Assignment 2: Compute Empirical P-value

You are free to choose whatever programming language you like, but in general you may not use any bioinformatics toolkits or libraries. You are allowed to use libraries that implement data structures (lists, arrays, etc), but you must fully cite the source. For example, if you are programming in C++, the use of the Standard Template Libraries is allowed. If you are programming in python, you may use scipy and the numpy multidimensional array library. If there are any questions as to what is allowed, please send an email to mazen.alborno@ucdenver.edu.

This assignment is to test of your individual algorithmic design and programming skills. If you want clarification about anything else, you may email mazen.alborno@ucdenver.edu. Answers to clarifying questions will be sent out to all students.

You must submit your assignment, which consists of the following components by the due date:

1. Written report
2. Algorithm pseudocode
3. Documented program code and program outputs.

Provide in your submission a README.txt file that includes detailed, concise instructions on how to install your program and its dependencies within a basic environment.

Grading

Your work will be graded based upon three components: a written report (25%), algorithm pseudocode (25%), and implementation (50%). The written report should describe and justify your strategy, define the input files, present the algorithm overview, define any scoring methods, and detail the expected output files. The report should also contain an analysis of the final results and discussion. You should cite the appropriate scientific literature where appropriate. The pseudocode should describe your algorithm in a code-agnostic manner. The implementation must include a working program with extensive comments describing each step, as well as describing all inputs and outputs.

Description

Over the semester, the goal is to Implement Prix Fixe from Tasan et al. 2015 (https://www.nature.com/articles/nmeth.3215).

In this assignment, compute the statistical significance of a population of subnetworks, given the population of subnetworks, the full network and a set of FA loci, as in Tasan et al.

Note that since the genetic algorithm (GA) is not yet implemented, stub out a function for GA which just generates an initial random population of 5000 networks, as described in Methods (Tasan et al.), and returns that population, i.e., no further optimization by GA.

Use “Input.gm.txt” as the set of FA loci.

In your written report, you must have descriptions for the following sections: Motivating Problem from domain, Computational Problem formulation, Specific Approach to the problem (i.e. choice of algorithm), Specific Implementation of approach.