Molecular Phylogenetics and Evolution 56 (2010) 1115–1118

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Short Communication

FASconCAT: Convenient handling of data matrices

Patrick Kück *, Karen Meusemann

Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany

ARTICLE INFO

Article history:
Received 2 February 2010
Revised 14 April 2010
Accepted 15 April 2010
Available online 21 April 2010

Keywords:
Concatenation
Software
Sequence data
Structure data
Perl

ABSTRACT

FASconCAT is a user-friendly software that concatenates rapidly different kinds of sequence data into one supermatrix file. Output files are either in FASTA, PHYLIP or NEXUS format and are directly loadable in phylogenetic programs like PAUP*, RAxML or MrBayes. FASconCAT can handle FASTA, PHYLIP and CLUSTAL formatted input files in one single run. It provides useful information about each input file and the concatenated supermatrix. For example, the program provides the range information of each concatenated gene (partition) and delivers a check list of all concatenated sequences (taxa). Information about the base composition of single input files and the resulting supermatrix is supplied for nucleotide data. For given structure strings (e.g. secondary structures) it displays single unpaired (loop) and paired (stem) positions after the concatenation process. Optionally, FASconCAT generates NEXUS files of concatenated sequences, either with MrBayes commands directly executable in PAUP* and MrBayes, or without any specific commands. If favoured, FASconCAT dispenses output files in PHYLIP format with relaxed (unlimited signs) or restricted taxon names (up to ten signs) while sequences are printed in non-interleaved format. FASconCAT is implemented in Perl and freely available from http://software.zfmk.de. It runs on UNIX and MS Windows operating systems.

1. Introduction

Today, data concatenation into supermatrices is a frequently used task in phylogenetic approaches. Data concatenation has been employed in rRNA analyses (Letsch et al., 2009; von Reumont et al., 2009), in analyses using 'mixed' nucleotide alignments combining rRNA sequences like 18S and 28S as well as protein coding genes (Brady et al., 2006; Moreau et al., 2006; Dinapoli and Klussmann-Kolb, 2010), in analyses based on nucleotide and amino acid alignments or in phylogenomic studies (Dunn et al., 2008; Philippe et al., 2009; Simon et al., 2009). The handling of different required file formats is often extensive and time consuming and different scripts or programs are often necessary. Most common formats are FASTA (Pearson and Lipman, 1988), NEXUS (Maddison et al., 1997), CLUSTAL (Higgins and Sharp, 1988) and PHYLIP (Felsenstein, 1989). To consider structure information of unpaired (loop) and paired (stem) regions using e.g. ribosomal RNA genes, most programs like RNAsalsa (Stocsits et al., 2009), MrBayes (Ronquist and Huelsenbeck, 2003), PHASE (Gowri-Shankar and Jow, 2006) and RAxML (Stamatakis, 2006) accept structure information in 'dot-bracket' format. Recent concatenation tools like CONCATENATOR (Pina-Martins and Paulo, 2008) can only concatenate and convert sequence data from FASTA to NEXUS and vice versa and are unable to handle additional structure information. Moreover, concatenation is mostly restricted to a limited number of gene alignments. With FASconCAT, we provide a new software tool for easy and fast data handling.

2. Methods

FASconCAT is implemented in Perl and runs on Windows PCs, Mac OS and Linux operating systems. It can be used via command line or by terminal menu options. The main menu of FASconCAT is subdivided into two parts, separated by a dashed line (Fig. 1). The upper component constitutes of a list of all possible options and their associated commands for adjustment. The lower part shows the actual parameter settings of FASconCAT. All default parameters can be optionally changed, and the new setting configuration will be displayed in the lower part of the menu.

3. Results

The software is designed to concatenate different data formats of nucleotide and amino acid alignments (sequence or artificially, e.g. RY coded) as well as 'dot-bracket' structure information of identical taxa into one supermatrix file. It can also be used as a simple data converter if just one file is provided. FASconCAT can

* Corresponding author. Address: Zoologisches Forschungsmuseum Alexander Koenig, Molecular Bioinformatic Unit, Adenauerallee 160, 53113 Bonn, Germany. Fax: +49 228 9122 295.
E-mail addresses: patrick_kueck@web.de (P. Kück), mail@karen-meusemann.de (K. Meusemann).

