BMI/CS 576 Fall 2012 Final

Instructor: Sushmita Roy

Name:

Question	Full points	Secured points
1	25	
2	25	
3	25	
4	25	
5	25	
Total	125	

Question 1. Phylogenetic trees

1a. (12 points) You are given the following distance matrix for four species, {H, C, G, M}. Use the Neighbor joining algorithm to create a phylogenetic tree for these four species. Recall here you need to correct the pairwise distances for each pair of species, $D_{ij} = d_{ij} - r_i - r_j$, where $r_i = \frac{1}{|L|-2}\Sigma_{k \in L} d_{ik}$, where L is the set of leaves. Furthermore, a new node k's distance from all other nodes, m is $d_{km} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$. Fill in all intermediate values for the distance d, r, and corrected distance D, and show the corresponding trees. The expected format of the matrices for the first iteration is given below.

L = 4

	H	C	G	M
H	0	1	2	4
C	1	0	1	3
G	2	1	0	2
M	4	3	2	0

Table 1: d matrix

H	3.5
C	2.5
G	2.5
M	4.5

Table 2: Iteration 1: r matrix

	H	C	G	M
H	-7	-5	-4	-4
C	-5	-5	-4	-4
G	-4	-4	-5	-5
M	-4	-4	-5	-9

Table 3: Iteration 1: D matrix

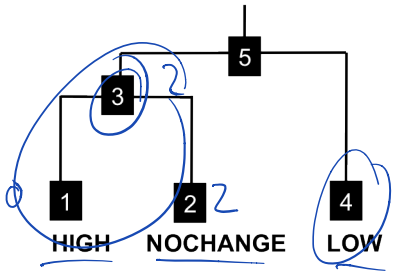
1b. (3 points) Does the algorithm infer a rooted tree or an unrooted tree? How can we convert an unrooted tree to a rooted tree?

find outgroup

1c. (points). An evolutionary biologist is interested in inferring the phylogenetic relationships for a set of three species. She has measured the expression level of a gene in these three species and has found that the gene is either in HIGH, LOW or NOCHANGE states. Using the weighted parsimony algorithm and the following cost matrix determine to infer the cost of the tree and reconstruct the ancestral assignment. In particular populate the S entries in the matrix below.

	HIGH	LOW	NOCHANGE
HIGH	0	3	2
LOW	3	0	1
NOCHANGE	2	1	0

Table 4: Cost Matrix



Recall, here you need to estimate integrate over all scores, $S_i(x)$ which specifies the cost of assigning a character to node i, and is given by the following recursion:

- if i is a leaf node

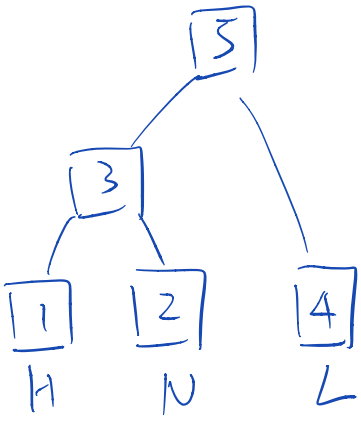
$$S_i(x) = \begin{cases} 0, \text{if } x \text{ is observed at } i, \\ \infty, \text{otherwise} \end{cases}$$

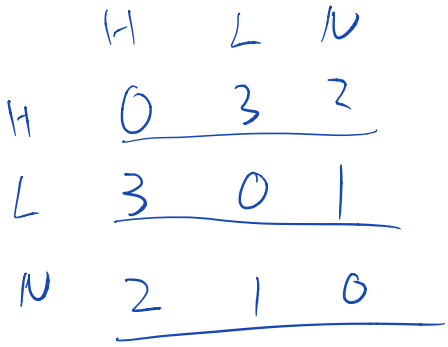
- if i is not a leaf node

$$S_i(x) = \min_y(S_k(y) + C(x,y)) + \min_y(S_l(y) + C(x,y))$$

  k and l are the child nodes of node i, and $x \in \{\texttt{HIGH}, \texttt{LOW}, \texttt{NOCHANGE}\}$ and $y \in \{\texttt{HIGH}, \texttt{LOW}, \texttt{NOCHANGE}\}$

