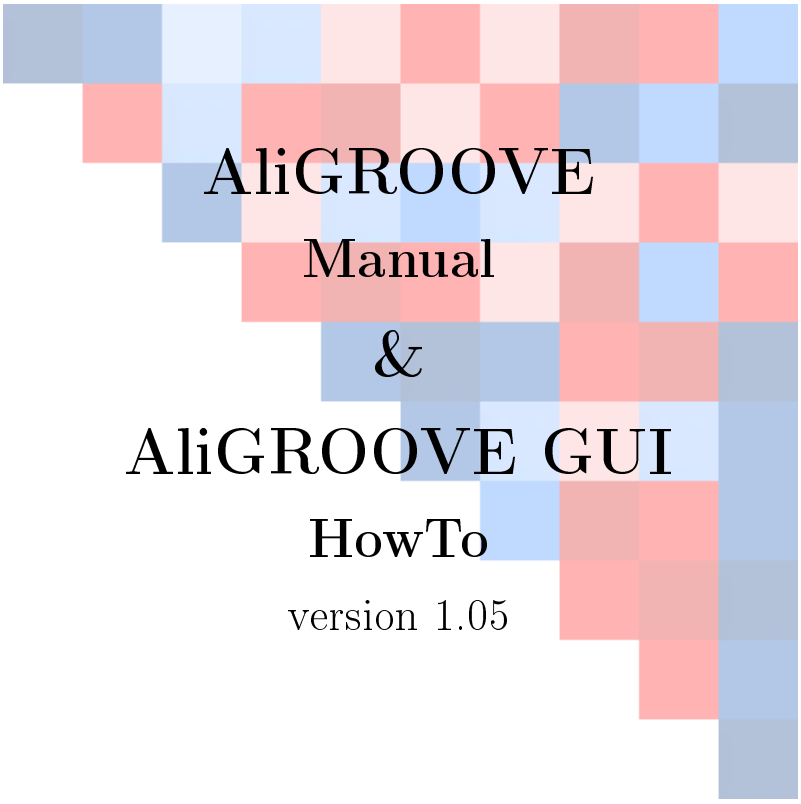
Zentrum
für molekulare
Biodiversitätsforschung



AliGROOVE
Manual
&
AliGROOVE GUI
HowTo

version 1.05

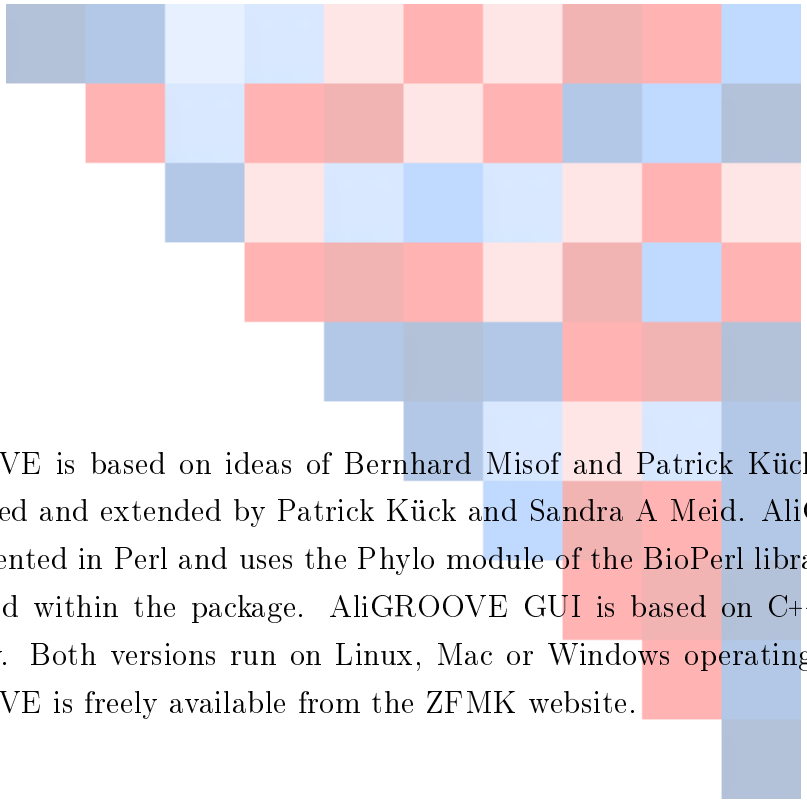
Sandra A Meid,
Patrick Kück, and
Bernhard Misof

2013

Zoologisches Forschungsmuseum
Alexander Koenig

AliGROOVE and AliGROOVE GUI

version 1.05



AliGROOVE is based on ideas of Bernhard Misof and Patrick Kück and has been refined and extended by Patrick Kück and Sandra A Meid. AliGROOVE is implemented in Perl and uses the Phylo module of the BioPerl library, which is delivered within the package. AliGROOVE GUI is based on C++ and the Qt library. Both versions run on Linux, Mac or Windows operating systems. AliGROOVE is freely available from the ZFMK website.

