Assignment 2

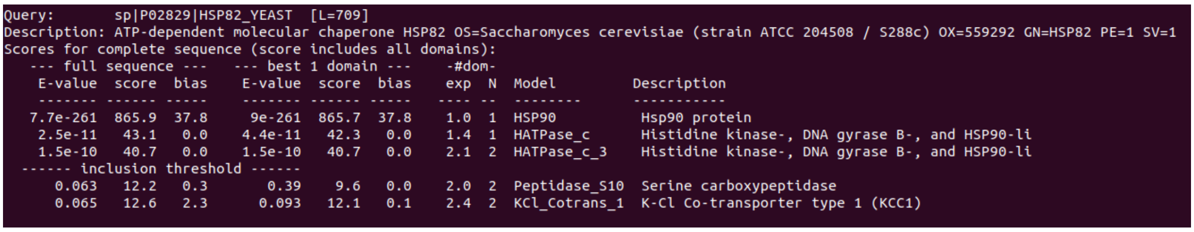
Nico Ares, Dante Aviñó, Miguel Borge, Isaac Soul

Structural bioinformatics Project – Assignment 2: Sequence analysis.

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

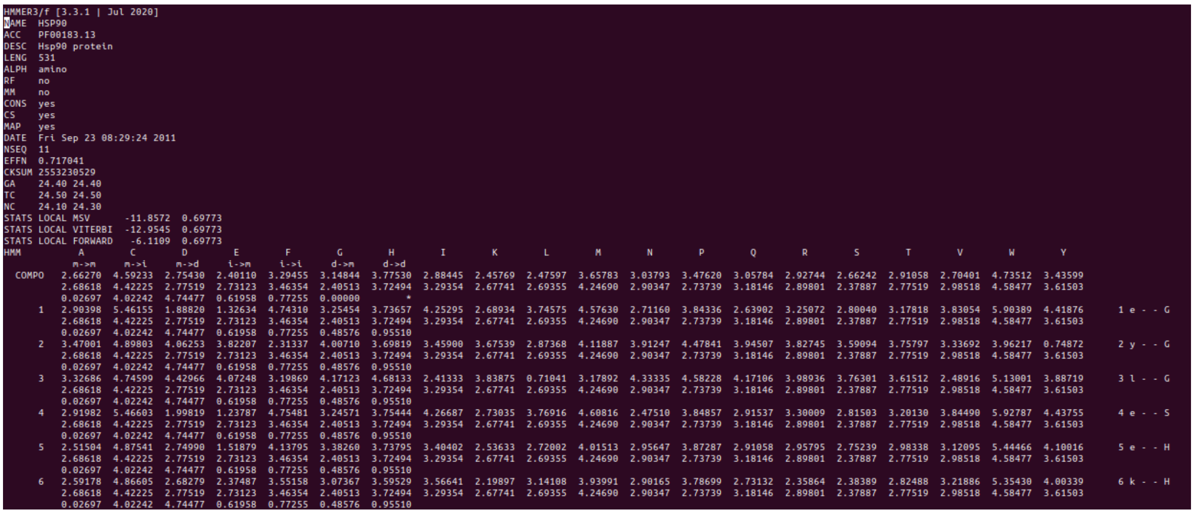
Our 2CG9 protein didn't have a valid fasta file in the pdb database since it had no gaps. To deal with this issue we obtained a valid fasta file from uniprot through the following accession name: UniProtKB - P02829 (HSP82_YEAST). We then used the pfam database to obtain the best sequence alignment for our desired protein through the hmmscan command from the HMMER package. Once we had the output file for the hmmscan, we observed that the best model for our protein was the "HSP90", which we used to obtain the hidden markov model with the hmmfetch command.

```
# Searching in Pfam database HMMs fitting 2CG9 protein sequence:
hmmscan /shared/databases/pfam-3/Pfam-A.hmm 2CG9_uniprot.fasta > 2CG9_uniprot.out
```



```
Query:       sp|P02829|HSP82_YEAST  [L=709]
Description: ATP-dependent molecular chaperone HSP82 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) OX=559292 GN=HSP82 PE=1 SV=1
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
    ------- ------ -----    ------- ------ -----   ---- --  --------      -----------
   7.7e-261  865.9  37.8     9e-261  865.7  37.8    1.0  1  HSP90          Hsp90 protein
    2.5e-11   43.1   0.0    4.4e-11   42.3   0.0    1.4  1  HATPase_c      Histidine kinase-, DNA gyrase B-, and HSP90-li
    1.5e-10   40.7   0.0    1.5e-10   40.7   0.0    2.1  2  HATPase_c_3    Histidine kinase-, DNA gyrase B-, and HSP90-li
  ------ inclusion threshold ------
      0.063   12.2   0.3       0.39    9.6   0.0    2.0  2  Peptidase_S10  Serine carboxypeptidase
      0.065   12.6   2.3      0.093   12.1   0.1    2.4  2  KCl_Cotrans_1  K-Cl Co-transporter type 1 (KCC1)
```

```
# Extracting profiles from Pfam corresponding to HSP90 protein family domains:
hmmfetch /shared/databases/pfam-3/Pfam-A.hmm "HSP90" > HSP90_domain.hmm
```



```
HMMER3/f [3.3.1 | Jul 2020]
NAME  HSP90
ACC   PF00183.13
DESC  Hsp90 protein
LENG  531
ALPH  amino
RF    no
MM    no
CONS  yes
CS    yes
MAP   yes
DATE  Fri Sep 23 08:29:24 2011
NSEQ  11
EFFN  0.717041
CKSUM 2553230529
GA    24.40 24.40
TC    24.50 24.50
NC    24.10 24.30
STATS LOCAL MSV      -11.8572  0.69773
STATS LOCAL VITERBI  -12.9545  0.69773
STATS LOCAL FORWARD   -6.1109  0.69773