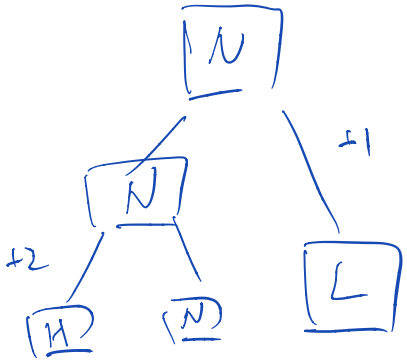
- Termination: cost of tree=$\min_y S_{2n-1}(y)$





	HIGH	LOW	NOCHANGE
$S_1(x)$	0	∞	∞
$S_2(x)$	∞	∞	0
$S_3(x)$	2	4	2
$S_4(x)$	∞	0	∞
$S_5(x)$	2+3=5	3+0=3	2+1=3

Table 5: Cost



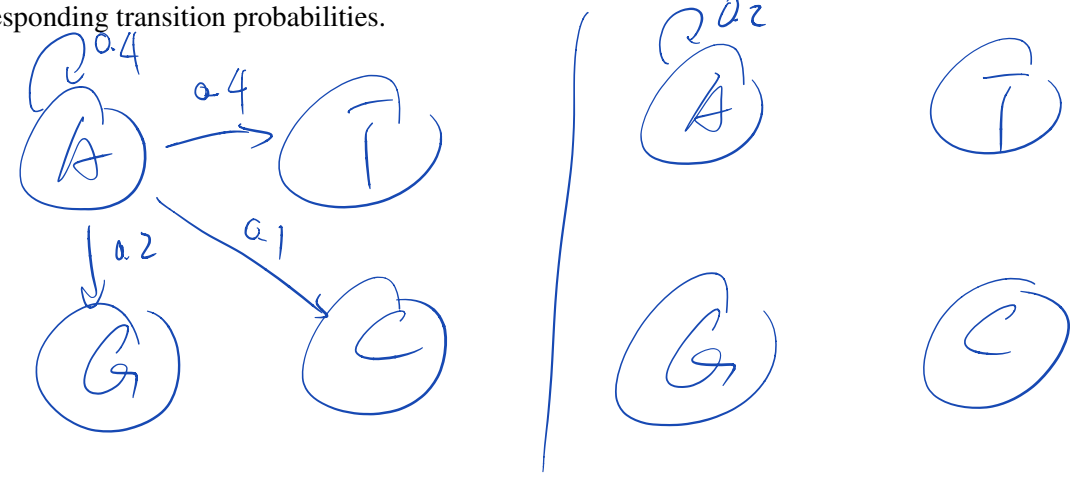
Question 2. Markov models

Biologists have discovered a new of genomic region, *long chromosomal domains* which is associated with some extra-ordinary properties of organization of chromosomes. Now they would like to study the extent to which such specific genomic regions occur elsewhere and also how conserved they are across other species. To do this they have collected a set of positive and negative examples of sequences and estimated the counts of number of times a base X at time t is followed by base Y, where $X, Y \in \{\text{A,T,G,C}\}$, (Table 6).

		$s(t+1)$			
		A	T	G	C
	A	40	40	20	10
	T	20	40	10	30
	G	5	10	25	60
	C	15	15	40	30
	POSITIVE				
		A	T	G	C
$s(t)$	A	20	20	30	30
	T	10	10	20	20
	G	20	20	20	20
	C	25	20	30	25
	NEGATIVE				

Table 6: Counts of transitions of bases for positive and negative examples

2a. (10 points) How would you use this data with a Markov model to learn the parameters of the Markov chain and classify other DNA regions? In particular, draw out the Markov model(s) and label the arcs with the corresponding transition probabilities.



