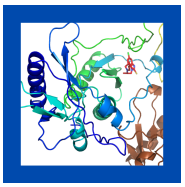

ESCI - UPF

Sequence analysis

Structural Bioinformatics

Marta Ortigas - Mònica Torner - Núria Mitjavila - Xavier Crespo

Second year
17/02/2023

Contents

1 Assignment 2 **3**
1.1 Sequence analysis 3

1 Assignment 2

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

1.1 Sequence analysis

1.- Does your protein have an HMM available in the PFAM database?
Yes, we find the best HMMER profile in PFAM for our target sequence that we got it from the 1REV PDB entry. To do this, we will select the best profile of our choice and fetch it from the database. Then, we will run a search of sequences in pdb_seq using this profile. To find the templates for our target in PFAM, we have 3 programs involved, hmmscan (finds what HMMs from a database match the input sequence), hmmfetch (extracts a HMM from a database) and hmmsearch (finds what sequences from a database match the input HMM). The following steps will be run in cluster, after importing the HMM module (HMMER/3.2.1-foss-2020a) and our sequence. First, we will execute hmmscan on the pfam database, assigning the best profile(s) to the target sequence:

```
hmmscan /shared/databases/pfam-3/Pfam-A.hmm 1rev_uniprot.fasta > 1rev.out
```

```
# hmmscan :: search sequence(s) against a profile database
# HMMER 3.2.1 (June 2018); http://hmmer.org/
# Copyright (C) 2018 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query sequence file:             1rev_uniprot.fasta
# target HMM database:             /shared/databases/pfam-3/Pfam-A.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       sp|P04585|POL_HV1H2  [L=1435]
Description: Gag-Pol polyprotein OS=Human immunodeficiency virus type 1 group M subtype B (isolate HXB2) OX=11706 GN=gag-pol PE=1 SV=4
Scores for complete sequence (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Model        Description
    ------- ------ -----    ------- ------ -----   ---- --  --------     -----------
    9.5e-80  266.8   0.0    2.5e-79  265.4   0.0    1.8  1  Gag_p24       gag gene protein p24 (core nucleocapsid protein
    2.3e-65  218.5   1.6    4.9e-65  217.4   0.6    2.2  2  Gag_p17       gag gene protein p17 (matrix protein)
    5.9e-45  151.2  10.2    5.9e-45  151.2  10.2    2.7  2  RVT_connect   Reverse transcriptase connection domain
      2e-37  126.4   1.2      2e-37  126.4   1.2    2.5  2  RVT_thumb     Reverse transcriptase thumb domain
    3.5e-34  117.8   0.6    2.1e-33  115.3   0.1    2.5  2  RNase_H       RNase H
      5e-33  112.8   0.3    1.9e-32  111.0   0.3    2.1  1  RVP           Retroviral aspartyl protease
    9.4e-32  109.9   0.0    2.2e-31  108.7   0.0    1.7  1  RVT_1         Reverse transcriptase (RNA-dependent DNA polyme
      5e-24   84.5   0.0    1.1e-23   83.5   0.0    1.6  1  rve           Integrase core domain
```

Figure 1: Portion of the output file 1REV.out.

Then inspect the output of hmmscan to find which HMM from PFAM fits the best with our target sequence we can observe that the models are Gag_p24, Gag_p17, RVT_connect, RVT_thumb, ... being Gag_p24 the one that fits the best with our sequence. Once we have found it, we copied the name that this HMM has in PFAM, and we will use it to fetch the HMM from PFAM using the hmmfetch program. We will extract the profile from PFAM that correspond to the Gag_p24 of the target sequence.

```
hmmfetch /shared/databases/pfam-3/Pfam-A.hmm Gag_p24 > Gag_p24.hmm
```

```
HMMER3/f [3.2.1 | June 2018]
NAME  Gag_p24
ACC   PF00607.15
DESC  gag gene protein p24 (core nucleocapsid protein)
LENG  206
ALPH  amino
RF    no
MM    no
CONS  yes
CS    yes
MAP   yes
```

Figure 2: Portion of the output file Gag_p24.hmm.