HMM          A        C        D        E        F        G        H        I        K        L        M        N        P        Q        R        S        T        V        W        Y
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.66270  4.59233  2.75430  2.40110  3.29455  3.14844  3.77530  2.88445  2.45769  2.47597  3.65783  3.03793  3.47620  3.05784  2.92744  2.66242  2.91058  2.70401  4.73512  3.43599
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02697  4.02242  4.74477  0.61958  0.77255  0.00000        *
      1   2.90398  5.46155  1.88820  1.32634  4.74310  3.25454  3.73657  4.25295  2.68934  3.74575  4.57630  2.71160  3.84336  2.63902  3.25072  2.80040  3.17818  3.83054  5.90389  4.41876      1 e - - G
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02697  4.02242  4.74477  0.61958  0.77255  0.48576  0.95510
      2   3.47001  4.89803  4.06253  3.82207  2.31337  4.00710  3.69819  3.45900  3.67539  2.87368  4.11887  3.91247  4.47841  3.94507  3.82745  3.59094  3.75797  3.33692  3.96217  0.74872      2 y - - G
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02697  4.02242  4.74477  0.61958  0.77255  0.48576  0.95510
      3   3.32686  4.74599  4.42966  4.07248  3.19869  4.17123  4.68133  2.41333  3.83875  0.71041  3.17892  4.33335  4.58228  4.17106  3.98936  3.76301  3.61512  2.48916  5.13001  3.88719      3 l - - G
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02697  4.02242  4.74477  0.61958  0.77255  0.48576  0.95510
      4   2.91982  5.46603  1.99819  1.23787  4.75481  3.24571  3.75444  4.26687  2.73035  3.76916  4.60816  2.47510  3.84857  2.91537  3.30009  2.81503  3.20130  3.84490  5.92787  4.43755      4 e - - S
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02697  4.02242  4.74477  0.61958  0.77255  0.48576  0.95510
      5   2.51504  4.87541  2.74990  1.51879  4.13795  3.38260  3.73795  3.40402  2.53633  2.72002  4.01513  2.95647  3.87287  2.91058  2.95795  2.75239  2.98338  3.12095  5.44466  4.10016      5 e - - H
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02697  4.02242  4.74477  0.61958  0.77255  0.48576  0.95510
      6   2.59178  4.86605  2.68279  2.37487  3.55158  3.07367  3.59529  3.56641  2.19897  3.14108  3.93991  2.90165  3.78699  2.73132  2.35864  2.38389  2.82488  3.21886  5.35430  4.00339      6 k - - H
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355  4.24690  2.90347  2.73739  3.18146  2.89801  2.37887  2.77519  2.98518  4.58477  3.61503
          0.02697  4.02242  4.74477  0.61958  0.77255  0.48576  0.95510
```

