Homework 2: Multiple Sequences and Phylogenetics

Due 5pm, Weds, November 11

Please solve the following problems and email your answers to me as a zip file. You may discuss problems, but work individually on the programming and write up of solutions. Make sure to read the entire problem set before attempting the individual problems.

1. In this problem you will develop an branch and bound algorithm for the *median string problem*. Suppose you have m DNA sequences $s_1, s_2, \ldots s_m$ of length n, the goal is to find the "best" motif of length k. Now, for a pair of strings x and y of the same length, let $d(x, y)$ be the number of positions where x and y differ (also known as the *Hamming distance*). Then, for strings x (of length k) and s (of length n, $n > k$), let,

$$A(x, s) = \min_{1 \leq i \leq n-k+1} d(x, s[i \ldots (i + k - 1)])$$

where $s[i \ldots j]$ represents the substring of s from position i to position j. Then the goal of the median string problem is to find a string x of length k that minimizes

$$score(x, s_1, s_2, \ldots, s_m) \quad = \quad \sum_{\ell=1}^{m} A(x, s_\ell).$$

Note that x naturally corresponds to a motif, and that we are essentially trying to find a motif that "best" matches some substring of length k in every given string s_i. A naive algorithm could simply iterate through all possible strings of length k. This enumeration would yield an algorithm that requires $O(4^k \cdot mn)$ time. Let's develop a way to eliminate motifs without necessarily looking at them.

   (a) Suppose we have a motif x such that $score(x, s_1, s_2, \ldots, s_m) = S$. Consider a string y of length $\ell < k$. How many strings of length k begin with y?

   (b) Give a simple criterion using S by which you could eliminate any string that begins with y.

2. In this question, we will continue in our efforts to gather more information about TranscriptX that was provided to us by our friends at the experimental cancer lab, so that we may be able to generate hypotheses regarding its role in cancer. Since little is known about the protein, X, that this gene expresses in humans, we need to find other sources of information that may help us find out more about its function. Typically, this involves finding proteins that have global or local similarity to the query protein (the one we are interested in) using a tool such as BLAST. We saw in question 2 of HW #1 that there were indeed several proteins in humans that were similar to our query. Here, will take our analysis further by attempting to find common trends in similar proteins, in other animals, which are also seen in X, by using the multiple sequence alignment generated by CLUSTAL (Steps: a-b). Of course, since there might be features in our protein X that are not observed in the homologous proteins selected, we will also look for motifs using Pfam that may generate some hypotheses about its function (Step c). To this end, please perform the following exercises:

   (a) Remember that in order to make evolutionary comparisons on proteins we need to create a list of homologous proteins to derive information from. Let us find 10 proteins similar to X in each of three species - Bos taurus (Cattle), Rattus norvegecus (Brown Rat) and Mus musculus (Common House Mouse) using BLAST. This is similar to the procedure we followed in assignment #1, except that this time we will be searching against three different organisms, where we had previously only searched against humans. Perform the following steps to create a FASTA file - 'sequences_msa.fasta'.

      i. *Run BLAST.* Go to the BLAST website at `http://blast.ncbi.nlm.nih.gov/Blast.cgi` and select 'blastX' to match our DNA sequence against protein databases. In the blastx search page provide the following information and then run BLAST:

A. *Provide the query sequence.* In the section labeled 'Enter Query Sequence' select fasta file provided in HW #1 (sequence_q3.fasta) as the source for the query by clicking the button labeled 'Browse...' and selecting the fasta file.

B. *Specify the database to search against.* In the section labeled 'Choose Search Set', select 'Non redundant protein sequences (nr)'. In the text box for 'Organisms' type 'Bos taurus' to search against Cattle. Note: Non-redundant databases are databases where no exact duplicate sequences can be found. Such duplicates can exist because of several reasons such as new sequencing projects re-adding previously seen sequences into public databases.

