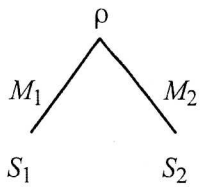
MATH 660 Midterm
March 9, 2012

Name : Solutions

1. (10 pts.) A 2-state model of sequence evolution describes a substitution process on the tree below.



Assuming the two states are R and Y (in this order), the root distribution is $\vec{p} = (0.3, 0.7)$, and the Markov matrices are

$$M_1 = \begin{pmatrix} - & 0.1 \\ 0.05 & - \end{pmatrix}, \quad M_2 = \begin{pmatrix} - & 0.12 \\ 0.08 & - \end{pmatrix}.$$

Give a numerical expression for the probability

$$P(S_1 = Y, S_2 = R).$$

(Do not perform any simplification; your answer may include sums and products.)

$M_1 = \begin{pmatrix} .9 & .1 \\ .05 & .95 \end{pmatrix}$, $M_2 = \begin{pmatrix} .88 & .12 \\ .08 & .92 \end{pmatrix}$

$P(S_1=Y, S_2=R) = P(\rho=R)P(S_1=Y|\rho=R)P(S_2=R|\rho=R) + P(\rho=Y)P(S_1=Y|\rho=Y)P(S_2=R|\rho=Y)$

$= (0.3)(0.1)(.88) + (.7)(.95)(.08)$

2. (10 pts.–5 pts. each) The following distance table exactly fits a metric tree:

	a	b	c	d
a		.4	.6	.9
b			.8	.9
c				1.3

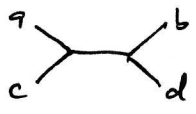
(a) What taxa would UPGMA join first? Draw the *unrooted topological* version of the full tree UPGMA would produce. a+b would be joined first, so the unrooted topological tree would be

(b) Compute the three sums in the 4-point condition, and use them to determine the unrooted topological tree this data fits. (Do not determine edge lengths.)

$d_{ab} + d_{cd} = .4 + 1.3 = 1.7$

$d_{ac} + d_{bd} = .6 + .9 = 1.5$

$d_{ad} + d_{bc} = .9 + .8 = 1.7$



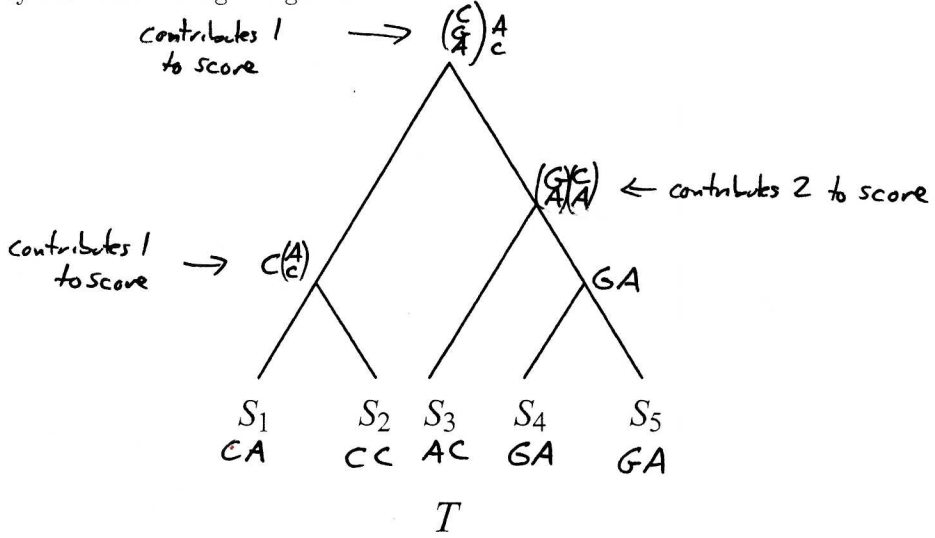
3. (13 pts.) Consider the following aligned sequences for five taxa.

	1
	1234567890
S1	CTGCACGCCA
S2	TTACCCGCCC
S3	CTGCAAGGGC
S4	CTGCAGTTCA
S5	GTGGAGGCCA

(a) (4 pts.) In site 1, the pattern CTCCG occurs. Explain why this pattern is considered parsimony non-informative. (Do not give the definition of non-informative, but instead explain what motivated the definition.) Since there are 3 different bases, on any tree this site requires at least 2 changes. Also on any tree we can achieve only 2 changes by putting a C on every internal node, so there are 2 changes on pendant edges.
Thus this site increases the parsimony score of every tree by 2, & hence has no effect on picking the tree(s) with the lowest score.

(b) (2 pts.) Which sites are parsimony informative? (List the site numbers.)
6, 10

(c) (7 pts.) Using only the sites listed in part (b), compute the unweighted parsimony score for tree T below, by the Fitch-Hartigan algorithm.



T

Total score is 4

4. (6 pts.) Suppose a weighted parsimony analysis with a 1:3 transition:transversion weighting scheme is undertaken, and that the figure below is taken from part of a larger tree during the computation of the parsimony score. Use the Sankoff algorithm to compute the vector of weighted parsimony scores for the parent node of the two child nodes shown. (Use the 1:3 weighting scheme, and assume the order of the DNA nucleotides is A, G, C, T.)

