CONCATENATOR: SEQUENCE DATA MATRICES HANDLING MADE EASY

Authors: F. Pina-Martins[1] and O. S. Paulo[1]

[1]Centro de Biologia Ambiental, Departmento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal.

Keywords: Sequence data; concatenation; Nexus; FASTA; software; perl

Corresponding author:

Name: Francisco Rente de Pina Martins;

Address: Centro de Biologia Ambiental, Departmento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal;

E-mail: f.pinamartins@gmail.com

Running Title: Sequence data conversion & concatenation

Accepted for publication in *Molecular Ecology Resources*. The final citation is:

Pina-Martins, F., & Paulo, O. S. (2008). Concatenator: Sequence data matrices handling made easy. *Molecular Ecology Resources*, *8*(6), 1254–1255. doi: 10.1111/j.1755-0998.2008.02164.x

ABSTRACT

Concatenator is a simple and user friendly software that implements two very useful functions for phylogenetics data analysis. It concatenates Nexus files of several fragments in a single NEXUS file ready to be used in phylogenetics softwares, such as PAUP and MrBayes and it converts FASTA sequence data files to NEXUS and *vice-versa*. Additionally, concatenated files can be prepared for partition tests in PAUP. It is freely available in http://cobig2.fc.ul.pt/.

THE PROGRAM

Sequence data files can be organized in many different formats. Different sequence analysis software require differently formatted input files. The FASTA format has become very popular due to its simplicity and the capacity to quickly compare sequences (Pearson & Lipman 1988); these characteristics made this format one of the NCBI default outputs. The Nexus format became popular due to its modular format which is at the same time flexible and standardized (Maddison *et al.* 1997).

The existence of different file formats for the same data types require investigators to know how to handle them since they are not shared by some of the most common phylogenetic analysis software.

Concatenator's main purpose is to turn data matrix handling into a simple task, by allowing intuitive format conversions and concatenations of data matrices.

Concatenator is written in Perl using the Perl/TK module in order to give it a GUI for simplifying usage. The software was compiled for Windows using the PAR module. It is available in the Win32 binary version and source code version at the authors' group website.

The only requirements are either a system with a Perl interpreter and the Tk module installed (source code version) or a system running Microsoft Windows XP (not tested on other versions). The software was developed with a very specific aim – the simple handling of data matrices from one program to another and the concatenation of several of these data matrices. All the functions that the program performs can be accomplished manually provided the user has some knowledge about the involved file formats; however, even in such case this process is very error prone due to complex data organization such as in the interleave Nexus format.

The user interface is very simple (**Fig. 1**) and consists of a window that accommodates essentially the input and output entry boxes; these files can be selected from the File menu, a browse button located on the right of every entry box or by entering the path and filename directly on the entry box.

Concatenator can be used to accomplish 2 essential tasks chosen from the welcome window buttons or from it's "File" menu.

(1) Fasta-Nexus-Fasta Converter – It converts files from FASTA to Nexus format and from Nexus to FASTA format. When converting from Nexus to FASTA, there are no options available to chose from, however, when converting from FASTA to Nexus the user can choose whether to include a Taxa block, a leave or interleave organization, the type of data, the character for missing data and the gap character. File comments are ignored when converting.

(2) Matrix Concatenator – This function takes 2 to 5 Nexus formatted matrices and concatenates them into a single file. Two output formats are possible, one formatted to be used with *PAUP** (Swofford 2003), and the other prepared to input to *MrBayes* (Ronquist & Huelsenbeck 2003). Several parameters are customizable such as the inputs' data type, the gap character, the missing character, whether or not to include the Taxa block and a pre input for performing a "partition test" in *PAUP** excluding constant characters. Should the input files contain different

taxa, the program will prompt the user about what to do with the "unpaired" taxa – whether to ignore them in the final concatenated matrix, or to fill the missing parts with the character used in the file for "missing data".

Each function has a help file. The whole program is simple to use, but the help files are nevertheless as descriptive as possible.

An example of the program usage could be the following:

The user downloads two arrays of sequences (e.g. two different genes from the same species) from the NCBI database using a program such as *BioEdit* (Hall 1999). After a proper alignment session, the program outputs two FASTA files – one for each gene.

The user then wants to analyze these files using *PAUP**, *MrBayes*, *TCS* (Clement *et al*. 2000) or Network (Bandelt *et al.* 1999). *Concatenator* is useful in this step, because it provides a simple way to convert these FASTA files into Nexus files, ready to use in the analysis programs.

If after this first analysis the user decides to analyze both genes as a single block, *Concatenator* can join the two Nexus files in a single data matrix, ready to input on software such *MrBayes* or *PAUP**; the built in function for data partitioning will automatically add the required commands for partitioning data required for a "Partition Test" in *PAUP**.

ACKNOWLEDGMENTS

I would like to thank Joana Morais for Beta testing as well as all the other people who in one way or another helped make this possible. This work was supported by the grant from FCT, Portugal and FEDER (POCTI/BSE/48365/2002) to OSP.

REFERENCES

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. In: Molecular Biology and Evolution 16:37-48.

Clement M, Posada D, Crandall K (2000) TCS: a computer program to estimate gene genealogies. In: Molecular Ecology 9(10): 1657-1660.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: Nucleic Acids Symposium Series 41:95-98.

Maddison DR, Swofford DL, Maddison WP (1997) NEXUS: An Extensible File Format for Systematic Information. In: Systematic Biology 46 4: 590-621.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. In: Bioinformatics 19:1572-1574.

Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. In: Biochemistry 85: 2444-2448.

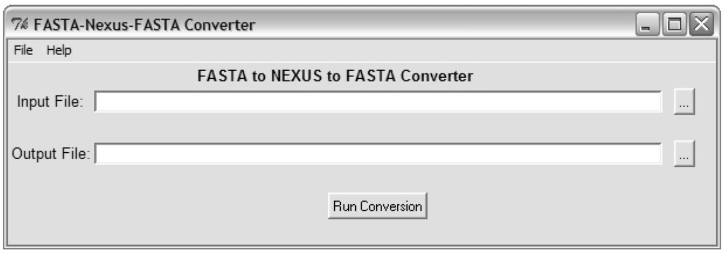
FASTA-Nexus-FASTA Converter
File
Help
FASTA to NEXUS to FASTA Converter
Input File:
...
Output File:
...
Run Conversion

Fig. 1. *Concatenator* interface on Win XP when converting matrices from Nexus to FASTA. Should the input be a FASTA file, the interface changes to accommodate the extra options to outputting a Nexus file.