Finally, we will search for sequences that contain the same domain as our target using hmmsearch:

```
hmmsearch Gag_p24.hmm /shared/databases/blastdat/uniprot_sprot >
↪  Gag_p24_search.out
```

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.2.1 (June 2018); http://hmmer.org/
# Copyright (C) 2018 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  Gag_p24.hmm
# target sequence database:        /shared/databases/blastdat/uniprot_sprot
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       Gag_p24  [M=206]
Accession:   PF00607.15
Description: gag gene protein p24 (core nucleocapsid protein)
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence              Description
    ------- ------ -----    ------- ------ -----   ---- --  --------              -----------
    9.1e-81  275.4   0.0    1.8e-80  274.4   0.0    1.5  1  sp|Q02843|GAG_SIVG1   Gag polyprotein OS=Simian immunodefici
    4.5e-80  273.1   0.0    1.2e-79  271.7   0.0    1.8  1  sp|Q02836|POL_SIVG1   Gag-Pol polyprotein OS=Simian immunode
    1.6e-79  271.3   0.5    4.3e-78  266.6   0.0    2.1  2  sp|P03349|GAG_HV1A2   Gag polyprotein OS=Human immunodeficie
    5.8e-79  269.5   0.2    1.4e-78  268.2   0.0    1.7  1  sp|P04591|GAG_HV1H2   Gag polyprotein OS=Human immunodeficie
      6e-79  269.4   0.2    1.4e-78  268.2   0.0    1.7  1  sp|Q70622|GAG_HV1LW   Gag polyprotein OS=Human immunodeficie
      6e-79  269.4   0.2    1.5e-78  268.1   0.0    1.7  1  sp|P03347|GAG_HV1B1   Gag polyprotein OS=Human immunodeficie
```

Figure 3: Portion of hmmsearch for Gag_p24_search.out.

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?
From the output of the hmmsearch we took the following accession numbers, making sure that they were not from the same species:

[Q02843 Q02836 P03349 P04591 Q70622 P03347]

The database we used to retrieve this sequences is UniProt, as it contains a large set of cured sequences and non-redundant data.

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.

```
hmmalign Gag_p24.hmm 1rev_family.fasta > 1rev_family.sto
# Now transform it to Clustal:
perl /shared/PERL/aconvertMod2.pl -in h -out c
↪  <1rev_family.sto>1rev_family.aln
```

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.

In the previous assignment, we choose Ribonuclease H (RNase H) as the region that we thought was essential, and now we see that it is not conserved for all sequences.

The corresponding part of the RNase for the reverse transcriptase can be seen in lowercase letters and only followed by Q02836. Going to its UniProt page, we see that it expresses the RNase, among other things. The others do not express it.

Also, we can see that all the entries do not share the same functions and contain the same information. For this, maybe RNase it is not very important in the protein family, but very important in those proteins that are capable of retro-transcribing.

```
sp|Q02843|GAG_SIVG1  knettappggesrnypvvnqnnawvhqplsPRTLNAWVKCVEE-KRWGAEVVPMFQALSE
sp|Q02836|POL_SIVG1  knettappggesrnypvvnqnnawvhqplsPRTLNAWVKCVEE-KRWGAEVVPMFQALSE
sp|P03349|GAG_HV1A2  aagtgnssqvsqnypivqnlqgqmvhqaisPRTLNAWVKVVEE-KAFSPEVIPMFSALSE
sp|P04591|GAG_HV1H2  aadtghsnqvsqnypivqniqgqmvhqaisPRTLNAWVKVVEE-KAFSPEVIPMFSALSE
sp|Q70622|GAG_HV1LW  aadtghssqvsqnypivqniqgqmvhqaisPRTLNAWVKVVEE-KAFSPEVIPMFSALSE
sp|P03347|GAG_HV1B1  aadtghssqvsqnypivqniqgqmvhqaisPRTLNAWVKVVEE-KAFSPEVIPMFSALSE
sp|P04585|POL_HV1H2  aadtghsnqvsqnypivqniqgqmvhqaisPRTLNAWVKVVEE-KAFSPEVIPMFSALSE

