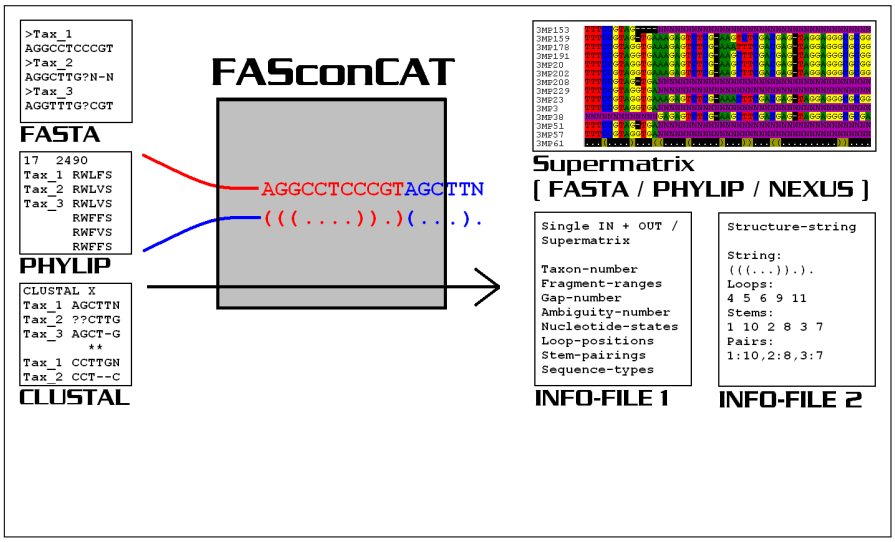
Patrick Kück

FASconCAT v1.0 - Manual



Contents

List of Figures

List of Tables

1 Introduction

FASconCAT is designed to concatenate different nucleotide, amino acid and structure sequence fragments of same taxa to one supermatrix file in FASTA, PHYLIP or NEXUS format which can be used for phylogenetic purposes. Because of the fact that sequence concatenation is normally used to multiple sequence alignments, sequences must have equal length within each file. FASconCAT can handle input files in PHYLIP, CLUSTAL and FASTA format. There has to be no unique input format setting. FASconCAT can handle all different formats in one single run. FASconCAT extracts taxon specific associated gene- or structure sequences out of given input files and links them to one string. Missing taxon sequences in single files are replaced either by 'N', 'X' or by '.' (dots), dependent on their taxon associated data level (nucleotide, amino acid or "dot-bracket" structures). It is possible to concatenate nucleotide and amino acid files to one supermatrix file. FASconCAT can read sequences in interleaved and non-interleaved format. For given FASTA files FASconCAT tolerates line breaks in sequences but not in taxon names. Sequence names may only include alphanumeric signs, underscores (_) and blanks, everything else is not allowed. FASconCAT will issue an error prompt and die if any non-alphanumeric sign is encountered in taxon names.

FASconCAT was written on Linux and works on WindowsPCs, Macs and Linux running systems. If input files are coming from Windows CRLF line feeds should be converted into Unix (LF) line feeds. This can be done in several editors like e.g. Bioedit or Notepad++. FASconCAT usually replaces them, but might not succeed in every instance.

Ambiguities and indels are allowed. Any other sign in sequences except for those covered by the universal DNA/RNA or amino acid code will also lead to an unacceptable error prompt. Structure information of ribosomal sequences are also recognized, analyzed and concatenated by FASconCAT if they are present once in each file and associated with equal names. Otherwise, FASconCAT will interrupt with a specific error prompt. Beside the concatenation process, FASconCAT delivers additional sequence information about each input file and the new concatenated supermatrix. The extent of information depends on the chosen setting. Additionally to the supermatrix file (FASTA format) FASconCAT delivers a second file in .xls format including single range information of each sequence fragment and a check list of all concatenated sequences. Under further options, FASconCAT can generate NEXUS files of concatenated sequences, either with MrBayes commands which can be directly executed in PAUP or without any specific commands. It is also possible to generate output files in PHYLIP format with relaxed (unlimited signs) or strict (limited up to ten signs) taxon names while sequences are always printed non-interleaved. The information file also includes reports about e.g. base composition of single and supermatrix files for nucleotide data and lists single loop– as well as stem pairing positions if structure sequences in dot-bracket format are found. Concatenated structure sequence as well as loop and stem positions are printed in a separate .txt file if the extended information option is chosen. For a more detailed report about additional information see section 'Usage/Options'. FASconCAT can be started directly via command line or by menu options.

