CS612 Homework Assignment 2

Due Tue. October 13 2020 on Gradescope

1. **Multiple Sequence Alignment Using PSI-BLAST:** This question is based on

   http://www.cbs.dtu.dk/courses/humanbio/2009/exercises/ExPsiBlast/PSIBLAST.php

   Go to BLAST at http://blast.ncbi.nlm.nih.gov. Now select the protein blast option. Paste the following query to the query window or upload the file http://www.cs.umb.edu/ nurith/cs612/query_hw2.fasta

   >QUERY1
   MKDTDLSTLLSIIRLTELKESKRNALLSLIFQLSVAYFIALVIVSRFVRYVNYITYNNLV
   EFIIVLSLIMLIIVTDIFIKKYISKFSNILLETLNLKINSDNNFRREIINASKNHNDKNK
   LYDLINKTFEKDNIEIKQLGLFIISSVINNFAYIILLSIGFILLNEVYSNLFSSRYTTIS
   IFTLIVSYMLFIRNKIISSEEEEQIEYEKVATSYISSLINRILNTKFTENTTTIGQDKQL
   YDSFKTPKIQYGAKVPVKLEEIKEVAKNIEHIPSKAYFVLLAESGLRPGELLNVSIENID
   LKARIIWINKETQTKRAYFSFFSRKTAEFLEKVYLPAREEFIRANEKNIAKLAAANENQE
   IDLEKWKAKLFPYKDDVLRRKIYEAMDRALGKRFELYALRRHFATYMQLKKVPPLAINIL
   QGRVGPNEFRILKENYTVFTIEDLRKLYDEAGLVVLE

   Set the algorithm to blastp and the database to "Protein Databank (PDB)". Under "algorithm parameters" change max target sequence to 1000 and expected threshold to 0.001. Hit search. It may take a minute or two for the results to be ready.

   (a) Did you get any results?

   (b) Now go back to the query window by clicking "Edit Search" on the top left. Change your database to nr (non-redundant protein sequences) and your program to PSI-BLAST. Set max target sequence again to 1000 and expected threshold 0.001. Hit search. How many significant sequences (below threshold) did you find now? Attach the first page of the results to your submission.

   (c) How large a fraction of the query sequence does the significant hits match (just the best Query coverage)? This can be seen by hitting "Graphic Summary" above the search results. Are the coverage from different significant hits evenly distributed across the entire length of query sequence? Explain.

   (d) You will probably not find any sequence with a known structure this time. To know whether there is a sequence with a known structure, look at the "Accession" column in the results section. If there is a known protein structure, the accession code will look like a PDB code, as in <number><three other characters, mostly letters> _ <Optional character for chain>. Other accession codes are longer. Run PSI-BLAST iteration 2 with 1000 query sequences. You find the button just above the resulting sequences. What is the first new sequence (highlighted in yellow)? New sequences are newly discovered by this iteration of PSI-BLAST.

   (e) You may need another iteration to find a sequence with a PDB structure or it will appear in iteration 2. If not, run a 3rd iteration. What is the PDB code of the sequence with a known structure? What is its query coverage? To find the sequence with the PDB code, look at the accession code column for a 4 letter code.

   (f) Click on that accession code to see information about the protein. What is its name? On the right hand side of the screen you'll see an image of it. Attach a screenshot.

2. This is a followup for question 1. You have now (hopefully) identified a structural relationship between the Query sequence and a protein sequence in the PDB database of protein structures. Say you would like to validate this relationship. This one could do by mutating (substituting) essential residues in the query sequence and test if the protein function (or structure) is affected by these mutations.

   The protein sequence of the query is large (more than 400 amino acids) and a complete mutation study including all residues would be extremely costly. Instead one can use PSI-BLAST and sequence profiles to identify conserved residues that are likely to be essential for the protein structure and/or protein function.

   (a) This is a fun one – go to the blast2logo server at http://www.cbs.dtu.dk/biotools/Blast2logo-1.1/ to create a logo which shows the conservation pattern along the sequence. The amino acid letter codes are depicted in sizes according to their conservation. Paste the query sequence to the window, change the database to NR, the iterations to 3 and the E-value to 0.001 and hit "submit". It may take a while and you can leave your e-mail and get a notification. Attach the resulting logo plot to your submission.

   Note: it may not display due to java settings. You can resubmit and set the output format to pdf.

   (b) Can you understand why the logo is so flat for the first 100 residues (how large a fraction of the query section did the Blast search cover)?

   (c) Out of the following amino acids, select the 4 most conserved according to the logo (the notation is <Amino acid code> <position in the sequence>): L280, R287, E290, Y334, F371, R380, R400, Y436. If it is difficult to detect these position you can zoom into parts of the sequence: Click on Customize Visualization. Scroll down and show advanced settings. You can select a segment (sequence range) to show.

   (d) Newer homology methods use Hidden Markov Models (HMM) to identify homologous sequences. Go to https://toolkit.tuebingen.mpg.de/tools/hhpred. Paste or upload the query sequence and submit. As before, this may take a while. Attach a screenshot of the results page (just the front of it). Look at the first hits. In what place does the Psi-BLAST identified PDB entry appear in HHPred?

   (e) Scroll down and find the sequence alignment and attach a screenshot of it.

   (f) Which of the eight residues listed above are most conserved and hence most likely to be essential for the protein stability and/or function? It may not be easy to track them down because while there used to be a histogram option it seems to not exist anymore. You can see if a residue is conserved by a vertical line between the sequence and the query, and the one letter code appearing in the consensus sequence, either as a capital letter (more confident) or small (less confident).

3. **Protein structure search and classification:** Search the PDB with the entry 2BAA. Download the pdb file as text, and download FASTA sequence as text.

   (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 2BAA in SCOPe (at scop.berkeley.edu) – what is the class, fold and family of this protein?

4. **Protein visualization:** Install Pymol and Chimera, both available for free in the links given in class (handouts for this HW). **Do not buy anything! Use the Pymol evaluation version**. Upload 2BAA.pdb (from the last question) in each viewer.

   (a) On Pymol – upload 2BAA.pdb (from last question). The protein will display as cartoon. On the right hand panel select the colorful "c" button near the "all" and select "color by ss" with the top

option. 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. You can now save the display using File→Export image. It will look nicer to convert the background to white. You can do it with Display→Background→White on the top menu. Select png and save. Submit it with the homework.

(b) On chimera – upload the protein with the PDB code 1BWW (using File→fetch by id). Select all the beta strands (Select → structure → secondary structure → strand), and color them yellow by action → color and select yellow. Attach a printout by either saving a screenshot or File → save image. To save ink you can change the background color to white by selecting favorites → preferences and change ”categories” to background. Click on color and change from black to white.

Programming projects

As before, it is an optional section. You can attach the code to your submission but it won't be graded. Write code that gives you the answers to parts a–d in question 3 above. You can use the PDB module in Biopython. As before, you can also use R but I have less experience with it.