

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
- \* pileup (a program used in the GCG software package [Genetics Computer Group, Wisconsin])
- \* Microsoft Excel
- \* PHYLIP (only DNADIST required)
- \* LIBSHUFF

NOTE: This tutorial uses programs to which not everyone may have access. This guide is an example of one possible way to make a 'sample' file for use with LIBSHUFF.

1. Use the 'pileup' program to create an alignment containing all of the sequences in the two libraries to be compared. Make sure that you list of sequences are sorted by library. To keep the sequences in the input order and not interleaved, use the '-NOSORT' option of pileup:

```
!!NA_MULTIPLE_ALIGNMENT 1.0
PileUp of: @example.list

Symbol comparison table: GenRunData:pileupdna.cmp  CompCheck: 6876
                GapWeight: 5
                GapLengthWeight: 1

rubro.msf  MSF: 1377  Type: N  October 30, 2000 16:28  Check: 4315
..

Name: Soilsam1      Len: 1377  Check: 4120  Weight: 1.00
Name: Soilsam2      Len: 1377  Check: 1299  Weight: 1.00
Name: Soilsam3      Len: 1377  Check: 2893  Weight: 1.00
Name: Soilsam4      Len: 1377  Check: 847   Weight: 1.00
Name: Soilsam5      Len: 1377  Check: 9396  Weight: 1.00
Name: Soilsam6      Len: 1377  Check: 6873  Weight: 1.00
Name: Soilsam7      Len: 1377  Check: 4136  Weight: 1.00
Name: Lakesam1      Len: 1377  Check: 9770  Weight: 1.00
Name: Lakesam2      Len: 1377  Check: 5135  Weight: 1.00
Name: Lakesam3      Len: 1377  Check: 5114  Weight: 1.00
Name: Lakesam4      Len: 1377  Check: 7157  Weight: 1.00
Name: Lakesam5      Len: 1377  Check: 5067  Weight: 1.00
Name: Lakesam6      Len: 1377  Check: 4911  Weight: 1.00
Name: Lakesam7      Len: 1377  Check: 7597  Weight: 1.00

//

                1                                50
Soilsam1  CGCTGGCGGC GTGCCTAACA CATGCAAGTG GAGCGACGAA CCG.GGCTTC
Soilsam2  CGCTGGCGGC GTGCCTAACA CATGCAAGTA GAGCGACGAA CCG.GGCTTC
Soilsam3  ~~~~~GCGGC GTGNTTAAACA CATGCAAGTG GAGCGACGAA CCA.GGGTTT
Soilsam4  ~~~~~ GTGCTTAAACA CATGCAAGTG GAGCGACGAA GCG.GACTTC
Soilsam5  ~~~~~GC GTGCTTAAACA CATGCAAGTG GAGCGACGAA CCA.GGGCTT
Soilsam6  ~~~~~ ~CGCTTAAACA CATGCAAGTC GAG.CGATAA CCATGGCTTC
```

```

Soilsam7 ~~~~~GC GCGCTTAACA CATGCAAGTC GAGCCGATAA CCATGGCTTC
Lakesam1 ~~~~~~AACa CATGCAAGTC gAG.CGATAA CCATGGCTTC
Lakesam2 ~GCTGGCGGC GCGCTTAACA CATGCAAGTC GAG.CGAGAA CCTTTCCTTC
Lakesam3 ~~~~~~CGCTTAACA CATGCAAGTC GAA.CGAGAA CCTTGCCTTC
Lakesam4 ~~~~~~ ~~~~~~ ~~~~~~ ~AGCGAGAAC CTGCCCTTCG
Lakesam5 ~~~~~~ ~~~~~~ ~ATGCAAGTC GAGCGAGAAC CCAGGCTTCG
Lakesam6 ~~~~~~ ~~~~~~ ~ATGCAAGTT GAGCGAGAAC CAGCGCTTCG
Lakesam7 ~~~~~~ ~~~~~~ ~~~~~~ ~~~~~~ ~~~~~~