2 Usage/Options

For using FAScon CAT open the terminal of your running system. Move through your directory path to the folder where FAScon CAT and input files are placed. Type the name of your FAScon CAT version, followed either by a blank and your demand options in one row to start FAScon CAT directly or followed by pressing enter to get into the FAScon CAT menu. Notice that all input files have to be located in the FAScon CAT including folder. To execute FAScon CAT a Perl interpreter must be installed on the current run system. Linux and Mac systems do normally not need a subsequent installation because the interpreter is a standard tool included in that systems in advance. Unfortunately, Windows users have to install a Perl interpreter ex post. We would recommend the ActivePerl interpreter which can be downloaded for free under:

- http://activeperl.softonic.de/

2.1 Start FAScon CAT via menu

2.1.1 Open the menu under Windows

Open a prompt (DOS) terminal on your Windows system and navigate to the folder where FAScon CAT and files are located <cd your_path...>. Then open FAScon CAT:

- C:\FAScon CAT_Folder> FAScon CAT_v1.0.pl <enter>

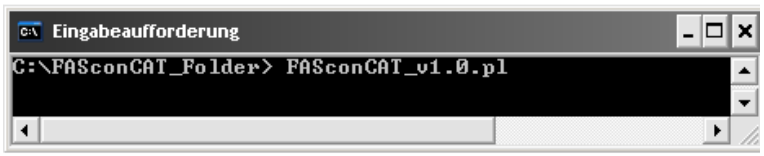


Figure 1: Open FAScon CAT menu under Windows

2.1.2 Open the menu under Linux/Mac

Open a terminal and navigate to the folder where FAScon CAT and files are located <cd your_path...>. Then open FAScon CAT:

- user@user:\~/FAScon CAT_Folder> perl FAScon CAT_v1.0.pl <enter>

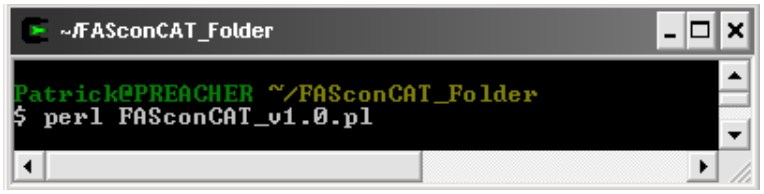


Figure 2: Open FAScon CAT menu under Linux

2.1.3 Menu handling

The main menu of FASconCAT is subdivided into two parts separated by a dashed line. The upper component constitutes all possible options and their associated commands for adjustment. The lower part shows the current parameter setting of FASconCAT.

