MATH 660
Midterm

Name : Solutions

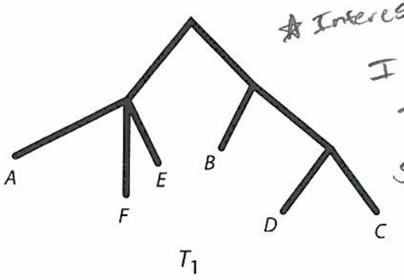
March 9, 2018

Instructions:

1. (10 pts.) Consider the two trees shown below.

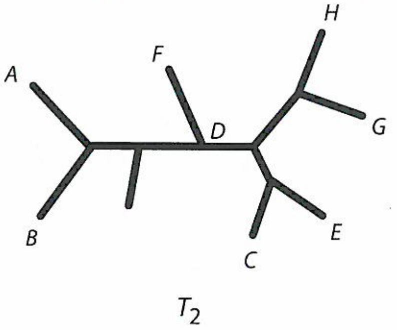
   (a) (8 pts.) Describe these trees using the following adjectives as appropriate. Circle all the adjectives that are correct and cross out the wrong ones. This list has been repeated for your convenience.

   [✓ *rooted,*] [✓ *leaf-labelled,*] [✓ *no-molecular-clock-at-work,*] ~~*metric*~~, ~~*binary*~~, ~~*unrooted*~~, [✓ *multifurcating,*] ~~*with-cherry-AB*~~, [✓ *topological,*] ~~*ultrametric*~~, [✓ *with-sister-taxa-C-and-D*]



★ Interestingly, I intended this to satisfy no-MC and topological !!

   ~~*rooted*~~, ~~*leaf-labelled*~~, [✓ *no-molecular-clock-at-work,*] ~~*metric*~~, [✓ *binary*] [✓ *unrooted,*] ~~*multifurcating*~~, [✓ *with-cherry-AB*] [✓ *topological,*] ~~*ultrametric*~~, ~~*with-sister-taxa-C-and-D*~~



   (b) (2 pts.) List all the cherries in tree T_2.

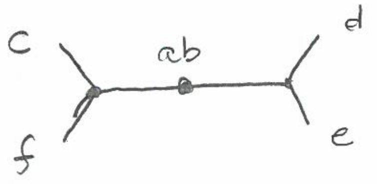
   AB CE GH

2. (7 pts.) Below is a collection $\mathcal{S}$ of compatible splits on the taxon label set $X = \{a, b, c, d, e, f\}$. Give the X-tree that tree popping constructs from this set $\mathcal{S}$.

   $\mathcal{S}$ contains the splits:

   $c \mid abdef$
   $d \mid abcef$
   $e \mid abddf$
   $f \mid abcde$

   $cf \mid abde$
   $de \mid abcf$



3. (10 pts.) Consider the ten characters and the rooted tree T shown below.

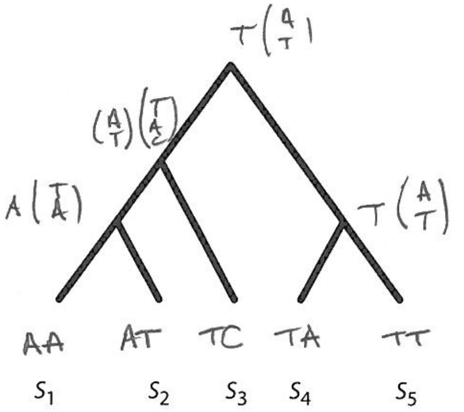
(a) Which sites are the *informative* sites?

↓ ↓

S_1: CCCAA AAAAG
S_2: CCAAA ATAAG
S_3: CCGTA ACAGG
S_4: CCGTA AACCA
S_5: CCGTA ATTTT

4, 7

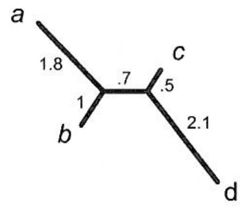
(b) Compute the unweighted parsimony score for this tree, using *only* informative sites.



||||

ps(T) = 4

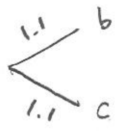
4. (8 pts.) Consider the tree and dissimilarity table shown below for the next questions. Complete answers must include a brief justification.



	a	b	c	d
a		2.8	3.0	4.6
b			2.2	3.8
c				~~2.5~~ 2.6

(a) Which pair of taxa would UPGMA join first? (Break ties at random, if needed.) Then draw the 2-edge *metric* tree that UPGMA would construct to join these two taxa.

b and c since they are closest.

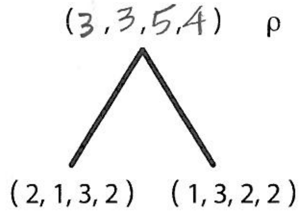


(b) Without much work, from the pairwise dissimilairities on the right you could complete the NJ algorithm and draw the NJ tree for the four taxa a, b, c, d. Justify why this requires essentially no work.

