Bruno Álvarez, Jan Izquierdo, Jaume Jurado & Marc Trujillo

Structural bioinformatics Project – Assignment 3: Structural analysis.

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?

This exercise can be done in different ways that yield slightly different results, we will do it using different methods to confirm results.

This exercise can be solved using psiblast, by taking one of the sequences resulting from the second exercise of the previous assignment we can use psiblast with the pdb database and get a list of PDB IDs from similar proteins that belong to the same family.

The commands used were:

$psiblast -query A3CS71.fasta -num_iterations 5 -out_pssm pssm_A3CS71.pssm -comp_based_stats 1 -out sprot_A3CS71.out -db ~/Documents/databases/uniprot_sprot.fasta

$psiblast -db ~/Documents/databases/pdb_seq -in_pssm pssm_A3CS71.pssm -comp_based_stats 1 -out pdbsprot_A3CS71.out

We picked the PDB ids from proteins that had an E-value of 0 or slightly above 0. As if all E-values were to be 0 all the structures would have been too similar. We used this on a random protein from exercise 2 assignment 2, it doesn't matter which protein we picked as they all belong to the same family.

The results for subunit F1 alpha, using protein A3CS71:

E-value 0:

1vdz_A, 1fx0_B, 1kmh_B

E-value>0:

2obm_A, 2obl_A, 2dpy_B

Now we will repeat the process for subunit F1 beta, we used protein A0RL95:

E-value 0:

2qe7_F, 1sky_E, 2jj2_E

E-value>0:

3b2q_B, 2rkw_B, 2c61_A

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.

For subunit F1 alpha:

We will use PDB IDs 1vdz_A, 1fx0_B, 1kmh_B, 2obm_A, 2obl_A and 2dpy_B

Aligned against the PDB of A3CS71 that we found using uniprot, that led us to a structure created by alphafold (https://alphafold.ebi.ac.uk/entry/A3CS71), we had to use the one from alphafold since it has no PDB structure. This .pdb file will be used to align all the other structures against.

We open the downloaded pdb with pymol and rename it to A3CS71

fetch 1vdz_A 1fx0_B 1kmh_B 2obm_A 2obl_A 2dpy_B

remove resn hoh

super 1vdz_A, A3CS71, object=aln1

RMSD = 1.185 (2648 to 2648 atoms)

super 1fx0_B, A3CS71, object=aln1

RMSD = 1.961 (1962 to 1962 atoms)

super 1kmh_B, A3CS71, object=aln1

RMSD = 1.994 (1965 to 1965 atoms)

super 2obm_A, A3CS71, object=aln1

RMSD = 1.149 (1227 to 1227 atoms)

super 2obl_A, A3CS71, object=aln1

RMSD = 1.172 (1282 to 1282 atoms)

super 2dpy_B, A3CS71, object=aln1

RMSD = 1.686 (1908 to 1908 atoms)

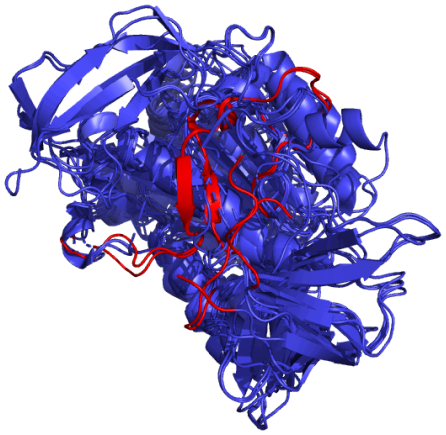
save aln1.aln, aln1

Initially we used the structure with ID 2qe7_D but it had a very high RMSD (2.562),so we replaced it with 1fx0_B.

To identify the regions with higher variability we can use the sequence display in pymol or the saved alignment, using either of them it can be seen that the region that has the most variance is located at the **beginning of the sequences**, this matches with the fact

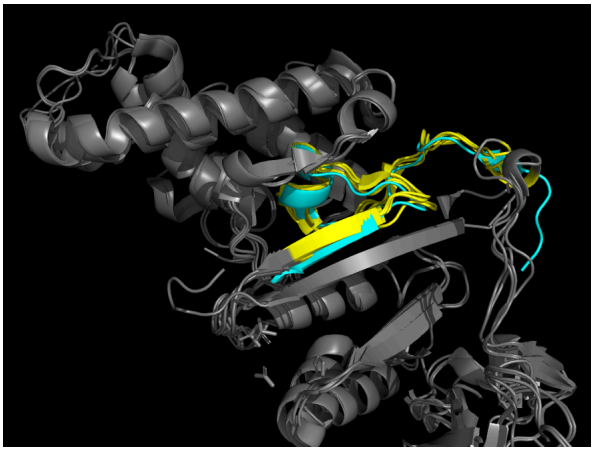
that alpha chain does not have any protein-protein interaction or ligands in the first amino acids, it is from position 49 that it starts to have some relevant interactions and bindings.

The region in red is the most variable section of each structure.



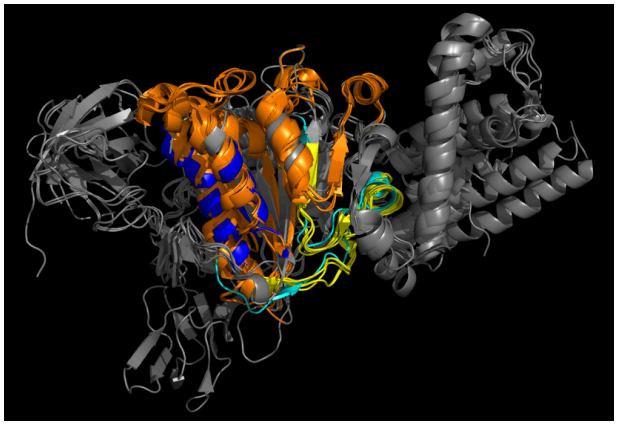
Based on the observations of this practical it can be said that the least variance regions (most conserved) are located around amino acids **145 to 171**, 225 to 361, except 1vdz_A, that goes from **291 to 321** and from 361 to 371 and 1obl_A, which presents a significant non-matching aligned region from 306 to 326.

These conservation zones might exist because there are interactions with the Beta and Gamma chain around those positions (protein protein interaction), also with ADP/ATP and Magnesium ions. Those regions are relevant since they are not only structurally important but they actually do have a function meaning that they are more likely to be conserved since a mutation there would have a greater impact.

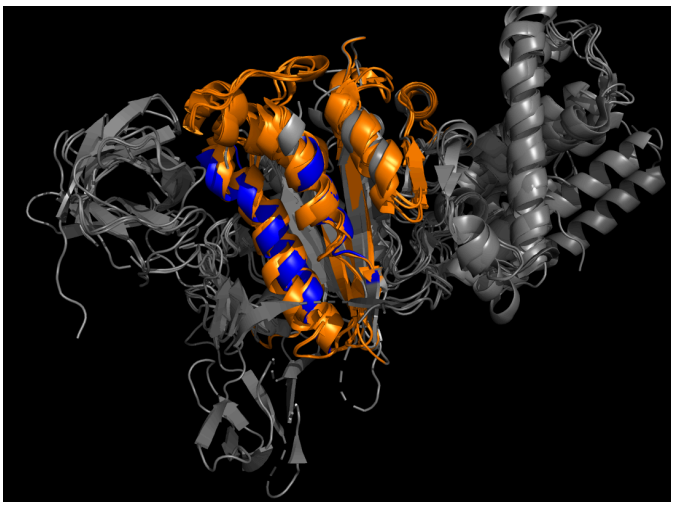


The first conserved region(145 to 171), the blue region is from 1vdz_A structure and the yellow is from the other ones.