sp|Q02843|GAG_SIVG1  GCLSYDVNQMLNVIGDHQGALQILKEVINEEAAEWDRTHRPPaGPLPagqlrdpTGSDIA
sp|Q02836|POL_SIVG1  GCLSYDVNQMLNVIGDHQGALQILKEVINEEAAEWDRTHRPPaGPLPagqlrdpTGSDIA
sp|P03349|GAG_HV1A2  GATPQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHaGPIApgqmrepRGSDIA
sp|P04591|GAG_HV1H2  GATPQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHaGPIApgqmrepRGSDIA
sp|Q70622|GAG_HV1LW  GATPQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHaGPIApgqmrepRGSDIA
sp|P03347|GAG_HV1B1  GATPQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHaGPIApgqmrepRGSDIA
sp|P04585|POL_HV1H2  GATPQDLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHaGPIApgqmrepRGSDIA

sp|Q02843|GAG_SIVG1  GTTSSIQEQIEWTFNanprIDVGAQYRKWVILGLQKVVQMYNPQ-KVLDIRQGPKEPFQD
sp|Q02836|POL_SIVG1  GTTSSIQEQIEWTFNanprIDVGAQYRKWVILGLQKVVQMYNPQ-KVLDIRQGPKEPFQD
sp|P03349|GAG_HV1A2  GTTSTLQEQIGWMTNn-ppIPVGEIYKRWIILGLNKIVRMYSPT-SILDIRQGPKEPFRD
sp|P04591|GAG_HV1H2  GTTSTLQEQIGWMTNn-ppIPVGEIYKRWIILGLNKIVRMYSPT-SILDIRQGPKEPFRD
sp|Q70622|GAG_HV1LW  GTTSTLQEQIGWMTNn-ppIPVGEIYKRWIILGLNKIVRMYSPT-SILDIRQGPKEPFRD
sp|P03347|GAG_HV1B1  GTTSTLQEQIGWMTNn-ppIPVGEIYKRWIILGLNKIVRMYSPT-SILDIRQGPKEPFRD
sp|P04585|POL_HV1H2  GTTSTLQEQIGWMTNn-ppIPVGEIYKRWIILGLNKIVRMYSPT-SILDIRQGPKEPFRD
```

Figure 4: Portion of the alignment between 1REV and its homologous.

You can find the whole alignment in the following link:
`https://github.com/XavierUPF/Structural-Bioinformatics-Project-BDBI/blob/main/generated_files/1rev_family.aln`

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.

