Homework #3 Selected solutions

3   Exercises

562. General comments and/or things to review:

- Please use the words *character* and *state* precisely. A *character* χ corresponds to a column in a data matrix, and the character is filled with *states*. For example, *long* and *short* might states in a character.
- (MATH) 11e. When counting the number of informative sites, don't forget to subtract the constants sites too.

$$4^n - 4 - 12n - 12n(n-1) - 4n(n-1)(n-2) = 4^n - 4n^3 - 8n$$

10. If characters are given for 3 terminal taxa, there can be no informative ones. Explain why this is the case, and why it doesn't matter.

    Response: An informative character is one in which two states appear at least twice, which would require at least 4 taxa. It doesn't matter, because there is only one unrooted tree on 3 taxa.

11.

$$ps(T) = \sum_{i=1}^{k} f_i ps_{pi}(T).$$

12. Parsimony scores can be calculated even more efficiently by using the fact that several different patterns always give the same score. For instance, in relating four taxa, the patterns `ATTA` and `CAAC` will have the same score.

    a. Using this observation, for 4 taxa how many different informative patterns must be considered to know the parsimony score for all?

    3

    b. Repeat part (a) for 5 taxa.

    $25 = 10 + 15$

13. Use the Sankoff algorithm to compute the weighted parsimony score, with 1:2 weights for transitions:transversions, for the following aligned sequences on the tree which shows A and B most closely related.

    A: TGA
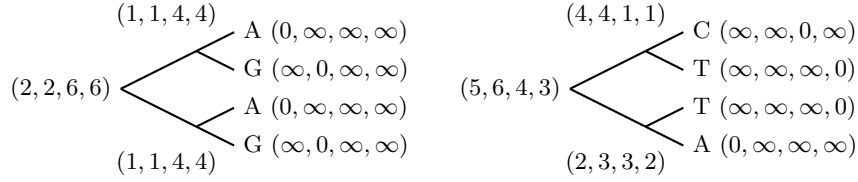   B: TAT
    C: TGT
    D: TAC

$(1,1,4,4)$ A $(0,\infty,\infty,\infty)$

G $(\infty,0,\infty,\infty)$

$(2,2,6,6)$

A $(0,\infty,\infty,\infty)$

G $(\infty,0,\infty,\infty)$

$(1,1,4,4)$

$(4,4,1,1)$ C $(\infty,\infty,0,\infty)$

T $(\infty,\infty,\infty,0)$

$(5,6,4,3)$

T $(\infty,\infty,\infty,0)$

A $(0,\infty,\infty,\infty)$

$(2,3,3,2)$

Figure 1: Trees on which we applied the Sankoff algorithm.

Response: The trees with internal weights for the second and third characters (the first is omitted, as it is constant) are shown in Figure 1 below.

From this, we see that the total weighted parsimony score is $2 + 3 = 5$.

14. For DNA data, what weight matrix makes weighted parsimony the same as unweighted parsimony?

Response: The following weight matrix makes weighted parsimony the same as unweighted parsimony:

	A	G	C	T
A	0	1	1	1
G	1	0	1	1
C	1	1	0	1
T	1	1	1	0

This weights transitions and transversions equally, so every state change has weight 1, as in unweighted parsimony.

18. Create an example of 4 sequences where unweighted parsimony leads to a different optimal tree than weighted parsimony with the transversion weight twice that of transitions.

Response: The following 4 sequences are an example:

$$
\begin{array}{ll}
\text{a:} & \texttt{AAAAA} \\
\text{b:} & \texttt{AAACC} \\
\text{c:} & \texttt{GGGAA} \\
\text{d:} & \texttt{GGGCC}
\end{array}
$$

as can be verified by computing weighted and unweighted scores on each of the 3 possible trees. Informally, in unweighted parsimony the first 3 sites overwhelm the last 2, leading to the tree $((a,b),(c,d))$. With a 1:2 transition:transversion weight, the last 2 sites count more, and overwhelm the first 3, leading to the tree $((a,c),(b,d))$.

21. When applying weighted parsimony to sequence data such as for DNA, it is reasonable to require the weight matrix satisfy the condition

$$w_{ij} + w_{jk} \geq w_{ik}, \text{ for all } i,j,k.$$

2

Explain why, and how this relates to the possibility of multiple mutations along an edge.

Response: The sum on the left requires two mutations to get from i to k, namely, $i \to k \to j$. Since under parsimony, the philosophy is that mutations are rare, it is quite reasonable that it should cost at least as much to undergo two mutations to get from i to j as going directly from state i to state j.

4   Exercises

1. List all the splits displayed on the tree in Figure 4.1, including trivial ones.

   Response: The splits are:

   $$
   \begin{array}{ccc}
   a|bcdef & b|acdef & c|abdef \\
   d|abcef & e|abcdf & f|abcde \\
   ab|cdef & abd|cef & ef|abcd
   \end{array}
   $$

   I can be confident that this is all of them, because the tree has 9 edges, and each split corresponds to an edge.

3. Explain why a trivial split of X is compatible with any other split.

   Response: Given any trivial split $x_0|x_1$, we have a single taxon (let us call it a) as x_0, and all the other taxa as x_1. Because splits are bipartitions, in any other split $y_0|y_1$, a will appear in exactly one of the sets y_0, y_1. Thus, we can intersect x_0 with the set y_i that does not contain a, and the resulting set will necessarily be the empty set.