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

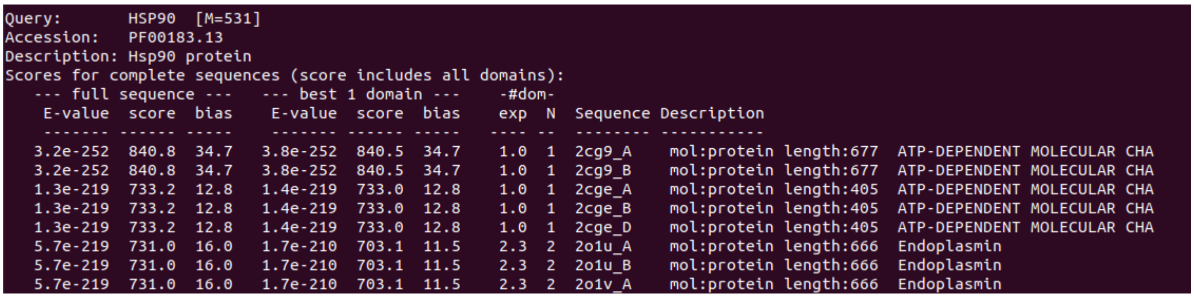
sequence alignment. From what database should you retrieve these sequences? Why?

We have used the hmmsearch command, using the hidden markov model for our protein target (2CG9), and we've compared our results for pdb and uniprot.

Before acknowledging the results, we know that PDB is very redundant since it has the same proteins repeated several times and that it is very biased. Some protein families are overrepresented due to medical significance, easier crystallography, etc. and other families are underrepresented. Also if we create a PSSM with a biased database our PSSM will be biased too.

