iTaxoTools - tools for integrative taxonomy

Addendum to iTaxoTools manual: Concatenator, MAFFTpy, FastTreePy

Version: May 2, 2022

Previous excecutables of iTaxotools 0.1 are available at https://github.com/iTaxoTools/iTaxoTools-Executables/releases

Both the previous and new tools can be downloaded from the project's website http://itaxotools.org which also features a section with useful links, news, and FAQs.

How to cite: When using one of the programs included in the iTaxoTools 0.1.1. release in your study, please cite the main paper as follows. For Concatenator, MAFFT, FastTree, and SequenceBouncer, cite also the original papers:

Vences, M., Miralles, A., Brouillet, S., Ducasse, J., Fedosov, A., Kharchev, V., Kumari, S, Patmanidis, S., Puillandre, N., Scherz, M. D., Kostadinov, I., Renner, S. S. (2021). iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. *Megataxa* **6**: 77-92.

Vences, M., Patmanidis, S., Kharchev, V., Renner, S.S. (2022). Concatenator, a user-friendly program to concatenate DNA sequences, implementing graphical user interfaces for MAFFT and FastTree. XXXX.

Katoh, K., Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30: 772–780.

Price, M.N., Dehal, P.S., Arkin, A.P. (2010). FastTree 2 -- Approximately Maximum-Likelihood trees for large alignments. PLoS ONE 5: e9490.

Dunn, C.D. (2020). SequenceBouncer: A method to remove outlier entries from a multiple sequence alignment. BioRxiv Preprint, doi: https://doi.org/10.1101/2020.11.24.395459

Disclaimer: The programs included in iTaxoTools are free software. All code specifically programmed for iTaxoTools can be redistributed and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 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.

All tools not specifically programmed for iTaxoTools but constituting a modification or extension of the original code are also free software and most of them licensed under GNU v3, but partly, other licences apply. Please check the original GitHub pages (see table at the end of Introduction below) for licensing of the original code of PTP. GMYC, tr2, DELINEATE, ABGD, ASAP, LIMES 2.0 and MolD.

We welcome all suggestions, comments, questions and bug reports. Please use the GitHub platform to submit those: for this, you just need to sign up at GitHub (a quick and easy procedure), navigate to the respective repository (see table with links below), and create "New Issue". The team of developers regularly checks all the issues and will deal with them asap (bug reports will be treated as priority).

1. Concatenator

Example files provided:

Concatenator_examplefile.tsv
Concatenator_examplefile.nex
Concatenator_examplefile_fasta1.fas
Concatenator_examplefile_fasta2.fas

All example files are nucleotide sequences. They can be inspected in a text editor.

The first file is a tab-separated value (.tsv) file. Concatenator also recognizes alternatively the extension .tab for such files. This file is best opened and inspected in a spreadsheet editor such as Microsoft Excel. It contains, besides the sequence metadata, sequences of each marker in a separate column.

The second file is a multi-marker file in interleaved Nexus format. Each block of sequences corresponds to one marker. The nucleotide positions correspomnding to each marker are specified in the "sets" block.

The third and fourth example files are two files in fasta format, of which the first is aligned and the second is unaligned.

To test the functions of Concatenator, important either the tsv, or the Nexus, or one or both of the fasta files, and let the program guide you through the concatenation process.

Concatenator is a tool to manage files of multi-marker nucleotide sequences of different formats (although it can also be executed to convert single-marker files). It can deal with already aligned sequence files as well as with unaligned sequences (which it optionally can align using MAFFT). As output, the user can choose a variety of formats, including Nexus and tab-delimited table, but also single-marker fasta files. Therefore, the program can also be used for de-concatenation.

Note that the current version of Concatenator is designed to concatenate nucleotide sequences only. Nevertheless, the program will also be able to concatenate protein sequences, but several of its options will in that case not work or cause errors, in particular codon subsetting and MAFFT alignment (as we have only implemented nucleotide-alignment strategies from this latter program).

Note: For most of the options in the graphical user interface, Concatenator offers additional explanations and tips that show up when hovering with the cursor over the respective menus or checkboxes.

Input/output formats

tab-separated value file (.tsv or .tab). The following example shows the structure of this format. The seqid column is included to maintain consistency with tab files used in other iTaxoTools programs. None of the metadata fields are not compulsory. Each marker is in its own column, and

the name preceded by "sequence_" separated from the marker (gene) name by an underscore. The program will recognize this naming and use the string after "sequence_" as marker name.
This file format allows users

seqid	**species**	**specimen-voucher**	**locality**	**sequence_16S**	**sequence_cytb**	**sequence_coi**
sample1	Mantella aurantiaca	ZCMV1234	Andasibe	ACGTTTTTZC	GATTTAGA	TAAGGCTGC
sample2	Mantella aurantiaca	ZCMV9876	Andasibe	ACGTTTTAC	GATCTAGA	TAAGGGTGA
sample3	Mantella aurantiaca	FGZC345	Andasibe	ATGTTTTAC		TAAGGCTGA
sample4	Mantella crocea	FGZC346	Ranomafana	ACGTAATAC	GATTTATA	TATGGCTGA
sample5	Mantella crocea	MNHN1991	Ranomafana	ACGTAATAG	AATTTATA	TATGGCTGA

Nexus (.nex). The following example shows the same data as in the tsv file example above. In this variant of the interleaved Nexus file, each marker is in a separate block, but this is not mandatory as the information on the boundaries between markers is specified by the charset commend in the "sets" block.

```
#NEXUS

begin data;

format datatype=DNA missing=N missing=? Gap=- Interleave=yes;

dimensions Nchar=27 Ntax=5;

matrix

Mantella_aurantiaca_ZCMV1234_Andasibe    ACGTTTTTC
Mantella_aurantiaca_ZCMV9876_Andasibe    ACGTTTTAC
Mantella_aurantiaca_FGZC345_Andasibe     ATGTTTTAC
Mantella_crocea_FGZC346_Ranomafana       ACGTAATAC
Mantella_crocea_MNHN1991_Ranomafana      ACGTAATAG

Mantella_aurantiaca_ZCMV1234_Andasibe    GATTTAGA
Mantella_aurantiaca_ZCMV9876_Andasibe    GATCTAGA
Mantella_aurantiaca_FGZC345_Andasibe     NNNNNNNN
Mantella_crocea_FGZC346_Ranomafana       GATTTATA
Mantella_crocea_MNHN1991_Ranomafana      AATTTATA

Mantella_aurantiaca_ZCMV1234_Andasibe    TAAGGCTGC
Mantella_aurantiaca_ZCMV9876_Andasibe    TAAGGGTGA
Mantella_aurantiaca_FGZC345_Andasibe     TAAGGCTGA
Mantella_crocea_FGZC346_Ranomafana       TATGGCTGA
Mantella_crocea_MNHN1991_Ranomafana      TATGGCTGA
;
end;


begin sets;

charset 16S = 1-10;
charset cytb = 11-18;
charset coi = 19-27;

end;
```

Fasta (.fas). If using fasta as input format, note that each file can only contain one marker, and the name of the fasta file will be used as marker name. You can add multiple markers to the program, of which each will be interpreted as including a different marker. You can also upload a ZIP-compressed archive containing multiple fasta files; such a ZIP file is also exported when selecting "Multifile Fasta" in the final step. The program allows to specify a single-fasta export of the concatenated data set (option "Concatenated Fasta"), but in this case the information of start/end of each marker will get lost.

Phylip (.phy). Similar to Fasta files, each Phylip file used as input will be interpreted as including sequences from a single marker. Concatenator accepts as input (and outputs) strict and extended phylip format (i.e., with sequence names of exactly 10 characters, or with an extended number of characters). You can also upload a ZIP-compressed archive containing multiple Phylip files; such a ZIP file is also exported when selecting "Multifile Phylip" in the final step. The program allows to specify a single-phylip export of the concatenated data set (option "Concatenated Phylip"), but in t his case the information of start/end of each marker will get lost.

