

MCS255 - Lab 1 - Maximum Parsimony

Getting the .m files and the data files:

You should map your home directory (otherwise your work will be on the local machine you are using at the time), make a subdirectory for this course and maybe this lab, and save any files you want to save there.

The files you will need are available at

<http://www.gustavus.edu/~mmcdermo/mcs255/j06/lab/>.

Right click on the files (`seqdata.mat`, `informative.m`, and `primatedata.m`) and save them either to the subdirectory of your home directory that you just created or to `C:\MATLAB7\work`.

Changing the current directory. MATLAB needs to know where the mfiles are stored to access them. You should change the current directory to where the files are located either by typing `cd z:\mcs255\lab` in the command window or by editing the Current Directory in the MATLAB toolbar.

1. Use the maximum parsimony method to construct an unrooted tree for the simulated sequences `a1`, `a2`, `a3`, and `a4` in the data file `seqdata.mat`. First, put the sequence into the rows of an array/matrix with `a=[a1;a2;a3;a4]`. Then find the informative sites with `ainfosites=informative(a)`. Finally, extract the informative sites with `ainfo=a(:,ainfosites)`.
 - (a) What percentage of the sites are informative? You may find the command `size()`, e.g. `size(ainfo)` useful. The output of `size(ainfo)` is two numbers. The first number is the number of rows and the second is the number of columns. Also, you can perform ordinary arithmetic operations with `+`, `-`, `*`, `/`.
 - (b) How many different trees must be considered to find the most parsimonious one relating the four taxa?
 - (c) You may find it too difficult to use all informative sites for a hand calculation. If so, use at least the first 10 informative sites to pick the most parsimonious tree. You can get the first 10 informative sites by entering `ainfo10=ainfo(:, [1:10])` in the command window. You can also print `ainfo10` by highlighting it and choosing *Print selection* under the *File* menu.
2. In this problem, you will attempt to use the maximum parsimony method to construct an unrooted tree for the simulated sequences `d1`, `d2`, `d3`, `d4`, `d5`, and `d6` in the datafile `seqdata.mat`. Begin by finding the informative sites as in the last problem.
 - (a) What percentage of the sites are informative?
 - (b) Compute the number of unrooted trees that must be examined if we really consider all possibilities.
 - (c) Using only the first 10 informative sites, compute parsimony scores of at least 5 candidate trees that you think are likely to be most parsimonious.
 - (d) How confident are you that the most parsimonious tree you found is actually the most parsimonious? What percentage of the possible trees did you compute parsimony scores for? What percentage of the informative sites did you use?

3. In this problem you will examine some of the primate data. Type `primatedata` in the command window. You can see the names corresponding to the data by entering `Names_hominoids=Names_primates(1:5)`. Note that MATLAB is case sensitive. Select out the hominoid data with `h=Seq_primates([1:5],:)`. As in the previous problems, you can find the informative sites as follows:

```
hinfosites=informative(h),  
hinfo=h(:,hinfosites), and  
hinfo10=h(:,[1:10])).
```

Isolate the first 10 informative sites. Use these to compute the parsimony score of each of the following trees, as well as the one with neighbor pairs (chimpanzee, gorilla) and (orangutan, gibbon).