BMI/CS 776 Spring 2015
Optional Homework #5

Prof. Colin Dewey

Due Sunday, May 10th, 2015 at 10:05am (before the final exam)

1. Consider the reference DNA sequence `CACTACGTACG`.

   (a) Draw the suffix tree for this sequence.

   (b) Show the path taken through the suffix tree to match the query sequence `ACG` and give the matching start position(s) for this query string.

   (c) Draw the threaded suffix trie for all k-mers of the reference sequence, with $k = 3$ (be sure to include the threading pointers).

   (d) Show the path taken through the suffix trie to find all matching k-mers of the query sequence `CGTACGT` and give the matching position of k-mers in the query and reference.

2. Suppose you wish to find a conserved network module between the protein networks of two species (A and B), each of which contains exactly four proteins. The alignment of the nodes of the two networks is the trivial one, with node i in species A corresponding to node i in species B. For each species we have yeast two-hybrid data for each possible protein-protein interaction. The data are in the form of light intensities, as we might get from a luciferase reporter gene. We assume that if an interaction is present, that the light intensities follow a normal distribution with $\mu = 1$ and $\sigma = 1$, whereas if an interaction is *not* present, that the light intensities follow a normal distribution with $\mu = 0$ and $\sigma = 1$. The intensity data are given in the table below.

Edge	Species A intensity	Species B intensity
(1, 2)	-0.36	2.42
(1, 3)	0.65	0.41
(1, 4)	0.31	-0.23
(2, 3)	-1.26	-0.37
(2, 4)	1.72	0.16
(3, 4)	1.70	0.35

Use the conserved network module algorithm described in lecture, with a clique target graph, $\beta = 0.8$, and $p_{uv} = 0.1$ (the null model probability of an edge existing) for all edges.

(a) Compute the weights for all six possible edges. Show your work.

(b) Given your computed edge weights, find the heaviest induced subgraph. Show your work.

3. Suppose we perform a GWAS for a disease of interest (e.g., Type I diabetes) and obtain the data found in Table 1 for one particular site in the genome. We notice that the subjects in the study come from one of two distinct populations, A and B.

population	disease status	genotype	count
A	case	CC	76
A	case	CT	314
A	case	TT	410
A	control	CC	251
A	control	CT	946
A	control	TT	1203
B	case	CC	976
B	case	CT	929
B	case	TT	1295
B	control	CC	1068
B	control	CT	1085
B	control	TT	1447

Table 1: GWAS data for one genomic site

(a) Using a χ^2-test, determine if there is an association between the genotype at this site and disease status within population A.

(b) Using a χ^2-test, determine if there is an association between the genotype at this site and disease status within population B.

(c) Suppose that we did not record which population each subject came from. Using a χ^2-test, determine if there is an association between the genotype at this site and disease status within the entire set of subjects.

(d) Explain your result in (c) in light of your results from (a) and (b).

4. Suppose in a GWAS you perform tests for association at 10 different sites, giving you 10 p-values. Consider two different p-value thresholding procedures: (i) FWER < 0.05 using Bonferroni correction, and (ii) q-value < 0.05 using the FDR approach of Storey and Tibshirani (fix $\pi_0 = 1$).

(a) Give an example list of 10 p-values for which both procedures pick the 5 sites with the smallest p-values.

(b) Give an example list of 10 p-values for which procedure (i) does not pick any sites whereas procedure (ii) picks all of the sites.

5. Consider a simple threading problem in which we have a template with three segments (i, j, k). We are given a sequence for which there are two possible starting positions for each segment. Given the following values for the scores of the individual segments and the scores for segment interactions, show how the branch-and-bound method would find the optimal threading.

$$\begin{array}{lll} g1(i,2) = 4 & g1(j,8) = 2 & g1(k,13) = 1 \\ g1(i,3) = 3 & g1(j,9) = 5 & g1(k,14) = 10 \end{array}$$

$$\begin{array}{llll} g2(i,j,2,8) = 6 & g2(i,k,2,13) = 1 & g2(i,j,2,9) = 0 & g2(i,k,2,14) = 0 \\ g2(i,j,3,8) = 1 & g2(i,k,3,13) = 9 & g2(i,j,3,9) = 0 & g2(i,k,3,14) = 0 \\ g2(j,k,8,13) = 3 & g2(j,k,8,14) = 12 & g2(j,k,9,13) = 5 & g2(j,k,9,14) = 11 \end{array}$$

Use the simple lower bound presented in class. When splitting a threading, split the segment having the minimal $g1$ value for some position (e.g., split on k first since $g1(k, 13) = 1$). To split a selected segment, divide it into two intervals of length one.