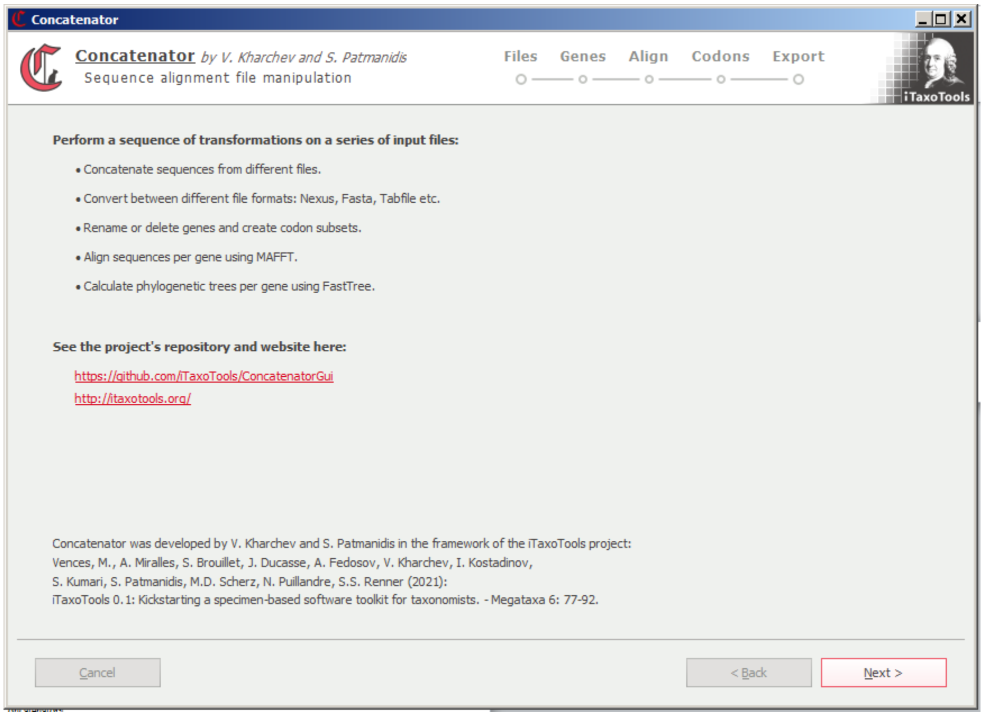
Partitionfinder. Although the program Partitionfinder (which determines the best partition and substitution model for the phylogenetic analysis of multi-marker nucleotide data sets) has been deprecated in favor of IQtree, it is still widely used in taxonomic studies that analyse a limited amount of markers, for instance with the program MrBayes. We have therefore included this legacy format in the output options. This option will produce a ZIP-archive which contains the concatenated alignment in Phylip format as well as a generic preference file for Partitionfinder (partition_finder.cfg) containing the information on the character sets for each marker in the alignment. Note that the preferences in this cfg file are set for running Partitionfinder to find the best partitions for MrBayes; for other purposes, you can adjust the settings as instructed in the Partitionfinder manual.

IQtree. Similar to the deprecated Partitionfinder, IQtree can be used to determine the best partition and substitution models for each partition subset, for downstream partitioned phylogenetic analysis. With the "IQtree" option set for export, Concatenator produces a ZIP archive containing a concatenated Phylip file, and a Nexus file containing only the "sets" block with character set definitions. These two files are used as input for IQtree.

Running Concatenator

The program is distributed as standalone executable for Windows and Linux. It should run in different Windows environments, including Windows 10 and Windows 7. Upon double-clicking the program icon, the starting window appears which gives some general information on the program.

The program is designed in form of a pipeline, and the progress bar in the upper right shows the completed steps and the current step. The Next and Back buttons on the lower light allow to move along the pipeline.


Concatenator
Concatenator by V. Kharchev and S. Patmanidis
Sequence alignment file manipulation
Files
Genes
Align
Codons
Export
iTaxoTools
Cancel
< Back
Next >

Perform a sequence of transformations on a series of input files:

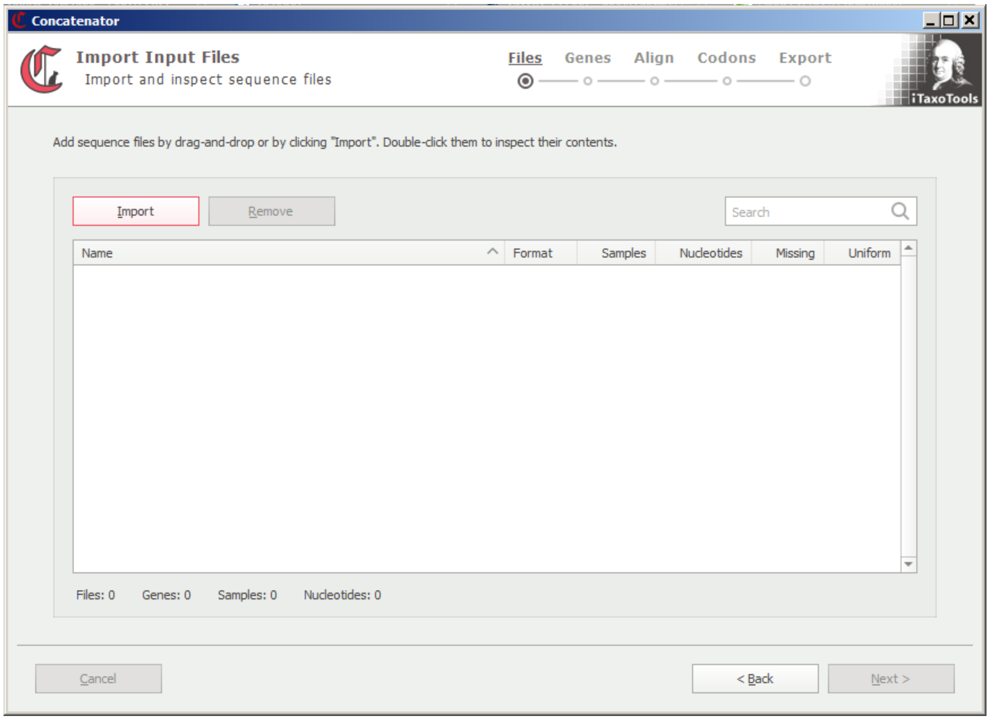
- Concatenate sequences from different files.
- Convert between different file formats: Nexus, Fasta, Tabfile etc.
- Rename or delete genes and create codon subsets.
- Align sequences per gene using MAFFT.
- Calculate phylogenetic trees per gene using FastTree.

See the project's repository and website here:

https://github.com/iTaxoTools/ConcatenatorGui
http://itaxotools.org/

Concatenator was developed by V. Kharchev and S. Patmanidis in the framework of the iTaxoTools project:
Vences, M., A. Miralles, S. Brouillet, J. Ducasse, A. Fedosov, V. Kharchev, I. Kostadinov,
S. Kumari, S. Patmanidis, M.D. Scherz, N. Puillandre, S.S. Renner (2021):
iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. - Megataxa 6: 77-92.

Upon pressing the Next button, the "Files" window appears. Here, data can be imported, either using the "Import" button or by drag-and-drop.

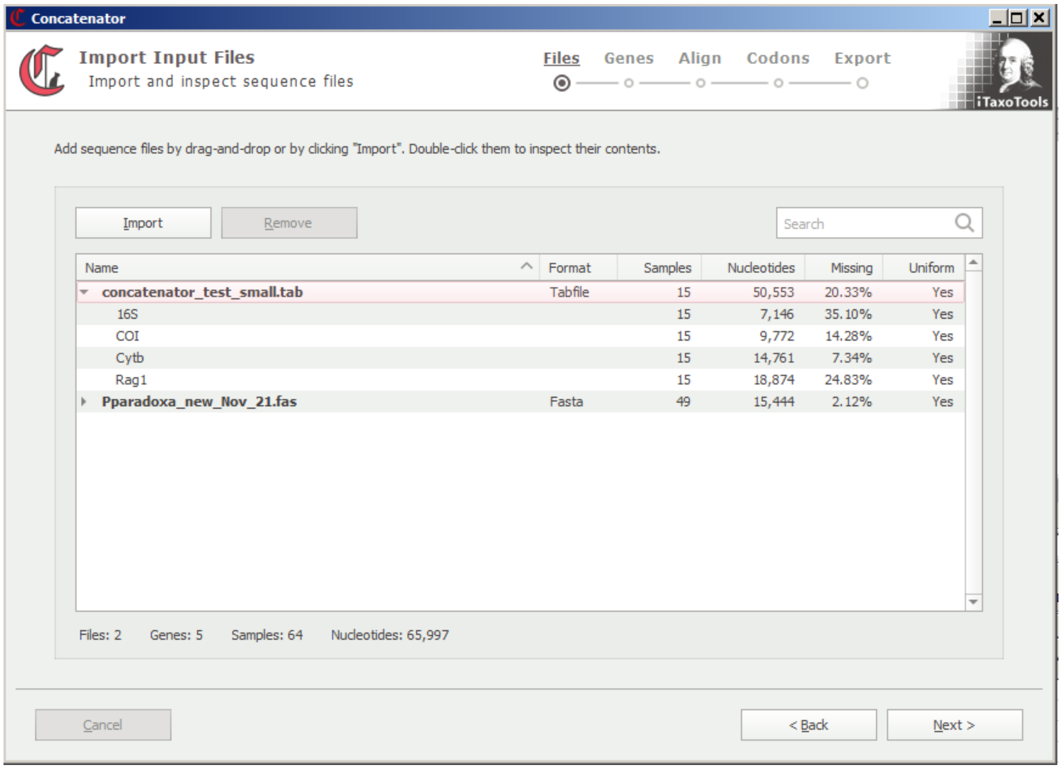


Add sequence files by drag-and-drop or by clicking "Import". Double-click them to inspect their contents.

Name	Format	Samples	Nucleotides	Missing	Uniform

Files: 0 Genes: 0 Samples: 0 Nucleotides: 0

After importing, the name of the input file(s) will be shown. Upon clicking on the small triangle next to the file name, the markers included in each input file will be shown, along with some general information such as the total number of samples (specimens), nucleotides, proportion of missing data, and uniform vs. variable length for each marker. At this stage, it is also possible to remove input files (but not single markers from one input file).



Add sequence files by drag-and-drop or by clicking "Import". Double-click them to inspect their contents.