The dissimilarity table is from a tree metric and NJ will reconstruct the tree on the left since it is a consistent estimator (i.e. returns true tree from tree metric dissimilarity table)

5. (10 pts.) In performing a weighted parsimony analysis, you find yourself close to the end, with only one iteration of the Sankoff algorithm left to compute. On the tree below, compute the weighted parsimony score for the tree shown. Use the symmetric weight matrix given below and **order the nucleotides** 'A','G','C','T'. A complete answer will fill in the vector of costs at the root ρ of the tree shown as well as find the parsimony score.

$$\begin{array}{c c} & \begin{array}{cccc} \mathtt{A} & \mathtt{G} & \mathtt{C} & \mathtt{T} \end{array} \\ \begin{array}{c} \mathtt{A} \\ \mathtt{G} \\ \mathtt{C} \\ \mathtt{T} \end{array} & W = \begin{pmatrix} 0 & 1 & 3 & 3 \\ 1 & 0 & 3 & 3 \\ 3 & 3 & 0 & 1 \\ 3 & 3 & 1 & 0 \end{pmatrix} \end{array}$$



C_A : $C_A = 3$

7 pts. ↙

2∧1

A	2 + 0 = $\boxed{2}$	1 + 0 = $\boxed{1}$
G	1 + 1 = 2	3 + 1 = 4
C	3 + 3 = 6	2 + 3 = 5
T	2 + 3 = 5	2 + 3 = 5

3 pts ↙

ps(T) = 3

C_G:

A	2 + 1 = 3	1 + 1 = $\boxed{2}$
G	1 + 0 = $\boxed{1}$	3 + 0 = 3
C	3 + 3 = 6	2 + 3 = 5
T	2 + 3 = 5	2 + 3 = 5

1∧2 3

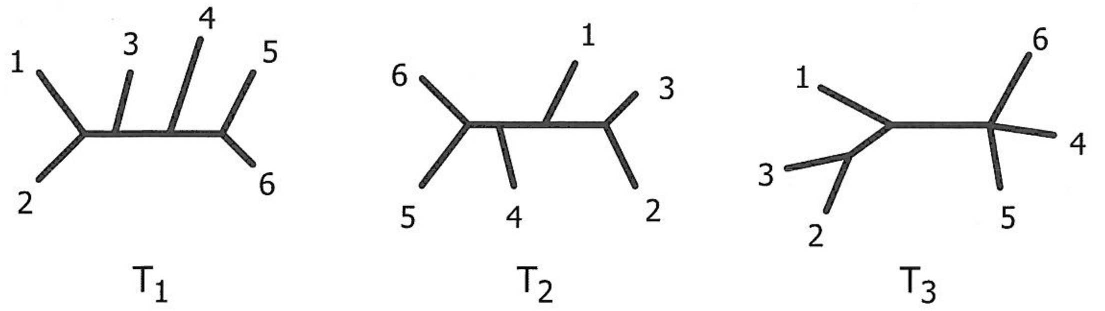
C_C:

A	2 + 3 = 5	1 + 3 = 4
G	1 + 3 = 4	3 + 3 = 6
C	3 + 0 = 3	2 + 0 = $\boxed{2}$
T	2 + 1 = $\boxed{3}$	2 + 1 = 3

C_T:

A:	2 + 3 = 5	1 + 3 = 4
G:	1 + 3 = 4	3 + 3 = 6
C:	3 + 1 = 4	2 + 1 = 3
T	2 + 0 = $\boxed{2}$	2 + 0 = $\boxed{2}$

6. (15 pts.) Below are three trees and a partial listing of the splits on them.

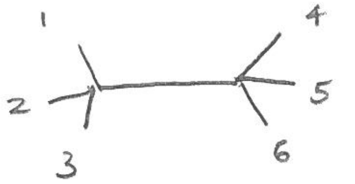


Splits on T_1	Splits on T_2	Splits on T_3
All trivial splits. *	All trivial splits. *	All trivial splits. *
$12 \mid 3456$	$23 \mid 1456$ ✓	$23 \mid 1456$ ✓
$123 \mid 456$ *	$123 \mid 456$ *	$123 \mid 456$ *
$1234 \mid 56$ ✓	$1234 \mid 56$ ✓	

(a) (4 pts.) Complete the listing of the splits on the three trees in the space provided above.

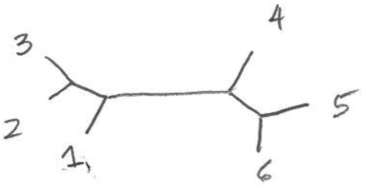
(b) (4 pts.) Compute and draw the strict consensus tree T_{strict} for these trees.

(*) in strict consensus tree



(c) (4 pts.) Compute and draw the majority rule consensus tree $T_{majority\ rule}$ for these trees.

