

Biostatistics & Medical Informatics / Computer Sciences 776
Advanced Bioinformatics
Spring 2002 Exam

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 11).

Problem	Score	Max Score
1.	_____	10
2.	_____	38
3.	_____	28
4.	_____	24
Total	_____	100

May 9, 2002

Name _____

1. Sequence Alignment:

1a. (10 points) Briefly compare and contrast the *Gapped BLAST* (sequence database searching) and the *MUMer* (whole genome alignment) algorithms. Focus on the most essential similarities and differences.

2. Probabilistic Sequence Models:

2a. (10 points) Suppose that we wanted to use a *second-order, inhomogeneous* Markov chain model to represent protein coding regions (i.e. genes) in DNA. The model is inhomogeneous because we want to take into account the position within a codon of each base. Draw a picture that shows the states of such a model. Your model should include a *start* state, but need not include an *end* state. Do not show all of the transitions, but instead show only those that are used by the first two codons of a gene that starts: “**ATG TCA...**”.

May 9, 2002

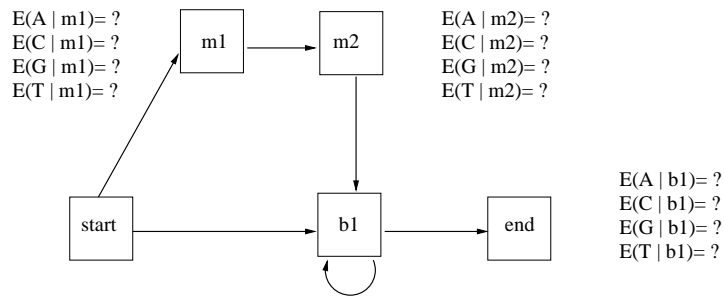
Name _____

2b. (6 points) How does an *interpolated Markov model* differ from a fixed-order Markov model. What is the rationale for using an interpolated Markov model instead of a fixed-order model?

2c. (10 points) Suppose we wanted to use the simple HMM below to detect DNA sequences that start with a two-base motif of interest. The states $m1$ and $m2$ provide a probabilistic representation of the bases in the motif, and the state $b1$ represents other bases in the sequence. In addition to transition parameters, the model has emission parameters in states $m1$, $m2$, and $b1$. Using the sequences below, estimate the transition and emission probabilities for this model.

Use *Laplace estimates* (i.e. pseudocounts of 1) for emission parameters, but *maximum likelihood estimates* for transition parameters.

The sequences in the **positive** column are known to start with the motif, and the sequences in the **negative** column are known *not* to start with the motif.



positive (with motif)	negative (without motif)
ATCGA	GTAC
TAAT	CGTAG
TAC	TGCAT

2d. (12 points) Suppose we now want to use the HMM from **2c.** to decide if the sequence “**ATT**” is likely to start with the motif or not. Show how you would use one of the dynamic programming algorithms for HMMs to make this decision. Show the calculations made by this algorithm as well as your final prediction. Use the parameter estimates that you calculated in **2c.**

3. Clustering and Gene Expression Analysis

3a. (10 points) Given five genes $\{x_1, x_2, x_3, x_4, x_5\}$, and the following **similarity** matrix, show how an *average-link* hierarchical clustering method would cluster the genes. In the matrix, larger numbers represent higher levels of similarity. Show each step of the algorithm as well as the final clustering returned.

	x_1	x_2	x_3	x_4	x_5
x_1		6	6	4	7
x_2			9	5	3
x_3				2	2
x_4					8
x_5					

3b. (6 points) Show how a *single-link* hierarchical clustering method would cluster the genes using the same similarity matrix. Again, show each step of the algorithm as well as the final clustering returned.

3c. (4 points) Suppose we wanted to compare the clustering returned in **3b.** to a clustering obtained with k -means clustering where $k = 3$. Show how you would convert your hierarchical clustering into a partitional clustering with 3 clusters.

May 9, 2002

Name _____

3d. (8 points) Briefly describe how you might use gene expression data, sequence data, k -means clustering and the MEME algorithm together to predict new transcription factor binding sites in a given genome.

4. Stochastic Context Free Grammars for RNA modeling

4a. (6 points) Given the following SCFG, show a sequence that could be generated from the grammar and the parse tree for it. Nonterminals are in lowercase letters, terminals are in uppercase, and the probability for each production is to the right of it.

start	→	h	1.0
h	→	A i U	0.2
h	→	C i G	0.3
h	→	G i C	0.4
h	→	U i A	0.1
i	→	b j	1.0
j	→	A k U	0.3
j	→	C k G	0.3
j	→	G k C	0.3
j	→	U k A	0.1
k	→	b b b b	1.0
b	→	A	0.4
b	→	C	0.1
b	→	G	0.2
b	→	U	0.3

May 9, 2002

Name _____

4b. (8 points) For your sequence, show the RNA secondary structure suggested by the grammar. Use different line thicknesses to distinguish between bases that are adjacent in the sequence and bases that are paired together in the secondary structure.

4c. (10 points) For the following SCFG, and the sequence “**AAAUU**” show how you would calculate $\alpha(3, 4, k)$ using the Inside algorithm. Be sure to show the other *alpha* values that would factor into this calculation, and the actual value that you would get for $\alpha(3, 4, k)$.

start	→	b k	1.0
k	→	a l	0.5
k	→	a u	0.2
k	→	b b	0.3
l	→	k u	1.0
a	→	A	1.0
u	→	U	1.0
b	→	A	0.4
b	→	C	0.1
b	→	G	0.2
b	→	U	0.3