```

```

51 100
Soilsam1 GGCCCGGGGA AGAGCCGCGA ACGGGTGAGT AACACGTGGG TAACGTGCCC
Soilsam2 GGCCCGGGGA TAAGCCGCGA ACGGGTGAGT AACACGTGGG TAACATGCCC
Soilsam3 GCCCTGGGGN AGAGCCGCGA ACGGGTGAGT AACACGTGGG TTACCTGCCT
Soilsam4 GGTCCGTGGC AGAGCCGCGA ACGGGTGAGT AACACGTGGG TAACCCACCC
Soilsam5 GCCCTGGGGC AAAGCCGCGA ACGGGTGAGT AACACGTGAG TAACCTGCCC
Soilsam6 GGCCATGGGG AGAGCGGCGA ACGGGTGAGT AACACGTGAG CGATCTGCCC
Soilsam7 GGCCATGGGG AGAGCGGCGA ACGGGTGAGT AACACGTGAG CGATCTGCCC
Lakesam1 GGCCATGGGG AGAGCGGCGA ACGGGTGAGT AACACGTGAG CGATCTGCCC
Lakesam2 GGGATTGGGG ACAGCGGCGA ACGGGTGAGT AACACGTGGG TAATCTGCCC
Lakesam3 GGGCTTGGGG ACAGTGGCGA ACGGGTGAGT AACACGTGGG TAATCTGCCC
Lakesam4 GGCCTGGGGA .AAGCGGCGA ACGGGTGAGT AACACGTGGG TAACCCACCC
Lakesam5 GCCTGGGGGA .AAGCGGCGA ACGGGTGAGT ATCACGTGGG TaATCTGCCC
Lakesam6 GCTCTGGGGA CAAGCGGCGA ACGGGTGAGT AACACGTGGG TAATCCACCC
Lakesam7 ~~~~~~G AAAGCGGCGA ACGGGTGAGT AACACGTGAG TAACCTGCCC

```

plus many hundreds of more bases....

(NOTE: THIS ALIGNMENT IS AN EXAMPLE ONLY. AN ALIGNMENT FOR USE IN LIBSHUFF WILL MOST LIKELY CONTAIN >100 SEQUENCES [minimum 50 per library].

2. This alignment must then be formatted to run in PHYLIP. Dr. Jose Gonzalez of the University of La Laguna, Spain, has graciously allowed us to distribute a visual basic macro for use in Microsoft Word which will format the file for you. You can find the instructions for downloading the macro and installing it on the main LIBSHUFF webpage. If you are not using Word, you may need to do the formatting by hand (go to step 6 when the formatting is done).
3. Load the alignment file into Microsoft Word. This may entail moving the file from a remote server which contains the pileup program to your local machine.
4. In Word, go the toolbar menu item 'Tools', then 'Macros', and finally 'Macros...'. Select the option for the 'phylip' macro (after it has been installed, step 2).
5. The macro will ask for a starting base and an ending base. These numbers should be the regions in the alignment where all sequences contain nucleotide information. For example, in the alignment example given above, this would be at base 60. Because of the way that the macro is written, the starting base number should be '59'. The ending number will be later in the alignment and where sequences no longer contain information. After the macro has completed, it will look something like this:

```

14          541
Soilsam1   A AGAGCCGCGA ACGGGTGAGT AACACGTGGG TAACGTGCCC
Soilsam2   A TAAGCCGCGA ACGGGTGAGT AACACGTGGG TAACATGCCC
Soilsam3   N AGAGCCGCGA ACGGGTGAGT AACACGTGGG TTACCTGCCT
Soilsam4   C AGAGCCGCGA ACGGGTGAGT AACACGTGGG TAACCCACCC
Soilsam5   C AAAGCCGCGA ACGGGTGAGT AACACGTGAG TAACCTGCCC
Soilsam6   G AGAGCGGCGA ACGGGTGAGT AACACGTGAG CGATCTGCCC
Soilsam7   G AGAGCGGCGA ACGGGTGAGT AACACGTGAG CGATCTGCCC
Lakesam1   G AGAGCGGCGA ACGGGTGAGT AACACGTGAG CGATCTGCCC
Lakesam2   G ACAGCGGCGA ACGGGTGAGT AACACGTGGG TAATCTGCCC
Lakesam3   G ACAGTGGCGA ACGGGTGAGT AACACGTGGG TAATCTGCCC
Lakesam4   A -AAGCGGCGA ACGGGTGAGT AACACGTGGG TAACCCACCC
Lakesam5   A -AAGCGGCGA ACGGGTGAGT ATCACGTGGG TaATCTGCCC
Lakesam6   A CAAGCGGCGA ACGGGTGAGT AACACGTGGG TAATCCACCC
Lakesam7   G AAAGCGGCGA ACGGGTGAGT AACACGTGAG TAACCTGCCC

