BMI/CS 776
Advanced Bioinformatics
Spring 2010 Final Exam

Name _____

Write your answers on these pages and show your work. You may use the back sides of pages as necessary. Before starting, make sure your exam has every page (numbered **1** through **8**).

Problem	**Score**	**Max Score**
1.	_____	28
2.	_____	15
3.	_____	15
4.	_____	15
5.	_____	12
6.	_____	15
Total		100

1. Finding Seeds for Alignments (28 points):

(a) Show how MUMmer would use a suffix tree to find the MUMs in the following two sequences. Be sure to show the MUMs returned.

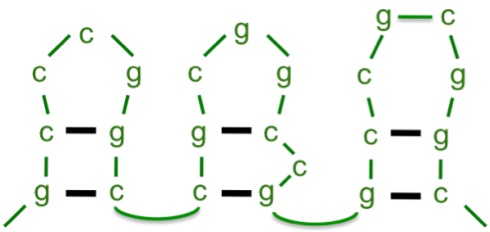
Genome A: **ttggctgg**
Genome B: **ttgggc**

(b) Show the threaded trie that would be constructed to index all 3-mers in Genome A: **ttggctgg**.

(c) Now suppose that you're using the threaded trie to find matching 3-mers in the query sequence **ttgggc**. Show the sequence of nodes in the trie that would be visited when processing this query. (You may want to label each node in your trie with a number, and then write down the visit sequence as a list of these numbers.)

2. RNA Secondary Structure Prediction (15 points): Show how the Nussinov algorithm would predict the secondary structure of the RNA sequence: **ccacugg**. Show the values computed by the dynamic program and the returned secondary structure. You do not need to show the details of the traceback procedure.

3. SCFGs and RNA Secondary Structure (15 points): Consider modeling a simple class of RNA secondary structures, as illustrated in the figure below.



Write down a set of productions and their associated probabilities for this class of RNAs. Assume that the alphabet uses only **c**'s and **g**'s. Your grammar should encode the following requirements

- All <u>three</u> stems always have <u>two</u> sets of paired bases.
- The stems may include a single unpaired base (a "bulge") between the base pairs in the stem (as shown in the right stem).
- The probability of a stem having a bulge is 0.1. The bulge will appear on either side of the stem with equal probability, and it is equally likely to be a **c** or a **g**.
- Each stem position is equally likely to have a **c-g** pairing or a **g-c** pairing.
- Each base in a loop position is equally likely to be a **c** or a **g**.
- The length of each loop is given by the following distribution:

$$P(l) = \begin{cases} 0 & \text{if } \; l < 2 \\ (0.1)^{l-2} \times 0.9 & \text{else if } l \geq 2 \end{cases}$$

4. Pairwise Hidden Markov Models (15 points): Draw the topology of a pairwise HMM that is designed for computing local sequence alignments (i.e., a pairwise HMM that is analogous to Smith-Waterman). Show all of the states and transitions in the model. Clearly indicate what each state in the model represents. You do not need to show transition or emission probabilities.

5. Undirected Graphical Models (12 points): We discussed the application of undirected graphical models to two different tasks: (i) determining breakpoints in the MERCATOR multiple-genome alignment method, and (ii) recognizing named-entities using conditional random fields. For both tasks, briefly describe the following:

(a) What is given as input and what is produced as output when the undirected model is applied to the task?
(b) What do the nodes in the model represent?
(c) What do the edges in the model represent?

6. Short Answer (15 points): Briefly define each of the following terms.

module network

back-off model

fragment-based protein structure prediction (Rosetta)

generalized dictionary (named entity recognition)

breakpoint graph