

This document is intended to give detailed instructions on how to create a sample file and run the LIBSHUFF program.

Programs used in this tutorial:

- \* Microsoft Word
- \* ClustalX
- \* PHYLIP (only DNADIST required)
- \* LIBSHUFF

1. All of the sequences in the libraries to be compared must be aligned by a multiple sequence alignment program (in this protocol, ClustalX). The program requires a single file containing the sequences in any number of common formats (e.g. FASTA). When preparing this information, make sure that the sequences in this input file are grouped by library and are not interleaved.
  - For example, consider a file containing a total of 200 sequences in the libraries to be compared. If library X has 90 sequences and library Y has 110 sequences, the first 90 sequences in the file should compose library X and the last 110 should belong to library Y.
  - Once the sequences are loaded into ClustalX, select the “Alignment – Output Format Option” menu and change the “Output order” to read “Input”. This will maintain the grouping of your libraries during the alignment. Also, make sure that the output format of the alignment is “PHYLIP” before initiating a complete alignment of the sequences.
2. Transfer the output of the ClustalX alignment (typically ends with the suffix “phy”) to the folder containing the PHYLIP suite of programs.
3. Open the program DNADIST. When prompted for the name of the input file, enter the name of the alignment file. Select the option “D” until “Jukes-Cantor” is selected as the distance algorithm. Select the option “Y” to run the program. Once completed, the file entitled “outfile” will contain the evolutionary distances of all of the sequences in your libraries and can be found in the PHYLIP folder.
4. Open the “outfile” using a word processing program such as Word. Make a note of the size of your libraries and the order in which they appear in the alignment. Find the option to reveal all formatting (e.g. paragraph marks, spaces, etc.) if the formatting marks are not already displayed. In order to format the data correctly for the program, remove sequence names, all paragraph marks except those just prior to the next set of sequence data, and all extraneous spaces. For example, if the first two entries in your DNADIST outfile appear as:

22	1153								
Seq1	0.0000	0.1614	0.1753	0.3578	0.3255	0.2745	0.3634	0.3664¶	
	0.3507	0.3589	0.3534	0.3620	0.3541	0.3611	0.3669	0.3706	0.3546¶
	0.3378	0.3699	0.2856	0.3793	0.4438¶				
Seq2	0.1614	0.0000	0.1786	0.3197	0.3139	0.2280	0.3119	0.2987¶	
	0.3083	0.3383	0.3247	0.3409	0.3612	0.3623	0.3695	0.3664	0.3626¶
	0.3151	0.3457	0.2735	0.3478	0.3483¶				

They should be reformatted to appear as:

0.0000	0.1614	0.1753	0.3578	0.3255	0.2745	0.3634	0.3664	0.3507	0.3589
0.3534	0.3620	0.3541	0.3611	0.3669	0.3706	0.3546	0.3378	0.3699	0.2856
0.3793	0.4438¶								
0.1614	0.0000	0.1786	0.3197	0.3139	0.2280	0.3119	0.2987	0.3083	0.3383
0.3247	0.3409	0.3612	0.3623	0.3695	0.3664	0.3626	0.3151	0.3457	0.2735
0.3478	0.3483¶								

In this correctly formatted entry, the sequence names, all paragraph marks except for the last one just prior to a value for a new sequence, and all unnecessary spaces have been removed. The remaining paragraph marks are important in letting the program know where one column of data ends and another begins. It is also extremely important that the sequences are still clustered by library. Once the reformatting is complete, save the file as “text-only”, with the name “sample”.

*Note: Use the “Find and Replace” command to find double spaces and replace with single spaces, as well as the utilization of macros to remove paragraph marks and sequence names can make this process much faster. See method 1, step 8 for some detail on creating macros.*