```

```

CGATGATTGG GACAACCCGA GGAAACTCGG GCTAATACCA AATGTGCCCT
CGATGATTGG GACAACCCGA GGAAACTCGG GCTAATACCA AATGTGCCCT
CGATGACCGG GACAACCCGA GGAAACTCGG GCTAATACCG GATGTGCCCG
CAATGACCGG GACAACCCGA GGAAACTCGG GCTAATACCG GATGTTCTGT
CGATGACCGG GACAACCCGN GGAAACTCGG GCTAATACCG GATGTGGTGC
TCGACACTGG GATAGCCCGG GGAAACCCGG ATTAATACCG GATAGCCTCT
TCGACACTGG GATAGCCCGG GGAAACCCGG ATTAATACCG GATAGCCTCT
TCGACACTGG GATAGCCCGG GGAAACCCGG ATTAATACCG GATGGCCTCT
TCGACATCGG GATAGCCCGG GGAAACCCGG ATTAATACCG GATAGCCTTA
TTGACATCGG GATAGCCCGG GGAAACCCGG ATTAATACCG AATAGCCTCC
TTGGTACTGG GATAGCCCGG GGAAACCCGG ATTAATACCG GATGGCCCAA
TCGACATCGG GATAGCCCGG GGAAACCCGG ATTAATACCG AATGGCCCAT
TCGGCACC GGATAGCCCGG GGAAACCCGG ATTAATGCCG GATGGCCCGT
TCGGCACC GGATAGCCCGA GGAAACTCGG ATTAATACCG GATAGCCATT

```

```

ATGGCCGTAA GGCTTGTGGG GAAAGGAAGC TTCGGCCTCC GTATCGGGAT
CCGACCATAA GGTTGTCTGGG GAAAGGAAGC TTCGGCCTCC GCATCGGGAT
C-----AAGGG GAAAGGAAGC TTCGGCCTCC GCATCGAGAT
ACTTTTCGTAA GGAAGTCCAG CAAAGATAGC TTCGGCCTTC GCATTGGGAC
ACATTCTTAA GTTTGTGTAC TAAAGGAAGC TTCGGCCTCC GCATTGGGAG
CGGGCCACG GGCTCGTGAG AAAAGATGGC TTCGGCTTTC GGTGAGGAG
CGGGCCACG GGCTCGTGAG AAAAGATGGC TTCGGCTTTC GGTGAGGAG
CGAGCCACG GGCTCGTCAG AAAAGATGGC TTCGGCTTTC GGTGAGGAG
CCGGTTTTCG GGCTGGTAAG AAAAGGTAGC TTTGGCCTCC GGTGAGGAG
GGAGCCTTCG GGCGCCGAG AAAAGGTAGC TTCGGCCTCT GGTCAAGGAT
CAGCTCTTCG GGGCGGTTGG AAAAGGTAGC TTCGGCCTCC GACCAAGGAC
CTGCTCTTCG GAGCGGCTGG AAAAGGTAGC TTCGGCCTCC GGTGAGGAG
CGACCCTTCG GGGCTGACGG AAAAGGTAGC TTCGGCCTCC GGCCGGGGAC
CGAGCTCTCG AGCGCGAATG AAAAGGTAGC TTCGGCCTCT GGCCGAGGAT

```

and many more bases....

Note the changes in the alignment. Two numbers begin the file. The first is the number of sequences, and the second is the total number of bases. Names of sequences only appear on the first line. All periods and '~' have been replaced with dashes. Save the file with the name 'infile' as a 'text-only' file. Now move the file into the folder containing the PHYLIP programs (a link to the PHYLIP website is provided on the main page).