Notice:

The authors are not aware of bugs that would cause the program to obtain incorrect results, but they could exist. Even though the authors try to make this program as reliable as possible, it could be that parts of the program do not work as intended. Please report any crashes, bugs, or problems you have with this program. This program is distributed in the hope that it will be useful, but without any warranty.

License:

AliGROOVE is Copyright (C) 2013 by Patrick Kück, Sandra A Meid and Bernhard Misof. You may use, distribute and copy AliGROOVE under the terms of GNU General Public License version 2.
http://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html

Contents

1 Manual

1.1 General Usage

The input must contain an infile, all other options are not mandatory.

```
AliGROOVE.pl -i infile.fas
```

A random input order of options is allowed. You can use e.g.

```
AliGROOVE.pl -i infile.fas -z treefile.tre -N
```

or

```
AliGROOVE.pl -N -i infile.fas -z treefile.tre
```

All options must proceed a '-' sign like $-i$ or $-z$ etc. Specifications of options like tree file or window size must directly follow option invokation with blanks inbetween like $-z$ *treefile.tre* or $-w$ 4. If you use no option, ALiGROOVE will switch to defaults, whereas indels are treated as 5^{th} character using a window size of 6.

1.2 Options

1.2.1 Multiple Sequence Alignment Infile ('-i' option)

```
-i filename.fas
```

Specify the full name of the input file including extension. The alignment input file must be in the same folder as AliGROOVE. AliGROOVE currently accepts multiple sequence alignments in FASTA (*.fas*) format, in simple ASCII text format. Avoid formating input files with text editors like MSWord or something comparable. The first line break is interpreted as taxon name separator following the FASTA file convention. Taxon names are allowed to contain alphanumeric signs and underscores. Sequences are allowed to contain nucleotide or amino acid coding signs, ?, indels and ambiguity characters. AliGROOVE identifies

the sequence data type from frequencies of A, C, G and T|U. RNA sequences will be recoded to DNA sequences. There is no restriction on sequence number and sequence length.

Example of a FASTA formatted alignment input file:

```
>Podura_aquatica_18S_1
aaagtctgtgacgttgtacggact
gcgtgtgcagctgtgacgggcgcc
>Sminthurus_sp
AUTGCTugccguuugaucgugugc
UUGGACUGCGUCGATCGUUGCGCG
```

1.2.2 Indel Treatment ('- N' option)

```
-N
```

Without invoking the '– N' option indels are treated as 5^{th} character state. With the '– N' option indels are treated as ambiguous characters. The '– N' option is ignored for amino acid sequences.

1.2.3 Dimension of the Sliding Window ('- w' option)

```
-w
```

Specifies dimension of the sliding window, default is w = 6. A window size below 4 does not make much sense since error rates for miscalling randomness and non-randomness will be much to high. Only window dimensions of even numbered sizes will be accepted, if an uneven number is called by the user it will be changed to the next larger even number. Larger window dimensions will make the profiling less sensitive to small alignment sections of randomness.

1.2.4 Define a Guiding Tree for Tagging

`-z treefile.tre`

Define a guiding tree in NEWICK format to tag potentially unreliable relationships. Specify the full name of the tree input file with file extension (*.tre* or *.txt*). Calculated reliabilities of single branches are shown colourized. Taxon names have to be identic between the sequence input file and the treefile. The tagging colour of each branch depends on the mean similarity score obtained between taxa connected by this branch. Red indicates that ambiguously aligned sequence positions dominate between two sequences while blue indicates the opposite. The more positive or negative the total similarity score between two sequences, the darker the corresponding colour.

1.2.5 Scoring

`-BLOSUM62, -PAM250, -PAM500`

Nucleotide Data For nucleotide sequences AliGROOVE uses a simple match/mismatch scoring matrix to calculate the observed window and resampled window scores. If ambiguities are present, AliGROOVE makes an optimistic estimate with a reduced match score according to degeneracy of ambiguities. Indels are either scored as 5^{th} characters or ambiguities.

