##
README for PhyloPipe v1.1
######################## License Information ##############################
#
This program is free software; you can redistribute it and/or modify it as you wish.
This program is developed by Parellel and Distributed Laboratory, Nankai Univeristy.
Program developer: Yixuan, Guo; Haoyang Wu; Yang Liu; Gang, Wang; Xiaoguang, Liu; Qiang Xie; Mingming Ren.
The PhyloPipe.v1.1 package is available at http://sf.net/projects/phylopipe/.
#
##

1 Preparation

You need to install a number of programs that are required to run the specific function of PhyloPipe v1.1 Installation on Linux would be straightforward.

1.1 HaMStR

To use the HaMStR function you need to install HaMStR first.

Download HaMStR from: http://sourceforge.net/projects/hamstr/.

To use hamstrsearch_local you need first to install a number of programs that are required to run the HaMStR search.

a) hmmsearch version 3 from http://hmmer.janelia.org/.

b) blastall from ftp://ftp.ncbi.nih.gov/blast/executables/release/. Alternatively, you can use the blast+ suite.

c) usearch from http://drive5.com/usearch/.

NOTE: installing usearch is optional. If you want to use HaMStR without the -ublast option configure the script with --noublast (see below)

d) genewise version 2.4.1 from ftp://ftp.ebi.ac.uk/pub/software/unix/wise2/

NOTE: genewise installation is optional. If you want to run HaMStR only on protein sequences, run the configure script with the option --protein_only (see below)

e) clustalw2 from http://www.clustal.org/download/current/

Then you need to install perl and bio-perl to support the scripts.

1.2 Mafft

To use the Mafft function you need to install Mafft first.
Download Mafft from: http://mafft.cbrc.jp/alignment/software/.

1.3 Muscle

To use the Muscle function you need to install Muscle first.
Download Muscle from: http://www.drive5.com/muscle.

1.4 Aliscore

To use the Aliscore function you need to install Aliscore first.
Download Aliscore from: https://www.zfmk.de/en/research/research-centres-and-groups/aliscore.

1.5 RaxMLGUI

To use the graphic software of RaxML, you need to install RaxMLGUI first. Then you need to install python and Tkinter. Note: raxmlGUI does not support Python 3.
Download RaxMLGUI from: sourceforge.net/projects/raxmlgui/.
Download Python from: https://www.python.org/downloads/.
Install Tkinter for Python via entering command in terminal, the command depends on the Operating System of your computer, use the instructions from http://tkinter.unpythonic.net/wiki/How_to_install_Tkinter as reference.

1.6 Ruby

To use the concatenate function, you need to install ruby script.
Download Ruby from: https://www.ruby-lang.org/en/downloads/.

1.7 QT Creator (Optional)

To make modification to PhyloPipe project, you need to install QT Creator development platform.
Dowload QT Creator development platform from: http://www.qt.io/download-open-source/.

2 Installation

2.1 Directory structure

Once you have unpacked the tar-file the following directory structure should be available:

PhyloPipe.v1.1

SystemFiles ##contains the concatenator.rb scripts for the concatenate function, SysExec file for alignment and matrixmasking function, Settings.ini for relevant program path parameters storage.

PhyloPipe_Project #contains the project files of PhyloPipe, you can build the project in QT creator and modify the software as you like.

PhyloPipe.v1.1 #the excutable file of the software.

2.2 Direct Execution

You can double-click the PhyloPipe.v1.1 executable file to run PhyloPipe software on Linux system. Our developing and running system is CentOS 6.5, on which PhyloPipe.v1.1 shows perfect performance.