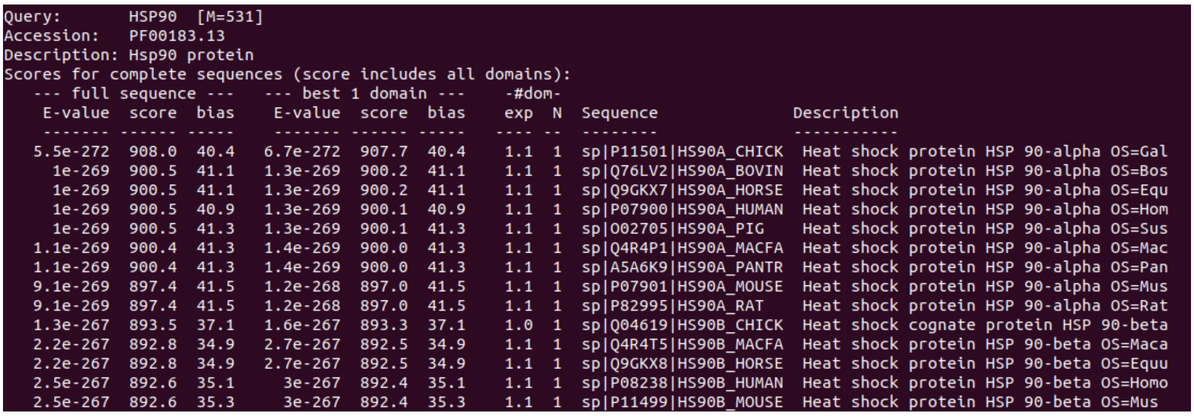
Now by looking at the obtained results we've seen that our knowledge on PDB being redundant and biased is certain since we have very high bias scores and lots of repeated proteins. On the other hand, uniprot gives a less biased percentage and also a higher score so therefore we've chosen to retrieve our proteins for which we're going to perform the multiple sequence alignment from uniprot.

```
# Query on the PDB Database
hmmsearch hsp90_pdb.hmm /shared/databases/blastdat/pdb_seq > hsp90_pdb_2.out
```



```
Query:       HSP90  [M=531]
Accession:   PF00183.13
Description: Hsp90 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
    ------- ------ -----    ------- ------ -----   ---- --  -------- -----------
   3.2e-252  840.8  34.7   3.8e-252  840.5  34.7    1.0  1  2cg9_A    mol:protein length:677  ATP-DEPENDENT MOLECULAR CHA
   3.2e-252  840.8  34.7   3.8e-252  840.5  34.7    1.0  1  2cg9_B    mol:protein length:677  ATP-DEPENDENT MOLECULAR CHA
   1.3e-219  733.2  12.8   1.4e-219  733.0  12.8    1.0  1  2cge_A    mol:protein length:405  ATP-DEPENDENT MOLECULAR CHA
   1.3e-219  733.2  12.8   1.4e-219  733.0  12.8    1.0  1  2cge_B    mol:protein length:405  ATP-DEPENDENT MOLECULAR CHA
   1.3e-219  733.2  12.8   1.4e-219  733.0  12.8    1.0  1  2cge_D    mol:protein length:405  ATP-DEPENDENT MOLECULAR CHA
   5.7e-219  731.0  16.0   1.7e-210  703.1  11.5    2.3  2  2o1u_A    mol:protein length:666  Endoplasmin
   5.7e-219  731.0  16.0   1.7e-210  703.1  11.5    2.3  2  2o1u_B    mol:protein length:666  Endoplasmin
   5.7e-219  731.0  16.0   1.7e-210  703.1  11.5    2.3  2  2o1v_A    mol:protein length:666  Endoplasmin
```

```
# Query on the UniProt Database
hmmsearch hsp90_pdb.hmm /shared/databases/blastdat/uniprot_sprot >
hsp90_uniprot.out
```



```
Query:       HSP90  [M=531]
Accession:   PF00183.13
Description: Hsp90 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
    ------- ------ -----    ------- ------ -----   ---- --  --------               -----------
   5.5e-272  908.0  40.4   6.7e-272  907.7  40.4    1.1  1  sp|P11501|HS90A_CHICK  Heat shock protein HSP 90-alpha OS=Gal
     1e-269  900.5  41.1   1.3e-269  900.2  41.1    1.1  1  sp|Q76LV2|HS90A_BOVIN  Heat shock protein HSP 90-alpha OS=Bos
     1e-269  900.5  41.1   1.3e-269  900.2  41.1    1.1  1  sp|Q9GKX7|HS90A_HORSE  Heat shock protein HSP 90-alpha OS=Equ
     1e-269  900.5  40.9   1.3e-269  900.1  40.9    1.1  1  sp|P07900|HS90A_HUMAN  Heat shock protein HSP 90-alpha OS=Hom
     1e-269  900.5  41.3   1.3e-269  900.1  41.3    1.1  1  sp|O02705|HS90A_PIG    Heat shock protein HSP 90-alpha OS=Sus
   1.1e-269  900.4  41.3   1.4e-269  900.0  41.3    1.1  1  sp|Q4R4P1|HS90A_MACFA  Heat shock protein HSP 90-alpha OS=Mac
   1.1e-269  900.4  41.3   1.4e-269  900.0  41.3    1.1  1  sp|A5A6K9|HS90A_PANTR  Heat shock protein HSP 90-alpha OS=Pan
   9.1e-269  897.4  41.5   1.2e-268  897.0  41.5    1.1  1  sp|P07901|HS90A_MOUSE  Heat shock protein HSP 90-alpha OS=Mus
   9.1e-269  897.4  41.5   1.2e-268  897.0  41.5    1.1  1  sp|P82995|HS90A_RAT    Heat shock protein HSP 90-alpha OS=Rat
   1.3e-267  893.5  37.1   1.6e-267  893.3  37.1    1.0  1  sp|Q04619|HS90B_CHICK  Heat shock cognate protein HSP 90-beta
   2.2e-267  892.8  34.9   2.7e-267  892.5  34.9    1.1  1  sp|Q4R4T5|HS90B_MACFA  Heat shock protein HSP 90-beta OS=Maca
   2.2e-267  892.8  34.9   2.7e-267  892.5  34.9    1.1  1  sp|Q9GKX8|HS90B_HORSE  Heat shock protein HSP 90-beta OS=Equu
   2.5e-267  892.6  35.1     3e-267  892.4  35.1    1.1  1  sp|P08238|HS90B_HUMAN  Heat shock protein HSP 90-beta OS=Homo
   2.5e-267  892.6  35.3     3e-267  892.4  35.3    1.1  1  sp|P11499|HS90B_MOUSE  Heat shock protein HSP 90-beta OS=Mus
```

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.