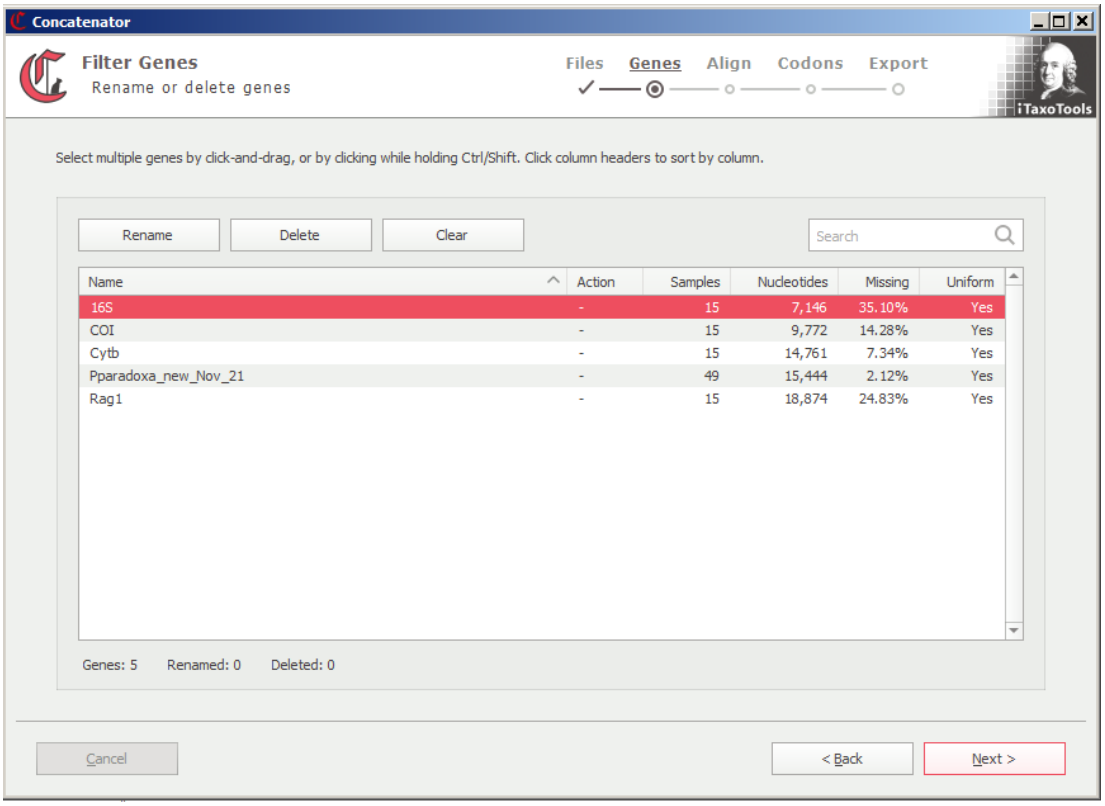
Name	Format	Samples	Nucleotides	Missing	Uniform
concatenator_test_small.tab	Tabfile	15	50,553	20.33%	Yes
16S		15	7,146	35.10%	Yes
COI		15	9,772	14.28%	Yes
Cytb		15	14,761	7.34%	Yes
Rag1		15	18,874	24.83%	Yes
Pparadoxa_new_Nov_21.fas	Fasta	49	15,444	2.12%	Yes

Files: 2 Genes: 5 Samples: 64 Nucleotides: 65,997

Note that in tsv files, the name of the markers is specified in the respective column header. In the case of single-gene fasta files, the filename will be interpreted as gene name. If (as in the example shown) the fasta file has a different name not reflecting the name of the included marker, it can be changed to the desired name in the next window.

The "Genes" window allows to rename or delete markers if needed. Markers (=Genes) can be re-ordered based on the number of samples, nucleotides, missing data, or aligned/unaligned ("Uniform" length) state. For this, click on the respective column header.

Note that markers cannot be manually re-ordered into a custom order. If this is needed, for a limited amount of genes it can be easily achieved by re-naming them by, for example, preceding gene name with a number according to the desired order, and then re-ordering according to "Name". For larger number of markers, it is preferable to adjust marker name already in the input files.



Select multiple genes by click-and-drag, or by clicking while holding Ctrl/Shift. Click column headers to sort by column.

Name	Action	Samples	Nucleotides	Missing	Uniform
16S	-	15	7,146	35.10%	Yes
COI	-	15	9,772	14.28%	Yes
Cytb	-	15	14,761	7.34%	Yes
Pparadoxa_new_Nov_21	-	49	15,444	2.12%	Yes
Rag1	-	15	18,874	24.83%	Yes

Genes: 5 Renamed: 0 Deleted: 0

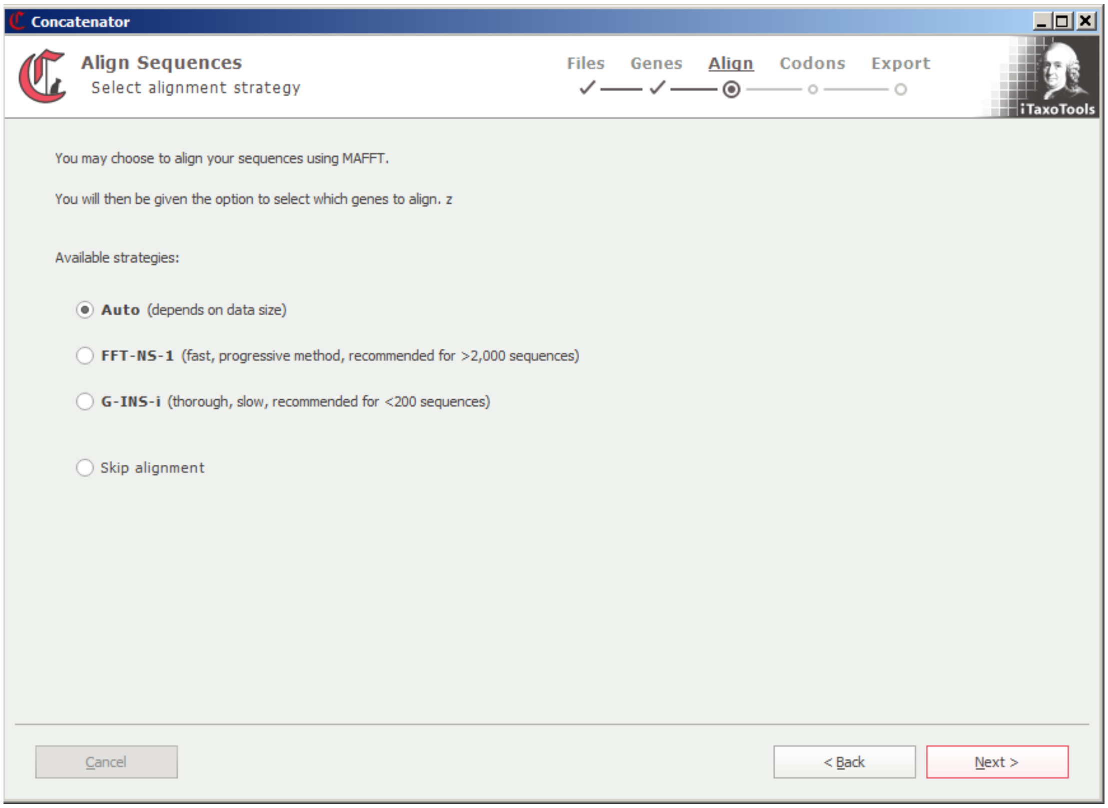
The next window allows choosing the alignment strategy to be applied to all or some of the markers (genes). Concatenator implements two of the nucleotide alignment options of MAFFT: FFT-NS1, a fast progressive method that is able to align large sets of sequences (>2000), and G-INS-i, a thorough method that performs well also with sequences such as rRNA genes with constant and hypervariable regions. Other alignment strategies of the original MAFFT are not enabled, in order to keep the program simple.

In the current version of Concatenator, the option (provided by the full Mafft program) to automatically adjust (reverse complement sequences) as necessary, has not been implemented. Sequences will therefore not be adjusted if some are mistakenly reverse-complemented.

It also is possible in Concatenator to choose "Auto". In this case, the program determines the most suitable of the two implemented alignment strategies. The alignment step can also be skipped altogether (e.g., if all sequences are already aligned).

For more detailed information on the MAFFT algorithms, consult the program's webpage: https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html

Note that in Concatenator it is only possible to choose one Alignment strategy (or "Auto") for all genes that should be aligned, i.e., the chosen strategy will be applied to all genes that in the next step will be selected.



You may choose to align your sequences using MAFFT.

