

MCS255 - Lab 2

Distance Methods: UPGMA and Neighbor Joining

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`, `compseq.m`, `distH.m`, `hamming.m`, and `nj.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.

Part I: UPGMA and Neighbor Joining

In this problem, you will construct both UPGMA and NJ trees for data that was simulated using MATLAB. There are two data sets to consider:

- a: This data relates four sequences, a_1, a_2, a_3 , and a_4 . These sequences were generated assuming a molecular clock.
- b: This data set relates five sequences, b_1, b_2, b_3, b_4 , and b_5 . These sequences were generated without assuming a molecular clock.

Step 1. Initialize data.

```
load seqdata      % load the sequence data
who               % list all variables
a=[a1;a2;a3;a4]   % creates a 'matrix' with sequences a1, a2, a3, a4
                  % Notice that the semicolon indicates the end of a row.
a                 % shows the contents of the sequences stored in a
```

Step 2. Compute pairwise Hamming dissimilarities between all pairs of sequences. Organize these computations into a table. Since the sequences are long, this should not be done by hand (unless you are a masochist). Instead, use MATLAB to compute the Hamming dissimilarity. Some useful commands are:

```
size(a)           % determine the size of the array a
help compseq      % view the help for the function compseq.m
F=compseq(a1,a2)  % compute a frequency table for the sequences a1 and a2
help distH        % view the help for the function distH.m
d=distH(F)        % compute the Hamming distance between a1 and a2
```

Repeat the above sequence of commands, modified appropriately, for the pairs a_1, a_3 , a_1, a_4 , etc.

Repeat Step 1, appropriately modified, for **b**. For the b_1, \dots, b_5 sequences, to compute the Hamming dissimilarities for all pairs at once, you can use the command `hamming(b)`.

```
b=[b1;b2;b3;b4;b5] % creates a 'matrix' with sequences b1, b2, b3, b4, b5
help hamming      % view the help for the function hamming.m
distB=hamming(b)  % compute pairwise Hamming distances for sequences in b
                  % Store these distances to a variable named distB
```

Step 3. By hand, compute the UPGMA trees that relate the sequences for **a** and **b**.

Step 4. Use MATLAB to create the neighbor joining trees. (Modify the commands below to construct the tree for the **b** sequences.)

```
help nj           % view the help for the function nj.m
distA=hamming(a)  % compute pairwise Hamming distances for sequences i
Names={'a1','a2','a3','a4'} % define a cell array with the names of the taxa
                  % Notice the braces in the syntax
nj(distA,Names{:}) % steps you through the NJ algorithm
```

Step 5. **Question:** Compare and contrast the UPGMA and NJ trees for both **a** and **b**. For each of these data sets, what do you notice? Does knowing that **a** was simulated using a molecular clock while **b** was not make you prefer one tree over another? Explain. Which, if any, method do you prefer? What biological circumstances might make you prefer one method to the other?

Part II: The NJ algorithm in detail

In class, some details of the computations involved in the NJ algorithm were omitted. In this problem, you will fill in these gaps and construct a NJ tree by hand for four taxa. First, let's recall some definitions and the NJ algorithm.

Definition. Let

$$\begin{aligned}d_{ij} &= \text{dissimilarity between } S_i \text{ and } S_j, \\R_i &= \sum_{j=1}^N d_{ij}, \text{ the total dissimilarity between taxon } S_i \text{ and all other taxa,} \\M_{nm} &= (N - 2)d_{nm} - R_n - R_m.\end{aligned}$$

As you saw in class, if S_n and S_m are neighbors, then

$$M_{nm} \leq M_{nk}$$

for all $k \neq m$. This is the criterion NJ uses to join neighbors.

Neighbor Joining Algorithm.

1. Given dissimilarity data for N taxa, compute a new table of values of M using the definition of M_{nm} . Choose the smallest value in the table for M to determine which taxa to join. (This value may be, and usually is, negative, so 'smallest' means the negative number with the greatest absolute value.)
2. If S_i and S_j are to be joined at a new vertex V , temporarily collapse all other taxa into a single group G , and determine the lengths of the edges from S_i and S_j to V by using the 3-point formulas for S_i, S_j and G as in the algorithm of Fitch and Margoliash.
3. Determine distances/dissimilarities from each of the taxa S_k in G to V by applying the 3-point formulas to the distance data for the three taxa S_i, S_j and S_k . Now include V in the table of dissimilarity data, and drop S_i and S_j .
4. The distance table now includes $N - 1$ taxa. If there are only 3 taxa, use the 3-point formula to finish. Otherwise go back to step 1.

Question: Before working through an example of Neighbor Joining, it is helpful to derive formulas for steps 2 and 3 of the algorithm. Suppose we've chosen to join S_i and S_j in Step 1.

- a. Show that for Step 2, the distances of S_i and S_j to the internal vertex V can be computed by

$$\begin{aligned}d(S_i, V) &= \frac{\delta(S_i, S_j)}{2} + \frac{R_i - R_j}{2(N - 2)}, \\d(S_j, V) &= \frac{\delta(S_i, S_j)}{2} + \frac{R_j - R_i}{2(N - 2)}.\end{aligned}$$

Then show the second of these formulas can be replaced by

$$d(S_j, V) = \delta(S_i, S_j) - d(S_i, V).$$

b. Show that for Step 3, the distances of S_k to V , for $k \neq i, k$ can be computed by

$$d(S_k, V) = \frac{\delta(S_i, S_k) + \delta(S_j, S_k) - \delta(S_i, S_j)}{2}$$

Question: Consider the dissimilarity data in the table below.

	$S1$	$S2$	$S3$	$S4$
$S1$.83	.28	.41
$S2$.72	.97
$S3$.48

Table 1: Taxon distances

Use the Neighbor Joining algorithm to construct a tree as follows:

- a. Compute R_1, R_2, R_3 and R_4 and then a table of values for M for the taxa $S1, S2, S3$ and $S4$. To get you started

$$R_1 = .83 + .28 + .41 = 1.52 \text{ and } R_2 = .83 + .72 + .97 = 2.52$$

so

$$M(S1, S2) = (4 - 2).83 - 1.52 - 2.52 = -2.38.$$

- b. If you did part (a) correctly, you should have a tie for the smallest value of M . One of these smallest values is $M(S1, S4) = -2.56$, so let's join $S1$ and $S4$ first.

For the new vertex V where $S1$ and $S4$ join, compute $d(S1, V)$ and $D(S4, V)$ by the formulas in part (a) of the previous problem.

- c. Compute $d(S2, V)$ and $d(S3, V)$ by the formulas in part(b) of the previous problem. Put your answers into the new distance Table 2.

	V	$S2$	$S3$
V		-	-
$S2$.72

Table 2: Group distances

- d. Since there are only 3 taxa left, use the 3-point formulas to fit $V, S2$, and $S3$ to a tree.
 e. Draw your final tree by attaching $S1$ and $S4$ to V with distances given in part (b).

Question: Would the UPGMA tree agree with this NJ tree topologically? Explain.