Structural bioinformatics Project – Assignment 2: Sequence analysis.

1. Does your protein have an HMM available in the PFAM database?

We were going to get the fasta of the protein sequence, name it ATP_target.fa and use the following command to get the PFAM entry of ATP_synthase.

$hmmscan ~/Documents/databases/Pfam-A.hmm ATP_target.fa > scanned_ATP.out

However, we couldn't find the full sequence to use as a query, so we used the Interprot database to search for ATP synthase in the PFAM database. The entry that better fit our needs was PF00231 (https://www.ebi.ac.uk/interpro/entry/pfam/PF00231/). We looked at its short name (ATP-synt), that is what we would be using as the name in the hmmfetch command.

$hmmfetch ~/Documents/databases/PFAM/Pfam-A.hmm ATP-synt_ab > alpha.hmm

As we could find an entry in the PFAM database for ATP Synthase, it means that we have an HMM available in the PFAM database.

2. Choose a set of 6 to 8 amino acid sequences that belong to the protein family you are studying. These sequences should represent the evolutionary history of your protein family, so you want them to have some diversity between them and avoid redundant or highly similar pairs of sequences. You will use these sequences to build a multiple sequence alignment. From what database should you retrieve these sequences? Why?

For the database that we choose to take our amino acid sequences, we've decided to go with uniprot. The reason behind this, is that compared to its closest contestant pdb, Uniprot offers us the available sequences for our desired protein (in this case ATP synthase) Non-biased and Non-redundant.

PDB is biased towards proteins easier to crystallize or study structurally. On the other hand, Uniprot represents proteins from diverse organisms and sources, including proteins from a wide range of species, which should be useful for determining evolutionary history ,and most importantly it attempts to eliminate redundancy by collapsing similar or identical protein sequences into a single entry. This feature will help us by saving us a lot of time in finding sequences that are similar/identical, and directly skipping to substantially different proteins.

Steps:

Go to UniProt and search for ATP synthase (homo sapiens), and look for the subunit F1 alpha, download the fasta sequence as target_alpha.fa do an alignment

$hmmscan ~/Documents/databases/PFAM/Pfam-A.hmm target_alpha.fa > alpha_HMM.out

We look into the .out file and pick the model with lowest E-value (highest score).

Model found: ATP-synt_ab

We use the ID of the model to fetch its corresponding hidden markov model.

$hmmfetch ~/Documents/databases/PFAM/Pfam-A.hmm ATP-synt_ab > alpha.hmm

Now that we have the model created we will use it to look for similar sequences from the same protein family , we will use the hmmsearch command.

$hmmsearch alpha.hmm ~/Documents/databases/UniProt/uniprot_sprot.fasta > prot_fam_ids.out

Results:
Q9UVJ8, P17255, P38078, P31406, P11592, Q5TTG1, P31400, P54647 → subunit F0 chain A, which we will not use

As Alberto said, to focus on the F1 subunit, we looked for results that appear on that chain, even if they have lower E-values (to account for diversity we didn't pick IDs that were next to each other, or were from the same species), these are the IDs we found.

Real results:

Q8TWL6, A3CS71, A8AUJ7, C6A5E8, A7IAU8, O29101, Q0W363, → for subunit F1 chain alpha.

We download the fasta files from the ID of the results and concatenate them into sequences_F1_chain_alpha.fa for use in the next exercises.

Now we do the same for the following parts of the protein:

subunit F1 beta:

B7IQV8, B7HFK1, Q814W2, A9VSA3, A0RL95, C1F0M8, C3LFH9, C3P1F4

3. Make a sequence alignment with the sequences you just obtained in the previous step. To create this alignment, use the HMM you found in PFAM and the programs from the HMMer package.

We will use the sequences in sequences.fasta that we created in the last exercise using

$cat Q8TWL6.fasta A3CS71.fasta A8AUJ7.fasta C6A5E8.fasta A7IAU8.fasta O29101.fasta Q0W363.fasta > sequences_F1_chain_alpha.fa

Then we will align the sequences against the hmm we created in exercise 1 using PFAM with the following command

$hmmalign alpha.hmm sequences_F1_chain_alpha.fa > MSA_F1_alpha.sto

Then we would transform the file from stockholm format to an alignment using perl

$perl ~/Documents/perl_scripts/aconvertMod2.pl -in h -out c <MSA_F1_alpha.sto > aln_F1_alpha.aln