- Once the 'infile' is in the PHYLIP folder on your computer, double-click on the DNADIST program. PC users may get a message saying something to the effect of "Infile not found". If this happens, simply type in 'infile.txt'

(newer versions of Windows append '.txt' at the end of text-only files). The only setting you need to change in the DNADIST program is the distance algorithm used. Hit 'D' followed by 'RETURN' three times, until 'Jukes-Cantor' is displayed as the algorithm used. Then hit 'Y' and 'RETURN' to run the program. Depending on the speed of your machine, this may take some time.

7. DNADIST creates a file called 'outfile' which contains the distances of all of your sequences to each other. Open this file in Microsoft Word.
8. In order to format the outfile so that a distance matrix can be formed, follow these steps:
  - (a) Remove the number at the top of the file which indicates how many sequences there are,
  - (b) Use the 'Find and Replace' command to find all double spaces " " and replace with a single space " ",
  - (c) Repeat b until no double spaces are found,
  - (d) Use the 'Find and Replace' command to remove any "-" symbols; find "-" and replace with nothing,
  - (e) You need to remove all paragraph marks from the ends of every line which ISN'T the last distance for a given sequence. You probably won't see these symbols on your document, but they are there. Since you will probably have a large number of sequences to modify, it's easier to create a macro just for this situation.
    - (I) Place the cursor before the first sequence name,
    - (II) Go to the "Tools", "Macros", "Record New Macro" tool,
    - (III) Assign the macro to the Keyboard, don't worry about the name,
    - (IV) Give the shortcut key (for example, "CTRL" and "Z" at the same time), and hit "ASSIGN", then "CLOSE",
    - (V) Once the macro is recording (a small window with symbols for 'STOP' and 'PAUSE' shows up), go to the 'Find and Replace' command, click on the button marked "MORE", under the menu for special, select "Paragraph Mark". Make sure that nothing is typed in the "Replace with..." field. Hit "Find Next." Hit "Replace" (the cursor should be at the end of the first line. Continue "finding next" and "replacing" until you get to the last line BEFORE the next sequence (where the **STOP** is in the following example).

**AT NO TIME HIT "REPLACE ALL".**

```
Soilsam1 0.000 0.001 0.049 0.022 0.026 0.029 0.193 0.199 0.194 0.199
          0.154 0.144 0.167 0.199 0.200 0.202 0.221 0.224 0.226 0.228 0.247
          0.299 0.401STOP
Soilsam2 0.01 0.000 0.050 0.024 0.030 0.040 etc ...
```

- (VI) Hit the square symbol (STOP) on the small floating 'macros record' window.
- (VII) From where the cursor now is (highlighting the paragraph mark at the end of the first series of distances), you should be

able to just run the macro (by hitting the shortcut you gave, e.g. CTRL-Z) again and again until all sequences in the list have been formatted correctly. Make sure you stop at the end of the list. If you hit the shortcut key one too many times, your first several sequences will be messed up.

(VIII) Select everything in the document and under the "Edit" menu hit "Copy".

- (f) Open Microsoft Excel. In a new window, hit "Paste".
- (g) Your data may or may not have been inserted nicely into a matrix. If it wasn't, you'll need to select the "Text to Columns" tool from the "Data" toolbar. Follow the wizard provided with this tool and you should have a nicely formatted distance matrix.
- (h) Since the "-NOSORT" option of pileup was used, the matrix should be sorted correctly with all of the sequences in the first library listed first and all of the sequences in the second library listed together second.
- (i) Make a note of which library appears first in the matrix and how many sequences it has. Now delete the names of the sequences. This is easily accomplished by selecting the entire row (either Row 1 or Column A), and selecting "Delete" from the "Edit" toolbar. Make sure all names are removed (on the top, and side).
- (j) Under the "File" toolbar, select "Save as...". Save this first as a Microsoft Excel Spreadsheet called "sample". Now, using the same "Save as..." command, save it as a "Tab delimited" file. If it asks if you want to overwrite the previous file called "sample", say yes. For some reason, some files which are saved directly as tab-delimited without first saving as an Excel document are not read correctly.

9. You now have a correctly formatted sample file! Move it into the same folder as LIBSHUFF and have fun!