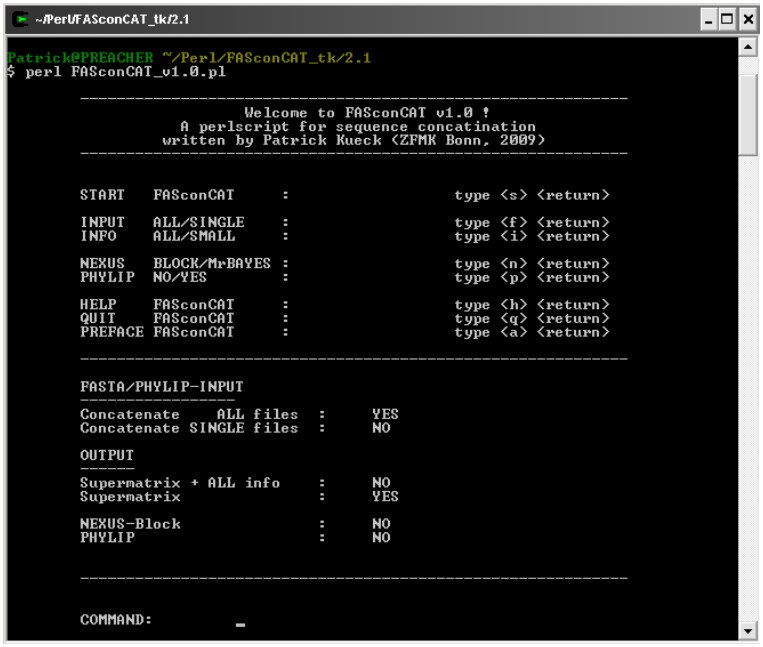


Figure 3: Menu of FASconCAT

To change the default parameter setting type the option associated command into the command line and press <enter>. The new setting configuration will be displayed in the lower part of the menu. After finishing parameter configuration FASconCAT can be started by typing “s” and pressing <enter>. For getting help type “h” and press <enter>, to return to FASconCAT type “b” and press <enter>, to quit the program type “q” and press <enter>.

2.2 Start FASconCAT via single command line

FASconCAT can directly started by command line commands in one row which simplifies the implementation of FASconCAT into complex process pipelines. Move through your directory path to the folder where FASconCAT and your files is located and type the name of the FASconCAT version, followed by a blank and the demand options with a minus (-) sign in front of each. Then press <enter>. Make sure you write the input options correctly, for example “-i” and not “- i”. Otherwise FASconCAT will not start working but instead open the menu.

- C:\FASconCAT_Folder> perl FASconCAT_v1.0.pl -h <enter> ↪ help menu
- C:\FASconCAT_Folder> perl FASconCAT_v1.0.pl -s <enter> ↪ start FASconCAT under default

Table 1: Overview of option codes via single command line start

Info options	**Command**		
Help menu		-help	
Preface		-a	
Start		-s	
Parameter options			**Default**
Defined input files		-f	none
Dispense all infos		-i	none
PHYLIP output (strict)		-p	none
PHYLIP output (relaxed)	-p	-p	none
NEXUS output (blank)		-n	none
NEXUS output (MrBayes)	-n	-n	none

2.3 Options

FASconCAT knows several options, it ignores commands if an unknown option is encountered. NOTE: Described commands are valid if the single command line is used. Working menu guided, type all options without "-", for example "i" instead of "-i".

2.3.1 -f option

FASconCAT asks before starting the concatenation process for defined input files. It will display a list of all files in FASTA (.fas), PHYLIP (.phy) and CLUSTAL (.aln) format which are found within its folder with an associated list number (Table 2). So the user can define specific files for concatenation. Type the file associated number of selected files separated by comma without blanks in one row and press <enter>. FASconCAT is able to concatenate a combination of different file formats. If only one input file is chosen, FASconCAT converts it to the selected output format. By typing b and <enter>, FASconCAT will skip back to the main menu.

Table 2: Example list of selectable files for specific file concatenation.

Listnumber	**Filename**
1	example_file_1.fas
2	example_file_2.phy
3	example_file_3.aln
4	example_file_4.aln
5	example_file_5.aln

- COMMAND: 2,3,4 <enter> ↪ only the PHYLIP and CLUSTAL files will access the concatenation process

2.3.2 -i option

FASconCAT provides useful additional information about the supermatrix file and all single input files, e.g. base composition of nucleotide sequence files or amount of gaps of each file. The evaluation of these additional information needs more computation time depending on the data set. Therefore this option is not included within the default setting. All additional information is listed in Table 3.