Sequences, obtained from the hmmsearch command, from which we create the multiple sequence alignment fasta file using the cat file.fa >> output_file.fa command:

```
# Joining all fasta sequences in a single file, putting our target fasta sequence (P02829.fasta) as the first one:
cat P02829.fasta > FINAL.fasta
cat B8IU50.fasta > FINAL.fasta
cat P04811.fasta > FINAL.fasta
...
```

sp|B8IU50|HTPG_METNO Chaperone protein htpG OS=Methylobacterium nodulans GN=htpG

sp|P04811|HSP83_DROVI Heat shock protein 83 (Fragment) OS=Drosophila virilis GN=Hsp83

sp|P11501|HS90A_CHICK Heat shock protein HSP 90-alpha OS=Gallus gallus GN=HSP90AA1

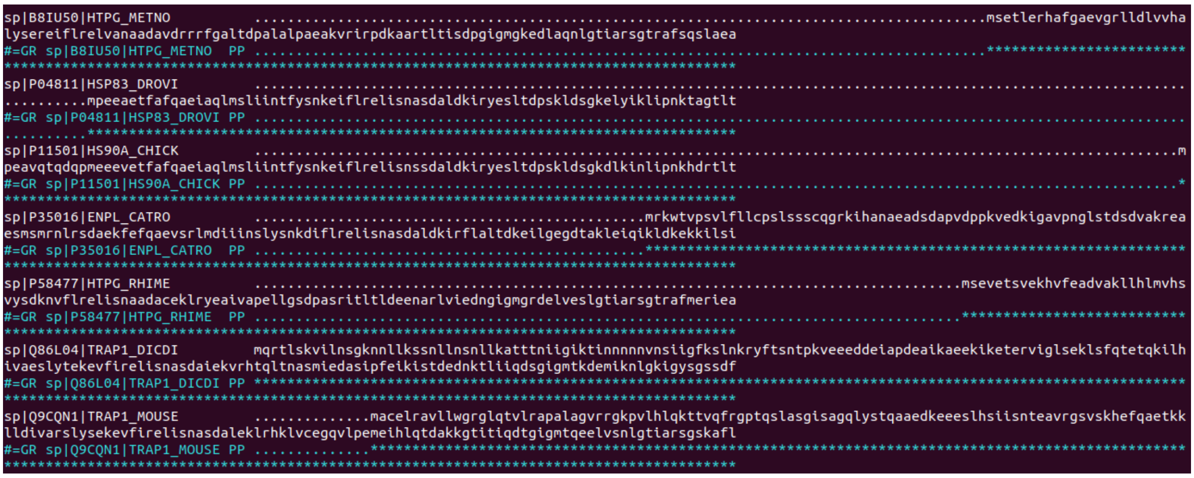
sp|P35016|ENPL_CATRO Endoplasmin homolog OS=Catharanthus roseus GN=HSP90

sp|P58477|HTPG_RHIME Chaperone protein htpG OS=Rhizobium meliloti (strain 1021) GN=htpG

sp|Q86L04|TRAP1_DICDI TNF receptor-associated protein 1 homolog, mitochondrial OS=Dictyostelium discoideum GN=trap1

