CISC 436/636 Computational Biology & Bioinformatics
Fall 2016
Homework 4

Due date: December 1, 2016

1. [25 points] **RNA secondary structure prediction**. For the following RNA sequence

AGACUGUCAC

a. [10 points] Fill out the DP table in the Nassinov folding algorithm to maximize the intramolecular Watson-Crick pairings.
b. [10 points] Show the optimal secondary structure predicted by the algorithm.
c. [5 points] Does the optimal structure have a bifurcation?

3. [15 points] **Lattice model**. A 2D HP model is given with the following energy function:

E_i = -1 if i-th and j-th residues are topological neighbor and both are hydrophobic
= 0, otherwise

a. [5 points] For a sequence HHPHHHPHPPH, find the configuration that minimizes the total energy. Draw the configuration and also represent it as a binary string using the following encoding convention: 00 = right, 11 = left, 01 = up, and 10 = down.
b. [10 points] For the same sequence in part a, draw the two configurations C1 = 00101110001011100010 and C2 =01010100001000010111. Calculate the total energy for each configuration. Form a new configuration C3 by crossing over C1 and C2 using the underlined parts. Draw the configuration C3 and calculate its total energy.

3. [30 points] **Gene expression analysis**. A hierarchical clustering algorithm is given as follows.

Step i. Start with m clusters, each contain one gene, and calculated the ***m x m*** symmetric distance matrix D1 (entries d_{IJ}).

Step ii. Determine from D1 which of the genes (or clusters in later iterations) are least distance. Suppose these happen to be genes (or clusters) I and J.

Step iii. Merge I and J into cluster IJ. From a new distance matrix D2 by deleting rows corresponding to I and J and columns I and J, and by adding a new row and column for distances IJ from all remaining genes (or clusters).

The distance between two clusters is calculated with the two schemes.

Single linkage: $d_{IJ} = \min_{a,b}\{d_{ab}: a \in I \text{ and } b \in J\}$

Complete linkage: $d_{IJ} = \max_{a,b}\{d_{ab}: a \in I \text{ and } b \in J\}$

Step iv. Repeat steps ii and iii a total of m-1 times until a single cluster is formed.

Apply the above hierarchical clustering algorithm to the following hypothetical gene expression data using Euclidean distance.

Gene	exp1	exp2
A	1.0	1.5
B	1.0	1.0
C	3.0	1.0
D	5.5	1.0
E	7.0	1.0
F	7.0	2.0
G	7.0	5.0
H	6.0	6.0
I	8.0	6.0

a. Use single linkage, and draw the dendrogram.
b. Use complete linkage, and draw the dendrogram.
c. Plot the data points, and explain with a diagram why the G, H, I clusters differ for single-linkage and complete-linkage clustering.

4. [30 points] **Regulatory network inference**. Given the following data matrix from DNA microarray experiments, use the Boolean network predictor as discussed in the class to infer the genetic network for genes.

x1	x2	x3	x4	x5	
0	1	1	1	+	P1
1	0	1	–	1	P2
0	1	–	1	1	P3
1	–	1	1	0	P4
+	1	1	1	0	P5

a. For each pair of experiments where the expression level of gene x5 is changed, give the set of genes that also changed their expression levels in the respective experiments.
b. Find a minimum set of genes such that it contains at least one gene from each set obtained in part a.
c. Give the truth table for gene x5's regulation that is compatible with the data matrix. Use * for undecided cases.