C. *Limit the number of sequences you will consider.* Since multiple sequence alignment is computationally expensive, we will, for the purpose of this assignment, select only 10 sequence from each organism. In the section 'Algorithm parameters' limit the number of results by selecting '10' in the drop-down list for 'Max target sequences'.

ii. *Save the FASTA sequences of the matched proteins.* BLAST returns the alignments of sequences in the database which are homologous to our query sequence. We are currently only interested in the sequences for these proteins, so that we may perform a multiple sequence alignment (MSA). In order to get the FASTA sequences for these result proteins, we need to use the NCBI website `http://www.ncbi.nlm.nih.gov/`. Fortunately, the BLAST result page allows us to do so. Scroll to the bottom of the results page from the blastx search and select the check-button 'Select All'. Click on the link 'Get selected sequences'. This will take us to the NCBI page showing a summary of the 10 proteins. Since we are interested in downloading the FASTA file with sequences of these 10 proteins, first select 'FASTA' from the dropdown list under 'Summary'. Then select 'File' under the drop down list 'Send to'. This will bring up a 'Save File' dialog box with the fasta file 'sequence.fasta' as the target file which will have the protein sequences from Bos taurus which have homology to X.

iii. *Repeat the above steps for the other two organisms - Mus musculus (Taxid: 10090) and Rattus norvegicus (Taxid: 10116)*

iv. *Create the fasta sequence file for the subsequent steps*. Combine the protein sequences from the three organisms into one fasta file by concatenating them into one file called 'sequences_msa.fasta'. Also translate the original query DNA sequence from 5'-3' in frame 3 using a translation tool such as `http://www.expasy.ch/tools/dna.html`. Add the translated protein sequence to 'sequences_msa.fasta'. Hint: the translated sequence should start with LLYK, end with SITK, and should be 140 amino acids long.

v. *Submit this file for grading*.

(b) Now that we have some proteins which may provide us some information about protein X, let us perform a multiple sequence alignment of these proteins using Clustal. Clustal is available at `http://www.clustal.org/` to run as a standalone. However, we will be using the Clustal webserver at the European Bioinformatics Institute (EBI) at the site `http://www.ebi.ac.uk/Tools/clustalw2/index.html`. Go to this site and do the following:

i. *Run Clustal*. Select the sequences that we want to perform a multiple sequence alignment on by clicking on the 'Browse' button and selecting the fasta file 'sequeces_msa.fasta' that we created in the previous step (containing the translated original sequence and the results of the three BLAST searches). Click on the 'Run' button.

ii. *View MSA*. In the Clustal results page, click on the button 'Start JalView' to look at the multiple sequence alignment result produced. This should bring up a java applet with the multiple sequence alignment (make sure the pop-up blocker on your browser is temporarily disabled for this). In the applet, each of the protein sequences is arranged horizontally and a consensus sequence for the entire alignment is noted at the bottom. Take a screen shot of the alignment window and submit. As you look along the multiple sequence alignment, do all the columns have the same amino acid in each of the proteins or does a trend manifest itself? Comment on the similarity of the aligned sequences.

iii. *Build the Phylogenetic tree*. Click on the menu item 'Calculate' ->'Calculate Tree' ->'Average Distance Using Percentage Identity'. In the phylogenetic tree that is created, do you notice a trend in the similarity of proteins? Are there groups of proteins that are more similar to each other for example? How does this relate to your observation in the previous question about the similarity of aligned sequences?

iv. *Splitting the tree*. Click on any open spot on the phylogenetic tree, to divide the tree into multiple partitions. The phylogenetic tree view is linked to the multiple sequence alignment view, and diving the tree into partitions should also cause the original sequences to be color coded according to their partitions. Choose a partition that breaks the proteins into two sets in such a way that proteins in a set look qualitatively similar (and well aligned) to each other in the MSA view. Report the set of proteins in each partition.

v. *Creating a better Phylogenetic tree*. Now, create two fasta files, one with sequences for first partition and another with sequences for second partition. Make sure that the query protein is in both fasta files, by copying it to the fasta file corresponding to the partition without protein X. Run Clustal on these two fasta files separately. Report the consensus sequences from these two separate Clustal runs. Are the multiple sequence alignments better with this method and how?

vi. (bonus) How would you modify the scoring parameters in Clustal to achieve a better alignment than was achieved in (3) above?

(c) Let us also see if we can find motifs in our sequence that correspond to protein families of known function. Go to the Pfam webserver at `http://pfam.sanger.ac.uk/` and select 'SEQUENCE SEARCH'. Enter the translated protein sequence into the textbox and click 'GO'. Report the Pfam domains (both significant and not significant) that are reported from the webserver. Is the Pfam domain sequence reported similar to the consensus sequences in the last part of question (b) above?