins is homologous to our protein of interest, could you identify what chain in the 2k79 entry is homologous to our protein of interest?

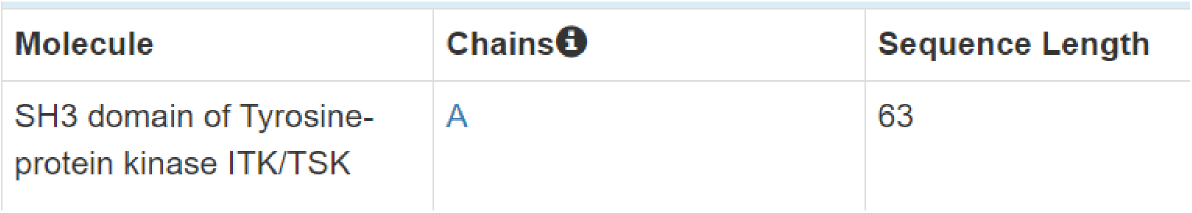
We can make an alignment using the hmm obtained in question 3.
This hmm is going to align with the region that is similar to the SH2 model.

```
perl /shared/PERL/PDBtoSplitChain.pl -i 2k79.pdb -o 2k79

cat 1.fa > quest_8.fa
cat 2k79B.fa >> quest_8.fa
cat 2k79A.fa >> quest_8.fa

hmmalign exam_3.hmm quest_8.fa > output.sto

perl /shared/PERL/aconvertMod2.pl -in h -out c <output.sto>output.clu
```

In the alignment, we can clearly see that 2k78B.fa is the homologous protein. In fact, we can just look at the PDB which of the proteins of the entry 2k79 has the SH2 domain (the same as our target protein).



Molecule	Chains	Sequence Length
SH2 domain of Tyrosine-protein kinase ITK/TSK	B	110

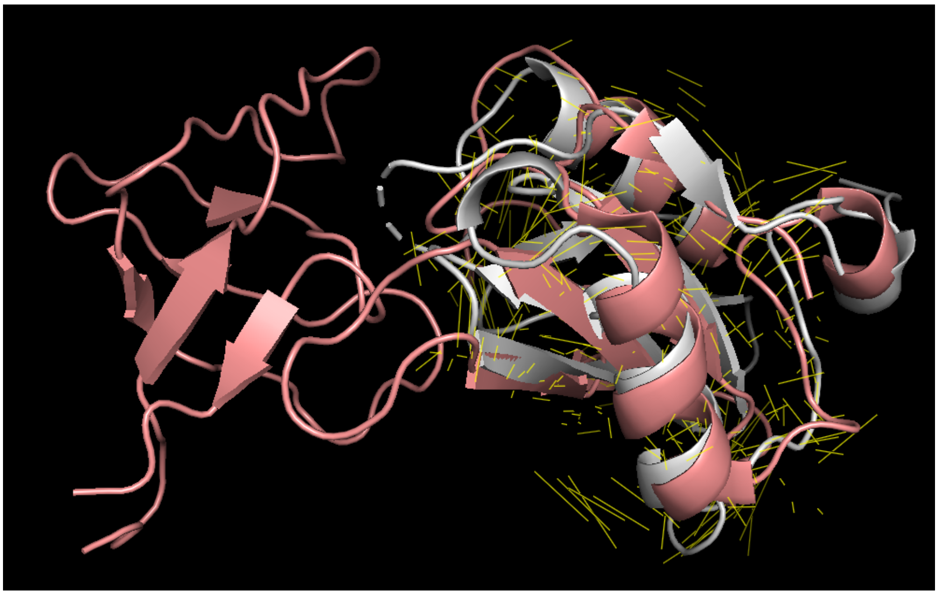


Molecule	Chains	Sequence Length
SH3 domain of Tyrosine-protein kinase ITK/TSK	A	63

9) Superimpose our protein of interest on top of the homologous chain of the 2k79 entry.
Provide the RMSD of the superimposition and a picture of the superimposition and name it 9.png.

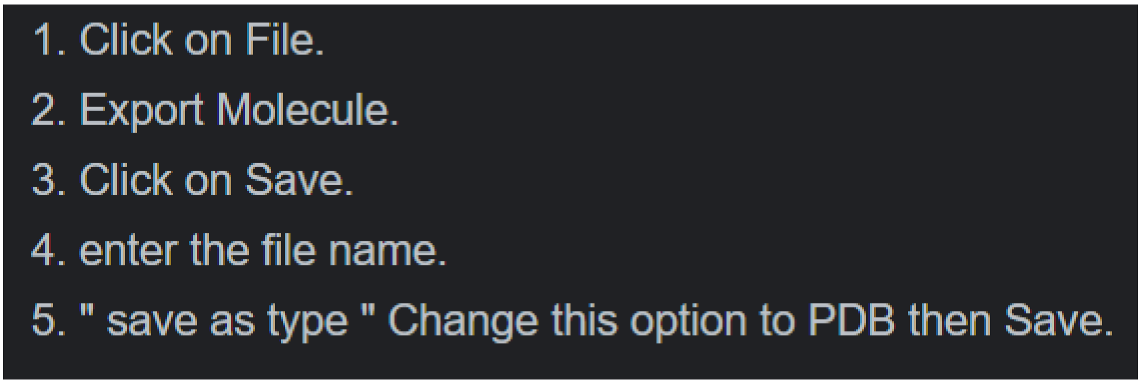
```
super 2k79, unknown, object=superimp, cutoff=5.0
```

RMSD = 2.371 (528 to 528 atoms)



10) Generate a new pdb file that only contains our protein of interest interacting with the non homologous chain in the 2k79 structure. Name it 10.pdb.

Open both files.
Make a superimposition
Select chains of interest → extract to object



1. Click on File.
2. Export Molecule.
3. Click on Save.
4. enter the file name.
5. " save as type " Change this option to PDB then Save.

Alternative question: Do you think the structure is correct? Can you prove it? Show an image of the energies that prove it.

Find if there is some structural problem and show the location with an image. What's the sequence fragment with this problem?

Do you think the protein PROBLEM can work as a tetramer? Show an image that can prove it. Even if it was a monomer, do you think it will work with the function you selected in question 1?

This protein can bind ligands such as drugs or hormones. Find an homologous structure that contains one of these ligands. Name it ligand_template.pdb.
Use the structure you obained to reconstruct the interaction between our protein of interest and the ligand included in that structure.

Pregunta pag 8.

We want to study the L253R mutation in SH2. Model the structure of SH2 containing this mutation. Be aware that the structure that you have is not complete, therefore the first amino acid corresponds with position 232. Name your model as mutant_model.pdb. (practical 4)