You will then be given the option to select which genes to align. z

Available strategies:

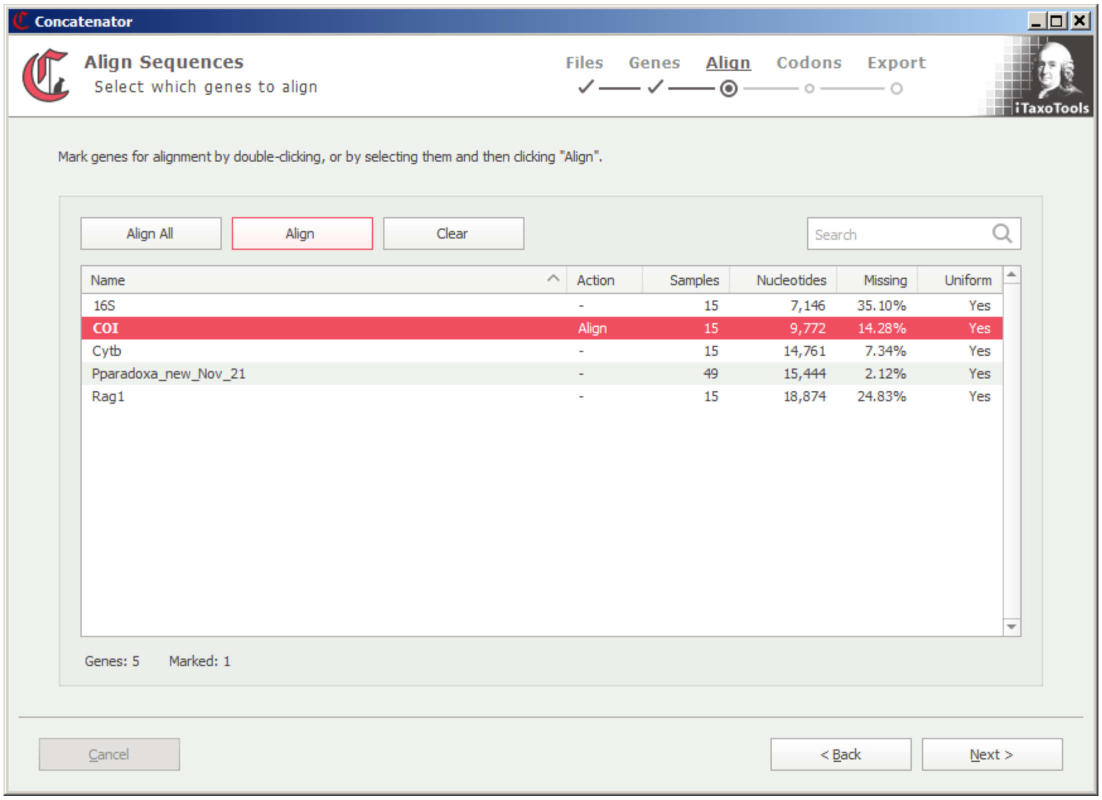
● **Auto** (depends on data size)

○ **FFT-NS-1** (fast, progressive method, recommended for >2,000 sequences)

○ **G-INS-i** (thorough, slow, recommended for <200 sequences)

○ Skip alignment

In the next window, users can choose sequences that should be aligned (or re-aligned). For this, select the respective gene(s) and click on "Align". Or click "Align All" if all genes should be aligned. Note that you will not be able to inspect and adjust the respective alignments. If you need to visually verify the alignments or want to manually adjust them, you need to do this in another program (e.g., in MEGA), and then use the aligned data sets for import in Concatenator.

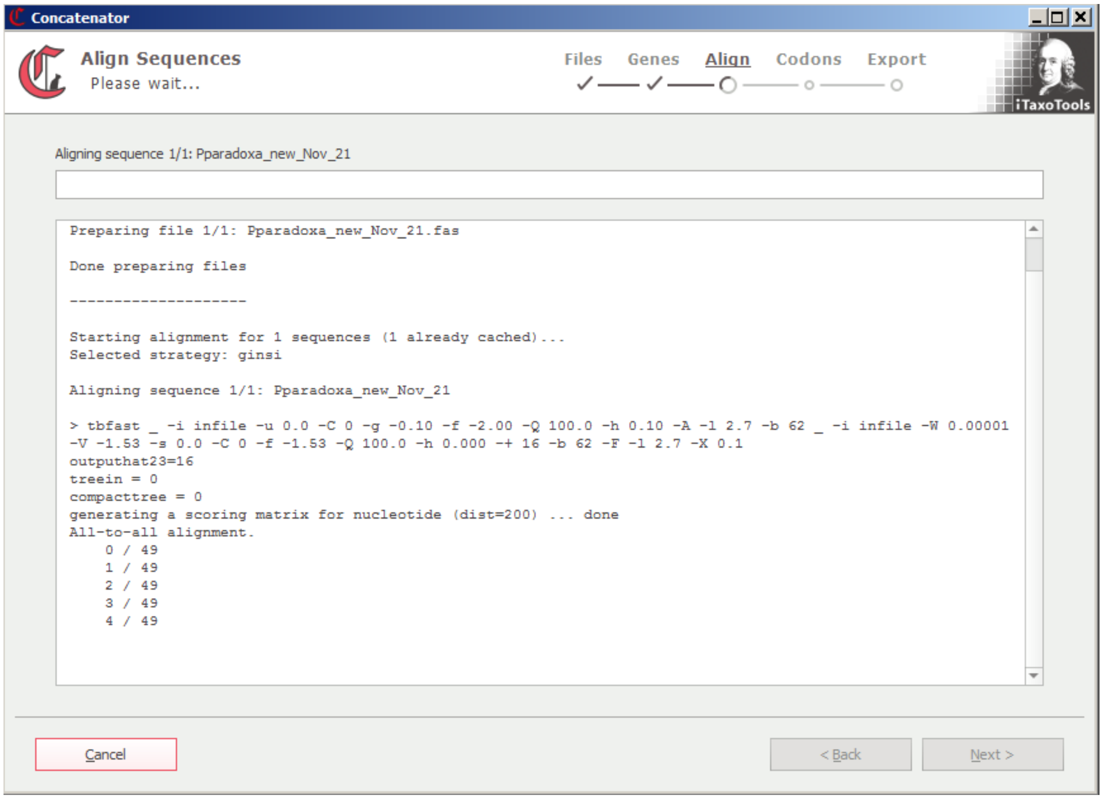


Mark genes for alignment by double-clicking, or by selecting them and then clicking "Align".

Name	Action	Samples	Nucleotides	Missing	Uniform
16S	-	15	7,146	35.10%	Yes
COI	Align	15	9,772	14.28%	Yes
Cytb	-	15	14,761	7.34%	Yes
Pparadoxa_new_Nov_21	-	49	15,444	2.12%	Yes
Rag1	-	15	18,874	24.83%	Yes

Genes: 5 Marked: 1

In the next step, the program will align (one by one) all of the selected genes, and inform about the progress.



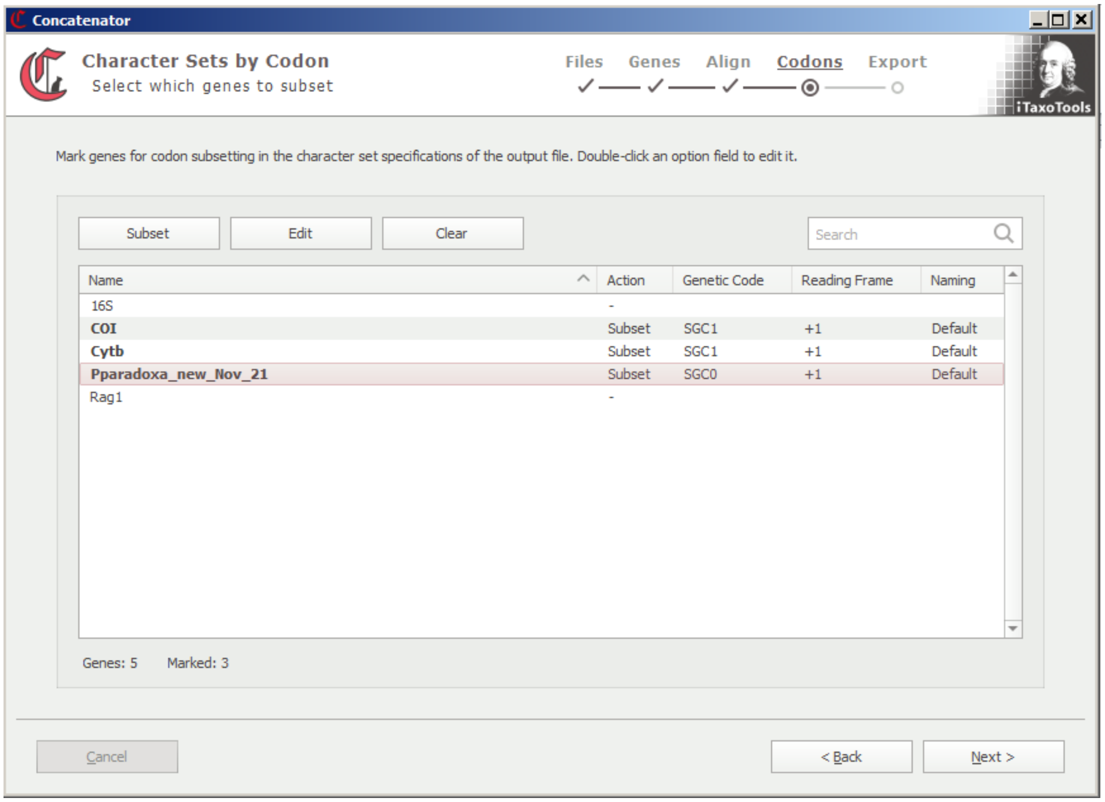
Aligning sequence 1/1: Pparadoxa_new_Nov_21

```
Preparing file 1/1: Pparadoxa_new_Nov_21.fas

Done preparing files

--------------------

Starting alignment for 1 sequences (1 already cached)...
Selected strategy: ginsi

Aligning sequence 1/1: Pparadoxa_new_Nov_21

> tbfast _ -i infile -u 0.0 -C 0 -g -0.10 -f -2.00 -Q 100.0 -h 0.10 -A -l 2.7 -b 62 _ -i infile -W 0.00001
-V -1.53 -s 0.0 -C 0 -f -1.53 -Q 100.0 -h 0.000 -+ 16 -b 62 -F -l 2.7 -X 0.1
outputhat23=16
treein = 0
compacttree = 0
generating a scoring matrix for nucleotide (dist=200) ... done
All-to-all alignment.
    0 / 49
    1 / 49
    2 / 49
    3 / 49
    4 / 49
```

After the alignment has been completed (or skipped), the next window allows users to subset protein-coding markers (genes) by codon. This is useful if for partitioned phylogenetic analysis if each codon position should be used as separate character set.