1055-7903/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2010.04.024

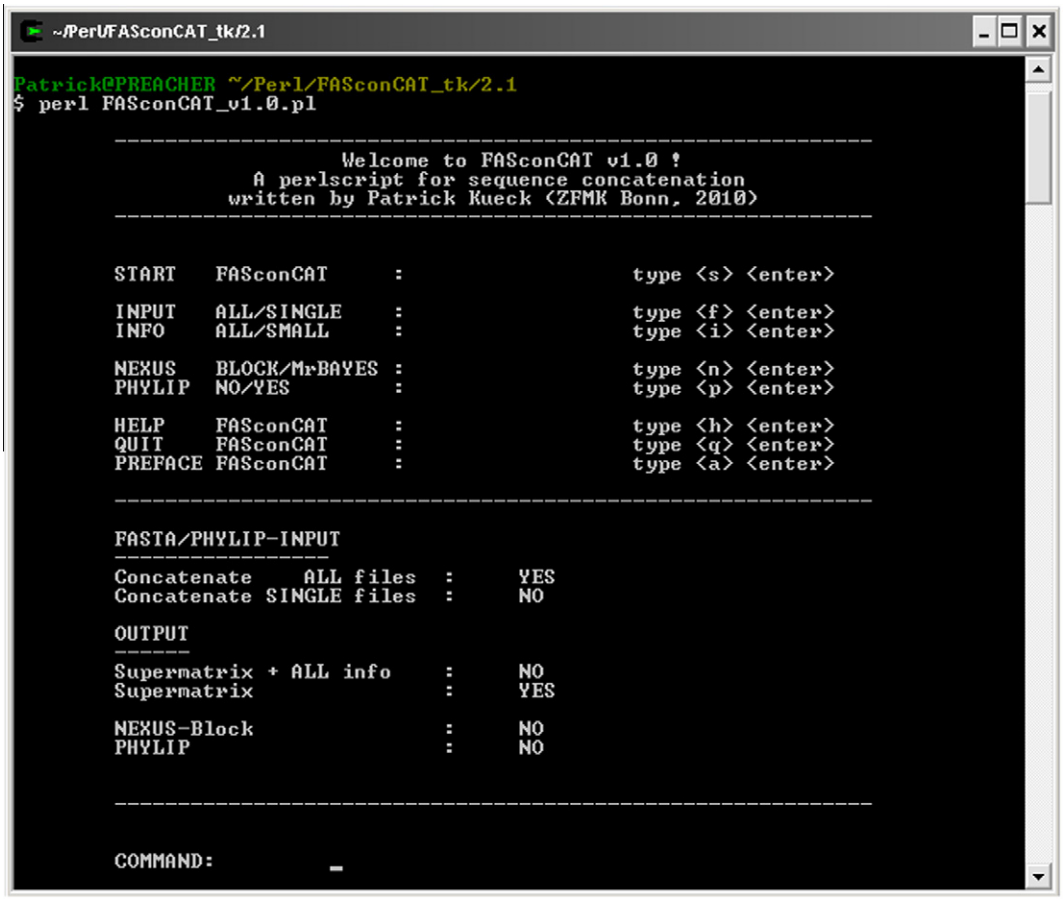

Fig. 1. Main menu of FASconCAT. The menu is subdivided into a command block (upper half) and a setting block (lower half). Users can specify their setting by using single commands via menu options or by typing multiple commands directly via the start command line of FASconCAT.

handle FASTA, CLUSTAL and PHYLIP input files. No unique input format is required. Sequences must have equal length within each file. FASTA is the standard output, additionally NEXUS or PHYLIP output can be chosen. The output files can be directly implemented into software like PAUP* (Swofford, 2003), MrBayes or RAxML. FASconCAT optionally creates NEXUS files with command blocks applicable in MrBayes. Among other things this option is very convenient for partitioned or mixed DNA/RNA analyses. Furthermore, it provides information about the supermatrix partitions (single ranges) which can be used in partitioned analyses.

3.1. Concatenation of data

Sequence data, with or without structure information, are concatenated either by taking all appropriate files in the folder where FASconCAT is located or by user specification. With FASconCAT, it is also possible to concatenate amino acid and nucleotide alignments into one supermatrix. Missing taxon sequences in single files are considered and replaced either by 'N' (nucleotide sequences), 'X' (amino acid sequences) or by '.' (dots, structure strings in 'dot-bracket' format), dependent on their associated data level. FASconCAT can read sequences in interleaved and non-interleaved format. The number of files for concatenation is not limited. The computation time rather depends on the computer hardware and the random access memory (RAM). For example, the concatenation of ten files comprising 108 taxa with a length of 1000 bp each requires between 0.5 (default option) and 3.4 s ('NEXUS' option) on a normal desktop computer (see manual for more information). Creating NEXUS files is the most time-consuming option. Every user can individually choose favoured options to optimise time performance. If no options are specified, FASconCAT runs under default which is the most time saving setting.