2b. (5 points) Use the above to determine whether sequence ATTAAT comes from a topological domain or not. Assume the initial probability of a base to be $P(A) = P(T) = P(G) = P(C) = 0.25$

ATTAAT

Pos : 0.25 × 0.4 × 0.4 × 0.2 × 0.4 × 0.4

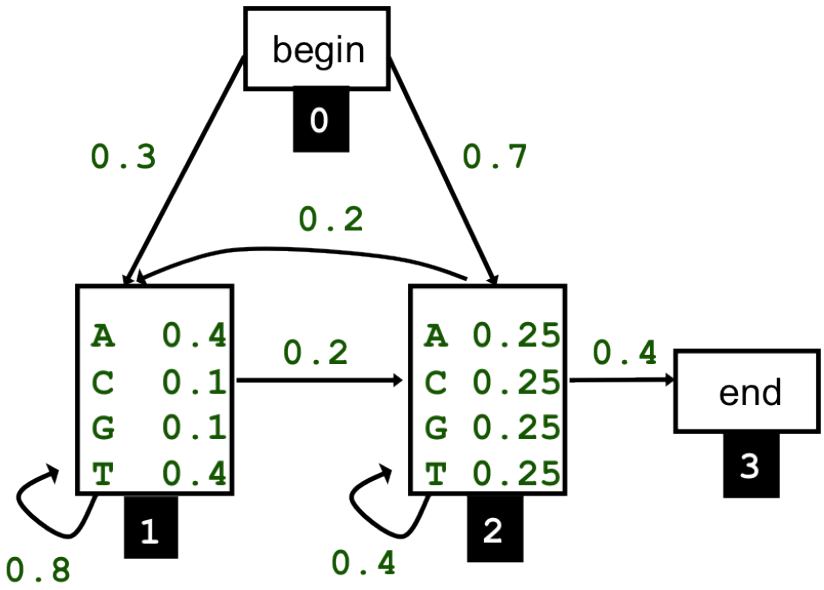
Neg : 0.25 × 0.2 × 0.1 × 0.1 × 0.2 × 0.2

Pos > Neg

Come from topological domain

2c. (10 points) Turns out the biologists are not satisfied with the above approach. Instead they strongly suspect that these regions are of variable lengths and they would like to use a Hidden Markov model to to automatically segment the genome. They have come up with the following Hidden Markov model with a start, end and two non-silent states. Use this HMM to infer the most likely segmentation of the DNA sequence ~~GAATA~~.

GAATA



Verbiti

		G	GA	GAA	GAAT	GAATA	
B	1	0	0	0	0	0	0
1	0	0.03	0.03×0.6×0.4 0.175×0.2×0.4				
2	0	0.175					
E	0	0	0	0	0	0	1

Question 3. Clustering

3a (10 points) Given the set of gene expression data in Table 7, our goal is to cluster these genes into two clusters. Assuming that each cluster is modeled by a Gaussian distribution, execute one expectation step of the EM clustering algorithm. Note, for a n-dimensional Gaussian, with no covariance, the probability of the vector x_1, x_2 is the product of two 1-dimensional Gaussians. Assume we start with an initial assignment of $\mu_1 = \{1, 1\}$ and $\mu_2 = \{4, 4\}$, σ=1 for both Gaussians, and the prior probabilities $p_1 = p_2 = 0.5$.

Gene	Exp 1	Exp 2
A	0	1
B	4	5
C	5	6
D	1	2

Table 7: Expression data

Recall in the expectation step we need to compute the probability of a gene given each of the Gaussian mixture components. The first row is populated. Fill the remaining entries in the table below for the remaining values.

Gene	$\mu_1 = \{1, 1\}, \sigma = 1$	$\mu_2 = \{4, 4\}, \sigma = 1$
A	$\frac{e^{-0.5}}{e^{-0.5}+e^{-12.5}}$	$\frac{e^{-12.5}}{e^{-0.5}+e^{-12.5}}$
B		
C		
D		

Table 8: Expectation step

3b (10 points) Assume our algorithm is in the last iteration and we ended up with the following expected values for our hidden cluster assignments. Use these to estimate the means of the Gaussians in Table 10, and also assign a gene to one of the clusters in Table 11.

