Problem Set 3, from the Midterm

Instructions (for everyone): Do all the problems in Problem Set 1 (relatively straightforward questions). Then, do either Problem Set 2 (harder mathematical problems) or the dataset analysis project in Problem Set 3. Problem Set 1 should take one hour, problem set 2 might take 1-2 hours, but problem set 3 will take substantially longer. Because of this, I am posting Problem set 3 now. Note - expanding on problem set 3 by modifying the input set in various ways could lead to an interesting research paper, and would be potentially the basis of your final project.

Problem Set 3: Do one of the following.

1. Write a paper in which you compare gene trees computed on a biological dataset with at least 50 unaligned sequences using at least two different techniques. You can use your own dataset or find a published dataset. Your paper should provide enough detail to be reproducible (e.g., software version numbers and commands, access to datasets), and should have some interesting discussion about what you observe, and if you were able to make comments about the methods you used. Your grade on this problem will be based on the content, writing, and scientific insight.

    - If you wish, you can use an "alignment-free" method (of your choice), in addition to a method that either co-estimates alignments and trees (e.g., PASTA) or a two-phase method. If you use two-phase methods, then use at least two different multiple sequence alignment methods, of which at least one must be from the following set – Clustal, MAFFT, Opal, Prank, PAGAN, PASTA, and UPP – and then compute a maximum likelihood tree (any software you like). If you compare ways of running PASTA, vary one of the following parameters: subset size, subset aligner, or decomposition strategy (longest branch vs. centroid).
    - Get bootstrap support on the branches of the tree you compute.
    - Compare the gene trees, taking bootstrap support into account. Where are they different? Are these differences interesting or important? What is your interpretation of these differences? If one method did particularly poorly, was there something about the data that was difficult for the method? What did you learn about the methods you used?

2. Write a paper in which you compare species trees computed on a biological dataset with at least 10 genes and between 10 and 100 species. It would be most interesting if you pick a dataset where gene tree heterogeneity has been observed or where it is expected. You can use your own dataset or find a published dataset. Your paper should provide enough detail to be reproducible (e.g., software version numbers and commands, access to

datasets), and should have some interesting discussion about what you observe, and if you were able to make comments about the methods you used. Your grade on this problem will be based on the content, writing, and scientific insight.

- Compute gene sequence alignments and gene trees using reasonable methods. (If you are using a dataset from a published study, these may already be computed for you!)
- Compute species trees using at least two coalescent-based methods and one concatenation analysis. Reasonably fast coalescent-based methods include SVDquartets, MP-EST, ASTRAL, and ASTRID.
- Compare the species trees that you obtain using different species tree estimation methods. Where are they different? Are these differences interesting or important? What is your interpretation of these differences? What does this tell you about the methods you used?