Amino Acid Data For amino acid sequences the BLOSUM62 matrix is used as default amino acid substitution matrix. Alternatively, one can choose between the PAM250 and the PAM500 matrix. Indels are scored as strongly penalized mismatches if matching and not penalized if matching an amino acid. As a result, if indels dominate in certain sections, these sections will be negatively scored in consensus profiles. However, in sections dominated by amino acid sequence information, indels will not have a negative effect. This scoring scheme accounts for missing data, which is typical for concatenated EST or phylogenomic data.

Blosum62 source: NCBI

```
Matrix made by matblas from blosum62.iij
* column uses minimum score
BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
Blocks Database = /data/blocks_5.0/blocks.dat
Cluster Percentage: >= 62
Entropy = 0.6979, Expected = -0.5209
```

PAM250 source: NCBI

```
This matrix was produced by "pam" Version 1.0.6 [28-Jul-93]
PAM 250 substitution matrix, scale = ln(2)/3 = 0.231049
Expected score = -0.844, Entropy = 0.354 bits
Lowest score = -8, Highest score = 17
```

PAM500 source: NCBI

```
This matrix was produced by "pam" Version 1.0.6 [28-Jul-93]
PAM 500 substitution matrix, scale = ln(2)/7 = 0.0990210
Expected score = -0.401, Entropy = 0.0803 bits
Lowest score = -9, Highest score = 34
```

1.3 Output

AliGROOVE prints two output files whithout additional tree specification and four output files if a topology has been specified via the *-z* option. The output files are saved to the same folder from which the input of the analysed data was loaded.

AliGROOVE_seqsim_matrix_infile.svg
→ Colour coded pairwise sequence similarity matrix

AliGROOVE_seqsim_matrix_infile.txt
→ Pairwise sequence similarity matrix scores

AliGROOVE_tagged_info_treefile.txt
→ Summarized branch reliability tagging information of a given tree file

AliGROOVE_tagged_tree_treefile.svg
→ Tagged branch reliabilities of a given treefile

1.4 Help

For detailed help on options type

```
AliGROOVE.pl help option
```

to get information on infiles, output and scoring.

AliGROOVE.pl help -i
AliGROOVE.pl help -z
AliGROOVE.pl help -output
AliGROOVE.pl help -scoring
AliGROOVE.pl help -commands

2 HowTo

After unpacking the *.zip* or *.tar.gz* package, the folder containing AliGROOVE GUI has a defined folder structure. The main folder contains a linkage or sh-script, which starts the application. The executables are stored within */bin/os*. The folder */data* contains test datasets. The Manual and How-To can be found within */doc*. */scipt* contains the AliGROOVE PERL scripts.

2.1 Start AliGROOVE GUI

Start the application. For Windows and MacOS double-click on the AliGROOVE linkage within the main folder (Microsoft Windows shortcut or the Apple Mac alias) or the executable (within */bin/win* or */bin/mac*), for Linux double-click or start the sh-script in the main folder or the executable (within */bin/ubuntu* or */bin/ubuntu*64).

