**COMP 7295/8295 – Assignment 2**

**Due: 9/26/2017 before class.**

In this assignment, you will put your understanding of pairwise alignment and phylogenetic tree construction together to build a phylogenetic tree using real data.

1. (30 pts) Write a Python function that reads a file containing a list of N genes and returns a dictionary, whose keys are gene names and values and DNA sequences that correspond to the keys.

   The input is a file, storing information about N genes. There are 2*N lines in the file. Corresponding to each gene is 2 lines. The first line starts with ">" followed by the gene ID, species name and other information, separated by "|". The second line is a DNA sequence that encodes the gene.

   The output is a dictionary as specified above.

   Use the 3 datasets provided together with this assignment as examples of the inputs. You can use Python function split() to get the species name from the first line.

2. (30 pts) Write a Python function that takes as input a dictionary, whose keys are gene names and values are DNA sequences corresponding to the genes. Your function should return a dictionary D that stores the edit distances between all pairs of genes. In other words, D[('gene1', 'gene2')] should return the edit distance between gene1 and gene2.

   To compute the edit distance between two DNA sequences, set the costs of substitution, insert and deletion to 1; and identical matches to 0.

3. (30 pts) Write a Python function that takes as input a file that has the same format as in problem 1. It should build a phylogenetic tree using the UPGMA method and draw the tree in ASCII format. You should use the distance matrix produced in problem 2.

   Use the 3 datasets provided together with this assignment as examples of the inputs.

4. (10 pts) Same goal and input as in problem 3. For the output, save the phylogenetic tree in an image instead of printing it out.

Turn in instruction:

- The name your solution file should be the same as your UID, plus a .py extension. For example, if your UID is jsmith (i.e. your email is jsmith@memphis.edu), then your solution file should be **jsmith.py**.
- In the file, put your full name, COMP 7295 or COMP 8295, and Assignment
- Send your solution to the TA (Diem-Trang Pham, dpham2@memphis.edu) with the subject line "**COMP 7295 Assignment 2**".