DUE: Monday March 23 at the BEGINNING of class.
Hand In: Answer the questions on paper, number your answers. Your work must be legible -- if your handwriting isn't great, type it up and print it.

For full credit you must identify key assumptions and provide reasoning (or show work) behind answers. Whenever possible, partial credit will be given if adequate work is shown. Remember I encourage working together, but you MUST indicate all collaborations and/or assistance received or given.

Note that showing work means that if you utilize software for assistance (programs you write or stock software such as Excel), you should indicate as such and provide sufficient details so that I can judge the work. That may mean sending me (by email) source code or associated Excel files.

Questions 1-6 are required of all sections. Questions 7 and 8 (marked Advanced) are required for the graduate sections (5520 and 7000). All questions are weighted equally in the overall grade. Those in the undergraduate section may do the advanced questions for extra credit.

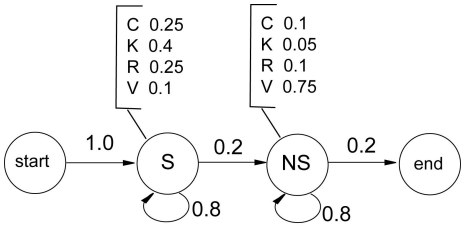
All problems have a maximum value of 10 points. Sub-problem values are marked when appropriate.

Questions

1. You consider using an HMM approach to model protein secondary structure prediction. The straight-forward approach uses three secondary structure confirmations: "**α**-helix", "**β**-strand", and "turn" as the hidden states emitting observable amino acids. It is assumed that the frequencies/probabilities of each of the twenty amino acids can be determined from experimental data for each of those confirmations.

a) (4pt) Draw the state diagram (circles and arrows) of the HMM.
b) (2pt) How many emission parameters are needed to describe this model?
c) (2pt) How many transition parameters are needed to describe this model?
d) (2pt) What is hidden in this hidden Markov model?

2. You suspect that there is a signal peptide in *PepY* and you will use an HMM to predict its position. The model and parameters are given in the graph below. Note in the figure 'S' stands for "signal peptide" state and 'N' (marked NS in the diagram) for "Non-signal peptide" state.



Emission of PepY (s) : **KKRKVRR**
State Path of PepY (π) : **SSSSNNN**

a) (4pt) You are given a sequence s and a path π (above), what is P(s, π)?

b) (6pt) Name the algorithm used for each of the following questions:

(i) Given a sequence, what is the most likely path through the model?

(ii) Given a sequence, how likely did it come from this model?

(iii) Given ***unlabeled*** training data, how do I determine the emission and transition parameters?

3. Consider a new algorithm for predicting whether a particular RNA binding protein binds to an exon. 10,000 exons are evaluated by the prediction method and a cutoff of 2 was selected. Everything scoring above a 2 was considered positive for the RNA binding protein whereas everything below this score was classified as negative. These results were then compared to a gold standard method of determining whether the RNA binding protein associates with the exon. The results are shown in the following table:

	"Gold Standard" Outcome		
Prediction Method	Positive	Negative	Total
Positive	125	25	150
Negative	375	9475	9850
Total	500	9500	10,000

Calculate:

