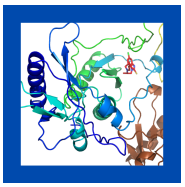

ESCI - UPF

Structure

Structural Bioinformatics

Marta Ortigas - Mònica Torner - Núria Mitjavila - Xavier Crespo

Second year
25/02/2023

Contents

1 Assignment 3

You can find the whole set of files generated in this project at:
https://github.com/XavierUPF/Structural-Bioinformatics-Project-BDBI

1.1 Structure

1.- Get a set of 4-6 structures from the PDB that belong to the family of your protein of interest. Try to get a set that is not biased, so avoid pairs of proteins that are identical or very similar. How would you do that? What programs would you use? What are the PDB IDs of the structures you have selected?
From the Uniprot database, we have the entry P04585 that is the one that we get from our PDB entry, 1REV. We download the sequence (we already have it) and we charge it in the Cluster. We load the module for the hmmer package to obtain a model using HMMs, and then we will try it again with BLAST to observe the different possible results.

First strategy is to find which HMM of PFAM can fit the best our target sequence (with hmmscan):

```
hmmscan /shared/databases/pfam-3/Pfam-A.hmm P04585.fasta > hmmscan.out
hmmfetch /shared/databases/pfam-3/Pfam-A.hmm RVT_1 > RVT_1.hmm
hmmsearch RVT_1.hmm /shared/databases/blastdat/pdb_seq > RVT_1.out
```

PDB ID: 1jkh, 1jla, 1jlb, 1jlc, 3drr, 2iaj, 2iaj, 2iaj, 1jkh...
Check that they are not biased to avoid identical proteins.

```
Selected PDB ID: 1JKH, 1N4L, 3DU5, 1XR7, 2D7S and 1MVP
```

The second strategy is to find it from the Blast database, the 4 or 6 structures from the PDB:

```
blastp -query P04585.fasta -db /shared/databases/blastdat/pdb_seq -out
↪  P04585.out

PDB ID: 3DI6, 3C6U, 3C6T, 2RF2, 3DRP, 3DRS, 3DLG, 3DLE, 2RKI...
```

Check that they are not biased to avoid identical proteins.
Selected PDB ID: In this case, these results are worse than the previous.

```
Selection: 1REV (original)
1JKH: HIV-1 reverse transcriptase in complex with Efavirenz.
1N4L: Moloney murine leukemia virus (reverse transcriptase)
3DU5: Structure of the catalytic subunit of telomerase
1XR7: RNA-dependent RNA Polymerase 3D from human rhinovirus
2D7S: Foot & Mouth Disease Virus RNA-dependent RNA polymerase
1MVP: Avian Myeloblastosis Virus
```

2.- Superimpose the structures you selected in question 1. Are they structurally similar? What is their RMSD? Can you identify some regions with higher variability? Why do you think these regions are more variable? What about the most conserved regions of your protein (the ones you described in assignment 1, question 6 and assignment 2, question 4), are they structurally variable or not? Can you relate this to the function of the protein? Include pymol images to support your explanation.
Here is the data obtained after the superimposition of the structures selected in question 1 with the 1REV structure (between brackets we have the RMSD value):

```
1JKH (1.287): HIV-1 reverse transcriptase in complex with Efavirenz.
1N4L (1.534): Moloney murine leukemia virus (reverse transcriptase).
3DU5 (6.504): Structure of the catalytic subunit of telomerase.
1XR7 (9.747): RNA-dependent RNA Polymerase 3D from human rhinovirus.
2D7S (10.132): Foot & Mouth Disease Virus RNA-dependent RNA polymerase.
1MVP (12.684): Avian Myeloblastosis Virus.
```

As we can see, the ones that have a better RMSD are 1JKH and 1N4L. However, as 1JKH is the same protein as the one that we are studying but just in complex with a drug, we are not going to select it and instead choose the 3DU5 structure for a better individual visualization of which parts are superimposing.

