BINF 8500 Assignment 2

a) **Write a computer program that implements the k-means clustering algorithm**
b) **Use the program to cluster prokaryotic genomes based on amino acid contents of their proteomes**
c) **Write a brief report (~1 page of text + tables and/or figures) on whether it gives the expected results**

Input: tab-delimited text file with point labels and coordinates (one point per line).

Output: Assignment of data points into clusters for multiple values of k (e.g., user-defined, or 1-10). Also provide a table of within-cluster sum of squares (WCSS) and Akaike information criterion (AIC) and/or Bayesian information criterion (BIC) for different values of k.

Notes:

Your program should work with any number (up to 1000) of data points and any number (up to 100) of dimensions.

Ideally, the program should determine the number of data points and dimensions by parsing the input file but if you find that difficult to do you can include the numbers of data points and dimensions in the input as command line parameters.

Sample input files are on eLC and on the cluster (instructor_data).

Try to figure out how many starting configurations you need to use to have a reasonable chance to find the global minimum or at least a minimum close to the global minimum. This might be different for different k's or numbers of dimensions.

Rewards and prizes:

- 35 points for writing a program that works; deductions may be applied for violations of the rules.
- 5 points for the written report
- 1 extra point if you correctly implement kmeans++

The assignment is due on September 20. Submit by email to mrazek@uga.edu and copy to shumeng.zhang25@uga.edu.