CS612 Homework Assignment 2

Due Fri. March 7, 2025 on Gradescope

1. **Multiple Sequence Alignment Using Clustal Omega and Muscle:** This question is loosely based on the following example: `https://www.ebi.ac.uk/jdispatcher/sss/fasta/summary?jobId=fasta-R20230421-121005-0548-43453433-p1m`, with only the human sequences selected.

   (a) First run Clustal omega from `https://www.ebi.ac.uk/jdispatcher/msa/clustalo`. Use the sequences (also available on the course webpage at `http://www.cs.umb.edu/~nurith/cs612/msa_human.fasta`) Leave other parameters as-is and run. It may take about 15-20 seconds. Download the results under "Tool output" and paste it to your submission.

   (b) Repeat with Muscle at `https://www.ebi.ac.uk/jdispatcher/msa/muscle?stype=protein`. Attach the alignment.

   (c) Do the results look similar?

2. **Structure based sequence alignment:** One of the main goals of multiple sequence alignment in structural bioinformatics is to model the structures of sequences whose structures are unknown, based on similar sequences with known structures (also called homology modeling, which we will dwell on later). Sometimes it is hard to find aligned sequences with a known structure. You can use Hidden Markov Models (HMM) to identify homologous sequences with known structures in a similar fashion to Psi-BLAST shown in class, but more easily.

   (a) Go to https://toolkit.tuebingen.mpg.de/tools/hhpred. Paste or upload the following query sequence from `http://www.cs.umb.edu/~nurith/cs612/query_hw2.fasta` and submit:

   ```
   >QUERY1
   MKDTDLSTLLSIIRLTELKESKRNALLSLIFQLSVAYFIALVIVSRFVRYVNYITYNNLVEFIIVLSLIM
   LIIVTDIFIKKYISKFSNILLETLNLKINSDNNFRREIINASKNHNDKNKLYDLINKTFEKDNIEIKQLG
   LFIISSVINNFAYIILLSIGFILLNEVYSNLFSSRYTTISIFTLIVSYMLFIRNKIISSEEEEQIEYEKV
   ATSYISSLINRILNTKFTENTTTIGQDKQLYDSFKTPKIQYGAKVPVKLEEIKEVAKNIEHIPSKAYFVL
   LAESGLRPGELLNVSIENIDLKARIIWINKETQTKRAYFSFFSRKTAEFLEKVYLPAREEFIRANEKNIA
   KLAAANENQEIDLEKWKAKLFPYKDDVLRRKIYEAMDRALGKRFELYALRRHFATYMQLKKVPPLAINIL
   QGRVGPNEFRILKENYTVFTIEDLRKLYDEAGLVVLE
   ```

   It will take about 1-2 minutes. Attach a screenshot of the results page to your solution (just the front of it). Look at the first hits. What is the PDB code and name of the first match?

   (b) Click on the PDB code of the first match and attach a screenshot of the structure.

   (c) Hit the "Aln" tab Scroll down and find the sequence alignment of this match and attach a screenshot of it. Notice that the query and match are not very similar.

   (d) The line in the middle of the alignment, that says "consensus", shows the consensus sequence of the multiple alignment. Only a small number of amino acids are listed there. You can see if an amino acid is conserved by a vertical line between the template and the query consensus, and the one letter code appearing in the consensus sequence, either as a capital letter (more confident) or small (less confident). Conserved amino acids are believed to be important for the structure or function of the

protein. An example of such a conserved amino acid is E290 (that is, the amino acid E, Glutamic Acid, in position 290 of the query sequence). You can see it in the sequence alignment from part b. It is easy to find because the second line of the alignment ends in position 294 of the query sequence. You can see that E290 shows in capital letters the both consensus sequences, the query and the first match. Find there other amino acids in positions where the consensus matches both the query and the match and in capital letters. Two of them are close to E290 and another one further down.

3. **Protein structure search and classification:** Search the PDB with the entry 3CHY.
   (a) How many atoms are there in the .pdb file?
   (b) What atom type is atom 289?
   (c) What is its amino acid 3 letter code?
   (d) What are the x,y,z coordinates of this atom?
   (e) Search for 3CHY in SCOPe (at scop.berkeley.edu) – what is the class, fold and family of this protein?

4. **Protein visualization:** Use the structure visualization on the PDB website. See guide here:
   https://www.rcsb.org/docs/3d-viewers/mol*/getting-started.
   (a) Visualize the 3CHY protein structure (from last question) by selecting the "structure" link from the visualization panel on the left of the page. The protein will display as cartoon. Color it by secondary structure (from the commands on the right). From the command panel on the right select "Polymer", then "Residue Property". then "Secondary structure". Below the polymer command there are two buttons - ion and water. Click on the eye icon next to them to make the crystallographic water and the ion molecules disappear. How many helices do you see? (refer only to those colored in red). How many beta strands (yellow)? Pressing the display and moving your left mouse allows you to rotate the view. Attach a screenshot of the protein structure. You can either screenshot from your computer or choose the camera icon on the structure window.
   (b) Hide the default cartoon representation and create a ball and stick representation. Attach its image too.

Programming projects

1. In the biopython template, fill in the missing parts: Implement clustal-omega and muscle for the three sequences in the code. Notice that even with three sequences there are differences in the results. You will need to use `ClustalOmegaCommandline` and `MuscleOmegaCommandline`.

2. Bonus: Implement Clustal omega with the parameters `distmat_full=True` and `distmat_out=dist-clustal.t` This will compute the distance matrix between the three sequences. Paste dist-clustal.txt into your submission. Which two sequences are closest?

3. Write code that gives you the answers to parts a–d in question 3 above. You can use the PDB module in Biopython.