Davide	*Ghiotto*	*1236660*

Midterm test No. 2

24 / 11 / 2020

Questions

1. **Paste below your assigned CATH superfamily identifier.**
   3.30.70.100
2. **Compare the sequences of your superfamily provided in the assignment file performing an all-vs-all pairwise sequence alignment.**

   Multiple sequence aligment using MUSCLE web services, with default parameters.

   a. **Paste below a 10 x 10 matrix where cells represent the pairwise sequence identity.**

1: 1lq9	*100.00*	*12.87*	*10.89*	*13.40*	*23.66*	*17.89*	*13.33*	*14.43*	*22.58*	*20.00*
2: 1vqs	*12.87*	*100.00*	*42.86*	*14.77*	*14.94*	*13.64*	*11.11*	*9.78*	*8.33*	*14.81*
3: 5k9f	*10.89*	*42.86*	*100.00*	*15.91*	*11.49*	*11.36*	*13.58*	*12.50*	*17.86*	*14.81*
4: 1sqe	*13.40*	*14.77*	*15.91*	*100.00*	*27.00*	*6.32*	*13.13*	*17.89*	*27.00*	*16.49*
5: 1tz0	*23.66*	*14.94*	*11.49*	*27.00*	*100.00*	*12.50*	*8.33*	*13.68*	*24.00*	*18.75*
6: 1y0h	*17.89*	*13.64*	*11.36*	*6.32*	*12.50*	*100.00*	*20.21*	*20.43*	*14.29*	*16.84*
7: 4npo	*13.33*	*11.11*	*13.58*	*13.13*	*8.33*	*20.21*	*100.00*	*18.75*	*14.14*	*15.31*
8: 3bm7	*14.43*	*9.78*	*12.50*	*17.89*	*13.68*	*20.43*	*18.75*	*100.00*	*21.05*	*23.16*
9: 1iuj	*22.58*	*8.33*	*17.86*	*27.00*	*24.00*	*14.29*	*14.14*	*21.05*	*100.00*	*35.64*
10: 3hx9	*20.00*	*14.81*	*14.81*	*16.49*	*18.75*	*16.84*	*15.31*	*23.16*	*35.64*	*100.00*

   b. **Which is the domain more similar to all other domains?**
   1iuj *with a total of 284.89 (summing the percentages)*

   c. **Based on sequence identity (e.g. 30% threshold), are there domains which can be grouped in the same family?**
   Possible family 1: ***1iuj*** + ***3hx9***
   Possible family 2: ***1vqs*** + ***1lq9***

3. **Download the PDB files associated with your CATH superfamily and answer the following questions considering the start/end positions of the domain fragment as provided in the assignment file.**

   *Downloaded the PDB files in one bulk operation using **PDB downloads** services (https://www.rcsb.org/downloads) checking **PDB format** as the option.*
   To compute the coverage of the domain fragments I used the script "exercise_result.py".

   a. **Which is the coverage of your domain fragments on the corresponding PDB chains (consider observed residues)?**

      4npo : 0.816
      3bm7 : 0.9217391304347826
      1y0h : 0.9901960784313726
      1sqe : 0.926605504587156
      1tz0 : 0.9473684210526315
      1iuj : 0.9622641509433962
      3hx9 : 0.8145161290322581
      1vqs : 0.9482758620689655
      1lq9 : 1.0
      5k9f : 0.9196428571428571

   b. **Which is the coverage of your domain fragments on the corresponding full length proteins (UniProt sequences)?**

      Q9RSM4 4npo_A 0.864406779661017
      Q9A6G2 3bm7_A 1.1041666666666667
      O86332 1y0h_A 1.0
      Q99X56 1sqe_A 0.9351851851851852
      Q81C15 1tz0_A 0.972972972972973
      P83693 1iuj_A 0.9622641509433962
      P9WKH3 3hx9_A 0.9619047619047619
      1vqs : Not found in UNIPROT
      Q53908 1lq9_A 0.9911504424778761
      Q13VQ7 5k9f_A 0.9903846153846154

4. **For each PDB create a new PDB with the coordinates of the domain fragment and perform an all-vs-all pairwise structural alignment using TM-align.**

   Used the script "structural_alignment.sh" and created a for loop in python to execute all-vs-all pairwise structural alignment using TM-align compiled from cpp (I'm using a Windows pc).
   To extract the best sequence identity score and the best RMSD score I just parsed the output of the TM-align using the ***domain1_domain2.out*** *file.*

   a. **Paste below a 10 x 10 matrix where cells represent the pairwise sequence identity obtained with the structural alignment (not sequence alignment).**