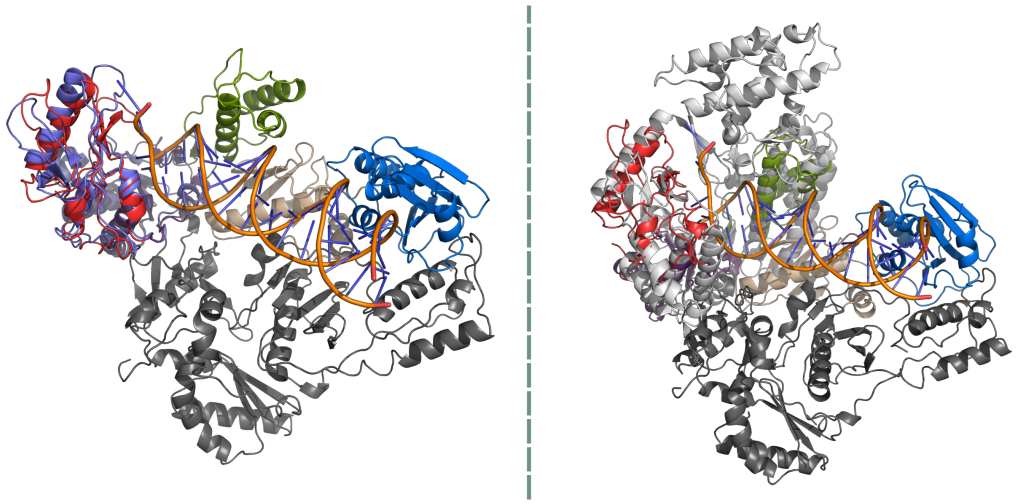

Figure 1: 1REV-1N4L (left), 1REV-3DU5 (rigth)

In the previous pictures, we can appreciate how the region which is superimposed is the one that, in the 1REV, corresponds to the fingers and sometimes also the thumb. This tells us that even though in assignment 1, we mentioned that one important region could be the RNase H, after performing this practical we localize a region that could also be important for the function of our protein. This region is the fingers and thumb domain, which will be properly explained in the next question.

3.- Choose the region (or regions) that you think are the most important for the protein function. [...] How is this active site interacting with its substrates? What contacts are made between substrate and enzyme? What amino acids are essential in this active site? How these amino acids contribute to catalyse a chemical reaction?
From what we've seen so far, the most important regions are:

- Fingers (maybe also thumb) [Polymerase active site]

  Reverse transcriptase clamps down on the incoming nucleotide with its fingers to the palm. This is because the union between fingers and palm is flexible because of the loops. This helps our protein to undergo a conformational change during the polymerization reaction. (We could try to simulate this with or without DNA in the cluster...)

  The thumb subdomain helps to stabilize the overall structure of the enzyme and facilitates movement of the enzyme along the RNA template during polymerization. How? Like with fingers, the thumb is connected to the palm with loops that allow flexibility.

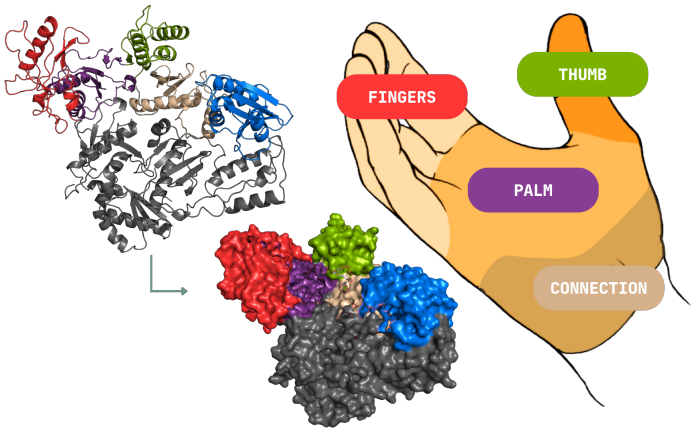

Figure 2: Here you see more clearly what we mean by naming the different regions as if it was a hand.Cartoon structure, surface and abstract diagram. In marine, RNase. In grey, p51 subunit.