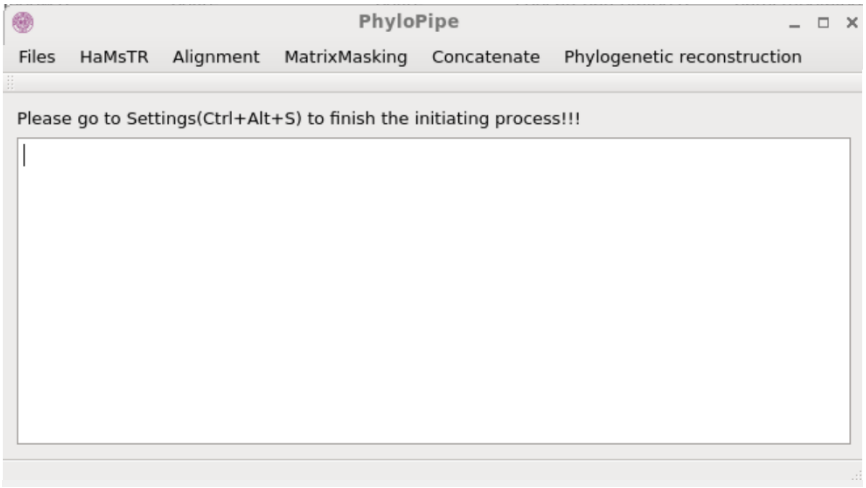
2.3 Project Modification

PhyloPipe_Project directory contains the PhyloPipe project files, you can modify and re-compile the PhyloPipe project on QT Creator development platform on Linux system.

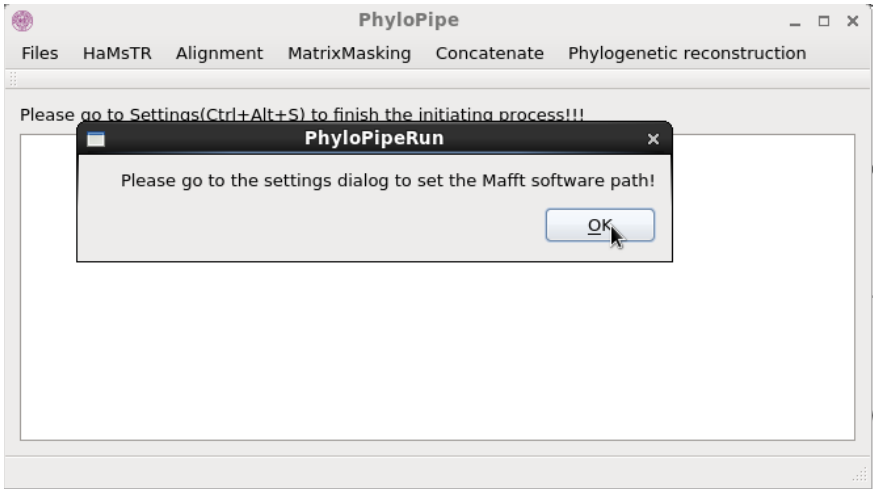
3 Instruction

3.1 PhyloPipe Main Window

The main window is like this.

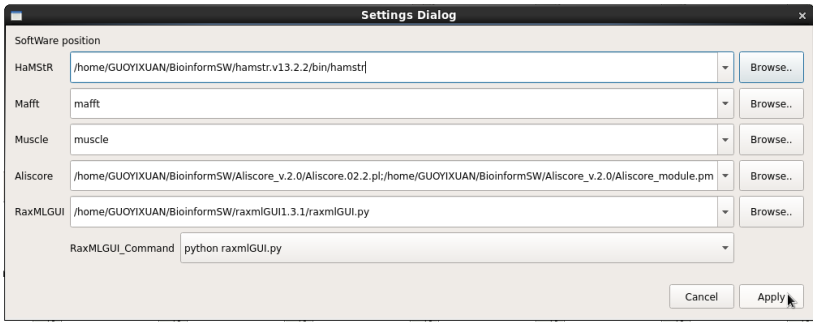


Notice: there is a sentence that guide you to set the Setting dialog before you start your analysis. If you do not follow the guidance and click a function, there will be a dialog popped out and remind you to set the Setting dialog and refuse to show the configuration dialog for a certain function.



Use the Ctrl+Alt+S combine keys to open the Settings Dialog, it looks like this. Use “Browse..” button to choose the path of those softwares. Users are free to edit the path by keyboard.

