

Homework #4

Due Tuesday Feb. 10 at the beginning of class. Assignments turned in more than 5 minutes after the beginning of class will be penalized 10 points, with an additional 10 points every 24 hours thereafter. You may discuss the homework assignment with other students, but do not share your work. **All Python programs should be run before being turned in. Even experienced programmers can seldom write a program perfectly on the first try.**

For problems 1-4, suggest a good phylogeny method to try for each of the following cases, and briefly explain your choice. (I will accept answers different from mine if they are accompanied by sound reasoning.)

1. (8 points) An insect researcher has a mixture of living and fossil species, and is scoring traits such as presence or absence of wings, number of legs, and type of eyes (simple or compound). She is interested in the overall relationships among her species.
2. (8 points) A microbiologist is sampling water for previously unknown organisms and sequencing their rRNA genes. He wants to roughly classify his new organisms by building trees relating them to his 1500 previously surveyed bacteria.
3. (8 points) An HIV researcher has developed a very precise mutational model, and now wants a detailed tree of the relationship of a handful of strains sampled in Africa, as she is attempting to date the spread of the virus to different countries. She has DNA sequence for one full-length gene.
4. (8 points) A microbiologist is trying to classify bacterial genomes based on the presence or absence of specific genes. He does not have a solid model of the gene gain/loss process, and no gene sequences are available.

For problems 5-7, briefly state any problems you see with the experiment described. Consider overall experimental logic as well as suitability of the phylogenetic algorithm.

5. (8 points) A researcher is trying to make a phylogeny of human chromosome 21 (an autosome) using DNA sequences spanning the length of the chromosome.
6. (8 points) A researcher wants to clarify the relationship among five closely related species of mouse lemur. She sequences a very slow-evolving histone gene and constructs a maximum likelihood tree.
7. (8 points) A researcher wants to find the historical relationship among species of rabbits from forest and desert habitats. He measures traits related to coat color, ear size, heat tolerance, and urine concentration, and constructs a parsimony tree.

Python problems:

8. (15 points) Write a Python function to return the number of positions at which two DNA sequences, passed in as strings, differ.
9. (15 points) Write a Python program that reads in a file of DNA sequences and produces a table showing the number of positions at which each pair of sequences differ (a distance matrix). Import and use the function you wrote for problem 8 (that is, the function should be in a separate file; do not copy it into this program). A sample DNA input file is available as infile.txt; it uses the same (PHYLIP) format that we saw for HW3.
10. (14 points) Modify the program in problem (9) to print only an upper-triangular distance table, like this:

	rat	mouse	cat	dog
rat	0.0	0.3	0.6	0.65
mouse		0.0	0.55	0.6
cat			0.0	0.38
dog				0.0