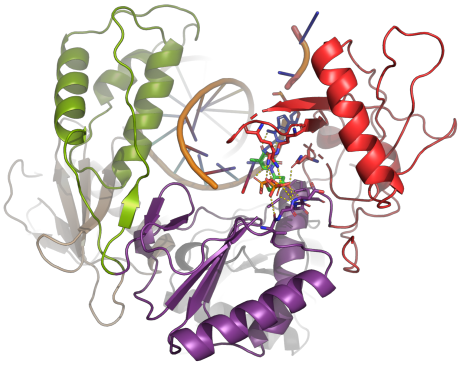

Figure 3: Overall image of the region. You can clearly see the loops that join the subregions that allows this handy flexibility.

The contacts between our protein and the TTP shown below, are polar contacts:

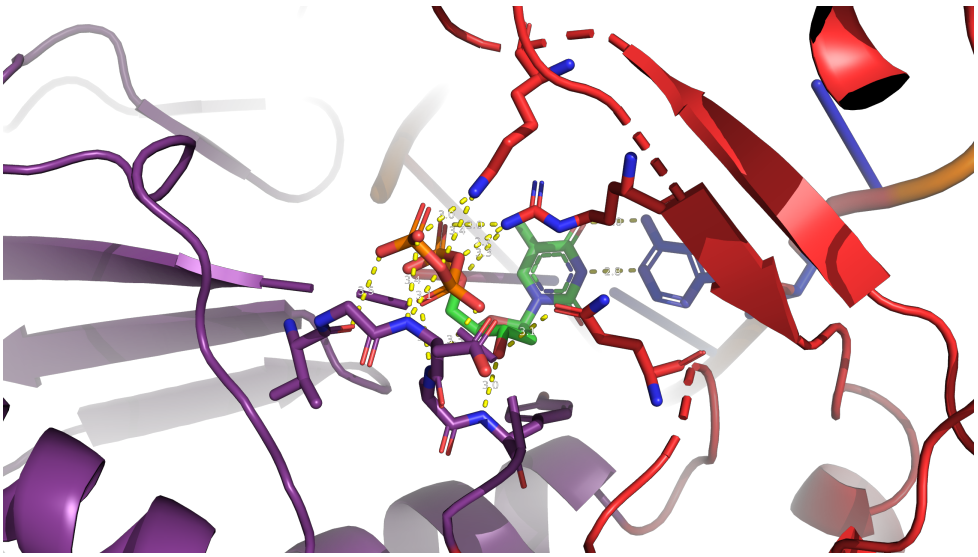

Figure 4: Polar interactions with TTP. Here we see how the fingers are clamping TTP with the palm.

- RNase [Nuclease]

The active site of the RNase is a nuclease. There is no structure where it is clearly visible the way it works to retrieve some images from pymol. But, as we know, the nuclease activity of RT helps to degrade the RNA template that the enzyme is copying, allowing the reverse transcription process to proceed. The nuclease domain cleaves the RNA strand, releasing small fragments of RNA that can be recycled into the synthesis of the new DNA strand.