a) (4pt) Sensitivity

b) (3pt) Specificity

c) (3pt) Positive predictive value

4. Consider the following multiple sequence alignment (spaces included for ease of reading) for the proto-insulin gene:

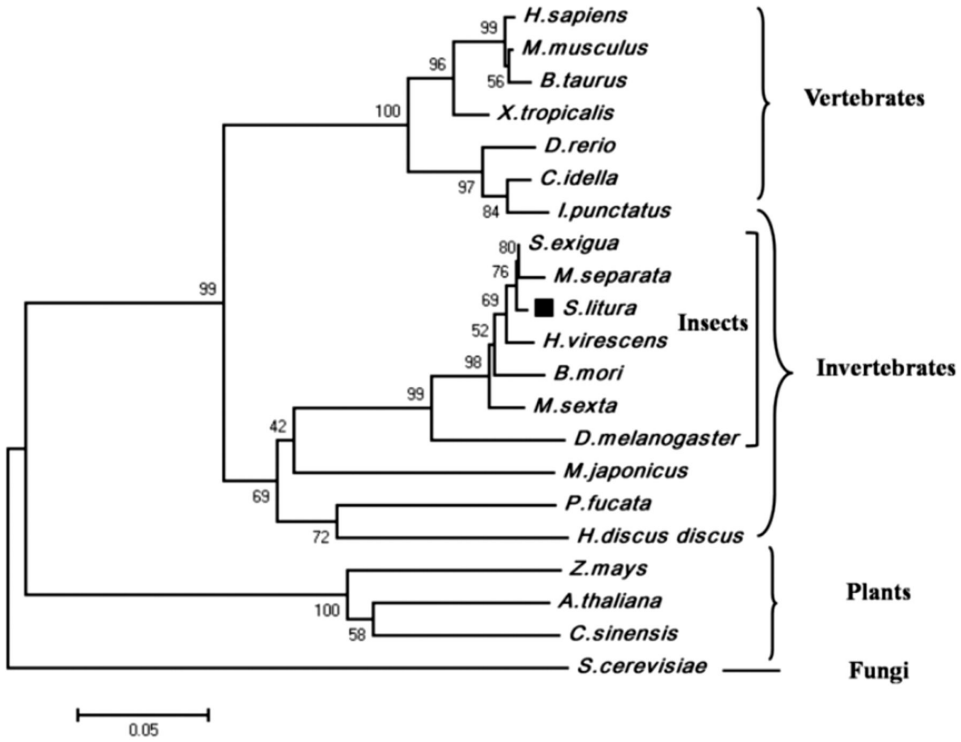
Human: ATGGCCCTGT GGATGCGCCT CCTGCCCCTG CTGGCGCTGC TGGCCCTCTG
Sheep: ATGGCCATGT GGACACGCCT GGTGCCCCTG CTGGCCCTGC TGGCACTCTG
Chick: ATGGCTCTAT GGACACGCCT TCTGCCTCTA CTGGCCCTGC TAGCCCTCTG

a) (4pt) You are considering the Jukes-Cantor model of sequence evolution, which is a single parameter model of evolution (typically described simply as α). Given only the comparison between Human and Sheep as training data, what is your best estimate of α?

b) (3pt) Would the mutation rate be greater or less than the observed substitution rate for mammals? Why?

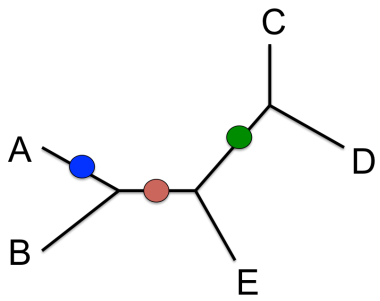
c) (3pt) From the standpoint of constructing a phylogenetic tree, how many positions (columns) in this alignment are informative?

5. Consider the following phylogenetic tree:



a) (2pt) Is this a cladogram or a phylogram?
b) (2pt) Which sequence(s) is/are presumably the outgroup?
c) (2pt) Which sequence is most closely related to A.thaliana?
d) (2pt) Circle (on the tree above) the last common ancestor of M. musculus and D. rerio.
e) (2pt) Which branch(es) do you have the least confidence in? Why?

6. Consider this unrooted tree:



a) (4pt) (Ignore the colored dots for this part.) How many unrooted and rooted trees are possible for this many operational taxonomic units (OTUs)?

b) (6pt; 2pt per node/tree) Draw the three rooted trees that arise by placing the root at each of the three labeled colored dots (blue, red, green).

7. (Advanced) Consider the two state HMM describing DNA sequence that was discussed in class. Namely where one state was GC-poor (we will call this state L) and one state is GC-rich (we will call this state H).

Consider the following parameters of the model:
T(H,H) = 0.5 T(H,L) = 0.5 T(L,H) = 0.4 T(L,L) = 0.6
Emissions:

	A	C	G	T
H	.2	.3	.3	.2
L	.3	.2	.2	.3

The probability of starting in H or L is 0.5 => T(0,L) = 0.5 T(0,H) = 0.5

a) (2pt) Draw the HMM state diagram corresponding to this information.
b) (8pt) What is the most likely path for the sequence GGCACTGAA?

8. (Advanced) (10 pt) Consider the following distance matrix:

	A	B	C	D	E
A	-				
B	90	-			
C	20	100	-		
D	80	30	90	-	
E	50	40	60	50	-

Calculate a rooted tree using the UPGMA method of tree construction. For full credit you must show the final topology of the tree, the calculated branch lengths, the location of the root, and **ALL** intermediate matricies utilized in its construction.