.

HGEN 48800 (2024 spring)
Problem Set #3

Problem 1 (15 points) Suppose we have n professional wrestlers and we have a given list of r pairs of wrestlers for which there are rivalries. Give an $O(n + r)$-time algorithm (pseudocode) that determines whether it is possible to designate some of the wrestlers as part of the 'blue' team and the remainder as part of the 'red' team such that each pair of rivalry is between wrestlers from different teams. If it is possible to perform such a designation, your algorithm should produce it.

Problem 2 (20 points) Count the number of paths from a start vertex s to a destination d in a directed acyclic graph.

Problem 3 (30 points) Implements the De Bruijn graph approach to assembling a sequence.

(a) Simulate the genome: sample a sequence of length 1000bp. You could just use uniform distribution over four nucleotides.

(b) Simulate read data: suppose the read length is 25 bp, sample 400 reads from the genome, with uniform start position.

(c) Construct the De Bruijn graph with $k = 10$.

(d) Perform sequence assembly either by finding the Eulerian path of the graph, or by finding all maximum simple paths (contigs). Both approaches are acceptable. For the Eulerian path, you may find some references helpful, e.g. see this.

(e) Report the statistics of your result, comparing with the true genome sequences.