Alberto	Calabrese	2103405

Midterm Test 1

31 / 10 / 2024

Introduction

Please answer all the questions below and submit this document in **PDF format** by **10:30** on **November 14, 2024** (two weeks) to damiano.piovesan@unipd.it. Please rename the file as **midterm1_<surname>_<name>.pdf**.

Each student is assigned a different **DNA sequence** and a protein superfamily from the **CATH** database with 10 representative sequences. The first part of the test is based on the analysis of the DNA sequence, and the second part pertains to the analysis of the superfamily.

For each question, concisely explain all the steps (**maximum of 5 rows**) necessary to reproduce the results, including parameters, database queries, algorithms, etc. If relevant, you can provide source code, but it is not necessary.

Assignments (input)

1. DNA sequence and superfamily CATH identifiers are available in the Students' sheet.
2. Superfamily sequences are available here (Columns: PDB ID, chain ID, PDB domain start, PDB domain end, domain sequence).

Questions (part 1)

Paste below your assigned DNA sequence.

```
CAGCCGTGGTTTCACGGCCCCCTGAGCCGAGCAGAGGCCGAGAACCTTCTGTCCCTCTGCAAGGA
AGGCAGCTACCTCGTTCGGCTCAGCGAGACCAGGGCTCAGGACTGCATTCTGTCTCTCAGAAGCA
ACCAGGGTTCCATGCACCTGAAATTCGCAAGGACCCGGGAGAACCAGGTGGTACTGGGACAGCAC
AGTGGGCCCTTCCCCAGCATACCTGAGCTGGTCCTGCATTACAGTGCCCGCCCACTGCCCGTGCA
AGGGGCAGAGCACCTGGCCCTGCTCTATCCT
```

1. Paste below the correct translation of your DNA sequence.
   ***Hint**: Test all possible frames (3 shifts x 2 forward/complement x 2 directions) as the sequence could be a fragment, complement, or inversion of a gene. **Warning**: Making a mistake at this step can affect the entire exam.*

   ```
   QPWFHGPLSRAEAENLLSLCKEGSYLVRLSETRAQDCILSLRSNQGSMHLKFARTRENQVVLGQH
   SGPFPSIPELVLHYSARPLPVQGAEHLALLYP
   ```

2. Align the amino acid sequence against the SwissProt database using the BLAST service.

***Hint**: If the search against SwissProt does not provide any significant hits, try using a larger database like UniRef50.*

a. How many significant hits?

I found 56 significant hits.

b. What is the coverage of the query sequence (your input) with the best matched sequence?

The query coverage is 100%, indicating that the sequence aligns across the entire length with the protein.

c. What is the coverage of the best matched sequence with your input sequence?

As we can see in the BLAST results, the best matched sequence has a accession length of 343, our sequence has a length of 97, so the coverage of the best matched sequence with our input sequence is ~ 28%.

d. According to the BLAST results, is your input sequence a fragment or a full protein?

Given the 28% coverage with a known full-length protein sequence, my input sequence is for sure a fragment rather than a full or nearly complete protein.

3. Compare your amino acid sequence with the best match found in the previous BLAST by using Needleman-Wunsch and Smith-Waterman algorithms.
***Hint**: You can retrieve the full sequence of a matched protein from UniProtKB using the protein ID or Accession code available in the BLAST output.*

a. Provide identity, number of gaps, similarity, and score for the two alignments generated with the two different algorithms.

```
Needleman-Wunsch Alignment:
Identity: 97
Gaps: 246
Similarity (Score): 367.0

Smith-Waterman Alignment:
Identity: 97
Gaps: 246
Similarity (Score): 509.0
```

b. Which algorithm gives the best alignment? Why?

The Smith-Waterman algorithm gives the best alignment in this case, as indicated by the higher similarity score that is the only different result obtained. The Smith-Waterman algorithm is a local alignment algorithm, while Needleman-Wunsch algorithm is a global alignment algorithm.

c. Which algorithm between BLAST and Smith-Waterman gives the best alignment? Why?

Generally, BLAST provides a quicker approximate alignment, but Smith-Waterman is more precise for identifying the best local alignment. In this case, the Smith-Waterman algorithm gives the best alignment because it is designed to find the best local alignment between two sequences, which is essential for comparing the query sequence with a known protein sequence.

