

The COMBINE and REORDER programs by Edward L. Braun

The “combine” and “reorder” programs were written to simplify the manipulation of nexus files containing multiple complex regions. These programs have two major goals:

1. To make it easy to generate supermatrices that preserve CHARSET information associated with the individual partitions.
2. To extract subsets of taxa and subsets of sites, automatically removing any columns that are all gaps and missing data in the desired taxa, while preserving CHARSET information.

Installing the programs

These programs have been tested on OS X Macs and on some linux distributions. The executables should work on intel Macs running OS 10.5 and should compile on any Mac OS X machines with a C++ compiler. If you do not have the compiler installed, it can be obtained from <http://developer.apple.com/tools/xcode/>. The combine, reorder and auth_subset programs can be installed using the “install” script in the “combine-0.9” directory – simply cd to the “combine-0.9” folder and type:

```
$ ./install
```

The “reorder” program uses the nexus class library (Lewis 2003); this library is present in the src directory. reorder is based upon the ncltest program distributed with the NCL. The README is available in word format, pdf format, and as text (content is identical).

Using the programs

I. To generate a supermatrix from a set of individual nexus files

The ability to collect sequence data from multiple gene regions has increased the use of large-scale supermatrices. Ideally, it would be possible to annotate individual partitions and then combine them into a single nexus file that preserves this annotation. This can be especially important if the individual partitions are relatively complex. For example, a number of the 19 loci used in the recent “Early Bird” study of avian phylogeny (Hackett et al. 2008) include multiple introns and exons and have regions that are ambiguously aligned. We found that it was easier to work with and annotate the individual partitions and then combine the data, but maintaining the charsets was difficult. “combine” solves this problem. To use combine, follow these steps:

1. Prepare a set of nexus files. They do not need to have the same set of taxa and the taxa can be in any order.
 - a. At this point, combine only works with DNA sequence alignments.
 - b. CHARSTATELABELS cause reorder to crash. Simply comment CHARSTATELABELS out before using combine and/or reorder.
 - c. CHARSETs and TAXSETs should be in an ASSUMPTIONS block, not a SETS block.
 - d. Use only upper case characters in the DATA or CHARACTERS block.

The COMBINE and REORDER programs
by Edward L. Braun

2. Prepare an **authority** file listing the taxa you want to include in the merged file. The set of taxa in the authority file can be either a superset of the taxa in nexus files or a subset, although it should authority file taxa should be a set for which at least four taxa are present in each of the nexus files.
 - a. For use with combine, the authority file should only include the taxon names. Nothing else should be included.
 - b. All taxa should be present in at least one of the nexus files.
 - c. If a taxon is not present in a specific input file, it will be added as missing data.
3. Run combine as follows (as above, you should be in the “combine-0.9” folder):

```
$ ./combine <outfile> <authority> <infile1> <infile2> .. <infileN>
```

Note that at least two infiles must be specified. If combine is called with an incorrect number of arguments it will provide a brief summary of the usage. The output file will be a nexus format file (with a taxa block, a characters block, and an assumptions block). Each of the input files is presented as a single block with a comment indicting the source file. The CHARSETs and TAXSETs are listed in blocks with comments indicating the source file. If you use same CHARSET names in different input files you will want to edit those names in the final output file (note that PAUP* will only store the information about the last occurrence of the CHARSET name).

II. To generate a nexus file with a subset of taxa and/or sites

It may be desirable to focus on a subset of taxa for specific analyses. For example, when Harshman et al. (2008) analyzed the phylogeny of ratites we focused on only 16 taxa despite having sequences for 70-171 taxa from the individual loci. If you are using regions with large numbers of indels, simply deleting the taxa that are not included will result in a large number of columns that include only gaps and missing data. The reorder program can be used to solve this problem. To do this, you should use reorder as follows:

1. Have a single nexus files with the same format requirements as shown above (the subpoints to point 1 in part I).
2. Prepare an **authority** file listing the taxa you want to include in the output file.
 - a. The authority file will typically be a subset of the taxa in input nexus file.
 - b. If the authority file includes a taxon that is not present in the input nexus file that taxon will be added to the output file as missing data.
3. Run reorder as follows:

```
$ ./reorder <infile> <authority> <outfile>
```

“reorder” can also be used to extract a subset of the of taxa and sites. For example, if you had a supermatrix file and wanted to extract a single partition, you would prepare a modified authority file using a simple program called “auth_subset” and then “reorder”.

1. Prepare a text file (the “removal list”) listing the taxa that are not present in a single partition.

The COMBINE and REORDER programs
by Edward L. Braun

2. Run the auth_subset program:

```
$ auth_subset <infile> <removal_list> <outfile> <inclusion_set>
```

<infile> is the input authority file (e.g., the list of all taxa in the supermatrix file), <removal_list> is defined in part 1, <outfile> is the output authority file, and <inclusion_set> is the CHARSET of characters that will be retained (all characters that are not in <inclusion_set> will not be included when “reorder” is run).

