Biostatistics & Medical Informatics / Computer Sciences 776
Advanced Bioinformatics
Spring 2006 Final Exam

101 Agricultural Engineering Hall, 2:45pm May 9

Name: _____

Write your answers on these pages and show your work. You may use the back sides of pages as necessary. Before starting, write your name on every page of the exam. Also, make sure your exam has every page (numbered 1 through 9).

Problem	Score	Max Score
1.		14
2.		14
3.		18
4.		12
5.		14
6.		14
7.		14
Total		100

May 9, 2006 Name _____

1. Interpolated Markov Models (14 points): Show how you would use an Interpolated Markov Model to calculate the probability of the sequence **atat**. Assume that the alphabet has only two characters, **a** and **t**, and the highest-order model to be used is a second-order model. Assume the λ coefficients in the IMM are defined recursively as in GLIMMER. Shown below are (i) the equation for setting the λ's, (ii) the probabilities estimated from the data, and (iii) the number of times each history was seen in the training set. You don't need to show the numerical results of sums and products, but show the specific numbers that you are using in your calculations.

$$\lambda_n(x_{i-1},...,x_{i-n}) = \begin{cases} 1 & \text{if } c(x_{i-1},...,x_{i-n}) > 400 \\ 0.8 & \text{else if } c(x_{i-1},...,x_{i-n}) > 200 \\ 0 & \text{otherwise} \end{cases}$$

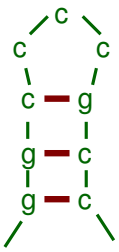
$c(a) = 500$ $c(aa) = 200$
$c(t) = 500$ $c(at) = 300$
$c(ta) = 100$
$c(tt) = 400$

$P(a) = 0.5$ $P(a \mid a) = 0.4$ $P(a \mid aa) = 0.4$
$P(t) = 0.5$ $P(t \mid a) = 0.6$ $P(t \mid aa) = 0.6$
$P(a \mid t) = 0.2$ $P(a \mid at) = 0.5$
$P(t \mid t) = 0.8$ $P(t \mid at) = 0.5$
$P(a \mid ta) = 0.1$
$P(t \mid ta) = 0.9$
$P(a \mid tt) = 0.5$
$P(t \mid tt) = 0.5$

2. Suffix Trees for Finding MUMs (14 points): 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: **ttacgc**
Genome B: **gcatac**

3. SCFGs and RNA Secondary Structure (18 points): Consider modeling a simple class of RNA secondary structures, as illustrated in the figure below. Suppose that the stem part of the RNA structure can be between two and four base pairs long (the stem in the figure is three), and the loop always consists of three **c**'s. Assume that the alphabet has only two characters, **c** and **g**.



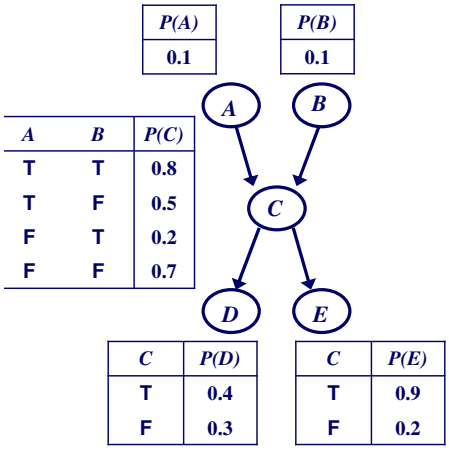
3a. (12 points): Write down the productions and probabilities for an SCFG for this class of RNA structures. The grammar should encode the following requirements:

- The stem length is two base-pairs with probability 0.5, three base-pairs with probability 0.1 and four base-pairs with probability 0.4.
- All base-pairs in the stem must be complementary (i.e., **c**'s and **g**'s paired together).
- It is equally likely to have a **c**-**g** pairing or a **g**-**c** pairing.

3b. (6 points): Now show how to to modify your grammar to encode the fact that it is three times more likely for the first two base-pairs in the stem to be the same (e.g., a **g**-**c** pairing stacked on a **g**-**c** pairing as shown in the figure) than different (e.g., a **c**-**g** stacked on a **g**-**c**).

4. Semi-Markov Models and SCFGs (12 points): We discussed how a semi-Markov model for a given class of sequences separately models its length distribution and its base-composition distribution. Suppose that we want to do something similar for models of RNA secondary structure. That is, we want to separate our representation of the length distributions of stems and loops from our representation of their base compositions. Describe how you would extend the semi-Markov idea to SCFGs for modeling RNA. Briefly discuss how you would (i) adapt the SCFG representation for this idea, and (ii) modify inference methods such as the Inside or the CYK algorithm.

5. Bayesian Network Inference (14 points): For the Bayesian network below, show how you would answer the query $P(a|b,e)$ (i.e., the probability that A is **True**, given that B is **True** and E is **True**) using *Inference by Enumeration*. You don't need to show the numerical results of sums and products, but show the specific numbers you are using in your calculations.



6. Short Answer (14 points): Briefly define each of the following terms:

named-entity recognition

Markov blanket

bootstrap method

sensitivity and specificity

recursive anchoring

parameter tying

multi-MUMs

7. Text Mining and Network Structure Learning (14 points): One challenge faced in inferring regulatory-network models from gene-expression data is that the search space is large (many genes and many possible dependencies among them) yet the available data sets are relatively small (on the order of 100 measurements per gene). Describe how you could use the biomedical literature as another source of evidence in the Bayesian network structure-search process. Assume that you are given a set of relevant abstracts for each gene. Discuss how you will (i) represent the text data, and (ii) use it as evidence during the structure search.