FASconCAT delivers useful accompanying information about the supermatrix and all single input files. As default, information is given for the partitions of the concatenated data set (fragment range) and the number of concatenated sequences per taxon. Additional information is provided by specifying several options, for example the number of sequence characters, sequence-type, number of gaps, a list of unpaired (loop) and paired (stem) positions (see manual for detailed instructions). A schematic overview is given in Fig. 2.

3.1.1. Default options

With standard options, FASconCAT takes all available input files (CLUSTAL, FASTA, PHYLIP) within the script placed folder and concatenates them into a supermatrix in FASTA format. Provided structure sequences in 'dot-bracket' format (one per file) are concatenated as well. Default information are accessorily provided (see above).

3.1.2. Additional options: -f, -i, -n and -p

With option -f, individual input files can be defined by the user. Additional information on the supermatrix and the input files, e.g. base composition of nucleotide sequences or the amount of gaps, can be activated by option -i. With -n, NEXUS files are generated that can be directly used in PAUP* or MrBayes. With typing -n -n, a complete setup for MrBayes is created. It can be easily modified as

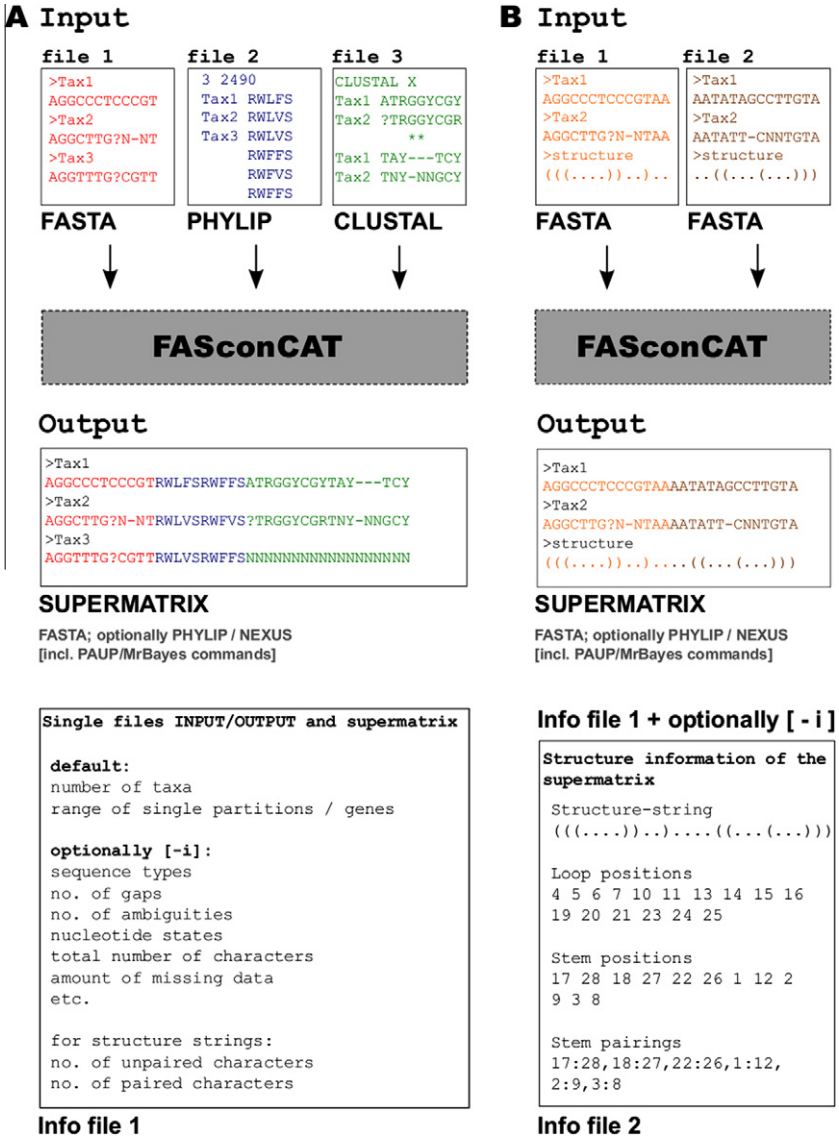

Fig. 2. Schematic overview of FASconCAT. (A) Three input files with different format (FASTA; PHYLIP, NEXUS), a nucleotide sequence alignment, an amino acid alignment and a nucleotide alignment with the third position RY recoded, are concatenated into a supermatrix (FASTA format, default). Additionally, an information file (Info file 1) is provided containing a list of concatenated sequences (taxa) and range information of single genes in the supermatrix (default). Optionally, additional information can be obtained by specific commands. (B) Two input files, nucleotide alignments with a structure string are concatenated into a supermatrix. Specifying the -i option, additional information about the percentage of unpaired (loop) and paired (stem) positions, is provided (Info file 1). A second information file is obtained, containing the concatenated structure string, loop and stem positions and related stem pairings (Info file 2).

favoured by the user. With option -p, FASconCAT additionally provides an output in PHYLIP format, either with non-interleaved sequences and restricted taxon names up to 10 signs (-p) or relaxed, with non-interleaved sequences and no restriction for taxon names (-p -p).