This is the second conserved region (225 to 361), in blue the one part belonging to 1vdz_A and in orange the part belonging to the rest. The grey overlapping zone belongs to the non-matching aligned region of 1vdz_A.



Here is an image of all the aligned structures, with the conserved regions colored using the same criteria as above:



For subunit F1 beta:

We will align the found PDB IDs against A0RL95, from which we have found the structure by searching in uniprot and downloading the PDB file created by alphafold (https://alphafold.ebi.ac.uk/entry/A0RL95).

We will use PDB IDs 2qe7_F, 1sky_E, 2jj2_E, 3b2q_B, 2rkw_B and 2c61_A

We open the downloaded pdb with pymol and rename it to A0RL95

fetch 2qe7_F 1sky_E 2jj2_E 3b2q_B 2rkw_B 2c61_A

remove resn hoh

super 2qe7_F, A0RL95, object=aln1

RMSD = 1.584 (3040 to 3040 atoms)

super 1sky_E, A0RL95, object=aln1

RMSD = 1.537 (3220 to 3220 atoms)

super 2jj2_E, A0RL95, object=aln1

RMSD = 1.512 (2927 to 2927 atoms)

super 3b2q_B, A0RL95, object=aln1

RMSD = 2.290 (1618 to 1618 atoms)

super 2rkw_B, A0RL95, object=aln1

RMSD = 2.162 (1666 to 1666 atoms)

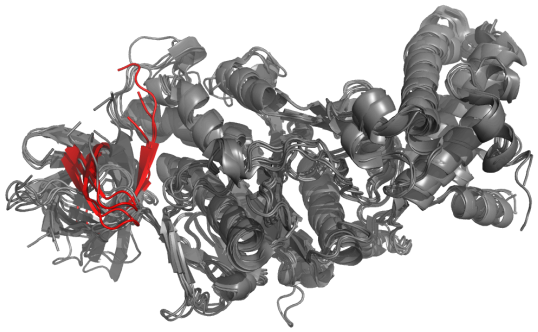
super 2c61_A, A0RL95, object=aln1

RMSD = 2.273 (1672 to 1672 atoms)

save aln2.aln, aln1

Initially we had chosen the structure with PDB ID 2jj2_M, but we changed it with 2jj2_E because the RMSD of the alignment was too high(2.544).

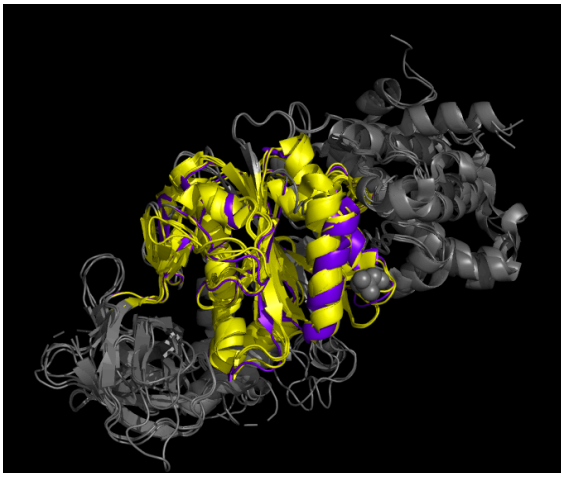
To identify the regions with higher variability we can use the sequence display in pymol or the saved alignment, using either of them it can be seen that the region that has the most variance is also located at the **beginning of the sequences**. The colored part is the one belonging to the region with highest variability



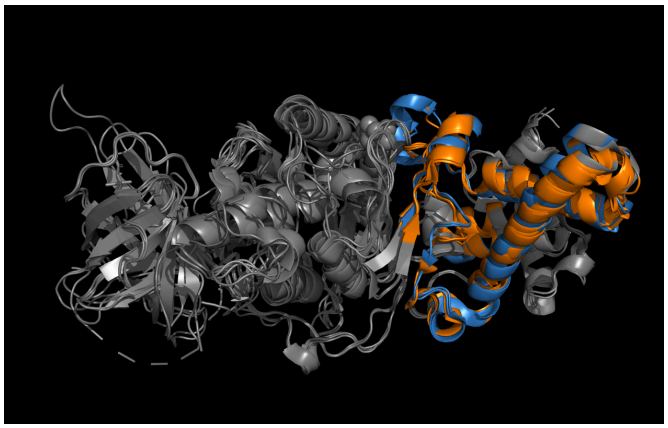
Based on the observations of the alignment we can conclude that the conserved regions are located around amino acid positions **81 to 251**, except for 2jj2_E that ends in position 261, 3b2q_B, 2rkw_B and 2c61_A also have some sequence fragments along the way that are not matching in the alignment.

There is another conserved region from around **326 to 426**, 3b2q_B, 2rkw_B and 2c61_A present some non-matching sequence fragments in the alignment and there is a bit of variance in the alignment of sequences 2jj2_E an 1sky_E, where the starting position is in amino acid 336 instead of 326 and the ending position is found around 436 while the others end around position 426.

The F1 beta subunit presents high sequence conservation because its the chain that presents more active sites that bind to ATP (apart from binding to Gamma and Beta chains) , and thus is more susceptible to changes.

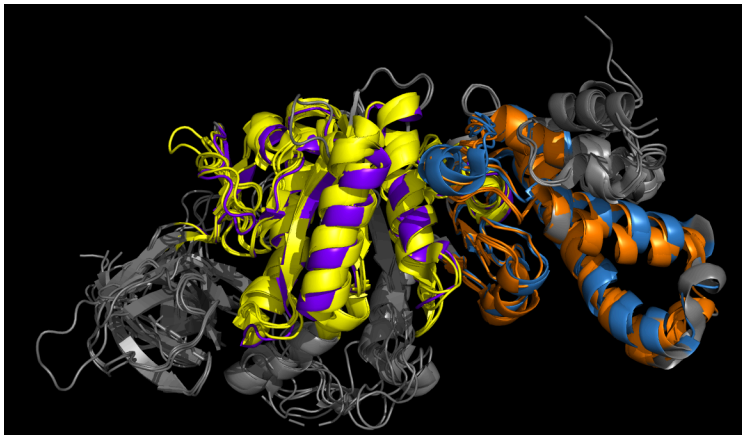


This is the first conserved region (81 to 251), 2jj2_E is colored in purple, the other sequences are in yellow.



This image belongs to the second conserved region (326 to 426), where the blue region belongs to 1sky_E and 2jj2_E and the orange belongs to the other structures. We can see at the end of the longer helix some grey-colored region which belong to non-matching aligned regions of sequences 3b2q_B, 2rkw_B and 2c61_A.