Note: the RaxmlGUI_Command option is for setting the command to call for RaxmlGUI in the terminal. There are three choices, of which to choose depends on the OS of user’s computer. Please read the RaxmlGUI program’s readme.txt document for more information.



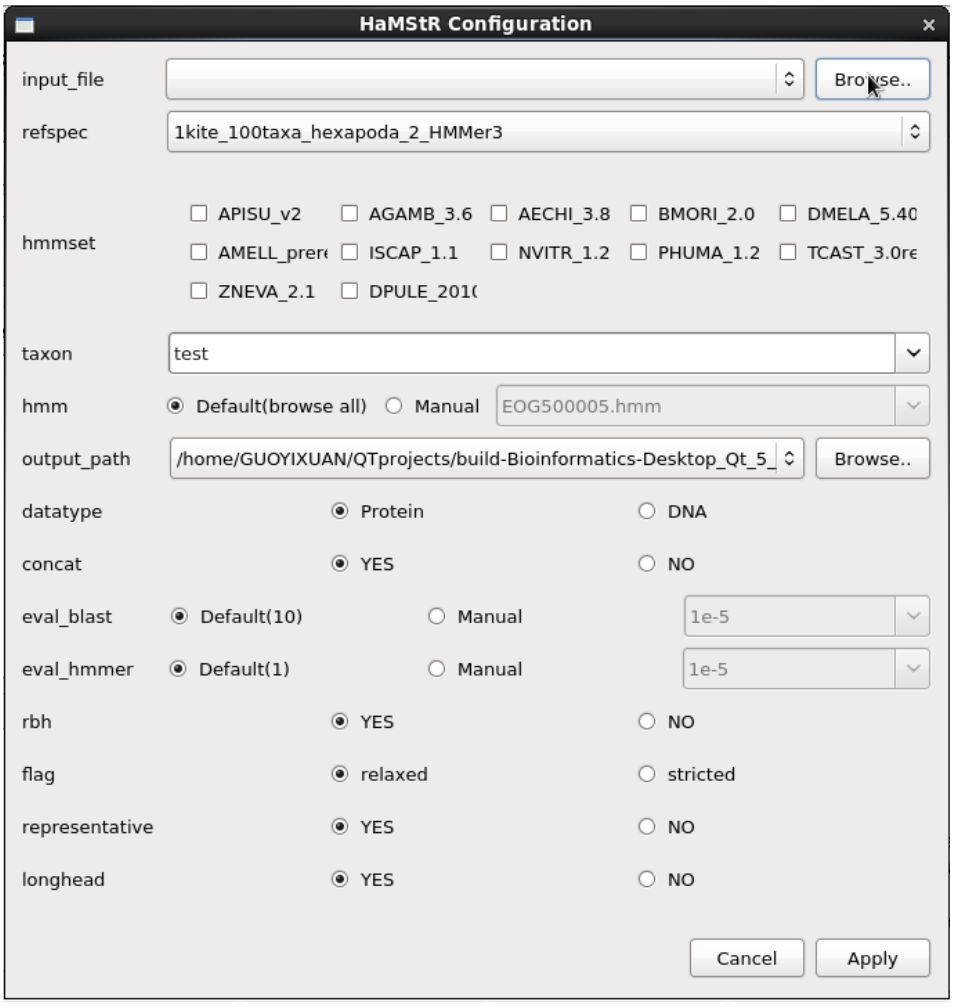
After finishing the settings, you can use the functions by clicking the menu button.

There are 7 blocks in this software.

