

## Homework #4

Due Tuesday Feb. 12 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. *Morphological traits almost force the use of parsimony, as we are unlikely to have a good general model for them. Distances might be possible but are unlikely to work well, as we will not be able to develop a good distance correction without a model.*
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. *The large number of species suggests a very fast method; distance matrix methods are the only ones capable of doing such a large data set quickly. We can use the 1500 previous organisms to build a good distance model.*
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. *This is a good situation for maximum likelihood or Bayesian analysis. We have a good mutational model and not too many sequences, so the slowness of these methods won't hurt us. Likelihood is possibly safer because it doesn't require a prior; a Bayesian method will give more information about the degree of certainty. Either one is very suitable.*
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. *This could be handled by parsimony. A distance method with no distance correction could also be tried, but without a correction it will likely be wrong for distantly related species. This is not an ideal situation, but sometimes bad data are the only data available.*

For problems 5-7, briefly state any problems you see with the experiment described.

5. (11 points) A researcher is trying to make a phylogeny of human chromosome 21 (an autosome) using DNA sequences spanning the length of the chromosome. *A recombining human chromosome doesn't have a phylogeny within humans; different parts will have followed different historical paths. Also, as many students pointed out, different parts of the chromosome will need very different mutational models, and the amount of data is probably impractically huge.*
6. (11 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. *Likelihood is not a bad method here, but no method is likely to work well. For a slow-evolving gene all of her closely related species will probably be identical, leaving no information to construct the tree.*
7. (11 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. *All of these traits are strongly selected by habitat, and are likely to group all desert rabbits together due to the shared environment. All desert rabbits will tend to be light colored with large ears, high heat tolerance, and concentrated urine, even if they are not closely related, as these are survival traits in the desert. Neutral traits would have a much better chance of getting the right tree. No phylogeny method will give the right answer on these data.*

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.

I put this function in a file named dnadistance.py:

```
def distance(seq1,seq2) :
    shortlength = len(seq1)
    if len(seq1) > len(seq2) :
        shortlength = len(seq2)
    count = 0
    for pos in range(0,shortlength) :
        if seq1[pos] != seq2[pos] :
            count += 1
    return count
```

9. (20 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.

```
import sys
filehandle = open(sys.argv[1],"r")

#discard the first line
garbage = filehandle.readline()

lines = filehandle.readlines()
filehandle.close()

names = []
seqs = []
for line in lines :
    names.append(line[0:10])
    seqs.append(line[10:].rstrip())

header = ""
for name in names :
    header += "\t"+name
print header

import dnadistance
for index1 in range(0,len(names)) :
    row = names[index1]
    for index2 in range(0,len(names)) :
        row += "\t" + str(dnadistance.distance(seqs[index1],seqs[index2]))
    print row
```