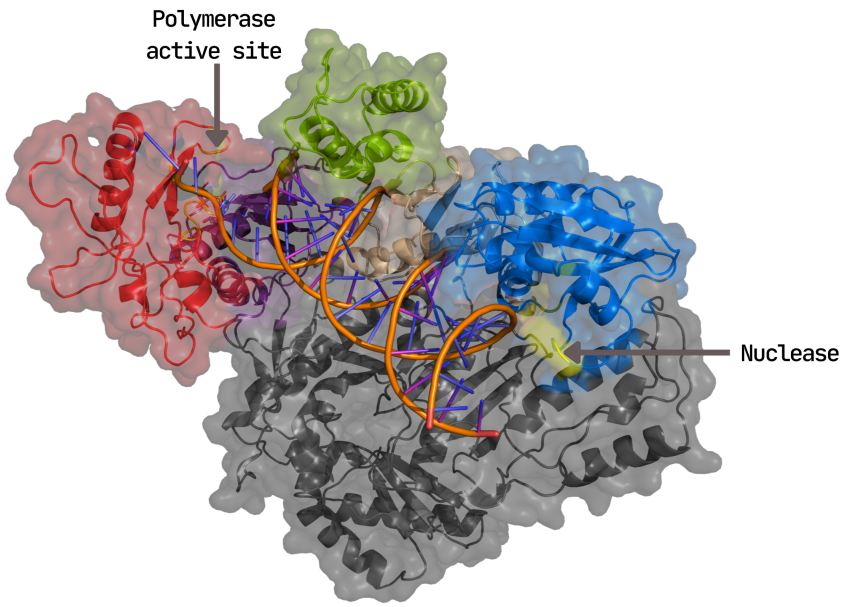

Figure 5: Both active sites indicated in yellow.

4.- Use MODELLER to create a model of your protein of interest that includes the mutation you chose in the first assignment (assignment 1, question 7). Show pymol images comparing the wild type structure of your protein and the structure of the mutant you just modeled. By comparing the structures, hypothesize why the mutation has an effect in the protein function.

Well, as we are working with drugs instead of mutations, It does not make sense to use modeller. But, we though: "What happens if we use modeller with the sequence of UniProt and the affected structure of Efavirenz? Will the output be affected as well, or corrected? What happens with Alphafold?" With this, we wanted to know if modeller relies on more in the template structure or the physicochemical properties of the sequence.

The answer is quite simple. Sequence is quite important to match with the structure. But, for Modeller, the template structure is what matter the most. It mostly bases the folding prediction on the structure. From this we can say that Efivarenz, the non-nucleoside drug we've covered, does not interfere with the sequence in any way, it just changes the conformation of the protein.