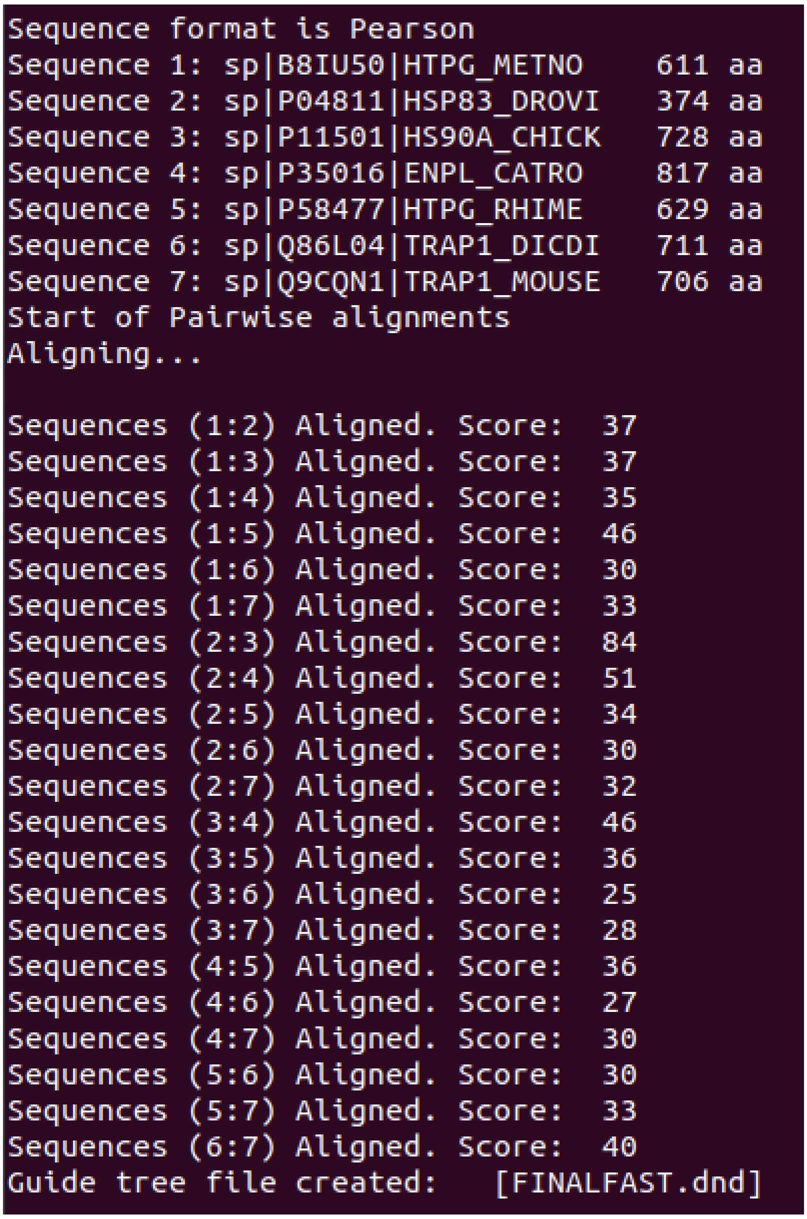
sp|Q9CQN1|TRAP1_MOUSE Heat shock protein 75 kDa, mitochondrial OS=Mus musculus GN=Trap1

With our MSA file (FINALf.fasta) we perform the hmmalign command (hmmalign HSP90_domain.hmm FINALf.fasta > HSP90_hmm.sto):



```
sp|B8IU50|HTPG_METNO          ........................................................................................msetlerhafgaevgrlldlvvha
lysereiflrelvanaadavdrrrfgaltdpalalpaeakvrirpdkaartltisdpgigmgkedlaqnlgtiarsgtrafsqslaea
#=GR sp|B8IU50|HTPG_METNO  PP ........................................................................................************************
****************************************************************************************
sp|P04811|HSP83_DROVI         ................................................................................................................
..........mpeeaetfafqaeiaqlmsliintfysnkeiflrelisnasdaldkiryesltdpskldsgkelyiklipnktagtlt
#=GR sp|P04811|HSP83_DROVI PP ................................................................................................................
..........******************************************************************************
sp|P11501|HS90A_CHICK         ...............................................................................................................m
peavqtqdqpmeeevetfafqaeiaqlmsliintfysnkeiflrelisnssdaldkiryesltdpskldsgkdlkinlipnkhdrtlt
#=GR sp|P11501|HS90A_CHICK PP ...............................................................................................................*
****************************************************************************************
sp|P35016|ENPL_CATRO          ...............................................mrkwtvpsvlfllcpslssscqgrkihanaeadsdapvdppkvedkigavpnglstdsdvakrea
esmsmrnlrsdaekfefqaevsrlmdiiinslysnkdiflrelisnasdaldkirflaltdkeilgegdtakleiqikldkekkilsi
#=GR sp|P35016|ENPL_CATRO  PP ...............................................*****************************************************************
****************************************************************************************
sp|P58477|HTPG_RHIME          .....................................................................................msevetsvekhvfeadvakllhlmvhs
vysdknvflrelisnaadaceklryeaivapellgsdpasritltldeenarlviedngigmgrdelveslgtiarsgtrafmeriea
#=GR sp|P58477|HTPG_RHIME  PP .....................................................................................***************************
****************************************************************************************
sp|Q86L04|TRAP1_DICDI         mqrtlskvilnsgknnllkssnllnsnllkatttniigiktinnnnnvnsiigfkslnkryftsntpkveeeddeiapdeaikaeekiketerviglseklsfqtetqkilh
ivaeslytekevfirelisnasdaiekvrhtqltnasmiedasipfeikistdednktliiqdsgigmtkdemiknlgkigysgssdf
#=GR sp|Q86L04|TRAP1_DICDI PP ****************************************************************************************************************
****************************************************************************************
sp|Q9CQN1|TRAP1_MOUSE         ...............macelravllwgrglqtvlrapalagvrrgkpvlhlqkttvqfrgptqslasgisagqlystqaaedkeeslhsiisnteavrgsvskhefqaetkk
lldivarslysekevfirelisnasdaleklrhklvcegqvlpemeihlqtdakkgtitiqdtgigmtqeelvsnlgtiarsgskafl
#=GR sp|Q9CQN1|TRAP1_MOUSE PP ...............*************************************************************************************************
****************************************************************************************
```