As in the previous assignment, we are working instead with a drug.

We discussed Emtricitabine in a theoretical basis. But, given there is no actual structure of reverse transcriptase in complex with Emtricitabine, we are now focusing on Tenofovir, a drug that works in the same way by mimicking a natural nucleotide [A] that is used by the reverse transcriptase enzyme to synthesize the viral DNA. Once incorporated into the viral DNA chain, tenofovir blocks the action of the reverse transcriptase enzyme, preventing further elongation of the DNA chain and ultimately inhibiting viral replication. The difference is that this drug has an available structure in PDB with our protein. [1T05].

Tenofovir is a nucleotide reverse transcriptase inhibitor. NRTIs, such as tenofovir, are designed to take advantage of the error-prone nature of reverse transcriptase. They mimic the natural nucleotides that are used to synthesize the viral DNA, but with a slight modification, as seen in the image above, that prevents further DNA elongation.

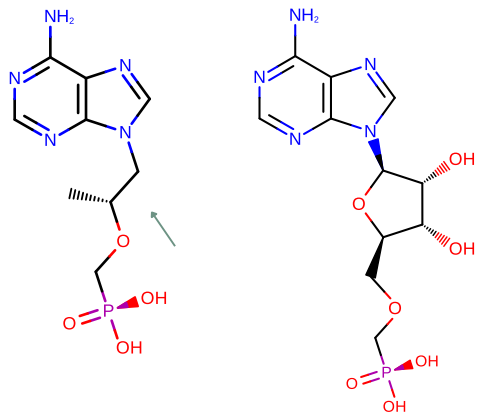

Figure 5: Here you can see that Tenefovir (the one in the left) lacks the pentose yet it has a posphate group. As a consecuence of lacking the pentose, nothing no other base can attach below it because there is no 3' hydroxyl group.

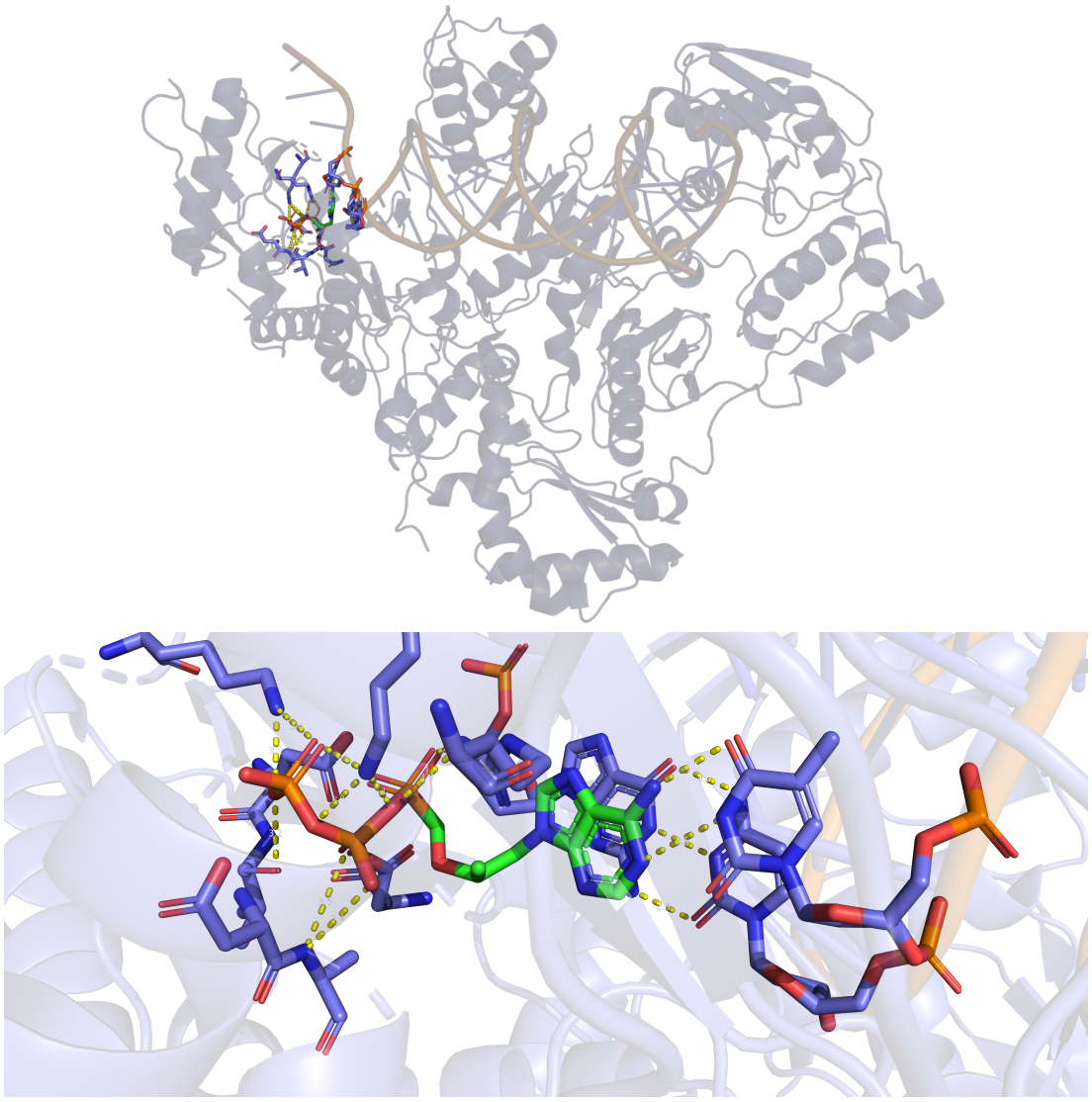

Figure 6: Here you can see that Tenefovir (the one in the left) lacks the pentose yet it has a posphate group. As a consecuence of lacking the pentose, nothing no other base can attach below it because there is no 3' hydroxyl group.