(*) + (✓)



(d) (3 pts.) Consensus trees correspond to picking a value of p where $.5 \leq p \leq 1$ for the cutoff criterion. (For instance, $p = 1$ in part (b), and $p = .5$ in part (c).) For the particular three trees given above, give the range of values $[p_{lower}, p_{upper}]$ so that if $p_{lower} \leq p \leq p_{upper}$ a consensus method with cutoff proportion p will reconstruct the strict consensus tree. (No need to justify. This is an 'all or nothing' question.)

$[2/3, 1]$

7. **Short answer.** (40 pts.) Solutions will be graded both for correctness and quality.

(a) (4 pts.) Why with an unweighted parsimony analysis is it justifiable to eliminate parsimony non-informative sites?

non-informative sites add exactly the same amount to each tree and do not affect the selection criterion of minimal score.

(b) (10 pts.) In choosing between the distance methods UPGMA and NJ, a practitioner should think about the specifics of the dataset to be analyzed.

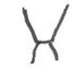
i. (4 pts.) Give, with brief justification, a scenario in which the practitioner might prefer to use UPGMA.

If one believes an MC is at work, or at least a reasonable assumption, the UPGMA is reasonable.

ii. (6 pts.) Why in general is NJ preferred to UPGMA? Give at least three good, distinct reasons.

1. Does not assume a MC

2. Correctly reconstructs a tree from a tree metric dissimilarity map.

3. Correctly joins neighbors in hard to infer trees like 

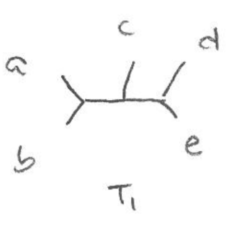
(c) (2 pts.) What is the difference between a *character* and a *state*? Give examples of each in your explanation.

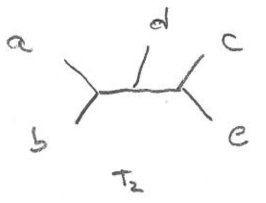
A state is, for example in DNA, 'A,'C,'G,'T' or a possible nucleotide at a site. A character is a pattern of states for n taxa.

(d) (4 pts.) Give an example of three quartets on the set $X = \{a, b, c, d, e\}$ such that these quartets are compatible in pairs, but all three are not compatible. In giving this example, make sure you communicate why they are (or are not) compatible. Use the notation $xy \mid wz$ for the quartets.

See 3a.

More formal definitions okay too.





Q_1, Q_2 are on $T_1 \Rightarrow$ compatible

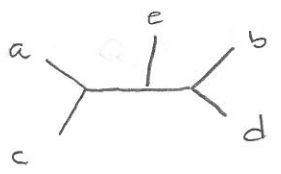
Q_1, Q_3 " " $T_2 \Rightarrow$ compatible

Q_2, Q_3 " " $T_3 \Rightarrow$ compatible

$Q_1 = ab \mid cd$

$Q_2 = ac \mid de$

$Q_3 = bd \mid ce$



But Q_3 is $\underline{\underline{\text{not}}}$ on T_1 since T_1 has $bc \mid de \Rightarrow$ not $\underline{\underline{\text{all}}}$ compatible.

(e) (10 pts.)

i. (6 pts.) Suppose you have a DNA dataset that consists of $k = 1000$ orthologous sites sequenced from a gene in $n = 25$ birds. Assume there are no gaps in the alignment. List three distinct reasons, with justification, for why you would choose to perform a NJ analysis over a parsimony analysis on these data. (You will be judged on the quality [and correctness of course] of your reasons. That is, some answers are *better* than others.)

1. fast, no requirement for searching all trees.

2. permits unequal branch lengths, no MC assumption

3. Consistent

- you want a metric tree to somehow quantify 'distances'
- you want a rough starting tree for further analysis ...

ii. (4 pts.) Give an example (or describe in words) a dataset for which you would choose to perform a parsimony analysis over a distance-based method.

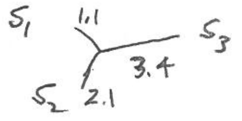
morphological data

(f) (5 pts.) Fit the dissimilarity data in the table below to a 3-taxon unrooted tree.

	S_1	S_2	S_3
S_1		3.2	4.5
S_2			5.5

$\frac{1}{2}(3.2 + 4.5 - 5.5)$

$= 1.1$



(g) (5 pts.) For four taxa a, b, c, d dissimilarity data is given in the table below.

	a	b	c	d
a		7	8	7
b			3	7
c				5

These data (circle one) **DO** / [✓ **DO NOT**] fit a tree with positive branch lengths?

Use the 4-point condition to justify briefly. (You must use the 4-point condition for any credit.)

$d(a,b) + d(c,d) = 12$

$d(a,c) + d(b,d) = 15$

$d(a,d) + d(b,c) = 10$

No.

Reasons: 1, 2 largest not equal

OR

2 $15 \neq \max\{10, 12\}$

Fails 4-point condition ⇒ no tree metric