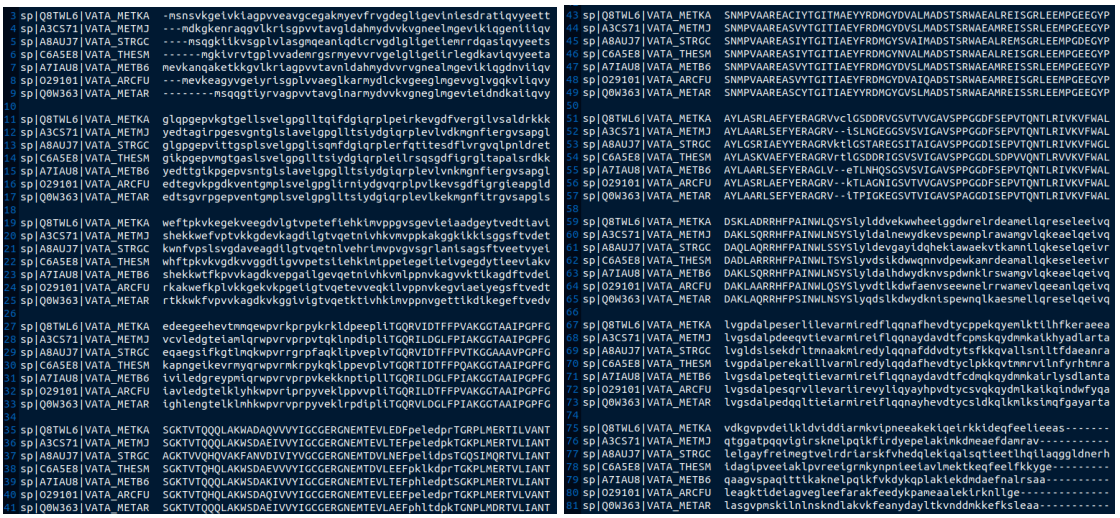
We repeat this process for all chains in all subunits and name each one in the appropriate manner.

4. Search for conserved regions in your alignment. Do these regions correspond with the essential regions you described in the previous assignment (question 6)? Why do you think this is happening? Provide images of your alignment to support your explanation. In these images, the alignments should be in clustalw format, use the perl script we learnt in practice 2 to change the format of the alignments produced by hmmer programs.

subunit F1 chain alpha:

For the alignment of subunit F1 chain alpha, we can observe that the beginning and ending of the sequence has 0 conservation, however the middle sequence has almost 100% conservation. From position 223 to 500 we can see that almost everything is conserved apart from very small parts of nucleotides that are not conserved (the biggest gap is of 7 aa and the rest are like 2 aa approximately).

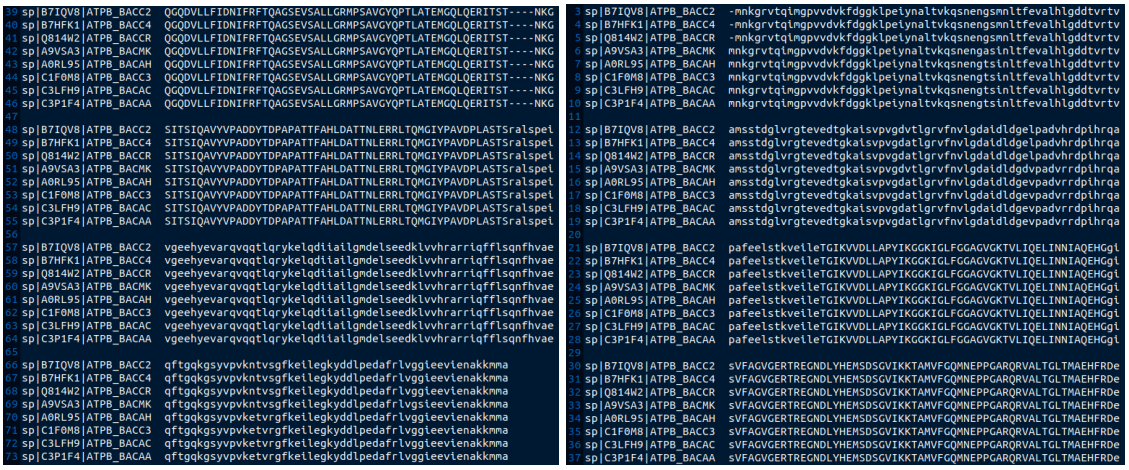
From that we can speculate that the region of F1 alpha is either the part of the protein that is involved in the synthesis of ATP from ADP and Pi, or a vital region in protein-protein interactions inside of the protein.



Subunit F1 chain beta:

After doing the alignment for subunit F1 beta we can see that the conserved region is in the middle, between position 144 and 383, being 1-144 and 383-503 being non conserved regions. From that we can speculate that the region of F1 beta 144-383 is either the part of the protein that is involved in the synthesis of ATP from ADP + Pi, or a vital region in

protein-protein interactions inside the protein. Essentially the same as alpha.



5. Work with the mutation you choose in the previous assignment (assignment 1, question 7). Find where this mutation would happen in the alignment you created in question 3. Compare the mutated amino acid with the amino acids that you find at that position in your alignment, do they share similar properties or not? Make a hypothesis of how this mutation is affecting the function of the protein. Provide images of your alignment to support your explanation.

In ex 7 we used subunit F0 a and for the alignments we used subunit F1 alpha and beta. Therefore, we would need to break our workflow in order to do this exercise. Also because we were recommended to focus on the workings of the protein we will put our efforts in that instead of looking for mutations. At the final project we might consider talking about mutations.

Disclaimer: in this part of the practice we mainly focused on learning how to deal with the whole protein. We had a bunch of problems because we didn't know how to deal with the fact that the protein subunits are partitioned).

Now that we have learned how to work correctly with our protein, we might include a study of how the proton gradient is done and how it affects our protein (subunit F0 C1, C2, C3)