By selecting one or several gene and clicking "Subset", it becomes possible to define the respective genetic code (in the example, SGC1 = Vertebrate Mitochondrial is selected for the mitochondrial genes COI and Cytb, and SGC0 = Standard is selected for the nculear encoded Rag1 gene), and the reading frame (in this case, it is supposed that each gene sequence in the data set starts with a first codon position). Future versions of Concatenator will also include the option to estimate genetic code and/or reading frame.

Note that subsetting in the current version of the program will only have effect on the output formats Nexus, Partitionfinder, and IQtree. In these formats, codons will be written as separate character sets into the "sets" block or into the cfg file. Codons will not be separated into separate files and cannot be automatically deleted.



Character Sets by Codon

Select which genes to subset

Mark genes for codon subsetting in the character set specifications of the output file. Double-click an option field to edit it.

Name	Action	Genetic Code	Reading Frame	Naming
16S	-			
COI	Subset	SGC1	+1	Default
Cytb	Subset	SGC1	+1	Default
Pparadoxa_new_Nov_21	Subset	SGC0	+1	Default
Rag1	-			

Genes: 5 Marked: 3

The final "Export Sequence Data" window allows specifying the output format (see above for possible output file formats). As a default, the filename of the output file will get extended by a timestamp to avoid the risk to overwrite previous files that may still be needed.

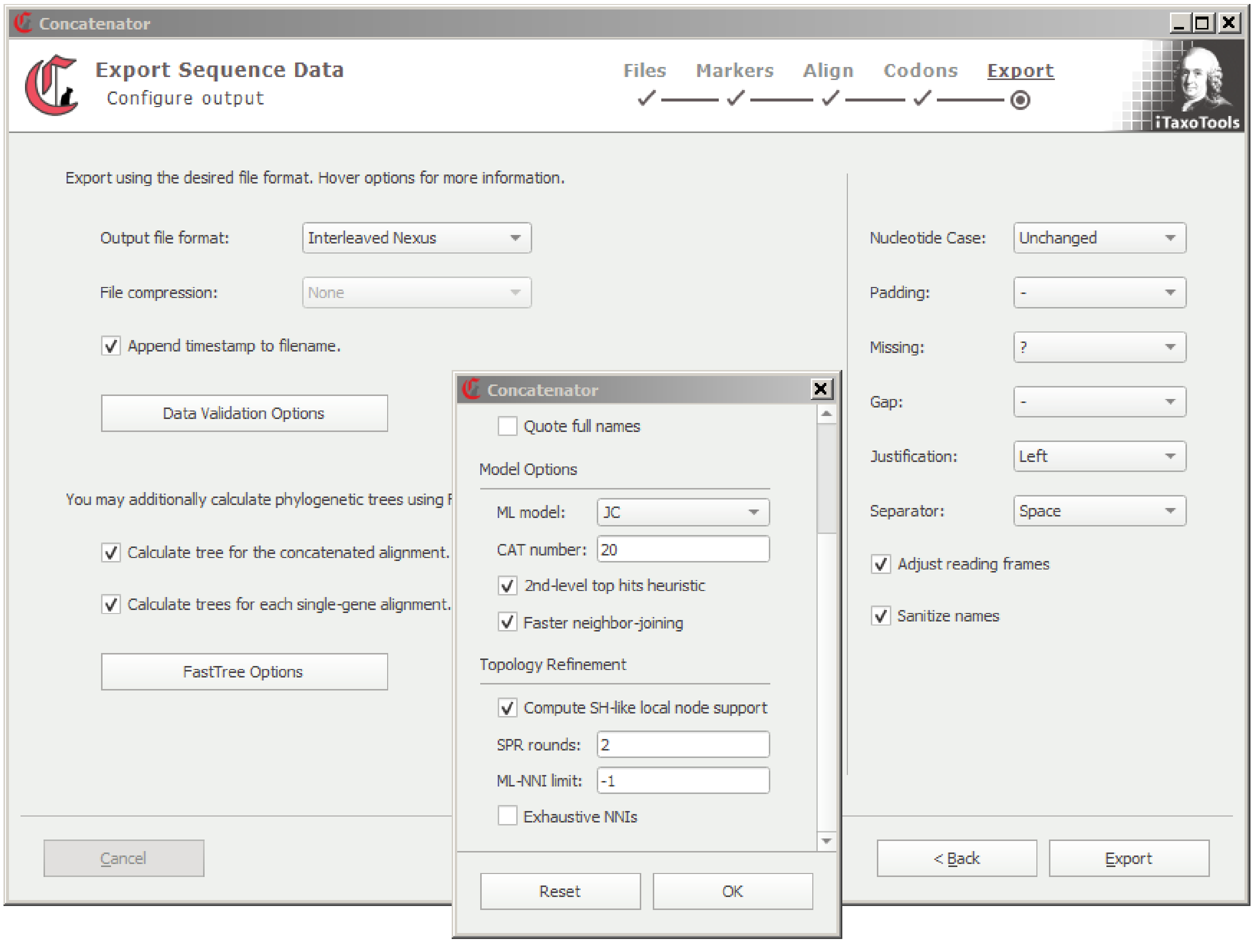
Several additional options can be selected:
►Users may choose to have the nucleotides in the alignment in upper or lower case, or unchanged (note that MAFFT after alignment changes all nucleotides to lower case).
►The "padding" option will add symbols for missing/ambiguous data (dashes,question marks, or N) to those genes not starting with a first codon position (only if selected in the codon subsetting window), so that all start with a first codon position. Padding also is used when concatenating sequences of different length, making them equal in length. This option is useful when you are sure the sequences are properly aligned, but some might be missing at the beginning or the end a few nucleotides.
►It also is possible which symbol should be used for missing data (N, n or question mark) and gaps (dash or asterisk).
►The "sanitize names" option will modify sequence names (sample names) so that all special characters, blanks and non-standard symbols are replaced by underscores.
►Finally, when converting to a multi-marker tsv (tab) file, an option is given to enforce spreadsheet compatibility. This means that in sequences starting with a gap symbol (dash), this first dash is replaced by a missing data symbol (N, n or question mark), so that spreadsheet editors may not interpret it as "minus" sign and thereby cause problems when viewing/editing the tsv file in Microsoft Excel or similar programs.

As shown in the screenshot below, Concatenator also implements a simplified version of Fasttree and allows user to export, along with the concatenated nucleotide alignment, a tree based on the combined (concatenated) data set as well as trees based on each nucleotide subset. Selecting "FastTree Options" opens a separate window with a limited amount of options of the FastTree program, as shown in the screenshot below. As a default, FastTree will calculate node support values as SH-like local supports (SH standing for the Shimodaira-Hasegawa test), but this option can be deselected. Trees will be saved in the Newick format.

Note that FastTree calculates "approximately-maximum-likelihood phylogenetic trees" using a specific algorithm described on the FastTree website and in the respective publications. For additional options, remember you can always hover over the respective parts of the menu to get additional explanations. Also consult the website for more detailed insights:
http://www.microbesonline.org/fasttree/

This algorithm is very fast and accurate, but for final analysis we recommend users to use specific programs using full likelihood calculations. The FastTree option is integrated to allow users to seamlessly obtain trees from their data, in order to inspect the outcome of different concatenation strategies or the phylogenetic signal in different gene trees.

FastTree is only available for aligned sequences (i.e., all sequences must be of the same length). If sequences of unequal length are used as input, these must be aligned with Mafft in order to be able to use the FastTree option.



Concatenator

Export Sequence Data

Configure output

Export using the desired file format. Hover options for more information.

Output file format: Interleaved Nexus

File compression: None

☑ Append timestamp to filename.

You may additionally calculate phylogenetic trees using F

☑ Calculate tree for the concatenated alignment.

☑ Calculate trees for each single-gene alignment.

Nucleotide Case: Unchanged

Padding: -

Missing: ?