Table 3: List of default & additional information content within the .xls outputfile

Default information	**Supermatrix file**	**input files**
Single fragment ranges	yes	no
Number of concatenated sequences per taxon	yes	no
Additional information		
Number of taxa	yes	yes
Number of sequence characters	yes	yes
Data type (nucleotide/amino acid)	yes	yes
Number of single nucleotide characters	yes	yes
Number of gaps	yes	yes
Number of ambiguity characters	yes	yes
Number of inserted replacement characters	yes	yes
Number of missing taxa per fragment	no	yes
Number of inserted replacement strings	yes	yes
Number of characters in total	yes	no
Number of amino acid characters	no	yes
Percent & total number of nucleotides	yes	no
Percent & total number of gaps	yes	no
Percent & total number of ambiguities	yes	no
Percent & total number of inserted replacements	yes	no
Percent & total number of loop characters	yes	yes
Percent & total number of stem characters	yes	yes
Percent & total number of missing data (?)	yes	yes
List of loop positions	yes	no
List of stem pairing positions	yes	no

2.3.3 -n option

With the single -n option, FASconCAT generates an additional NEXUS file (.nex) which can be directly loaded into PAUP, MrBayes or other NEXUS file using programs. With the -n -n option, FASconCAT generates not only a NEXUS file with implemented taxa sequence blocks, but rather an executable file for bayesian analysis with the software

MrBayes. For that reason I integrated a presetting of parameters which seems to be a good start point for bayesian analyses. This can be easily changed manually by using any text editor. If a structure string in dot-bracket format is given while dots code unpaired (loop) positions and brackets code pairings (stems) ("(" ↪ opening a stem, ")" ↪ closing a stem), FASconCAT compiles automatically a partition set for MrBayes with single charset for stem and loop regions. Table 4 gives an overview of the integrated set up for MrBayes. To choose the MrBayes option via the FASconCAT menu the "-n" command has to be selected twice. If FASconCAT is started directly via command line "-n" must be double typed ("-n -n").

Table 4: Overview of all MrBayes set up parameters integrated into the NEXUS output under the double "-n" option. Structure partition parameters are only printed out by given structure information.

MrBayes commands	**Set up**
Number of generations	2000
Print frequency	100
Sample frequency	100
Number of chains	4
Save branch lengths	yes
Set autoclose	yes
No warnings	yes
Unlink statefrequency	all
tratio	all
Shape	all
Number of substitution	6
Rates	gamma
Sump burnin	20
Number of sump runs	2
Sumt burnin	20
Number of sumt runs	2
Inputfilename	FcC_smatrix.nex
Structure partition	
Set partition	looms
partition looms	2: loops, stems
lset 1	nucmodel= 4by4
lset 2	nucmodel= dublet

2.3.4 -p option

With -p option FASconCAT additionally generates an output in PHYLIP (.phy) format. The PHYLIP format can be printed either in a strict PHYLIP format with non-interleaved

sequences and taxon name restriction up to 10 signs or relaxed with no restriction in sign number for taxon names. To choose the strict PHYLIP option via the FASconCAT menu the "-p" command has to be selected once, for the relaxed PHYLIP format twice. If FASconCAT is started directly via command line the "-p" option must be either single typed ("-p") or double typed ("-p -p").

3 Internals

3.1 Input/Output

FASconCAT is able to import three different file formats. The number and formats of the output files depend on the user given parameter settings. Table 5 gives a summary of possible input and output formats.

Table 5: Overview of possible input and output formats under given parameter options.

Input format	**Ending**				
FASTA	.fas/.fasta				
PHYLIP	.phy				
CLUSTAL	.aln				
Output files	**Content**	**Output options**			
FcC_smatrix.fas	Supermatrix in FASTA format	all			
FcC_smatrix.phy	Supermatrix in PHYLIP format	-p	&	-p	-p
FcC_smatrix.nex	Supermatrix in NEXUS format	-n	&	-n	-n
FcC_info.xls	Concatenation info	-i	&	-n	
FcC_structure.txt	Structure info	-i	&	-n	

