

Bioinformatics Computing
CSE40532/60532
Homework #6

Problems: (due 11/21)

1. Download the small sample set of sequences from the course website.
2. Modify your previous global alignment code to compute the Hamming distance (number of mismatches) between all pairs of sequences in the small sample ignoring indels. Output the resulting simple matrix to the screen, and place the program, the sample set, and instructions to compile/run this code in your dropbox (10 points)
3. Generate by hand the distance-based tree for the small set of sequences using the UPGMA algorithm. You may either hand in your answer as a hardcopy in Prof. Emrich's email box or as an ecopy in your dropbox (5 points).
4. Download and install the phylip package from Prof. Felsenstein (<http://evolution.genetics.washington.edu/phylip/>) and the larger set of sequences from the course website. All programs below are part of phylip.
5. Generate a distance-based tree using phylip for the larger dataset using both KITSCH and NEIGHBOR. You will need to compute a distance matrix first using DNADIST. Place the raw tree files into your dropbox and briefly comment on differences if any. (10 points)
6. Generate a parsimony tree using DNAPARS on the larger dataset. Also run DNAPENNY on this dataset (branch and bound approach). IF DNAPENNY takes a long time to run (more than 30 mins) note this in a file placed in your dropbox. Plot the tree(s) in a nice graphical format using one of the recommended tools and place at least the DNAPARS tree file in your dropbox (10 points).