CS5566 Fall 2016 Assignment III. 100 pts. NAME(S):

OBJECTIVES: To understand the basis of substitution matrices; Learning to program in a scripting language
RESOURCES: Textbook, Class website, Internet

Electronic submission on Blackboard is due by 11 pm on Wednesday, Oct 26th, 2016. Submissions received after the deadline will be graded only for effort and receive a maximum of only 70% of the grade (Refer to class syllabus for detailed grading policy). If you are submitting this jointly as a 2-member team, include all names on a single upload and remember that undergrads and grad students cannot partner with each other. Questions marked as G are optional for undergraduate students. All answers should be in your own words with all sources you referred to cited at the appropriate places (Refer to class syllabus for detailed policy on plagiarism). It is highly recommended to use Python for this assignment.

For programming assignments, the submission should include this page as the first page of your assignment, pseudo-code reflecting your understanding of the solution, and examples of actual input and output to validate your implementation. This report should be a standalone summary of your work. A plain text file version of your code should be submitted (include your name as comments at the beginning of the code file), together with test input file(s), and a README file that includes instructions on how to run the program. Combine all files into a single ZIP file before uploading them to Blackboard. State any assumptions you make and show work for maximum credit. In addition to completeness and accuracy, grading will also take into account the modularity and clarity of your implementation. For example, it is necessary to read and write input and output using filenames as arguments rather than hardcoding filenames. It should be possible to run programs from a command prompt or shell by typing "python programName InputFileName OutputFileName." This will help you reuse your code for subsequent assignments. It is necessary to write this from scratch rather than by using pre-existing code. **Make sure your program prints intermediate output to a file so that the underlying steps and reasoning are self-evident.**

1. (100UG/80G) Your goal is in this assignment is to create an amino acid substitution matrix (AASM) from a given multiple sequence alignment and compare it with the values in existing AASMs. Click on any of the domain links of the SMART database at http://smart.embl-heidelberg.de/browse.shtml. Then select "family alignment" in FASTA format to download the corresponding multiple sequence alignment to use as input for your program. Use the following pseudo-code to implement your matrix generator:

For every pair of sequences in the multiple sequence alignment
    Count the number of times each amino acid occurs, adding this to the appropriate counter
    Count the number of times each aligned pair occurs, adding this to the appropriate counter
Calculate P(a) for each amino acid
Calculate P(a,b) for each pair of amino acids
Calculate the AASM as lg [P(a,b) / P(a)P(b)]
Compare your AASM with an existing AASM (PAM or BLOSUM), and compute a correlation coefficient.
Comment on your results.

2G. (20G) Visit http://www.genome.jp/dbget-bin/www_bfind?aaindex and type in "substitution matrix" into the search bar. Examine a few of the matrices and briefly discuss the basis of their similarities and differences.