3.2 Computation time

The computation time of FASconCAT depends on the amount of data and of the chosen output options. Even for phylogenomic data sets like EST data, computation time will be in acceptable manner on a normal desktop computer. The generation of an additional PHYLIP output does not make any difference in computation time compared to an equal set up without it. The most time consuming step is the compilation of NEXUS output files. Choosing all possible information in .xls format is only little more time expensive than the default setting. The following examples may give you an impression about the computation time with different kind of data amount and usage options. I simulated two series of tests with different numbers of sequences using the INDELIBLE program. The first test includes 26 nucleotide sequences, the second 108. Each test was executed with seven distinct concatenation runs of ten data sets of same length whereas single runs differed in number of base positions from 100 to 100,000 and in favored output options.

I measured computation time for each single concatenation process and output option. Results are displayed in Table 6 and 7.

Table 6: Computation time of FASconCAT regarded to different sequence lengths and output processes for 26 sequences in each data set (test series 1).

	Distinct concatenation processes						
N data sets	10	10	10	10	10	10	10
Single lengths [bp]	100	500	1000	10000	25000	50000	100000
Supermatrix [bp]	1000	5000	10000	100000	250000	500000	1000000
Output options	**Computation time [sec]**						
Default	0.2	0.1	0.1	0.5	1.2	2.4	4.8
PHYLIP	0.1	0.1	0.2	0.6	1.2	2.4	4.9
Default + all info	0.2	0.3	0.5	4	9.7	19.7	40.1
PHYLIP + all info	0.1	0.3	0.5	3.9	9.7	19.8	40.1
NEXUS	0.2	0.4	0.9	16.1	75.8	281.9	1321.6

Table 7: Computation time of FASconCAT regarded to different sequence lengths and output processes for 108 sequences in each data set (test series 2).

	Distinct concatenation processes						
N data sets	10	10	10	10	10	10	10
Single lengths [bp]	100	500	1000	10000	25000	50000	100000
Supermatrix [bp]	1000	5000	10000	100000	250000	500000	1000000
Output options	**Computation time [sec]**						
Default	0.3	0.4	0.5	2.3	5.5	11.3	21.4
PHYLIP	0.3	0.4	0.5	2.3	5.5	11	21.9
Default + all info	0.5	1.1	1.9	16.6	42.8	89.1	180.5
PHYLIP + all info	0.4	1.1	1.9	16.8	43.2	88.7	156.8
NEXUS	0.5	1.7	3.4	69.3	320.5	1172.5	5583.4

3.3 Error reports

FASconCAT checks each input file according to correct format and forbidden sequence and structure characters. This subsection gives a short explanation for possible reasons to all implemented error reports. Notice that each error allocates FASconCAT to stop all running processes and to abort.

3.3.1 *Taxon* in *filename.fas* not in FASTA format!

The *file.fas* file is not in a FASTA file typical manner. FASconCAT is able to read sequences of FASTA files either if they are in one line or with line interruptions (blocks).

Sequence names have to be in one line and have to start with an ">"! Each line has to end with a line break. Table 8 gives an example of both acceptable FASTA formats.

Table 8: Known FASTA formats in non-interleaved (format 1) and interleaved format (format 2).

FASTA format 1
>Name_sequence_1
AGCTCCCGTCCTTTG–AGA–GTGTCCTTTCCT
>Name_sequence_2
AGCTCCGGCCCTTTG–AGA–GTGTCCTTTCCT
>Name_sequence_n
AGCTCCCGTCCTTTGGAGAGGTGTCCTTTCCT
FASTA format 2
>Name_sequence_1
AGCTGTCCTTTCTTG–AGA–GTGTCCTTTCCT
GGGGCCCTTTC–GGTTTTCCCCGTCCTTTCCT
>Name_sequence_n
AGCTGTCCTTTCTTGCAGACGTGTCCTTTCCT
GGGGCTTCAAGTTTTCCCCCGGGTCCTTTCCT

3.3.2 *filename.aln* is not a CLUSTAL format!