We can also perform the clustalw2 command with the FINALf.fasta file (clustalw2 FINALf.fasta):



```
Sequence format is Pearson
Sequence 1: sp|B8IU50|HTPG_METNO    611 aa
Sequence 2: sp|P04811|HSP83_DROVI   374 aa
Sequence 3: sp|P11501|HS90A_CHICK   728 aa
Sequence 4: sp|P35016|ENPL_CATRO    817 aa
Sequence 5: sp|P58477|HTPG_RHIME    629 aa
Sequence 6: sp|Q86L04|TRAP1_DICDI   711 aa
Sequence 7: sp|Q9CQN1|TRAP1_MOUSE   706 aa
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  37
Sequences (1:3) Aligned. Score:  37
Sequences (1:4) Aligned. Score:  35
Sequences (1:5) Aligned. Score:  46
Sequences (1:6) Aligned. Score:  30
Sequences (1:7) Aligned. Score:  33
Sequences (2:3) Aligned. Score:  84
Sequences (2:4) Aligned. Score:  51
Sequences (2:5) Aligned. Score:  34
Sequences (2:6) Aligned. Score:  30
Sequences (2:7) Aligned. Score:  32
Sequences (3:4) Aligned. Score:  46
Sequences (3:5) Aligned. Score:  36
Sequences (3:6) Aligned. Score:  25
Sequences (3:7) Aligned. Score:  28
Sequences (4:5) Aligned. Score:  36
Sequences (4:6) Aligned. Score:  27
Sequences (4:7) Aligned. Score:  30
Sequences (5:6) Aligned. Score:  30
Sequences (5:7) Aligned. Score:  33
Sequences (6:7) Aligned. Score:  40
Guide tree file created:   [FINALFAST.dnd]
```

```
# Run ClustalW to perform MSA using HSP90 sequences
clustalw2 FINAL.fasta

# Use hmmalign to make a MSA with HSP90 sequences
hmmalign HSP90.hmm fastas/FINAL.fasta > HSP90.sto
# Change format of the MSA using perl script:
perl /shared/PERL/aconvertMod2.pl -in h -out c <HSP90.sto>HSP90.clu
```

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.

Q6) Our protein presents a motif (regions of protein structure that may or may not be defined by a unique chemical or biological function) from position 723 to 732, essential for tetratricopeptide repeat (TPR) repeat-binding domains, which bind specific peptide ligands and are thought to mediate protein–protein interactions in a variety of biological systems, in this case the TPR repeat-binding motif mediates interaction with TPR repeat-containing proteins like the co-chaperone STUB1.

This motif consists of two protein regions: the first one (728-732), is essential for interaction with SMYD3, TSC1 and STIP1/HOP and the second one (729-732), is essential for interaction with SGTA and TTC1.

This region is also essential for other proteins of our family because since our essential region is found in the motif we can state that it is shared among other proteins.

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.

Q7) One of the mutations that can occur in position 598 of our protein, concretely in the endothelial cells, is the substitution of the C amino acid to either the A, N or D amino acids which causes the reduction of ATPase activity and client protein activation. This mutation is signaled by the S-nitrosylation of which our chaperone Hsp90 is a target. This S-nitrosylation affects the positive effect Hsp90 has on eNOS and limits the eNOS activation. This can derive into serious health issues like excess production of superoxide, hypertension, hypercholesterolemia, diabetes mellitus and many more since eNOS, an enzyme that generates the vasoprotective molecule nitric oxide (NO·), is a major weapon of endothelial cells to fight vascular disease.