   4npo 1.0 0.197 0.224 0.148 0.076 0.098 0.15 0.075 0.141 0.057
   3bm7 0.197 1.0 0.227 0.186 0.125 0.276 0.342 0.055 0.133 0.135
   1y0h 0.224 0.227 1.0 0.069 0.114 0.111 0.092 0.13 0.232 0.125
   1sqe 0.148 0.186 0.069 1.0 0.267 0.274 0.178 0.06 0.138 0.117
   1tz0 0.076 0.125 0.114 0.267 1.0 0.263 0.203 0.079 0.183 0.072
   1iuj 0.098 0.276 0.111 0.274 0.263 1.0 0.38 0.037 0.181 0.049
   3hx9 0.15 0.342 0.092 0.178 0.203 0.38 1.0 0.064 0.134 0.062
   1vqs 0.075 0.055 0.13 0.06 0.079 0.037 0.064 1.0 0.101 0.412
   1lq9 0.141 0.133 0.232 0.138 0.183 0.181 0.134 0.101 1.0 0.102
   5k9f 0.057 0.135 0.125 0.117 0.072 0.049 0.062 0.412 0.102 1.0

   b. **Paste below a 10 x 10 matrix where cells represent the pairwise RMSD.**

   4npo 0.0 1.48 1.93 2.91 2.9 2.52 2.95 2.95 2.61 2.72
   3bm7 1.48 0.0 1.43 2.48 2.9 2.07 2.68 2.66 2.09 2.48
   1y0h 1.93 1.43 0.0 2.49 3.17 2.33 3.03 2.75 2.48 2.62
   1sqe 2.91 2.48 2.49 0.0 1.92 2.0 2.79 2.38 2.16 2.39
   1tz0 2.9 2.9 3.17 1.92 0.0 2.39 2.75 3.25 2.42 3.51
   1iuj 2.52 2.07 2.33 2.0 2.39 0.0 2.33 2.62 2.08 2.54
   3hx9 2.95 2.68 3.03 2.79 2.75 2.33 0.0 2.38 2.89 2.39
   1vqs 2.95 2.66 2.75 2.38 3.25 2.62 2.38 0.0 2.73 0.85
   1lq9 2.61 2.09 2.48 2.16 2.42 2.08 2.89 2.73 0.0 2.79
   5k9f 2.72 2.48 2.62 2.39 3.51 2.54 2.39 0.85 2.79 0.0

   c. **Which is the domain more similar to all other domains looking at the sequence identity (calculated with the structural alignment)?**
   3bm7 2.676 (summing percentages)

   d. **Which is the domain more similar to all other domains looking at the RMSD?**
   3bm7 20.27 (summing values)

5. **Create a multiple sequence alignment (MSA) starting from the domain sequences available in the assignment file using EBI T-Coffee.**
   I used the web service offered by EBI (https://www.ebi.ac.uk/Tools/msa/tcoffee/) with default parameters to generate a multiple sequence alignment from the starting domain sequences. I selected "fasta" format as the output in order to reuse it for the next questions.

   a. **Which are the most conserved columns looking at the amino acid composition?**
      Using JalView, colouring with "clustalx" mode, I picked the columns with conserved amino acid composition value above 7.
      *Columns: 21, 27, 56, 81, 86, 87 (referring to the first sequence of the msa that is **1iuj**)*

   b. **Which are the most conserved columns looking at the column entropy?**
      Using the script "entropy.py" with a arbitrary threshold of 0.6820342019820005
      Which is the 95% percentile (computed with numpy)
      Columns:
      - *column 27 entropy 0.7223461442082889*
      - *column 52 entropy 0.7223461442082891*
      - *column 111 entropy 0.7223461442082889*
      - *column 115 entropy 0.7223461442082891*
      - *column 141 entropy 0.6820342019820005*
      - *column 142 entropy 0.6820342019820007*

      *The columns are still referring to the first sequence of the msa, that is **1iuj**.*

6. **Use the MSA generated before to perform a PSI-BLAST and a HMMER search against human proteins.**
   For PSI-BLAST I used this webservice: https://myhits.sib.swiss/cgi-bin/blast .
   I selected every sources and as optional parameter I entered "homo sapiens" as taxonomic restriction.
   For HMMER I used the webservice: https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch .
   The search was restricted to "homo sapiens" as well.

   a. **How many significant hits are returned by the two methods?**
      PSI_BLAST : 2 hits
      HMMsearch : 0 hits

7. **Which PFAM HMMs match your superfamily? Hint: you can use hmmscan EBI service.**
   I used the webservice: https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan .
   I uploaded a file containing a list of sequences in FASTA format (the domain sequences given) and the service retrieved a list of results, one for each entry.
   I selected as HMM Database only Pfam.

   The PFAM HMMs that match my superfamily are:
   - *ABM : 8 domains match*
   - *NIPSNAP: 2 domains match*