An example for FASconCAT usage could be: The user has three sequence alignment files in the same folder where FASconCAT is located, one in FASTA, the second in PHYLIP and the third in CLUSTAL format. The user wants to concatenate all alignments into a supermatrix in FASTA format and obtain all possible information via command line in a terminal on a LINUX system. FASconCAT has to be started as follows:

perl FASconCAT.pl -i -s <enter>.

3.4. Data conversion

Sequence formats can be simply converted by running FASconCAT just with one input file.

4. Conclusions

With FASconCAT, we deliver a new, convenient tool for concatenation of sequence files.

FASconCAT is easy to use and not limited in number of input files or input sequences.

Running on UNIX and Windows operating systems, the software reads several input formats, considers structure information, provides several output formats and optionally complete setup blocks at once, e.g. for analyses in MrBayes. It facilitates data handling, it is time saving in generating data matrices and in converting file formats and delivers many useful additional information about the input sequences. Detailed information and instructions are provided in the manual of FASconCAT. The manual also includes some tests about computation time of FASconCAT on a normal desktop computer. Help is provided for every option. FASconCAT is simple to use and freely available from http://fasconcat.zfmk.de or upon request from the corresponding author.

Acknowledgments

We thank Birthe Thormann and Dorothee Schillo for testing FASconCAT, Bernhard Misof and Johann Wolfgang Wägele for helpful comments and Francisco Hita Garcia for proofreading. This work was supported by DFG grants WA530/33 and MI 649/6-3. P.K. developed and implemented FASconCAT. Both authors tested FASconCAT and wrote the manuscript.

References

Brady, S.G., Schultz, T.R., Fisher, B.L., Ward, P.S., 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. Proc. Natl. Acad. Sci. USA 103, 18172–18177.

Dinapoli, A., Klussmann-Kolb, A., 2010. The long way to diversity–Phylogeny and evolution of the Heterobranchia (Mollusca:Gastropoda). Mol. Phylogenet. Evol. 55, 60–76.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sorensen, M.V., Haddock, S.H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q., Giribet, G., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.

Felsenstein, J., 1989. PHYLIP – phylogeny inference package (version 3.2). Cladistics 5, 164–166.

Gowri-Shankar, V., Jow, H., 2006. PHASE: A Software Package for Phylogenetics and Sequence Evolution. 2.0. University of Manchester.

Higgins, D.G., Sharp, P.M., 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237–244.

Letsch, H.O., Greve, C., Kück, P., Fleck, G., Stocsits, R.R., Misof, B., 2009. Simultaneous alignment and folding of 28S rRNA sequences uncovers phylogenetic signal in structure variation. Mol. Phylogenet. Evol. 53, 758–771.

Maddison, D.R., Swofford, D.L., Maddison, W.P., 1997. NEXUS: an extensible file format for systematic information. Syst. Biol. 46, 590–621.

Moreau, C.S., Bell, C.D., Vila, R., Archibald, B., Pierce, N.E., 2006. Phylogeny of the ants: diversification in the age of angiosperms. Science 312, 101–104.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Biochemistry 85, 2444–2448.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacalet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D.J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G., Manuel, M., 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19, 706–712.

Pina-Martins, F., Paulo, O.S., 2008. CONCATENATOR: sequence data matrices handling made easy. Mol. Ecol. Resour. 8, 1254–1255.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Simon, S., Strauss, S., von Haeseler, A., Hadrys, H., 2009. A phylogenomic approach to resolve the basal pterygote divergence. Mol. Biol. Evol. 26, 2719–2730.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Stocsits, R.R., Letsch, H., Hertel, J., Misof, B., Stadler, P.F., 2009. Accurate and efficient reconstruction of deep phylogenies from structured RNAs. Nucleic Acids Res. 37, 6184–6193.

Swofford, D.L., 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0. Sinauer Associates, Sunderland, MA.

von Reumont, B.M., Meusemann, K.A., Szucsich, N.U., Dell'Ampio, E., Gowri-Shankar, V., Bartel, D., Simon, S., Letsch, H.O., Stocsits, R.R., Luan, Y., Wägele, J.W., Pass, G., Hadrys, H., Misof, B., 2009. Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. BMC Evol. Biol. 9, 119.