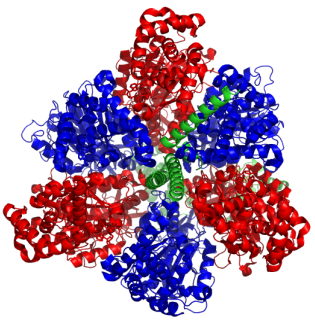
Here is an image of all the aligned structures, with the conserved regions colored using the same criteria as above:



3. This is the most important part of the submission: Choose the region (or regions) that you think are the most important for the protein function. Then, describe this region, why can it carry out the function that it does? What are the weak interactions that allow this function to happen? Include pymol images to support your explanation. You can inspire yourselves with the works of students from previous years, find them in: https://sbi.upf.edu/web/index.php/courses/undergraduatedprojects. Here you have some examples of how to orient this question:

- If your protein is an enzyme, you should describe its active site. 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 do these amino acids contribute to catalyzing a chemical reaction?
- If your protein needs to interact with another protein to carry out its function, describe the interaction between the two proteins. How are the two proteins interacting? What are the interactions that make the two proteins have chemical affinity for each other? Can you find any amino acid that is essential for that interaction to happen?

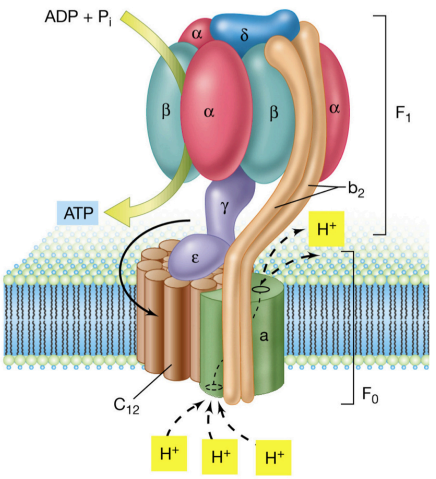
Our protein is ATP-synthase, since its sequence is splitted in different regions it is important to remark that there are many regions that are extremely important and essential for its function. However, for the sake of time management and this assignment, we are only studying F1 alpha and beta. These two regions are involved in the reaction of ADP and Pi into ATP. ATP synthase has three alpha beta complexes, when ATP-synthase spins it pushes one of the complexes (using gamma) by applying pressure, this provokes it to open and release an ATP while the previously opened can now close and catalyze the reaction.



This alpha beta complex has three main phases E, DB and TP. Each one corresponds to open (where it releases an ATP and obtains the ADP + Pi for a new synthesis), loose (it's an intermediate state the synthesis is being prepared), tight (where the synthesis is essentially made). When an ADP and Pi enters an open complex the tight complex with the synthesized ATP opens, this cycle is perpetuated thanks to the gamma protein spinning.

To be more specific on its functioning, let's explain how the whole protein works:

Bruno Álvarez, Jan Izquierdo, Jaume Jurado & Marc Trujillo



1- A proton gradient is generated due to the difference in pH between the two sides of the mitochondrial membrane (this gradient passes through the F0 subunit to the interior of the membrane). ATP synthase uses this gradient to function and generate the movement that will be used to create ATP from ADP.

2- A ring of subunits c (F0) forming a complex interacts with the gradient causing some conformational changes in itself.

3- This interaction with the proton gradient and conformational changes creates a rotation in the ring complex of subunits c that is transmitted to subunit gamma, which is connected to the complex of subunits c and acts as a central axis.

4- Then this movement is transmitted to subunits alpha and beta. This movement is transmitted from subunit gamma to alpha and beta through connector delta, which allows the complex formed by 3 beta and alpha subunits to rotate.

5- The rotation itself of subunits alpha and beta induces conformational changes that make alpha and beta switch between the 3 states previously mentioned (open, loose and tight). Being each one of the states present in one of the pairs of alpha and beta beta it means that the protein is capable of having 3 ADP (or ATP) and will eventually convert them into ATP one at a time.

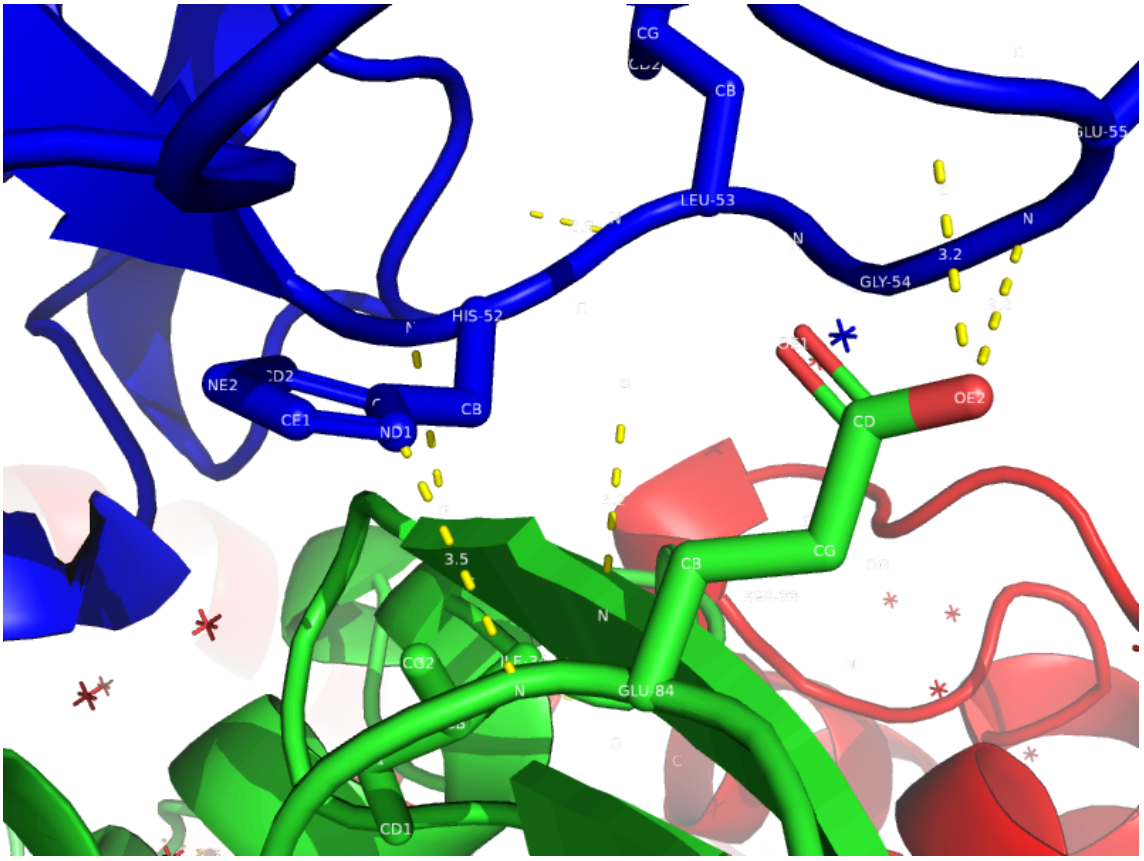
6- During open state the ATP previously formed is liberated and the catalytic site is empty in preparation for the next step. Then the alpha-beta pair enters in a loose state and is ready to accept ADP + P into its binding site. The next step is the tight state in which the conformation of beta and alpha makes more favorable the synthesis of ATP from ADP+P. Then we start again with the open state by liberating the formed ATP during the tight state.

Despite the epsilon subunit of F1 not taking place directly in the synthesis of ATP it has a vital role in our protein. Its main function is to stabilize the whole protein, to avoid inversion of ATP synthase and it is also involved in regulation.