Gene	P from 1	P from 2
A	0.7	0.3
B	0.2	0.8
C	0.4	0.6
D	0.9	0.1

Table 9: Expected values of cluster assignments

Mean	Exp1	Exp2
μ_1		
μ_2		

Table 10: Estimated means of Gaussians

Gene	Cluster (1 or 2?)
A	1
B	2
C	2
D	1

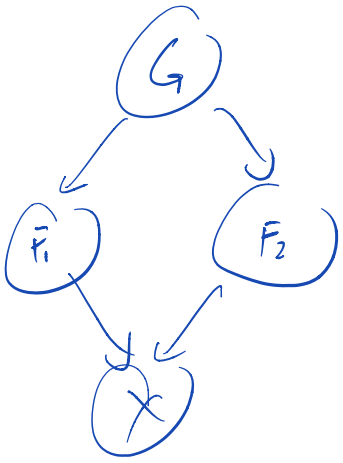
Table 11: Cluster assignments

3c (5 points) If we did not know how many clusters there are how would you estimate the number of clusters?

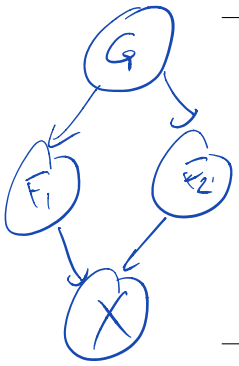
Question 4. Networks

The gene X can exist in three states $0, 1, 2$ which indicate low, no-change, and high states of activity. Whether gene X is in this state is controlled by the presence of two proteins F_1 and F_2. F_1 and F_2 can be themselves be in two states, ON or OFF. Furthermore, F_1 and F_2's state depends upon whether there glucose (G) is in the environment, which can be in two states, `present` or `absent`.

4a (10 points) Draw a Bayesian network which shows the dependencies described in this description.



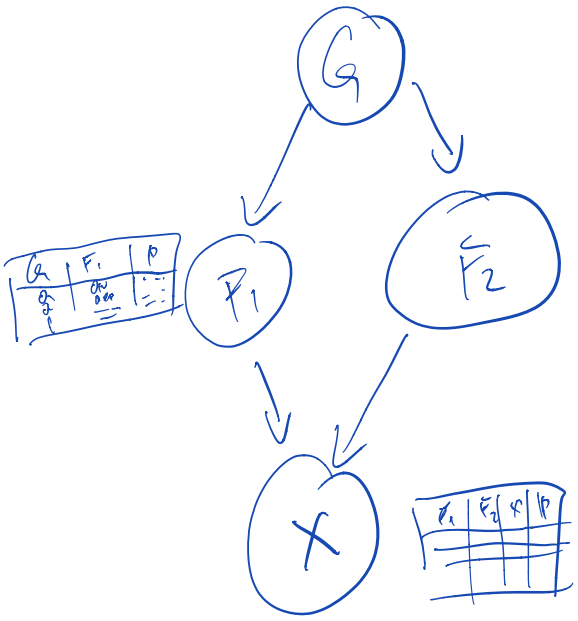
4b (12 points) Assume we make some measurements associated with these four variables and populate the following table Table 12. Use these measurements to compute the Laplace estimates of the conditional probability distribution of the variable X and F_1. In particular, you need to fill two tables, one for $P(X|\text{Pa}_X)$ and $P(F_1|\text{Pa}_{F_1})$, where Pa_X means parents of X.



X	F_1	F_2	G
0	OFF	OFF	absent
0	OFF	OFF	absent
1	OFF	ON	absent
1	OFF	ON	absent
1	ON	OFF	present
2	ON	ON	present
2	OFF	ON	present
2	ON	ON	present
2	ON	ON	present

Table 12: Joint assignments for different variables

4b (3 points) Use the Bayesian network above to compute the probability of the following joint assignment: {X=0,F_1=ON,F_2=OFF,G=`absent`}.



$$P(F_1 = \text{OFF} \mid G = \text{absent})$$

$$\propto P(F_2 = \text{OFF} \mid G \;)$$

$$\propto P(X = 0 \mid F_1 = \;, F_2 = \;)$$

Question 5. Short answers 5 points each

Compare and contrast each of the terms below saying (1) where you have come across these terms, and (2) give one defining characteristic that distinguishes one from the other. Please write no more than 2 sentences for each term.

1. Overlap consensus layout vs debruijn graphs

2. Regression tree vs conditional probability tables

3. Hierarchical clustering vs K-means clustering

4. Neighbor joining vs UPGMA

5. Transcriptional regulatory networks vs protein-protein interaction networks