

BIN506 Spring 2024-2025 Assignment 4

Due Date: 20 April, 23:59

Late Submission Policy: 10 points of deduction will be applied for each extra hour.

Uploading your assignment as a PDF is mandatory.

Questions

- 1) Discuss the following: Why does using protein sequences result in more accurate sequence alignments? Think about biological reasons, statistical reasons, and computational reasons.
- 2) Answer the following questions about PAM and BLOSUM matrices.
 - a) What kind of biological information are they built upon?
 - b) Explain the difference between PAM250 and PAM500 and the difference between BLOSUM62 and BLOSUM95.
- 3) Use “unknown.fasta” and search for the source using NCBI BLAST, UCSC BLAT, and Ensembl BLAST/BLAT.
 - a) Share the list of the results. Write down the name and the type (gene, transcript, etc.) of the source.
 - b) Select the top scoring hit (both in terms of E-Value and score). Click on the link that will direct you to the genome browser on each database. Share a screenshot from each genome browser.

4) Use “protein.fasta” and run a search using **blastp** and **psi-blast**.

- For both results, set the ‘Show’ value to 50 and make sure to keep it that way.

A screenshot of a web interface element. It consists of a light blue rectangular button with the word "Show" in black text. To the right of the text is a small white box containing the number "50" and a black downward-pointing arrow, indicating a dropdown menu.

- Set ‘Number of sequences’ to 50 in **psi-blast** results and make 5 iterations.

A screenshot of a web form. On the left, the text "Number of sequences" is in bold. To its right is a white input box containing the number "50". Further to the right is a blue button with the word "Run" in white text.

- Make sure the resulting hits are sorted by the E-value.
- Answer the following questions by comparing the blastp result to the 5th iteration of psi-blast.
 - a) Are the first 5 hits the same in both tools?
 - b) Select the last 5 hits in both tools and compare their scores, E-values, and query covers. Submit a screenshot of ‘Graphical Summary’ with an explanation for both graphs for the last 5 hits. What do the graphs tell? What is the difference between the two graphs?
 - c) Again, select the last 5 hits, go to the ‘Alignments’ tab, and change the ‘Alignment View’ to ‘Flat query-anchored with letters for identities’. Explain the differences you see with a few sentences. You may take hints from the graphs from 4b.

5) Submit the **msa.txt** file to [MEME Suite](#).

- a) How many motifs are these sequences sharing? How are these motifs positioned on the sequences? Add a screenshot of the graphical view of the motifs.
- b) Select the most significant motif and add **1) the seq logo, 2) the alignment, and 3) scores** for that motif. What does the seq logo and size of the letters represent? Explain using the screenshots you provided.
- c) Change the parameters to find the maximum possible number of motifs. How many of them are significant? Submit a screenshot of the graphical view of the motifs.