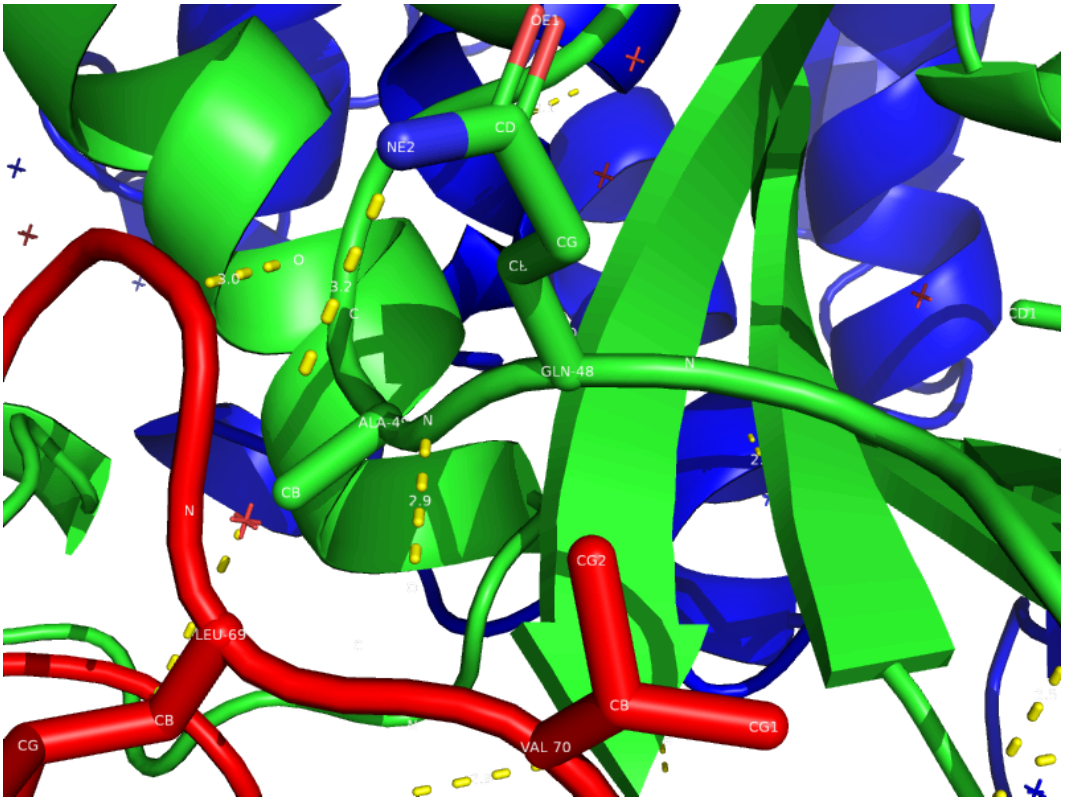
Subunit b has a similarity with epsilon subunit. Its function is to contribute to the structure and stability of the protein and to connect F1 and F0. It hasn't been fully investigated but it is believed to be involved in regulation of ATP synthase.

To study this complex let's first see what interactions do alpha and beta make with gamma (ɣ-α,β interactions)

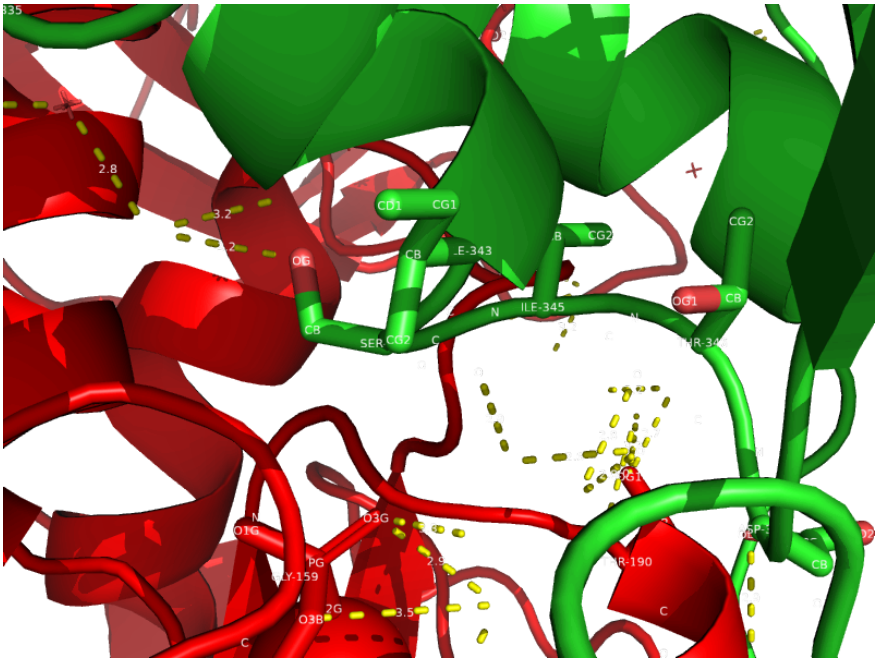
Now lets see some of all of the interactions alpha and beta make between them.



Here we can observe a Glutamic acid establishing a hydrogen bond (O-N) with the Beta, and also a Histidine interacting with the Glutamic acid.



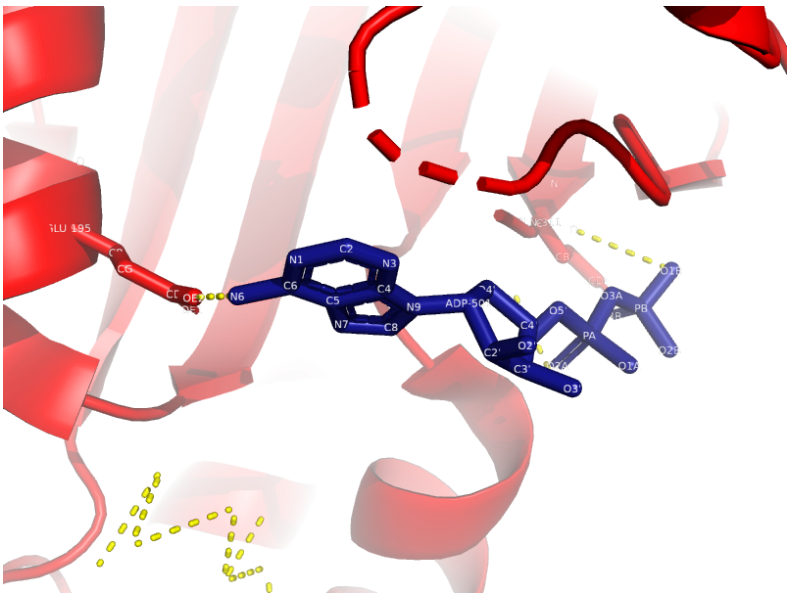
In this other case we observe a Leucine, an Alanine and a Glutamine bond.