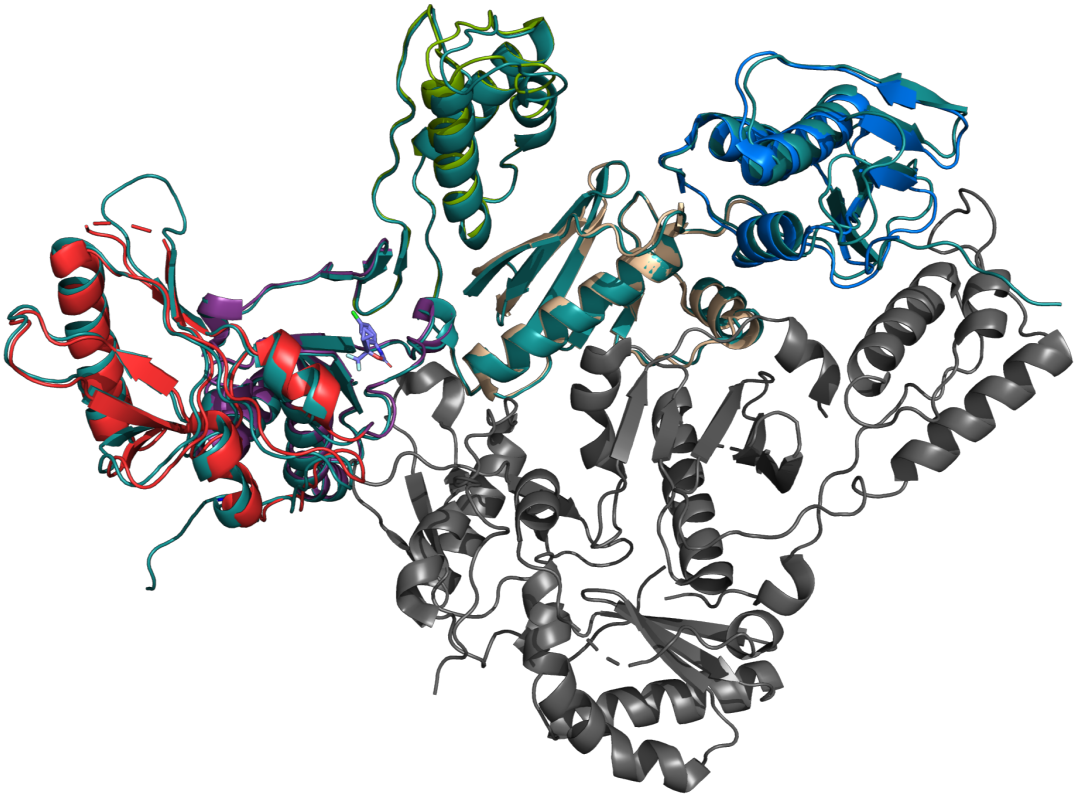

Figure 6: The Modeller output (cyan) and the affected open protein because of Efavirenz (1FK9).

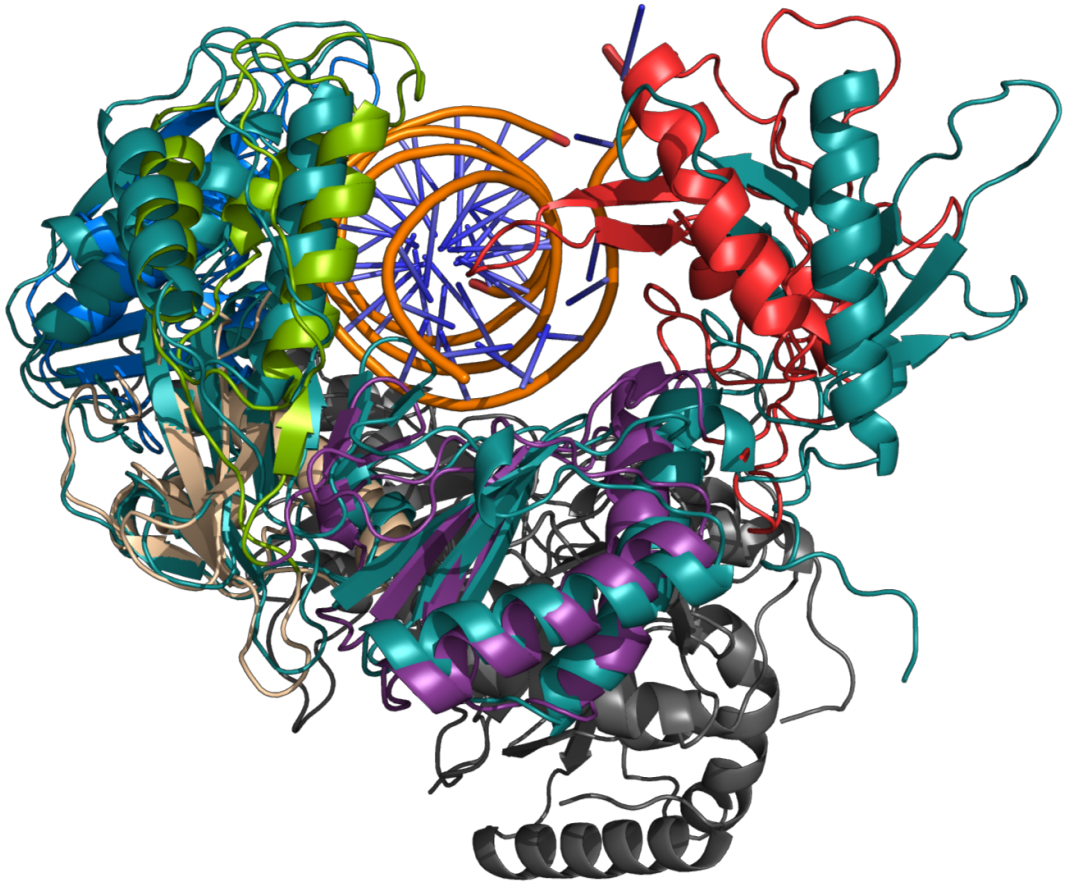

Figure 7: Here we can see the Modeller output (cyan) superimposed with normal state protein just to compare how open actually is.