The *filename.aln* file is not in a CLUSTAL file typical manner. Each line has to end with a line break. Table 9 shows a typical CLUSTAL format.

Table 9: Example of a CLUSTAL formatted input file.

CLUSTAL format	
CLUSTAL X (1.81) multiple sequence alignment	
<line break>	
<line break>	
Name_sequence_1	AGGGCCCTTGCGCTTGCTTCC
Name_sequence_2	AGGGCCCTTGCGCCTGCTTCC
Name_sequence_n	AGGCCCCTTGCGCCGGCTTCC
<line break>	
<line break>	
Name_sequence_1	ATTTCCCTTGGGCTTGCTTCC
Name_sequence_2	ATTTCCCTTGGGCCTGCTTCC
Name_sequence_n	ATCTCCCTTGGGCCGGCTTCC

3.3.3 *filename.phy* is not a PHYLIP format!

The *filename.phy* file is not in a PHYLIP file typical manner. Each line has to end with a line break. Table 10 shows a typical PHYLIP file in non-interleaved format in which sequence names are allowed to contain up to ten signs at maximum.

Table 10: Example of a interleaved PHYLIP formatted input file.

PHYLIP format (interleaved)		
6 40		
Name_sequence_1	AGGGCCCTTG	CGCTTGGCCC
Name_sequence_2	AGGGCCCTTG	CGCCTCCCCC
Name_sequence_n	AGGCCCCTTG	CGCCGCCCGG
<line break>		
	ATTTCCCTTG	GGCTTCCCCC
	ATTTCCCTTG	GGGGGCCTCC
	ATCTCCCTTG	GGCCGGGGGC

3.3.4 Sequence name missing in *file.fas*!

Maybe you have forgotten the whole sequence name or to start the name line with the ">" sign or your FASTA format is completely wrong. See also Table 8 for known FASTA formats.

3.3.5 Sequence missing in *file.fas*!

Either you have forgotten the sequence or an additional line-break sign in your FASTA file or your FASTA format is completely wrong. See Table 8 for known FASTA formats.

3.3.6 Sequence name *sequence_name* in *filename* involves forbidden signs!

Sequence names may only include alphanumeric signs, underscores (_) and blanks, everything else is not allowed. If sequence names are correct check the FASTA format in common. See Table 8 for known FASTA formats.

3.3.7 Sequences of *filename* have no equal length!

FASconCAT allows sequences within the same input file only if they have equal length.

3.3.8 Multiple sequence names of *sequence_name* in *filename*!

Identic sequence names are not allowed in same input files, because FASconCAT concatenates sequences on the basis of them. Two equal names in one file cannot be assigned correctly.

3.3.9 Sequence of *filename* involves forbidden signs in *sequence_name*!

Ambiguities and indels are understood. Any other signs in sequences except for these covered by the universal DNA/RNA or amino acid code are not allowed. If sequence signs are correct check the input format in common. See Table 8 for known FASTA formats.

3.3.10 *filename* involves multiple structure sequences

Multiple structure sequences in one input file are not allowed. FASconCAT can concatenate only one structure sequence which is sufficient for most phylogenetic analyses.

3.3.11 Additional structure sequence of sequence *sequence_name* in *filename* not allowed!

FASconCAT can handle only one structure sequence which is sufficient for most phylogenetic analyses. For that reason, single structure sequences must have identic sequence names.

3.4 License/Help-Desk/Citation

FASconCAT v1.0 is written/developed in Perl by Patrick Kück in 2009. It is implemented in Perl and a free software. It can be distributed and/or modified under the terms of the GNU General Public License as published by the Free Software Foundation; either 2 of the license, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details. You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139, USA.

If you have any problems, error-reports or other questions about FASconCAT feel free and write an email to fasconcat@web.de which is the official help desk email account for the software. For further free downloadable programs from our institute visit:
http://software.zfmk.de.

If you use FASconCAT please cite:

Kück P., FASconCAT, Version 1.0, Zool. Forschungsmuseum A. Koenig, Germany, 2009

4 Copyright