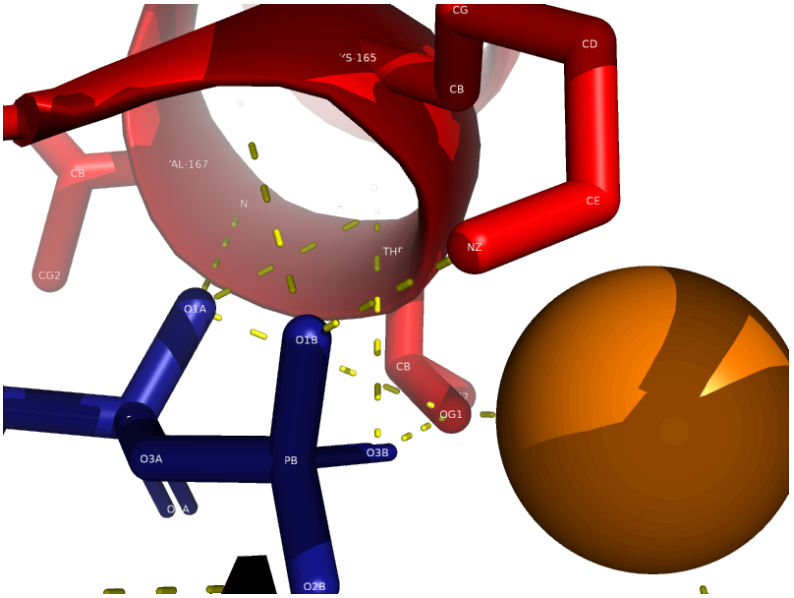
Finally, here we observe a Serine and an Isoleucine bonding.

Now let's see how alpha and beta interact with the ATP in its active site

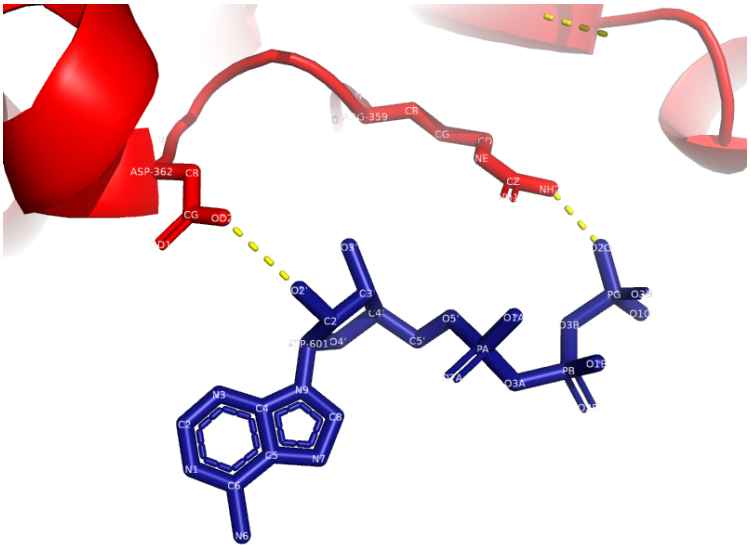
First of all let's have a look at the alpha chain.



Here we can see the ATP bounded through a N to an O of the GLu-195, one of the active sites, essentially capturing the ADP.

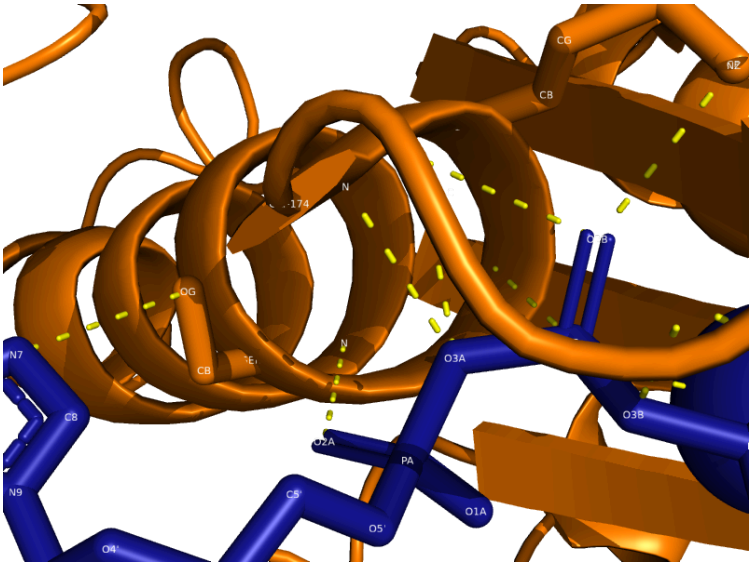


In this image we can see the ADP in another of the active sites of the alpha beta complex. In this case it is in the tight stage and the ATP is being synthesized, that is why we found a Mg (the orange ball) since it helps to stabilize the protein. THe ATP is bounded to the LYS-165 through an O-N bound, and to a THR-166 through an O-O bound. There are other bounds made with the main chain of the alpha chain and also with some waters.



Finally we see the ATP synthesized, bounded to an ARG-359 and an ASP-362

Lets now see how it interacts the beta chain with the ATP



Between different stages in the beta chain there is an aminoacid that is most of the time bound to the ATP. This is the LYS-175, here we can see it is bounded to the N of the ATP through its N. Also in this active site we can see GLY-174 and SER-177, these are used to stabilize and make the reaction easier, also we can observe a Mg molecule in the left since it helps to stabilize the protein.

So we can see that the essential amino acids in the active site are for the alpha chain GLU-195, LYS-165, THR-166, ARG-359 and ASP-362.

And for the beta chain LYS-175, GLY-174 and SER-177.

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.

The first step required to use the modeller to create a model of our protein of interest is to find the protein sequence with the mutation. In our case, since we change our chain of study we will choose a different mutation to create our model.

The mutation that we chose is the var_088542. It is in position 207 and it changes R>H.

We will need a good template though as the homology degree between target and template could affect the quality of the model. To find the templates we decided to use blast instead of HMM, as blast will allow us to search for proteins that resemble our query in all it extension, HMM works better for simple proteins, and if we take into account the complexity that ATP synthase has with it's different subunits, we've decided that blast was the better choice here. So we search for the sequence in PDB using blast

$blastp -query mutation.fa -db ~/Documents/databases/pdb_seq -out mutation.out

Obtaining as the best result: 2jdi

Now that we got the best hit we need to search for the template sequence in the rcsb pdb and we download it in pdb format

To avoid any possible problem for lack of understanding from any other program, it's highly recommended to use PDBtoSplitChain, to split our PDB file in different chains and correct the PDB format:

$perl ~/Documents/perl_scripts/PDBtoSplitChain.pl -i 2jdi.pdb -o Template

Then we need to do the execution of the modeller, we have our target file, our script file and we need our Alignment file:.

We use cat to add all the sequences of our templates and our mutation sequence into the same .fa so we can make a proper clustalw alignment, cat should work for this.