Gap: -

Justification: Left

Separator: Space

☑ Adjust reading frames

☑ Sanitize names

Concatenator

□ Quote full names

Model Options

ML model: JC

CAT number: 20

☑ 2nd-level top hits heuristic

☑ Faster neighbor-joining

Topology Refinement

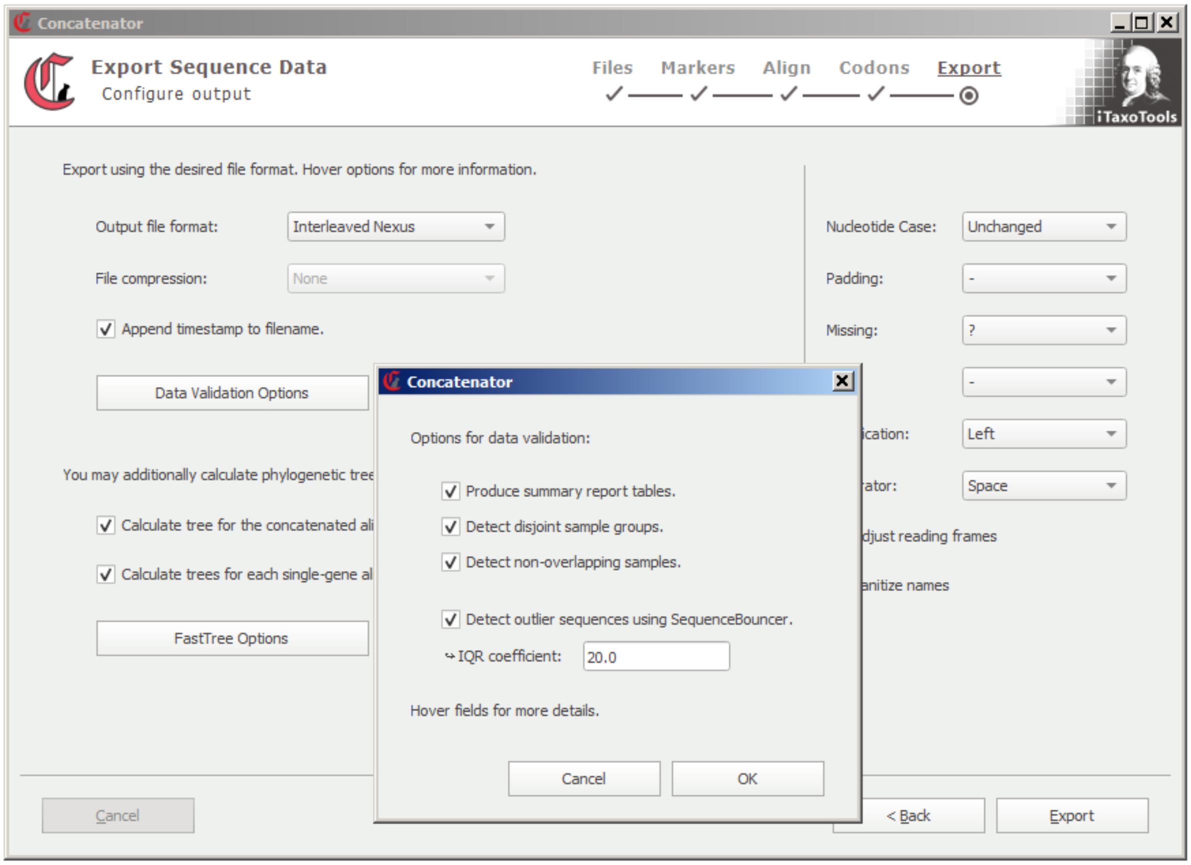
☑ Compute SH-like local node support

SPR rounds: 2

ML-NNI limit: -1

□ Exhaustive NNIs

The "Export Sequence Data" window also includes the option to activate various data validation algorithms via a dedicated button.



Export using the desired file format. Hover options for more information.

Output file format: Interleaved Nexus

File compression: None

☑ Append timestamp to filename.

You may additionally calculate phylogenetic tree

☑ Calculate tree for the concatenated ali

☑ Calculate trees for each single-gene al

Nucleotide Case: Unchanged

Padding: -

Missing: ?

-

ication: Left

ator: Space

djust reading frames

anitize names

Concatenator

Options for data validation:

☑ Produce summary report tables.

☑ Detect disjoint sample groups.

☑ Detect non-overlapping samples.

☑ Detect outlier sequences using SequenceBouncer.

↪ IQR coefficient: 20.0

Hover fields for more details.

Concatenator will produce summary reports of the total data set, content of input files, markers and species, including information on missing data, sequence length, and GC content. The "Total" report will also inform about markers and samples (taxa) that might have been entirely excluded from the output because they contained only missing data.

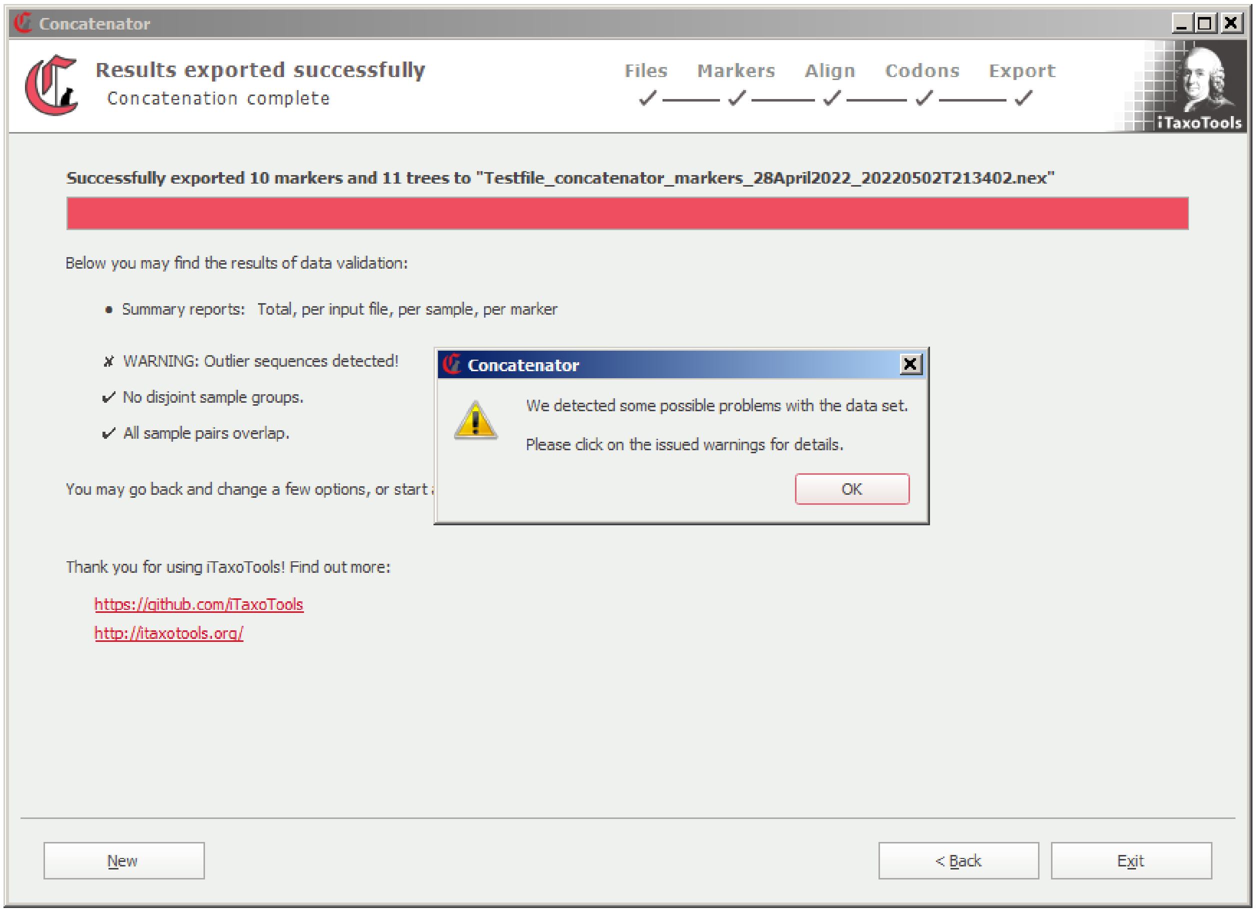
If sequences of unequal length were used for producing output files (i.e., sequences of one or several markers were of unequal length in the input and were not MAFFT-aligned), concatenator can adjust their length by padding (i.e., filling shorter sequences with missing data added terminally). This approach may automatically correct errors in an input sequence (e.g., a lacking terminal gap) but can also lead to a data set of basically unaligned sequences producing artifacts in downstream analysis. The reports therefore will inform which marker has been padded, and a warning will be issued if padding has been applied by the program.

The program furthermore includes three validation algorithms that can help to find problematic sequences or a problematic data structure:

►Concatenator implements the SequenceBouncer algorithm written by C.D. Dunn (https://github.com/corydunnlab/SequenceBouncer). This algorithm uses Shannon entropy values of alignment columns to identify outlier alignment sequences in a manner responsive to overall alignment context. See the original SequenceBouncer publication for details.
Concatenator does not implement all options of the original SequenceBouncer, but allows adjusting the IQR coefficient which makes the program more or less sensitive to the detection of outliers. These may represent misaligned, reverse-complemented or erroneous sequences, but can also simply represent outgroups or otherwise strongly divergent taxa. Concatenator submits sequences of each marker separately to a SequenceBouncer analysis. It uses IQR=20 as a default as in various test datasets of mitochondrial protein-coding DNA sequences, this value led to the identification of clearly misaligned sequences as outliers but did not flag all somewhat divergent taxa as such. However, for highly divergent sequences, such as ribosomal RNA sequences of phylogenetically diverse organisms, with numerous insertions and deletions leading to basically "unalignable" stretches, much lower IQR values (down to 1) can be appropriate. Finally, be aware that identification of outliers depend on them being an exception in the dataset; if an alignment contains a high proportion of misaligned or erroneously reverse-complemented sequences, the algorithm will likely fail identifying these as outliers.
In general, when sequences are flagged as outliers and these do not represent the outgroup or otherwise samples of expected high divergence, we recommend to carefully check the alignments.