4. Evaluate your amino acid sequence against the Pfam database of HMM models.
   ***Hint**: Use the HMMER web services to evaluate HMM models. Use the UniProtKB website (advanced search) to search Pfam proteins.*

   a. Which Pfam domain(s) match your sequence?

      The SH2 domain (Pfam ID: PF00017) matches the sequence.

   b. Is your sequence fully covered by Pfam domains?

      Our sequence has a length of 97 amino acids, the coverage with the SH2 domain is from 3 to 79 so it is partially covered.

   c. How many proteins in SwissProt have the same domain? (Consider only one Pfam ID if your protein is multi-domain.)

      401

Questions (part 2)

Paste below your assigned superfamily identifier.

```
>3.10.200.10

3ks3   A      4      261
       MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPLSVSYDQATSLRILNNGHAFNVEFDDSQ
DKAVLKGGPLDGTYRLIQFHFHWGSLDGQGSEHTVDKKKYAAELHLVHWNTKYGDFGKAVQQPDGLAVLGIFLKVGSAKP
GLQKVVDVLDSIKTKGKSADFTNFDPRGLLPESLDYWTYPGSLTTPPLLECVTWIVLKEPISVSSEQVLKFRKLNFNGEG
EPEELMVDNWRPAQPLKNRQIKASFK

5msa   A      3      263
       MSKWTYFGPDGENSWSKKYPSCGGLLQSPIDLHSDILQYDASLTPLEFQGYNLSANKQFLLTNNGHSVKLNLPS
DMHIQGLQSRYSATQLHLHWGNPNDPHGSEHTVSGQHFAAELHIVHYNSDLYPDASTASNKSEGLAVLAVLIEMGSFNPS
YDKIFSHLQHVKYKGQEAFVPGFNIEELLPERTAEYYRYRGSLTTPPCNPTVLWTVFRNPVQISQEQLLALETALYCTHM
DDPSPREMINNFRQVQKFDERLVYTSFSQ

3fw3   A      5      259
       AESHWCYEVQAESSNYPCLVPVKWGGNCQKDRQSPINIVTTKAKVDKKLGRFFFSGYDKKQTWTVQNNGHSVMM
LLENKASISGGGLPAPYQAKQLHLHWSDLPYKGSEHSLDGEHFAMEMHIVHEKEKGTSRNVKEAQDPEDEIAVLAFLVEA
GTQVNEGFQPLVEALSNIPKPEMSTTMAESSLLDLLPKEEKLRHYFRYLGSLTTPTCDEKVVWTVFREPIQLHREQILAF
SQKLYYDKEQTVSMKDNVRPLQQLGQRTVIKS

3jxg   A      58     320
       GPGSGDPYWAYSGAYGPEHWVTSSVSCGGSHQSPIDILDHHARVGDEYQELQLDGFDNESSNKTWMKNTGKTVA
ILLKDDYFVSGAGLPGRFKAEKVEFHWGHSNGSAGSEHSVNGRRFPVEMQIFFYNPDDFDSFQTAISENRIIGAMAIFFQ
VSPRDNSALDPIIHGLKGVVHHEKETFLDPFILRDLLPASLGSYYRYTGSLTTPPCSEIVEWIVFRRPVPISYHQLEAFY
SIFTTEQQDHVKSVEYLRNNFRPQQALNDRVVSKS

5ush   A      3      234
       MPQQLSPINIETKKAISNARLKPLDIHYNESKPTTIQNTGKLVRINFKGGYISGGFLPNEYVLSSLHIYWGKED
DYGSNHLIDVYKYSGEINLVHWNKKKYSSYEEAKKHDDGLIIISIFLQVLDHKNVYFQKIVNQLDSIRSANTSAPFDSVF
YLDNLLPSKLDYFTYLGTTINHSADAVWIIFPTPINIHSDQLSKFRTLLSLSNHEGKPHYITENYRNPYKLNDDTEVYYS
GHHHHHH

2w2j   A      23     290
       EEEGVEWGYEEGVEWGLVFPDANGEYQSPINLNSREARYDPSLLDVRLSPNYVVCRDCEVTNDGHTIQVILKSK
SVLSGGPLPQGHEFELYEVRFHWGRENQRGSEHTVNFKAFPMELHLIHWNSTLFGSIDEAVGKPHGIAIIALFVQIGKEH
VGLKAVTEILQDIQYKGKSKTIPCFNPNTLLPDPLLRDYWVYEGSLTIPPCSEGVTWILFRYPLTISQLQIEEFRRLRTH
VKGAELVEGCDGILGDNFRPTQPLSDRVIRAAFQ

4xfw   A      22     247
       KWDYKNKENGPHRWDKLHKDFEVCKSGKSQSPINIEHYYHTQDKADLQFKYAASKPKAVFFTHHTLKASFEPTN
HINYRGHDYVLDNVHFHAPMEFLINNKTRPLSAHFVHKDAKGRLLVLAIGFEEGKENPNLDPILEGIQKKQNFKEVALDA
FLPKSINYYHFNGSLTAPPCTEGVAWFVVEEPLEVSAKQLAEIKKRMKNSPNQRPVQPDYNTVIIKRSAETR
```