$cat mutation.fa > mutation_template.fa

$cat TemplateA.fa >> mutation_template.fa

Now that we have a single file containing both the wild type sequence and the mutated sequence we can align the sequences using clustalw

$clustalw mutation_template.fa

Finally, now that we have our clustalw successfully done, we need to pass the format of the alignment to pir format as is one requirement from the MODELLER

$perl ~/Documents/perl_scripts/aconvertMod2.pl -in c -out p <mutation_template.aln>mutation_template.pir

Finally once we have our .pir ready is time to execute the modeller. However we need it to be very precise in fulfilling all the requirements the modeller asked us. So we change a little bit the output of the .pir file to accommodate the requirements of the code(only the names of course as the sequence and alignment remains exactly the same). Once the changes in the modeling.py are done, we execute the modeller with the following command:

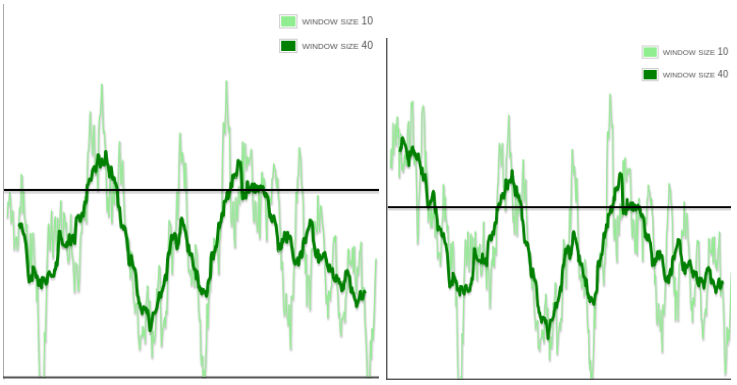
$mod10.5 modeling.py

Obtaining two different files in pdb format which are really similar

P25705.B99990001.pdb

P25705.B99990002.pdb

After passing our models through prosa we have found that the model is decently good, except for one extreme of the chain which peaks at a much higher level of energy than the rest. However we can work with this.

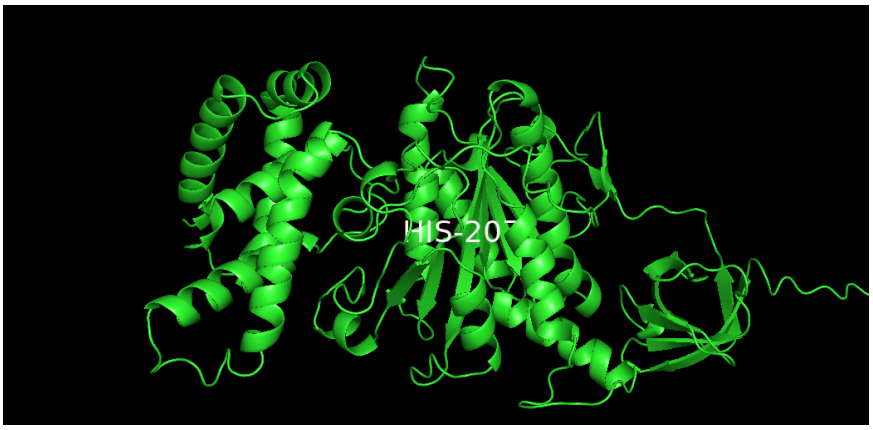


Template vs Model

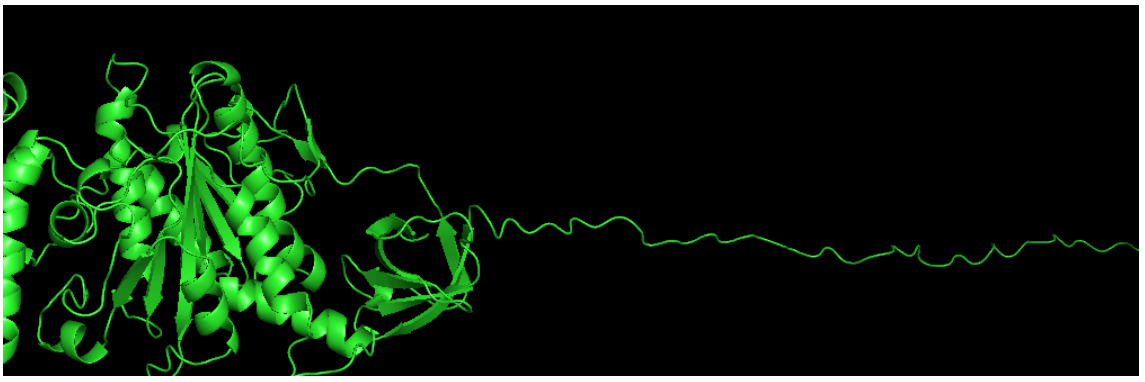
PYMOL COMPARISON:

Now that we have our protein model, we can load them into pymol, and start comparing them, even to see if we actually got a decent one, and try to make a hypothesis on how could this have an effect in protein function

First we will superimpose both models, so we can see the differences between them



This is our model, we used prosa to test its quality, the analysis result turned out quite good.
And after seeing the representation in pymol we can clearly see, what is the low quality zone.



(The long string that doesn't seem to belong with the rest of the structure)

To check for the differences between both models, the one where the mutation happened and the one where it didn't we decided to make a superimposition with the command:

super Template, mutation, object=aln1 with RMSD = 0.298

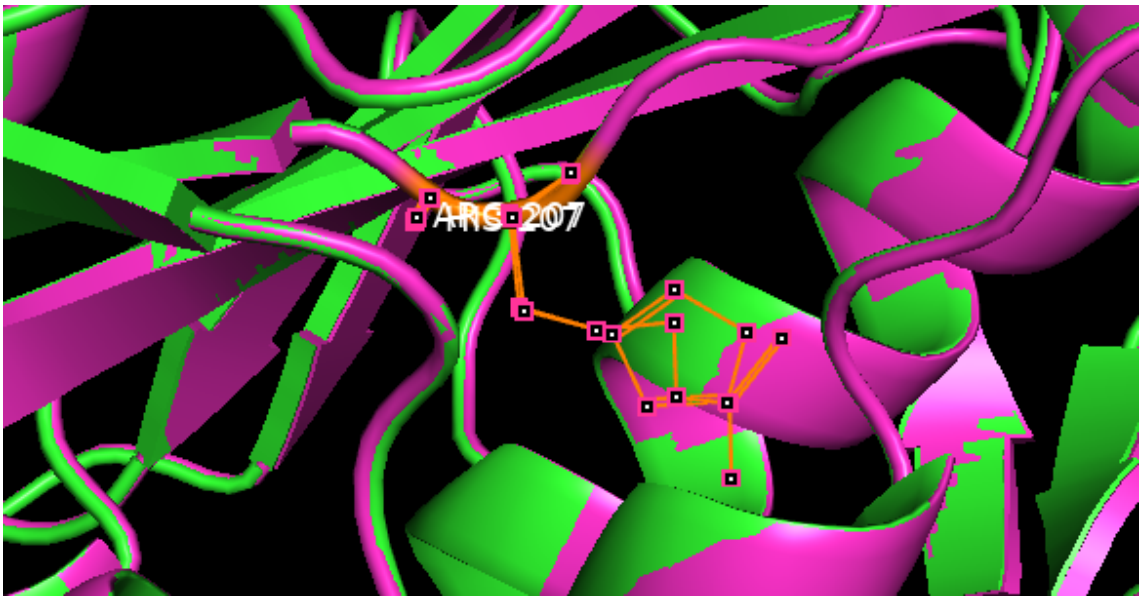
The fact that the RMSD is that low helps us see that our model was in fact good as the modeled mutation protein model still maintains a really low RMSD which needs to have taking into account that both structures should be nearly the same which gives us:

(pink = mutation, template = green)

Of course the mutation being only one amino acid makes the structure mostly the same. Which in general terms will mean that the protein should interact nearly the same.