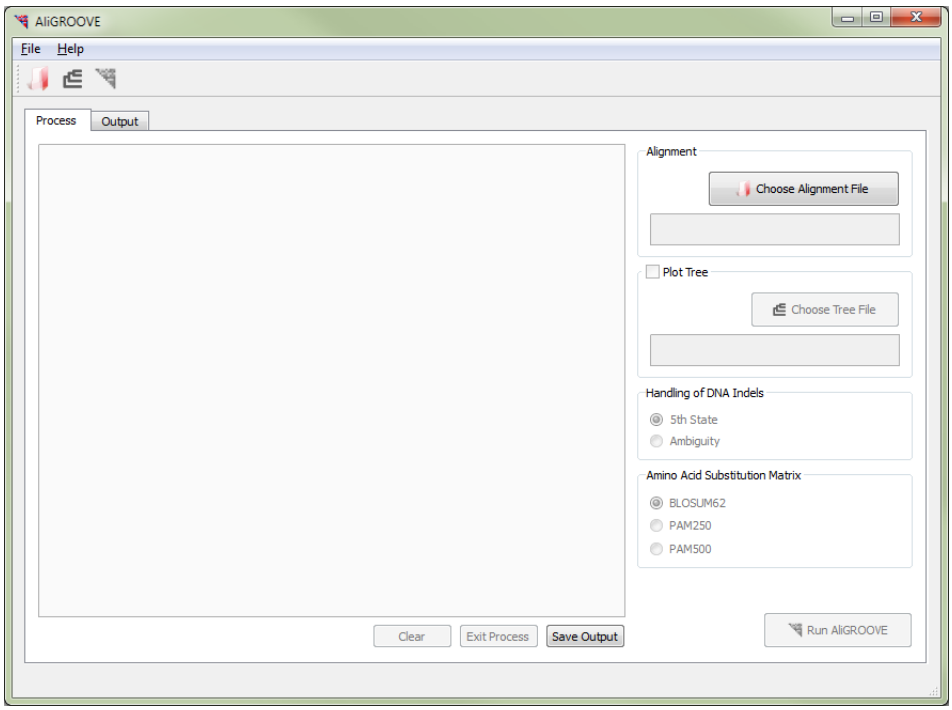


Figure 1: Screenshot of the AliGROOVE GUI.

AliGROOVE depends on Perl to run the analysis. If not installed, AliGROOVE GUI will offer to go to the website where you can find the required software.

2.2 Choose a Multiple Sequence Alignment File

By clicking on the <Choose Alignment File> button a window appears in which you can choose the file you want to process. AliGROOVE is capable of reading data files in the FASTA format and attempts to estimate the data type, either nucleotide or amino acid data. Therefore ALiGROOVE considers sequences whith an A, C, G, T|U content of > 0.8 (without counting indels and N) as nucleotide sequences, if less then 0.8 as amino acid data. It estimates data property from every sequence, if two sequences are considered of different data type, ALiGROOVE stops. ALiGROOVE might stop if a single nucleotide sequence contains more then 0.2 ambiguities. In almost every case, ALiGROOVE will correctly estimate data type, if it does not, it will stop and report on the problem. If the data contains sequences of more then 0.2 ambiguities, it might be advisable to recode ambiguities as N's or remove the particular sequence. RNA sequences will be recoded to DNA sequences. Nucleotide data can be a mix of RNA/DNA data.

Example of an input file:

```
>Podura_aquatica_18S_1
aaagtctgtgacgttgtacggact
gcgtgtgcagctgtgacgggcgcc
>Sminthurus_sp
AUTGCTugccguuugaucgugugc
UUGGACUGCGUCGATCGUUGCGCG
```

2.3 Plot Tree

If you choose the **Plot Tree** option, the button <Choose Tree File> is activated. After clicking this button a window appears, in which you can choose a tree file in Newick format, rooted or unrooted. The tree must correspond to the alignment file to be processed. It will then be plotted in a new register card with tree branches coloured corresponding to the results. Make sure that your tree does not contain CRFL linefeeds from Windows if working on Linux.

2.4 Handling of DNA Indels

The **Handling of DNA Indels** parameter controls how indels are treated in molecular sequence data. This option does currently not work for amino acid data.

default The default option is 5^{th} *State*. It can be switched to *Ambiguity*.

5^{th} State Gaps are treated as an additional state.

Ambiguity Gaps are interpreted as ambiguous characters.

2.4.1 Amino Acid Substitution Matrix

`-BLOSUM62, -PAM250, -PAM500`

The BLOSUM62 matrix is used as default amino acid substitution matrix. Alternatively, one can choose between the PAM250 and the PAM500 matrix. Indels are scored as strongly penalized mismatches if matching and not penalized if matching an amino acid. As a result, if indels dominate in certain sections, these sections will be negatively scored in consensus profiles. However, in sections dominated by amino acid sequence information, indels will not have a negative effect. This scoring scheme accounts for missing data, which is typical for concatenated EST or phylogenomic data.

Blosum62 source: NCBI

```
Matrix made by matblas from blosum62.iij
* column uses minimum score
BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
Blocks Database = /data/blocks_5.0/blocks.dat
Cluster Percentage: >= 62
Entropy = 0.6979, Expected = -0.5209
```

PAM250 source: NCBI

```
This matrix was produced by "pam" Version 1.0.6 [28-Jul-93]
PAM 250 substitution matrix, scale = ln(2)/3 = 0.231049
Expected score = -0.844, Entropy = 0.354 bits
Lowest score = -8, Highest score = 17
```

PAM500 source: NCBI

```
This matrix was produced by "pam" Version 1.0.6 [28-Jul-93]
PAM 500 substitution matrix, scale = ln(2)/7 = 0.0990210
Expected score = -0.401, Entropy = 0.0803 bits
Lowest score = -9, Highest score = 34
```

2.5 Run AliGROOVE

To start the analysis, click on the <Run AliGROOVE> button. AliGROOVE switches to the tab <Output> after processing the data. The matrix can be zoomed in and out by using the mouse wheel.

