Biol4559 Mid-term project – Due 5:00 PM, March 23

The goal of the project is to have you use many of the methods that were introduced in the first half of the course, in particular similarity searching, automated scripting of sequence downloads, multiple sequence alignment, and evolutionary tree reconstruction.

Overall, the project is to start with a human protein sequence, identify 20 - 30 homologs, including homologs from rat and mouse, but also homologs that share less than 30% sequence identity, and then build evolutionary trees using both protein and DNA sequences.

Specifically:

1. Identify a human protein family for searching and tree building. Do NOT use class-mu glutathione transferases. You should use a protein family that has a relatively constant length. Some of the families you might consider are:

   MAP Kinase: `sp|P45985|MP2K4_HUMAN`
   Lactoglutathione lyase: `sp|Q04760|LGUL_HUMAN`
   Trypsin-like serine Proteases: `sp|P07477|TRY1_HUMAN`
   Subtilisin-like serine proteases `sp|Q8NBP7|PCSK9_HUMAN` (isolate Peptidase_S8 domain)
   glucose transporter: `sp|P11166|GTR1_HUMAN`
   DS protein phosphatase: sp|P28562|DUS1_HUMAN
   cytochrome P450: sp|P04798|CP1A1_HUMAN
   proline isomerase: sp|Q9NWM8|FKB14_HUMAN

   If you wish to use your own protein family, the protein should be between 100 - 500 amino-acids, should contain a single domain (or multiple domains that are usually found together), and have distant homologs (<30% identity with E()<1e-6).

   Find a set of 20 - 30 clear homologs (E < 1E-6), some of which share less than 30% identity, preferably with homologs from human, mouse, and rat (as well as more distant organisms).

2. Build a multiple sequence alignment of the proteins (or homologous domains) using `muscle`. Look at the multiple sequence alignment to ensure that there are not large numbers of gaps across the alignment (there will be some gaps, but hopefully they will be clustered).

3. Download the set of mRNA (cDNA) sequences that code for your proteins. You may find this easiest to do if you work with RefSeq proteins, because every RefSeq protein (NP_) has a RefSeq mRNA sequence (NM_) that can be translated into the protein. Uniprot proteins do not always have cDNAs.

4. Use the `muscle` protein alignment to direct a mRNA (cDNA) DNA alignment using `tranalign`.

5. Build parsimony, distance, and maximum likelihood trees for both your protein and DNA alignments

6. Use the `consense` program to evaluate the consistency of the trees.

7. Use seqboot to generate a bootstrap for the protein sequences, and then evaluate the robustness of the `protdist/fitch` tree using the boostrap approach.

8. Compare the best consensus from your bootstraps to the other 6 trees.

IMPORTANT:

You should write a script that does every step of the analysis. I should be able to copy that script to an empty directory, run the script (and possibly an associated accession list) and reproduce your results. You do NOT have to script the initial BLAST/SSEARCH search, but, given a list of protein accessions, the script should reproduce your results.