To find where is our mutation located we need to go to:

Display → Sequence → and mark the position 207 in the sequence



After comparing both models in pymol, we can hypothesize how our mutation will affect the functionality of the protein.

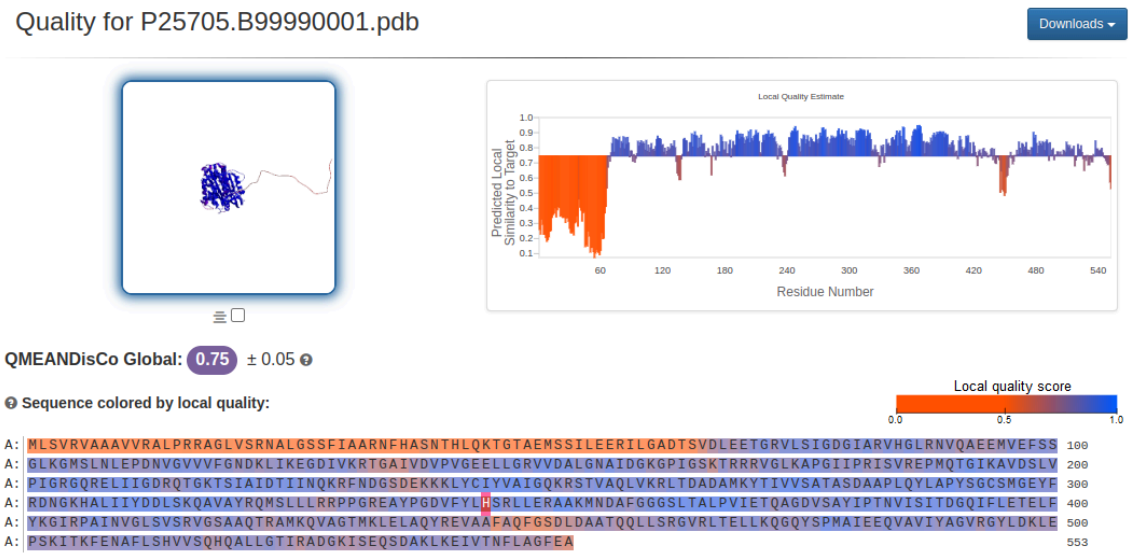
Since this mutation transforms an R to H, arginine to histidine, we can extract some conclusions. At first glance it could seem that nothing has really changed, due to the fact that both amino acids are positively charged, however as histidine has a ring, we can expect that this will affect the binding of DNA to the alpha helix which could cause some problems in the folding of the protein.

5. Use Qmeans to compare the energy profiles of the wild type protein with the structure of the mutant you modeled in the previous question. Is this mutation improving or worsening the energies of your protein? Make sure that the two proteins that you are comparing have similar lengths.

From the previous exercise we have obtained some pdb files as the output from the modeler. We will use Qmeans to evaluate the energy profile of the models we created using the modeler

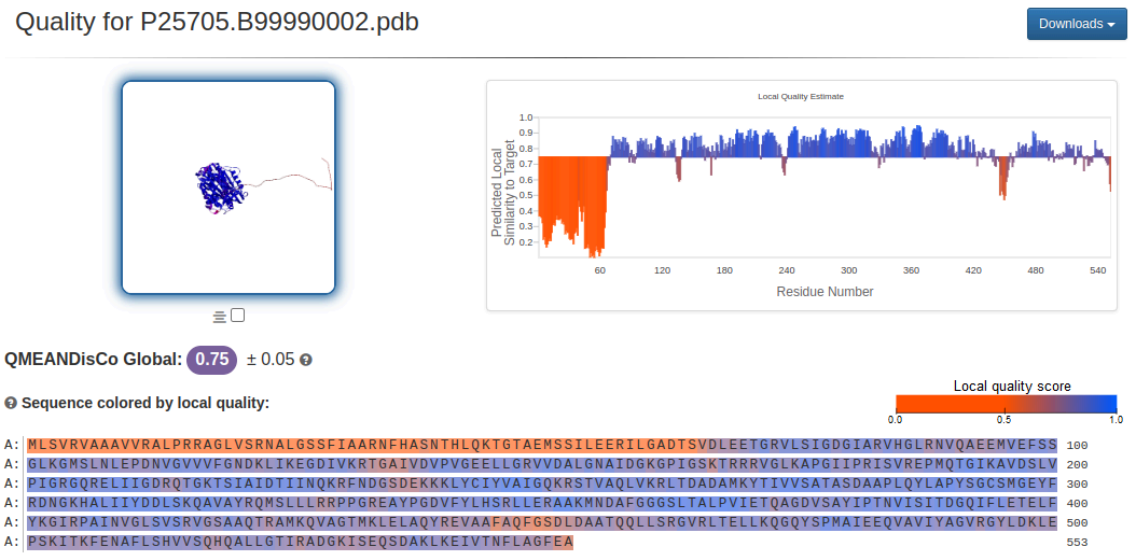
The first step is to upload the models into Qmeans and wait until we get the results.

Once we have the results let's start analyzing the wild-mutant model



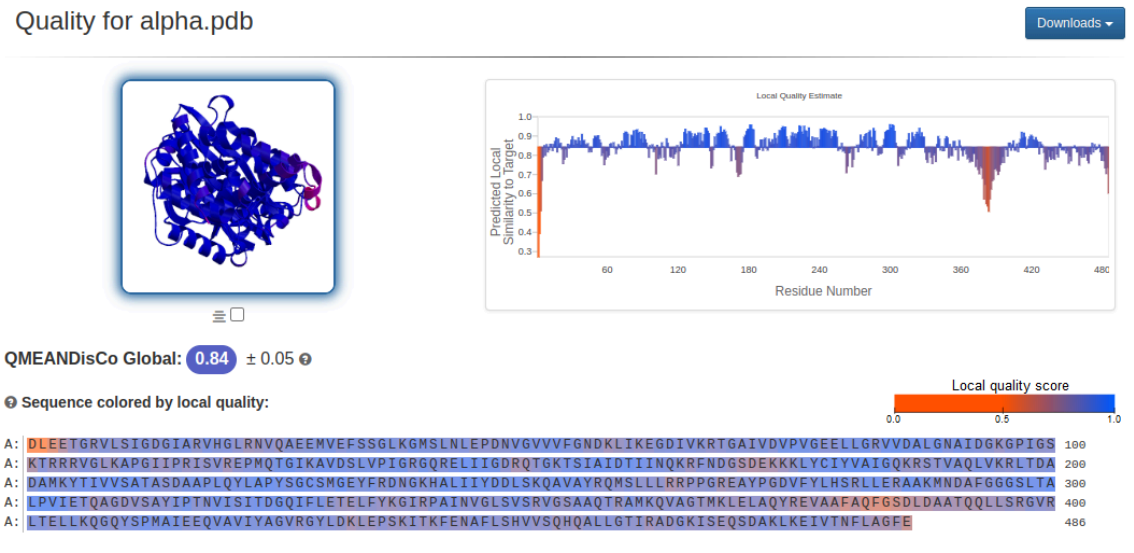
The global score of the correctness of the model according to statistical potential is 0.75, which despite not being the best possible score, we can consider our model to be quite successful. We can easily detect a part of the protein that is extremely badly modeled (between positions 0 and 60 aprox). Therefore if we wanted to create a better model an easy way to correct it would be to just eliminate the part that Qmean detects to be bad modeled from the input used for the modeler.