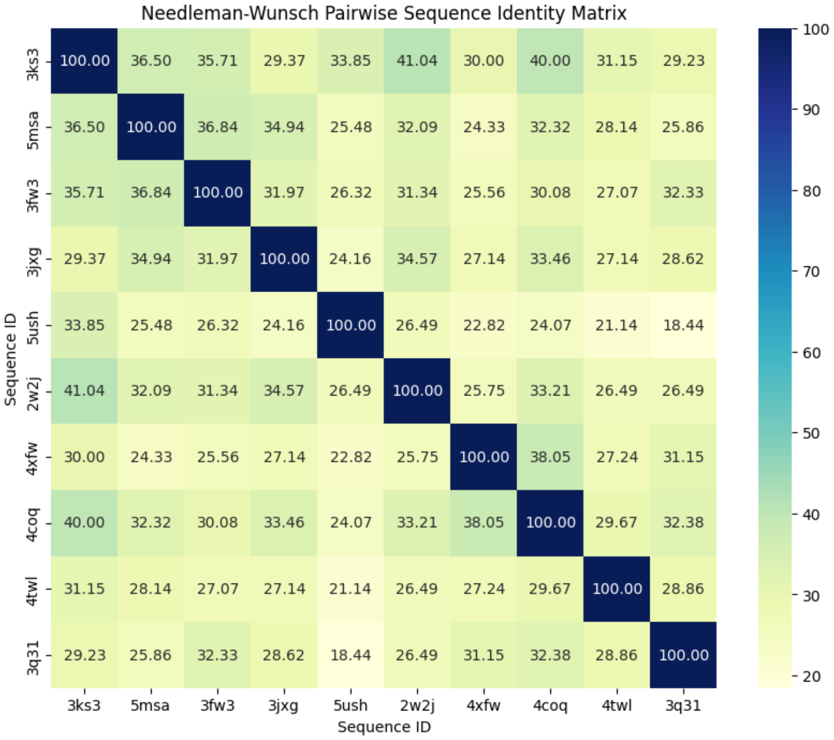
```
4coq   A      23     247
       GAHWGYSGSIGPEHWGDLSPEYLMCKIGKNQSPIDINSADAVKACLAPVSVYYVSDAKYVVNNGHTIKVVMGGR
GYVVVDGKRFYLKQFHFHAPSEHTVNGKHYPFEAHFVHLDKNGNITVLGVFFKVGKENPELEKVWRVMPEEPGQKRHLTA
RIDPEKLLPENRDYYRYSGSLTTPPCSEGVRWIVFKEPVEMSREQLEKFRKVMGFDNNRPVQPLNARKVMK

4twl   A      3      241
       VEDEFSYIDGNPNGPENWGNLKPEWETCGKGMEQSPIQLRDNRVIFDQTLGKLRRNYRAVDARLRNSGHDVLVD
FKGNAGSLSINRVEYQLKRIHFHSPSEHEMNGERFDLEAQLVHESQDQKRAVVSILFRFGRADPFLSDLEDFIKQFSNSQ
KNEINAGVVDPNQLQIDDSAYYRYMGSFTAPPCTEGISWTVMRKVATVSPRQVLLLKQAVNENAINNARPLQPTNFRSVF
YFEQLKSKLGVI

3q31   A      35     270
       AAGGLDDANKFNYTGLGGPLNWYGLDEANEACAKGKHQSPIVIDSAAIDYAASGSLKLDLPLADGSKLENLGFG
LQVTLTNGSLTANSKTYTLAQFHFHTPSEHHVNEEHFPMEVHFVFQTAAKETAVVGFFFQLSEVGDSVPLFDSVFAPIDN
IPDAGTSTTTGQLDFGGLLDHFNRHGVYQYTGSLTTPPCTEEVMWNLSTEPLPLTVQGYNKVKKIIKYNARYTQNALGQD
NLLEVAAQKL
```