Note that it is also straightforward to manually alter the authority file – “auth_subset” is simply provided because it may be easier to work with a list of taxa that were not obtained for a specific partition. The <inclusion_set> can be added manually at the beginning of the file using the syntax “# inclusion_set_name”. Note that <inclusion_set> has to be a CHARSET in the Assumptions block of the input nexus file used with “reorder”

3. Run reorder as shown above. This will produce an output nexus file that includes only the characters in <inclusion_set> and only the taxa in the modified authority file.

III. To generate a reorganized nexus file

As described above with part II, “reorder” can be used to generate an output file in any order. Rearranging taxon order can be useful for manually refining sequence alignments (the order of taxa has the potential to bias manual adjustment, so it may be desirable to use a “conservative” taxon order to avoid biasing the alignment).

Troubleshooting

The reorder program is sensitive to a number of aspects of the nexus format. Two of these issues are listed in part I of the “Using the programs” section. Potential issues with the nexus format include:

1. CHARSTATELABELS should not be used. If you use CHARSTATELABELS, simply comment them out before using combine and/or reorder.
2. CHARSETs and TAXSETs should be in an ASSUMPTIONS block, not a SETS block.
3. Taxon names should not be repeated – they will halt data reading.

The reorder program produces a text file called “output.txt”. This file can be useful for troubleshooting – it will tell you whether data was read by “reorder” and how. The “combine” program calls “reorder”, so its problems are often associated with “reorder”. The programs also echo a substantial amount of information to the screen, and this information can be used to determine whether specific taxa were found in the file and whether specific CHARSETs or TAXSETs were identified.

The COMBINE and REORDER programs by Edward L. Braun

To test the programs use the example files in the “combine_test” and “trim_tests”. The files in these folders also serve as examples that can be examined if you have questions.

To test the “combine” program, make sure you are in the “combine-0.9” directory and then type:

```
$ cp combine_test/* .
$ ./combine combine_test.nex EarlyBird_Auth.txt ALDOB.nex BDNF.nex
CLTC.nex CRYAA.nex
$ diff combine_test.nex EB_Quartet.nex
```

diff should indicate that “combine_test.nex” and “EB_Quartet.nex” are identical, and both should execute in PAUP* without problems.

To test the subset feature of the “reorder” program, use the files in “trim_tests”:

```
$ cp trim_tests/* .
$ ./auth_subset EarlyBird_Auth.txt no_ALDOB.txt ALD_authst.txt ALD_all
$ diff ALD_authst.txt ALD_trim_auth.txt
```

diff should indicate that “ALD_authst.txt” and “ALD_trim_auth.txt” are identical, and it should be possible to use either with “reorder” to produce trimmed files as follows:

```
$ ./reorder Hackett_etal2008.nex ALD_authst.txt ALDOB_trim.nex
$ diff ALDOB_trim.nex ALD_trim.nex
```

Similar subsets of taxa and sites can be obtained using the other files in the “trim_tests” folder. If it is unnecessary to remove taxa from the authority file and all that is needed is to limit the sites to a specific CHARSET, all that is necessary is to add “# CHARSET” to the beginning of the authority file (note that there must be whitespace between the “#” and the desired CHARSET).

References

Hackett SJ, Kimball RT, Reddy S, Bowie RC, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han KL, Harshman J, Huddleston CJ, Marks BD, Miglia KJ, Moore WS, Sheldon FH, Steadman DW, Witt CC, Yuri T (2008) A phylogenomic study of birds reveals their evolutionary history. *Science* **320**:1763-1768.

Harshman J, Braun EL, Braun MJ, Huddleston CJ, Bowie RC, Chojnowski JL, Hackett SJ, Han KL, Kimball RT, Marks BD, Miglia KJ, Moore WS, Reddy S, Sheldon FH, Steadman DW, Steppan SJ, Witt CC, Yuri T. (2008) Phylogenomic evidence for multiple losses of flight in ratite birds. *Proc Natl Acad Sci USA* **105**:13462-13467.

Lewis PO (2003) NCL: a C++ class library for interpreting data files in NEXUS format. *Bioinformatics* **19**:2330-2331.

The COMBINE and REORDER programs
by Edward L. Braun

Copying these programs

The NCL (Lewis 2003) is distributed under the Gnu public license (included as the text file “COPYING”). If you wish to use the NCL to write your own programs, I recommend you download the most recent version of the NCL. The reorder program was written with an older version of the NCL, and I recommend you use the most current version of the NCL possible.

You are free to copy, modify, and redistribute these programs as long as their origin (both their authorship by Edward Braun and their use of the NCL) is acknowledged. These programs are provided “as is” without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the program is with you. Should the program prove defective, you assume the cost of all necessary servicing, repair or correction.