Now let's see what conclusions we can get from the 2nd model extracted from the mutant and wild protein.



We can see that this second model is extremely similar to the previous one. Therefore we can extract the same exact conclusions: the first amino acids (between 0-60 aprox) are bad modeled and that the model, in average is pretty good. We can also say that if we wanted to correct the model we could editate the alignment that the model uses as input manually by elimination the amino acids that are not well modeled

Now let's compare the statistic potentials of the wild structure with the mutated model but first we will analyze the wild protein

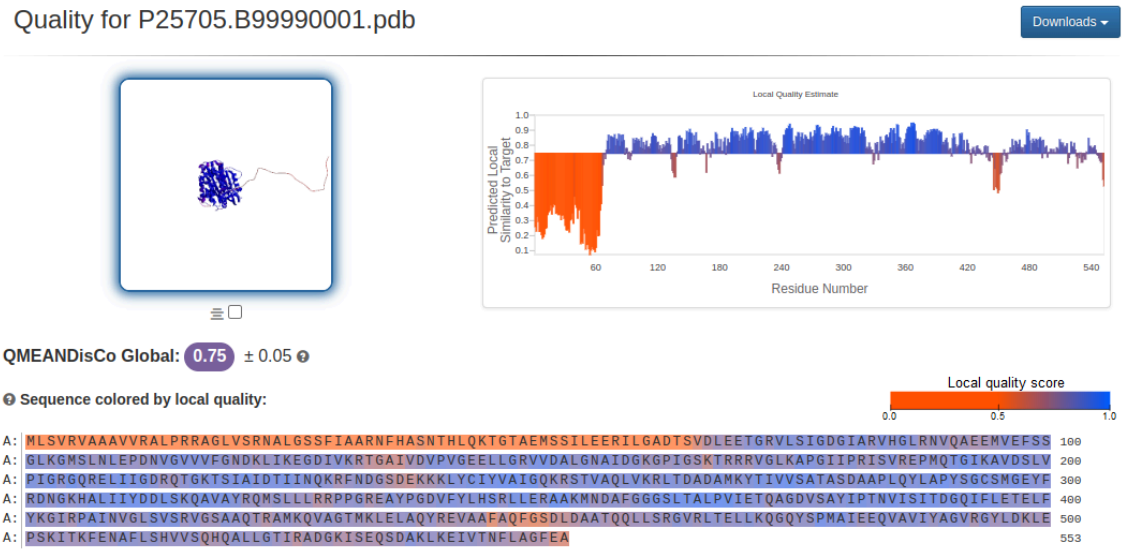


As we can see it's overall score is 0.84, which is better than the ones we obtained with the model. However, this can be due to the fact that our model has a part (between pos 0-60) that is obviously bad modeled. Therefore, if we eliminated that part of the model we would have a much better and similar model to it's original protein. Another difference between the model and the wild protein is that the models have little dawns

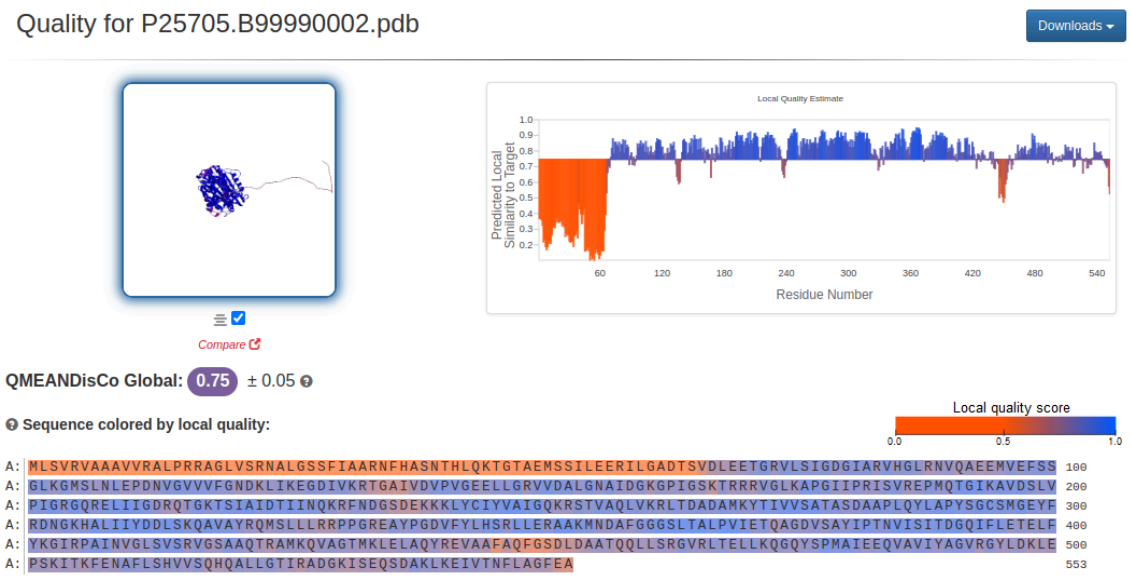
of energy that are slightly bigger than the wild type. This could be due to the fact that the mutation is slightly increasing the instability of the protein and therefore affecting it's potential energy.

An interesting observation we made is that most of the energy peaks of the wild type protein correspond to either one of the possible protein-protein interactions between some chains of ATP synthase and alpha or a ligand of alpha chain. This might be caused by the fact that when we study the energy profile of the alpha chain without having in account those interactions, the protein might be less stable and therefore Qmeans tells us there is a problem in a determined zone. It actually makes sense because the protein is prepared to be in it's most stable condition while maintaining all those interactions that make the protein do it's main function

Now let's take another different approach. We will try to create a model of the wild protein and then compare it to the mutant model. By doing so we might be able to discover more about the energy profile of our protein.



In the 1st model we can see that the model faced the same problem as in the mutant model, the 1st part of the protein until aa 60 approx is bad modeled. From it's overall score of 75 we can assume that the difference score between the mutant model and the wild protein is caused mainly by the bad modeling around positions 0-60 and not because of the mutation itself.



In this 2nd model created from the wild protein we can pretty much observe the same as in our previous model: A bad modeled part followed by a pretty good modeled part with some little bad modeling peaks.

The score ob 0.75 of both the mutated and wild model leads us to think that the mutation doesn't have a great impact on the protein function or structure.nucleotidos

Overall, we could conclude that our mutation doesn't have a huge impact on the 3d structure of our protein and the difference in between the scores of our mutant model and the wild protein are most probably caused by bad modeling and not the mutation we are studying. However, we have noticed something interesting: at around position 207, where the mutation is located there is a diminution of score on both the wild protein and the wild model but it still isn't below the threshold to consider it bad. Meanwhile, in our mutant model, at this exact position this diminution of score does indeed pass that threshold. This leads us to the final conclusion that despite the mutation not having a big impact on the overall protein structure it still does have some relatively small impact on its structure. Both amino acids involved in the mutation are basic and hydrophilic, and the main difference is R being slightly bigger and better at interactions with negatively charged molecules. This little difference in properties between both aminoacids also indicates us that the mutation has not a grat impact on the overall chain, which is congruent with the other conclusions we have made.