|  |  |  |
| --- | --- | --- |
| *Darko* | *Ivanovski* | *1243085* |

## Midterm test No. 2

### 24 / 11 / 2020

## Questions

1. **Paste below your assigned CATH superfamily identifier.**

*3.30.70.120*

1. **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 service (*[*https://www.ebi.ac.uk/Tools/msa/muscle/*](https://www.ebi.ac.uk/Tools/msa/muscle/)*) , with default parameters.*

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

*1: 4co3 100.00 42.06 18.00 11.88 11.43 9.52 16.50 8.82 7.84 8.57*

*2: 4ush 42.06 100.00 21.00 7.29 15.84 11.43 18.37 11.34 9.90 9.52*

*3: 5d4n 18.00 21.00 100.00 8.60 16.33 7.84 12.63 11.70 13.00 13.46*

*4: 1uku 11.88 7.29 8.60 100.00 32.35 32.35 40.40 36.36 21.21 36.63*

*5: 2nuh 11.43 15.84 16.33 32.35 100.00 30.84 40.78 34.31 30.10 34.91*

*6: 4e98 9.52 11.43 7.84 32.35 30.84 100.00 41.75 35.29 28.30 29.09*

*7: 1nza 16.50 18.37 12.63 40.40 40.78 41.75 100.00 44.12 38.24 38.83*

*8: 1p1l 8.82 11.34 11.70 36.36 34.31 35.29 44.12 100.00 27.45 31.37*

*9: 6gdx 7.84 9.90 13.00 21.21 30.10 28.30 38.24 27.45 100.00 39.81*

*10: 1naq 8.57 9.52 13.46 36.63 34.91 29.09 38.83 31.37 39.81 100.00*

* 1. **Which is the domain more similar to all other domains?**

***1nza*** *with a total of 391.62 ( summing the percentages )*

* 1. **Based on sequence identity (e.g. 30% threshold), are there domains which can be grouped in the same family?**

*Possible family 1:* ***4ush*** *+* ***4co3***

*Possible family 2:* ***2nuh*** *+* ***1uku*** *+* ***4e98*** *+* ***1nza*** *+* ***1p1l*** *+* ***6gdx*** *+* ***1naq***

1. **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*](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”.*

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

*4co3 : 0.9285714285714286*

*4ush : 0.6688311688311688*

*5d4n : 0.9722222222222222*

*1uku : 1.0*

*2nuh : 0.8813559322033898*

*4e98 : 0.7971014492753623*

*1nza : 1.0*

*1p1l : 1.0*

*6gdx : 0.8045112781954887*

*1naq : 0.9464285714285714*

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

*P70731 4co3\_A 0.9285714285714286*

*A8JI83 4ush\_A 0.5024390243902439*

*D5X329 5d4n\_A 0.9722222222222222*

*O58720 1uku\_A 1.0*

*Q9PFN8 2nuh\_A 0.9285714285714286*

*Q5CX58 4e98\_A 0.9401709401709402*

*Q7SIA8 1nza\_A 1.0*

*O28301 1p1l\_A 1.0*

*Q31KX8 6gdx\_A 0.9469026548672567*

*P69488 1naq\_A 0.9464285714285714*

1. **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.*

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

*4co3 1.0 0.439 0.246 0.159 0.044 0.101 0.119 0.103 0.101 0.072*

*4ush 0.439 1.0 0.181 0.123 0.13 0.117 0.136 0.121 0.092 0.081*

*5d4n 0.246 0.181 1.0 0.107 0.065 0.075 0.122 0.08 0.098 0.086*

*1uku 0.159 0.123 0.107 1.0 0.318 0.341 0.412 0.353 0.214 0.376*

*2nuh 0.044 0.13 0.065 0.318 1.0 0.333 0.437 0.326 0.33 0.378*

*4e98 0.101 0.117 0.075 0.341 0.333 1.0 0.448 0.337 0.281 0.323*

*1nza 0.119 0.136 0.122 0.412 0.437 0.448 1.0 0.442 0.388 0.368*

*1p1l 0.103 0.121 0.08 0.353 0.326 0.337 0.442 1.0 0.294 0.326*

*6gdx 0.101 0.092 0.098 0.214 0.33 0.281 0.388 0.294 1.0 0.426*

*1naq 0.072 0.081 0.086 0.376 0.378 0.323 0.368 0.326 0.426 1.0*

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

*4co3 0.0 0.95 2.43 2.23 2.04 2.54 2.22 2.32 2.28 2.49*

*4ush 0.95 0.0 2.11 1.73 1.85 1.86 2.02 1.87 1.89 2.07*

*5d4n 2.43 2.11 0.0 2.55 2.14 2.19 2.23 2.3 2.47 2.27*

*1uku 2.23 1.73 2.55 0.0 0.96 0.78 1.33 0.95 0.84 1.13*

*2nuh 2.04 1.85 2.14 0.96 0.0 0.91 0.94 0.73 0.96 1.15*

*4e98 2.54 1.86 2.19 0.78 0.91 0.0 0.98 0.87 0.96 1.13*

*1nza 2.22 2.02 2.23 1.33 0.94 0.98 0.0 0.9 1.29 1.43*

*1p1l 2.32 1.87 2.3 0.95 0.73 0.87 0.9 0.0 0.98 1.11*

*6gdx 2.28 1.89 2.47 0.84 0.96 0.96 1.29 0.98 0.0 1.05*

*1naq 2.49 2.07 2.27 1.13 1.15 1.13 1.43 1.11 1.05 0.0*

* 1. **Which is the domain more similar to all other domains looking at the sequence identity (calculated with the structural alignment)?**

*1nza 3.872*

* 1. **Which is the domain more similar to all other domains looking at the RMSD?**

*2nuh 11.680000000000001*

1. **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/*](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.*

* 1. **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: 24, 25, 34, 42, 51, 58, 62, 64, 73, 88, 122, 123, 126 ( referring to the first sequence of the msa that is* ***1naq*** *)*

* 1. **Which are the most conserved columns looking at the column entropy?**

*Using the script “entropy.py” with a arbitrary threshold of 0.6297948589443854*

*Which is the 95% percentile (computed with numpy)*

*Columns:*

*- column 35 entropy 0.6586040494376154*

*- column 69 entropy 0.6297948589443854*

*- column 95 entropy 0.6306168212798319*

*- column 120 entropy 0.6297948589443854*

*- column 130 entropy 0.7223461442082891*

*- column 134 entropy 0.6306168212798319*

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

1. **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*](https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch) *.*

*The search was restricted to “homo sapiens” as well.*

* 1. **How many significant hits are returned by the two methods?**

*PSI\_BLAST : 5 hits*

*HMMsearch : 6 hits*

1. **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*](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:*

* *P-II : 3 domains match*
* *CutA1 : 7 domains match*