►The program reports a list of samples (taxa) that do not share any common nucleotide position, e.g., in a dataset of two markers, one sample only has a sequence of marker 1 and the other sample has only a sequence of marker 2. If such non-overlapping samples occur in the data set, some downstream analyses (such as distance-based neighbor-joining or minimum evolution phylogenetic inference) of the concatenated data set become impossible. Occurrence of non-overlapping samples can also be an indication of wrongly concatenated sequences due to non-agreeing sequence names in input files. When non-overlapping samples are detected unexpectedly, users should check if these reflect the true data structure or are due to input errors, make sure the dataset is compatible with the planned downstream analyses, and be aware that such a data structure could lead to artifacts in inferred phylogenies.

►Concatenator includes a newly programmed algorithm that identifies sample groups of disjoint data (disjoint sample groups). These are groups of which each shares data for certain markers (among the samples included in the group), but does not share any marker with samples of the other groups. Such a distribution of missing data makes phylogenetic analysis among groups impossible and will necessarily create artefacts in the inferred phylogenetic trees. Disjoint sample groups are an artifact that can arise when sequence names in different input files are consistently different (e.g., missing underscores in sequence names in one input file). Upon data validation, Concatenator will issue a warning if disjoint sample groups are identified.



Results exported successfully

Concatenation complete

Successfully exported 10 markers and 11 trees to "Testfile_concatenator_markers_28April2022_20220502T213402.nex"

Below you may find the results of data validation:

- Summary reports: Total, per input file, per sample, per marker

✘ WARNING: Outlier sequences detected!

✓ No disjoint sample groups.

✓ All sample pairs overlap.

You may go back and change a few options, or start

Thank you for using iTaxoTools! Find out more:

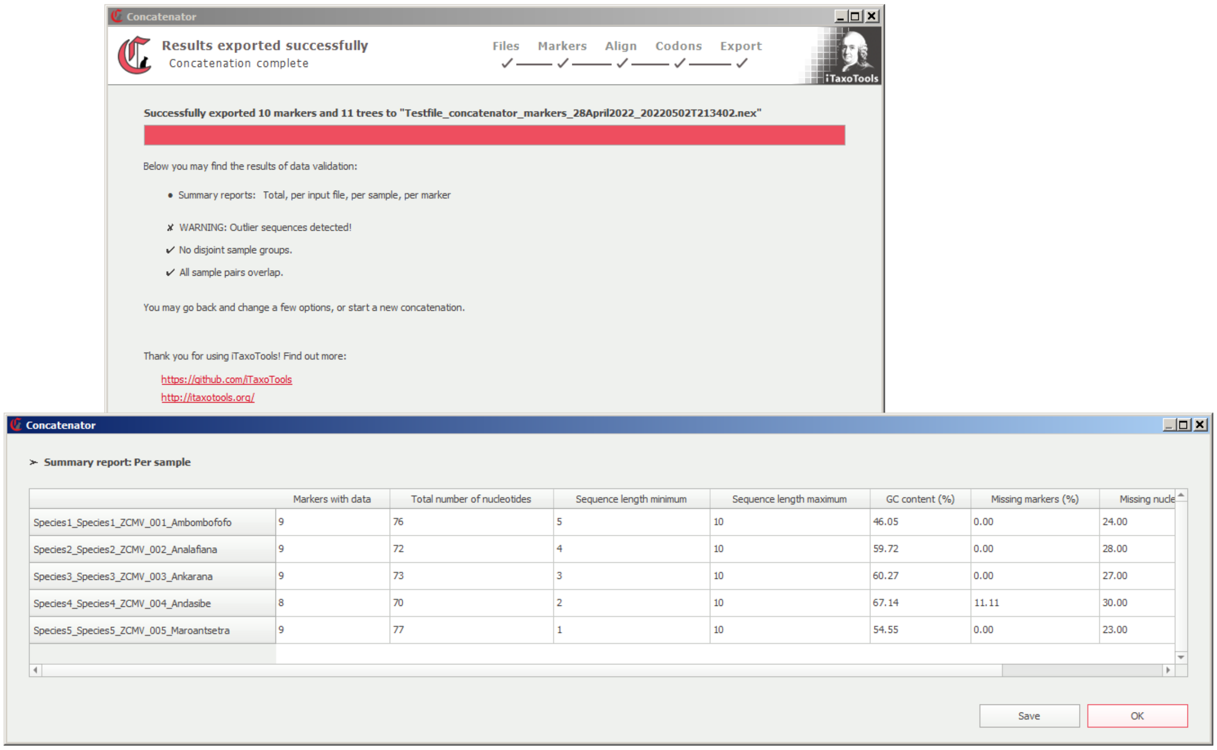
https://github.com/iTaxoTools

http://itaxotools.org/

Concatenator

We detected some possible problems with the data set.

Please click on the issued warnings for details.



Results exported successfully

Concatenation complete

Successfully exported 10 markers and 11 trees to "Testfile_concatenator_markers_28April2022_20220502T213402.nex"

Below you may find the results of data validation:

- Summary reports: Total, per input file, per sample, per marker

✘ WARNING: Outlier sequences detected!

✓ No disjoint sample groups.

✓ All sample pairs overlap.

You may go back and change a few options, or start a new concatenation.

Thank you for using iTaxoTools! Find out more:

https://github.com/iTaxoTools

http://itaxotools.org/

Concatenator

➢ **Summary report: Per sample**

	Markers with data	Total number of nucleotides	Sequence length minimum	Sequence length maximum	GC content (%)	Missing markers (%)	Missing nucle
Species1_Species1_ZCMV_001_Ambombofofo	9	76	5	10	46.05	0.00	24.00
Species2_Species2_ZCMV_002_Analafiana	9	72	4	10	59.72	0.00	28.00
Species3_Species3_ZCMV_003_Ankarana	9	73	3	10	60.27	0.00	27.00
Species4_Species4_ZCMV_004_Andasibe	8	70	2	10	67.14	11.11	30.00
Species5_Species5_ZCMV_005_Maroantsetra	9	77	1	10	54.55	0.00	23.00

2. MAFFTpy

Example files provided:

MAFFT_examplefile.fas

MAFFT accepts Fasta files with nucleotide (or protein) sequences and aligns them. The example file provided contains a series of 16S rRNA sequences that include a hypervariable region, showing the power of MAFFT to align such complex sequences.

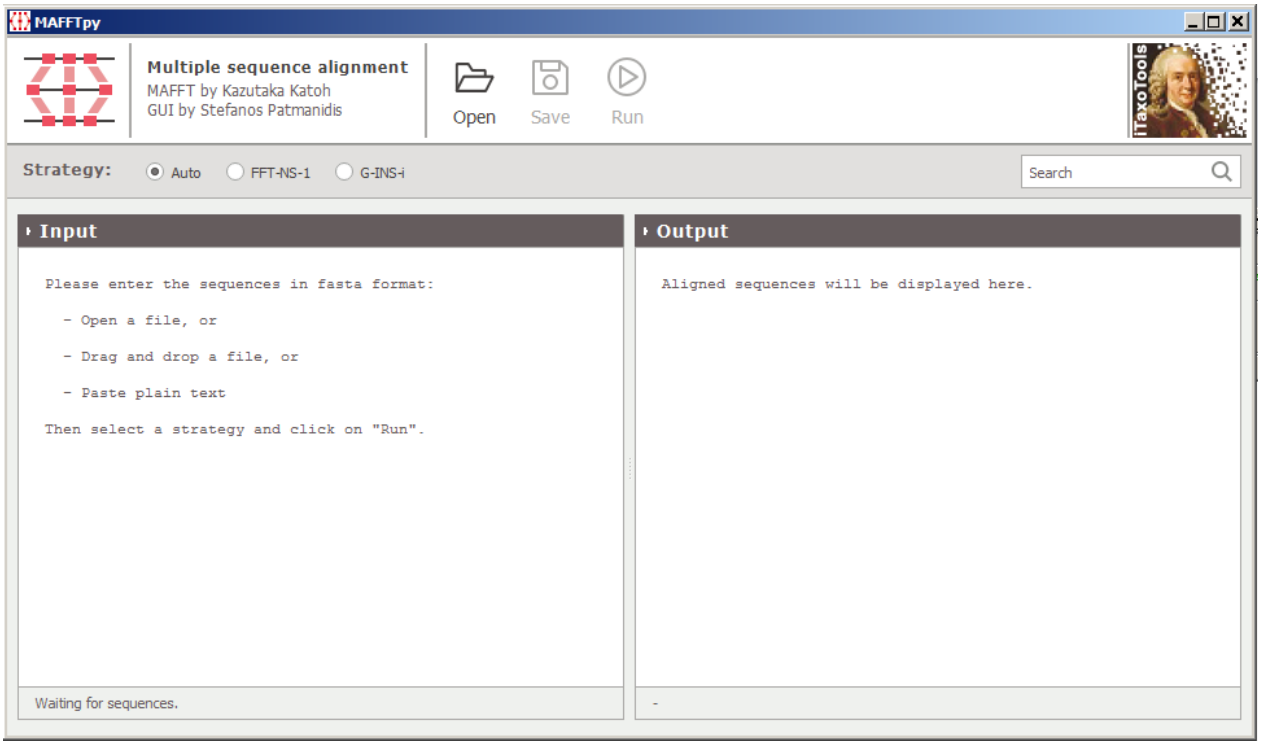
MAFFTPy is an offshoot of Concatenator, and provides in a dedicated executable a Python wrapper and GUI version of the original MAFFT program published by Katoh & Standley (2013).

