Homework #11

Selected solutions

9.7 Exercises

9. Consider the two caterpillar trees with n-taxa,

$$T_1 = (\dots((A_1, A_2), A_3), A_4)\dots, A_n),$$
$$T_2 = (\dots((A_1, A_n), A_2), A_3), \dots, A_{n-1}).$$

b) What is the splits metric distance between the two trees?

Answer: The two caterpillar trees have none of their $n-3$ interior edges in common so the Robinson's Found distance is $2(n-3)$.

10.5 Exercises

3. How many parameters are needed for a GTR+I model on a binary n-taxon tree, assuming we want the invariable sites to have the same distribution as the stable distribution of variable ones?

   Answer: Counting branch lengths you should get $(2n-3)+(3)+(6)+(1) = 2n+7$, or $2n+6$ if you normalize.

4. If data are produced according to by a GTR+I model, but analyzed according to a (misspecified) GTR model, one might expect that the edge lengths of a tree would be estimated to be shorter than they actually were. Why? Explain informally.

   Answer: For data produced under a GTR+I, but analyzed under a GTR model the edge lengths are likely to be estimated **shorter** than the true parameter values. This is because the data will look more similar due to the presence of invariable sites (i.e., some of the constant sites were invariable), but the analysis assumes those were free to vary, but did not display mutations.

5. Suppose a pattern distributions is produced from a JC+I model, with r the proportion of variable sites. With substitution rate $\alpha = 1$ and edge length t_e, what is the matrix giving the joint distribution of patterns on a 1-edge tree? If the Jukes-Cantor distance (for a model *without* invariable sites) is used to infer the edge length, does it give t_e? If not, is the estimate it gives larger or smaller than the true value?

   Answer: The joint distribution will have off-diagonal entries $\frac{ar}{12}$ and diagonal entry $1 - \frac{ar}{4}$. Thus the proportion of sites that show non-matching bases is $12(\frac{ar}{12}) = ar$. Since the mixing parameter r satisfies $0 < r < 1$, $ar < a$ and the distance estimate is *less* than it would be if you were to plug in a. To say this another way, the JC distance is an increasing function and $ar < a$ means that you underestimate the branch length when you fail to incorporate invariable sites in your model.