5. Compare the sequences of your superfamily provided in the assignment file by performing an all-vs-all pairwise sequence alignment.
***Hint**: You can use BioPython and the Align module.*

    a. Paste below a 10 x 10 matrix where cells represent the pairwise sequence identity. Provide sequence identifiers in the matrix tick labels.



	3ks3	5msa	3fw3	3jxg	5ush	2w2j	4xfw	4coq	4twl	3q31
3ks3	100.00	36.50	35.71	29.37	33.85	41.04	30.00	40.00	31.15	29.23
5msa	36.50	100.00	36.84	34.94	25.48	32.09	24.33	32.32	28.14	25.86
3fw3	35.71	36.84	100.00	31.97	26.32	31.34	25.56	30.08	27.07	32.33
3jxg	29.37	34.94	31.97	100.00	24.16	34.57	27.14	33.46	27.14	28.62
5ush	33.85	25.48	26.32	24.16	100.00	26.49	22.82	24.07	21.14	18.44
2w2j	41.04	32.09	31.34	34.57	26.49	100.00	25.75	33.21	26.49	26.49
4xfw	30.00	24.33	25.56	27.14	22.82	25.75	100.00	38.05	27.24	31.15
4coq	40.00	32.32	30.08	33.46	24.07	33.21	38.05	100.00	29.67	32.38
4twl	31.15	28.14	27.07	27.14	21.14	26.49	27.24	29.67	100.00	28.86
3q31	29.23	25.86	32.33	28.62	18.44	26.49	31.15	32.38	28.86	100.00

    b. Which sequence is the most similar to all other sequences?

    Based on the identity matrix values, the sequence 3ks3 is most similar to all other sequences with an average score of 34.09%.

    c. Based on sequence identity values, are there sequences that can be grouped in the same (sub)family?

    We can group the sequences in two different sub families by looking to an identity score greater than 30% between the sequences:

    - First Family: `2w2j, 4coq, 3ks3`

    - Second Family: `5msa, 3fw3`

6. Create a multiple sequence alignment (MSA) starting from the domain sequences available in the assignment file using EBI ClustalOmega.

    a. Which columns are the most conserved when looking at the amino acid composition?

    ```
    Most conserved columns by amino acid composition (excluding gaps):
    Position: 13, Amino Acid: Y, Frequency: 100.00%, Non-gap Coverage: 90.00%
    Position: 42, Amino Acid: S, Frequency: 100.00%, Non-gap Coverage: 100.00%
    Position: 43, Amino Acid: P, Frequency: 100.00%, Non-gap Coverage: 100.00%
    Position: 98, Amino Acid: G, Frequency: 100.00%, Non-gap Coverage: 50.00%
    Position: 101, Amino Acid: L, Frequency: 100.00%, Non-gap Coverage: 60.00%
    ```

    b. Which columns are the most conserved when looking at the column entropy?

    ```
    Most conserved columns by entropy (excluding gaps):
    Position: 13, Entropy: 0.000, Non-gap Coverage: 90.00%
    Position: 42, Entropy: 0.000, Non-gap Coverage: 100.00%
    Position: 43, Entropy: 0.000, Non-gap Coverage: 100.00%
    Position: 98, Entropy: 0.000, Non-gap Coverage: 50.00%
    Position: 101, Entropy: 0.000, Non-gap Coverage: 60.00%
    ```

7. Use the MSA generated before to perform a PSI-BLAST and an HMMER search against human proteins (or SwissProt if the search against human returns nothing).
    ***Hint**: For the PSI-BLAST search, you can use the NCBI web service and provide a PSSM generated with the PSI-BLAST command line. For HMMSEARCH you can provide directly the MSA.*

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

    PSI-BLAST: 54

    HMMSEARCH: 53