This program performs multiple sequence alignments and in our experience is particularly suitable to align complicated data sets, e.g. of ribosomal RNA genes containing hypervariable (loop) regions. For iTaxoTools, our Python wrapper includes only a part of the original functionality of MAFFT – if you need the full performance of the program, we recommend to use the original version. However, for most routine alignment applications, MAFFTpy should be a suitable tool.

Input files can be selected using the "Open" button in the upper lines of symbols of the GUI, or by drag-and-drop. To execute the program, press the "Run" button in the upper line of buttons. The program will show the progress of the analysis in the window, and upon completion will show the resulting alignment in the window on the right. Upon completion, press the "Save" button to save the alignment.

Of the various alignment strategies of MAFFT, this Python wrapper only implements two:
► FFT-NS1, a fast progressive method that is able to align large sets of sequences (>2000),
► and G-INS-i, a thorough method that performs well also with sequences such as rRNA genes with constant and hypervariable regions.
► It also is possible to choose "Auto" and let the program determine the most suitable alignment strategy.

For more detailed information on the MAFFT algorithms, consult the program's webpage:
https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html



Strategy: ● Auto ○ FFT-NS-1 ○ G-INS-i

```
Please enter the sequences in fasta format:

  - Open a file, or

  - Drag and drop a file, or

  - Paste plain text

Then select a strategy and click on "Run".
```

```
Aligned sequences will be displayed here.
```

Waiting for sequences.

-

2. FastTreePy

Example files provided:

FastTree_examplefile.fas

The example file is a Fasta file with a large number of pre-aligned nucleotide sequences that can be uploaded to FastTreePy to test the speed of the program.

FastTreePy is an offshoot of Concatenator, and provides in a dedicated executable a Python wrapper and GUI version of the original FastTree program published by Price et al. (2010). This program can be used to calculate at high speed phylogenetic trees using the "approximately-maximum-likelihood" approach described on the FastTree website and in the respective publications. For additional options, remember you can always hover over the respective parts of the menu to get additional explanations. Also consult the website for more detailed insights: http://www.microbesonline.org/fasttree/

FastTree can calculate node support values as SH-like local supports (SH standing for the Shimodaira-Hasegawa test), but this option can be deselected.

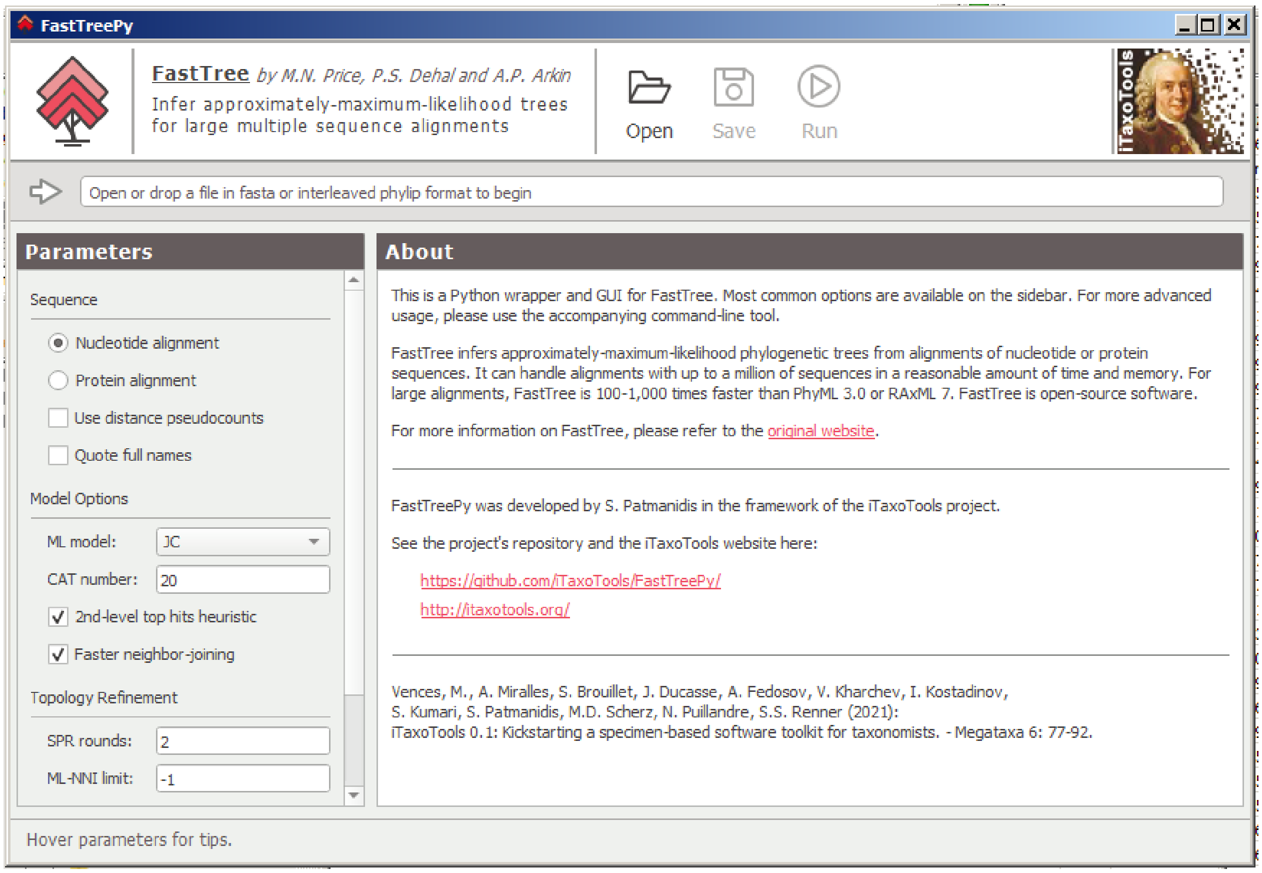
FastTree accepts input files in Fasta and Phylip formats. Sequences must be aligned, and sequence names sanitized (no blanks or special characters in sequence names – you can use Concatenator or DNAconvert to seamlessly adjust sequence files accordingly).

You can choose between the JC and GTR substitution models.

A main advantage of FastTree is its speed, potentially several orders of magnitude faster than programs relying on full likelihood calculations. The program can handle alignments of up to a million of sequences in a reasonable amount of time and memory.

Input files can be selected using the "Open" button in the upper lines of symbols of the GUI, or by drag-and-drop. To execute the program, press the "Run" button in the upper line of buttons. The program will show the progress of the analysis in the window. Upon completion, press the "Save" button to save the tree (in Newick format).

For more in-depth explanations of the FastTree algorithm and the different options, see the original website of the FastTree developers:
http://www.microbesonline.org/fasttree/



FastTree *by M.N. Price, P.S. Dehal and A.P. Arkin*

Infer approximately-maximum-likelihood trees
for large multiple sequence alignments

Open or drop a file in fasta or interleaved phylip format to begin

Parameters

Sequence

- ● Nucleotide alignment
- ○ Protein alignment
- □ Use distance pseudocounts
- □ Quote full names

Model Options

- ML model: JC
- CAT number: 20
- ☑ 2nd-level top hits heuristic
- ☑ Faster neighbor-joining

Topology Refinement

- SPR rounds: 2
- ML-NNI limit: -1

About

This is a Python wrapper and GUI for FastTree. Most common options are available on the sidebar. For more advanced usage, please use the accompanying command-line tool.

FastTree infers approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences. It can handle alignments with up to a million of sequences in a reasonable amount of time and memory. For large alignments, FastTree is 100-1,000 times faster than PhyML 3.0 or RAxML 7. FastTree is open-source software.

For more information on FastTree, please refer to the original website.

FastTreePy was developed by S. Patmanidis in the framework of the iTaxoTools project.

See the project's repository and the iTaxoTools website here:

https://github.com/iTaxoTools/FastTreePy/
http://itaxotools.org/

Vences, M., A. Miralles, S. Brouillet, J. Ducasse, A. Fedosov, V. Kharchev, I. Kostadinov,
S. Kumari, S. Patmanidis, M.D. Scherz, N. Puillandre, S.S. Renner (2021):
iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. - Megataxa 6: 77-92.

Hover parameters for tips.