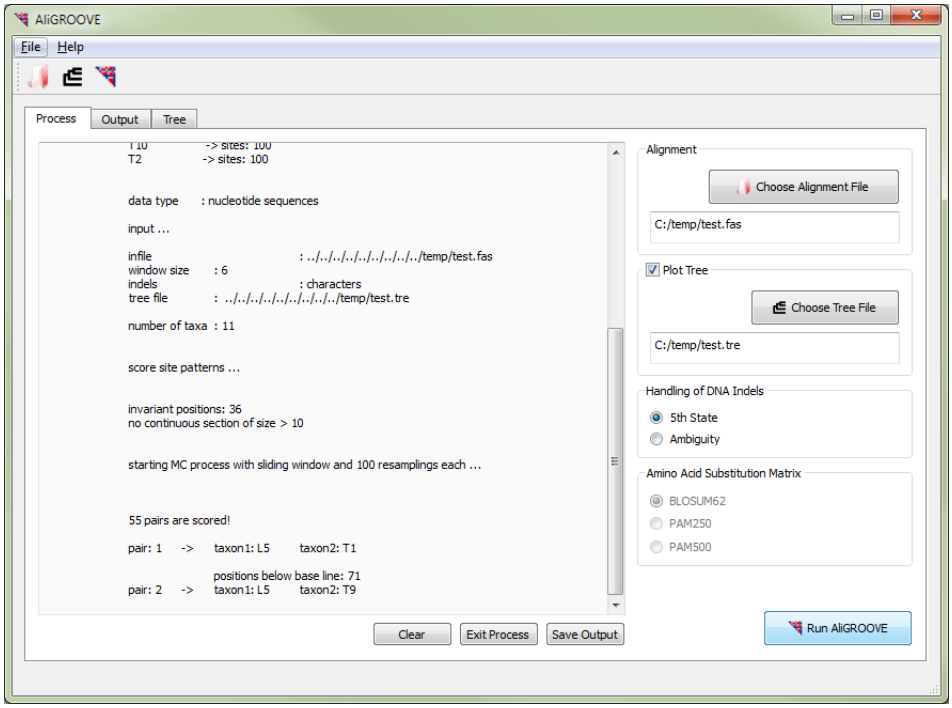


Figure 2: Screenshot of the AliGROOVE 'Process' tab while a process is running.

2.6 AliGROOVE Matrix

After the analysis has been finished the results are plotted as matrix within the tab <Output>.