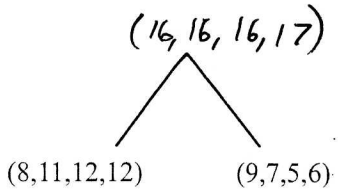
From L From R

A: (8+0) + (5+3) or (7+1) = 16

G: (8+1) + (7+0) = 16

C: (8+3) + (5+0) = 16

T: (8+3) + (5+1) or (6+0) = 17



5. (14 pts.) A collection of 7 unrooted trees relating the 6 taxa a, b, c, d, e, f are to be combined into a single consensus tree. All trees show the trivial splits, so only the *non*-trivial splits are listed in the following table.

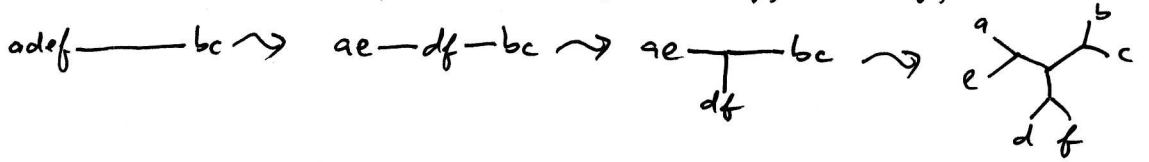
Split	Number of trees
$adef\|bc$	7
$ae\|bcdf$	5
$abce\|df$	4
$abc\|def$	2
$abcf\|de$	2
$aef\|bcd$	1

(a) (4 pts.) Draw the strict consensus tree. (Your drawing should display all trivial splits.)

Tree should have split adef|bc + trivial ones

(b) (6 pts.) Draw the majority-rule consensus tree. (Your drawing should display all trivial splits.)

Tree should have splits adef|bc, ae|bcdf, abce|df, + trivial ones



(c) (4 pts.) Give an example of two splits in the table that are not compatible, and indicate how you know that.

abcf|de + aef|bd are incompatible since neither of the sets in the first split is disjoint from either set in the second.

(Other pairs of splits can be given as examples.)

6. (17 pts.) Below is count data for the frequencies of patterns at the leaves of a 2-taxon tree. The total of all entries in the table is 1000.

		S_2				Row Sum
		A	G	C	T	
	A	146	51	21	24	242
	G	54	149	25	28	256
S_1	C	26	30	148	51	255
	T	28	24	40	155	247

(a) (4 pts.) Without doing any computations, does it appear reasonable to use a time-reversible model to describe this data? Explain briefly. Yes, the table is roughly symmetric about the diagonal (which also implies the row sums are roughly equal to the column sums, & hence there appears to be a roughly stable base distribution).

(b) (4 pts.) Estimate the base distributions $\vec{p}_1$ for S_1, using the order A,G,C,T for the bases. (You need not simplify your answer.) The row sums are shown above, so

$$\hat{\vec{p}}_1 = \left(\frac{242}{1000}, \frac{256}{1000}, \frac{255}{1000}, \frac{247}{1000}\right)$$

(c) (6 pts.) Assuming that the root is taken to be S_1, give an estimate for the entry p_{GC} of the Markov matrix that describes the evolution from S_1 to S_2. (You need not simplify your answer.)

$$\hat{P}_{GC} = \frac{25}{54+149+25+28} = \frac{25}{256}$$

(d) (3 pts.) Looking at the table and perhaps your previous answers, which of the models JC, K2P, GTR, or GM would you pick to describe the data? Explain.

K2P, since the base distribution for S_1 is roughly $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, all transversion entries in the table are roughly equal, transition entries are roughly equal, & the transition & transversion entries are fairly different.

7. (8 pts.–4 pts. each) Suppose that a phylogenetic analysis of morphological data is undertaken for 60 taxa using parsimony. To give an idea of how much work would be involved, answer the following.

(a) Give an expression for the number of unrooted trees that would have to be examined for a full parsimony analysis.

$$(2n-5)!! = 115!! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot 113 \cdot 115$$

(b) How many edges will each of these trees have?

$$2n-3 = 117$$

8. (10 pts.–5 pts. each) The algorithms UPGMA and Neighbor Joining both construct trees from table of dissimilarities.

(a) What assumptions, if any, are made by *both* of these algorithms about the nature of the dissimilarity data? Under what circumstances is this likely to be met for dissimilarities computed by the Hamming distance? Explain. Both assume the dissimilarity data is "tree-like" in the sense that it does correspond to distances along a metric tree. More formally, they assume it is additive, meaning if we have $S0 - S1 - S2$, then $\delta(S0,S1) + \delta(S1,S2) = \delta(S0,S2)$, at least approximately.

(b) What assumptions, if any, are made that are *different* between these methods?

UPGMA additionally assumes the dissimilarity data roughly fits an ultrametric tree, where all leaves are equidistant from the root.

NJ does not make that assumption

9. (12 pts.–4 pts. each) The probabilistic models for DNA evolution used in phylogenetics make many simplifying assumptions about the nature of the mutation process. Briefly explain the meaning of the assumptions below.

(a) Time-reversibility (of a continuous time model) The model parameters to describe evolution are the same regardless of the direction of time. Thus if $S0 \xrightarrow[P_0,\ M]{} S1$, then

$S1 \xrightarrow[P_0\ M]{} S0$ both produce the same expected data for $S0 + S1$.

(Alternately $\mathrm{diag}(p)M$ is symmetric, or $\mathrm{diag}(p)Q$ is symmetric)

(b) The Markov assumption (of all models we have discussed)

Any changes that occur along an edge occur with probabilities determined solely by the state at the immediate parent. The process is thus "memoryless."

(c) The i.i.d. assumption (of all models we have discussed)

i.i.d = independent & identically distributed

"independent" means each site is independent of all others, so what happens at one site has no impact on the others

"identically distributed" means all sites follow the same probabilistic process.