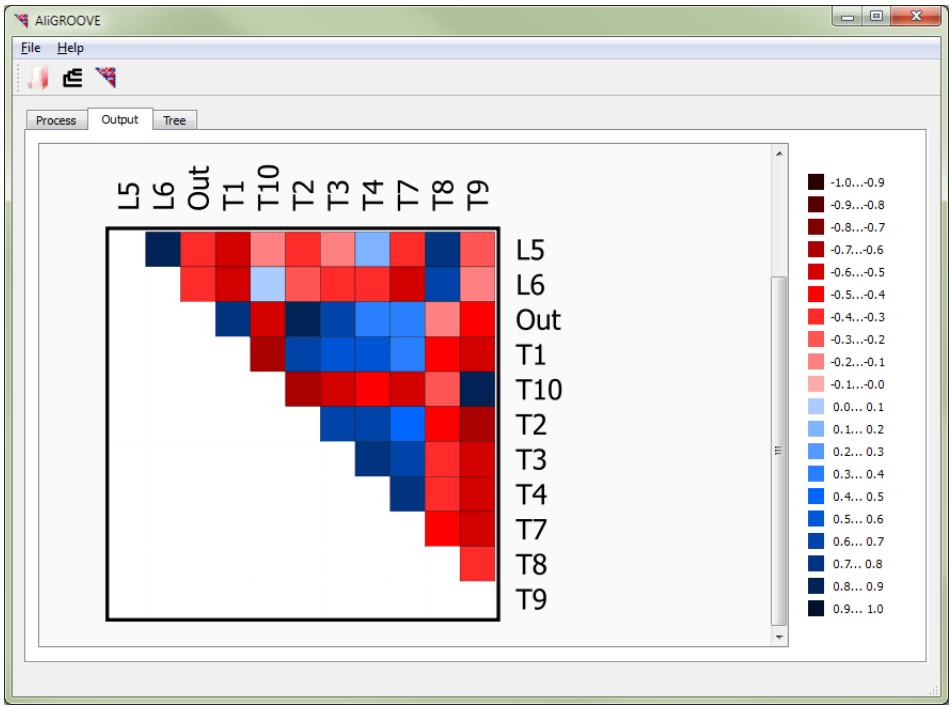


Figure 3: Screenshot of the AliGROOVE 'Output' tab.

All taxa names are listed on top and the right hand side of the matrix. The obtained mean similarity score between sequences is represented by a coloured square where the row of one and the column of the other taxon have an overlap. The site scores are ranging from -1, indicating full random similarity, to +1, non-random similarity. The corresponding colour code is listed on the right hand side of the matrix widget.

Red indicates that ambiguously aligned sequence positions dominate between two sequences while blue indicates the opposite. Increasing random similarity is indicated by darker red, decreasing random similarity is expressed by darker blue squares.

2.7 AliGROOVE Tree

If a tree file has been specified, the tree and the identifed branch reliabilities are shown colourized in a new tab <Tree> after the analysis has been finished. The tree can be zoomed in and out by using the mouse wheel.

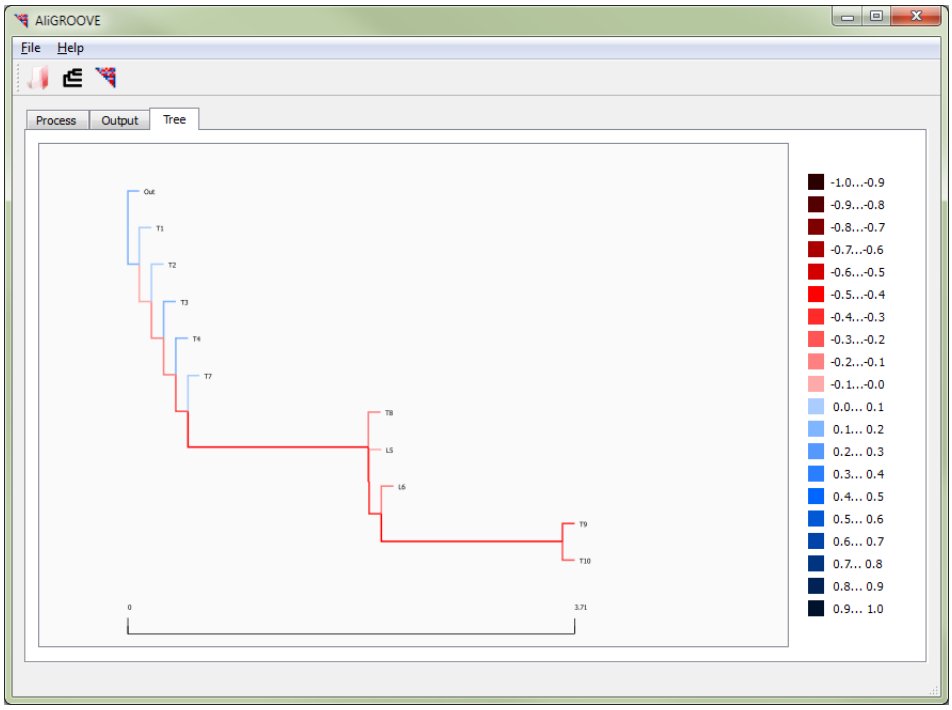


Figure 4: Screenshot of the AliGROOVE 'Tree' tab.

The terminal branches are coloured based on the mean score between the corresponding terminal taxon and all other taxa. Internal branch scores are calculated by the mean of the total pairwise similarity scores between taxa which are connected by a branch.

This tagging of branches is effectively an indirect estimation of reliability of a subset of all possible splits guided by a topology. Calculated reliabilities of single branches are colour coded. The corresponding colour key is listed on the right hand side of the tree widget.
Red coloured branches indicate a low mean similarity score obtained between taxa connected by this branch. Blue indicates a higher mean similarity score and therefore a higher reliability of the subset. The higher or lower the similarity score is, the darker is the colour of the branch.