HaMStR
DataFilter
Mafft
Muscle
MatrixMasking
Concatenate
RaxMLGUI

3.2 HaMStR

HaMStR Dialog:



The parameters used in HaMStR are as follows:

inputfile

path and name of the file containing the sequences hmmer is run against.

hmmset

specifies the name of the core-ortholog set. The program will look for the

files in the default directory 'core-orthologs' unless you specify a different path via the option -hmmpath. Setting this flag will list all available core ortholog sets in the specified path. Can be combined with -hmmpath.

refspec

sets the reference species. Note, it has to be a species that contributed sequences to the hmms you are using. NO DEFAULT IS SET! For a list of possible reference taxa you can have a look at the speclist.txt file in the default core-ortholog sets that come with this distribution. Please use the abreviations in this list. You can choose more than one reference species at one time. The lower-ranking reference species will only be used if a certain gene is not present in the preferred refspecies due to alternative paths in the transitive closure to define the core-orthologs.

CURRENTLY NO CHECK IS IMPLEMENTED!

NOTE: A BLAST-DB FOR THE REFERENCE SPECIES IS REQUIRED!

taxon

you need to specify a default taxon name from which your ESTs or protein sequences are derived.

datatype

select the type of input data.

concat

set this flag if you want hamstr to concatenate sequences that align to non-overlapping parts of the reference protein. If you choose this flag, no co-orthologs will be predicted.

eval_blast

this option allows to set the e-value cut-off for the Blast search. Default: 10

eval_hmmer

this options allows to set the e-value cut-off for the HMM search. Default: 1

hmm

option to provide only a single hmm to be used for the search.

longhead

set this flag in the case your sequence identifier contain whitespaces and you wish to keep the entire sequence identifier throughout your analysis. HaMStR will then replace the whitespaces with a '__'. If this flag is not set, HaMStR will truncate the sequence identifier at the first whitespace, however only if the sequence identifier then remain unique.

NOTE: too long sequence headers (~ > 30 chars) will cause trouble in the hmmsearch as the program will truncate the output!

outputpath

you can determine the path to the HaMStR output. Default: current directory.

rbh

set this flag if you want to use a reciprocal best hit criterion. Only the highest scoring hit from the hmmer search will be used for re-blast.

flag

set this option determines whether the stricted reciprocity criterion is

applied.

-relaxed:set this flag if the reciprocity criterion is fulfilled when the re-blast against any of the primer taxa was successfull. Note that setting this flag will substantially decrease the stringency of the ortholog assignment with the consequence of an increased number of false positives.

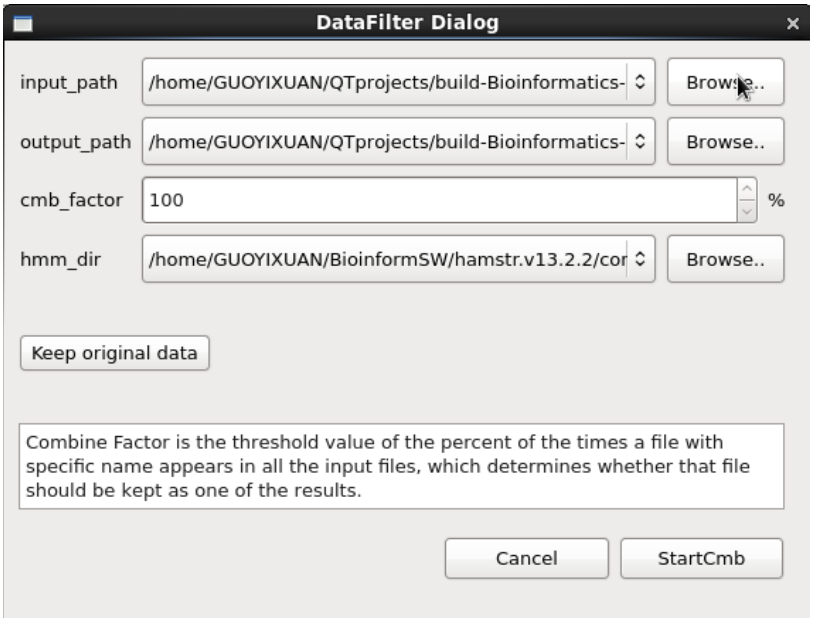
-strict:set this flag if the reciprocity criterion is only fulfilled when the re-blast against all primer taxa was successfull

representative

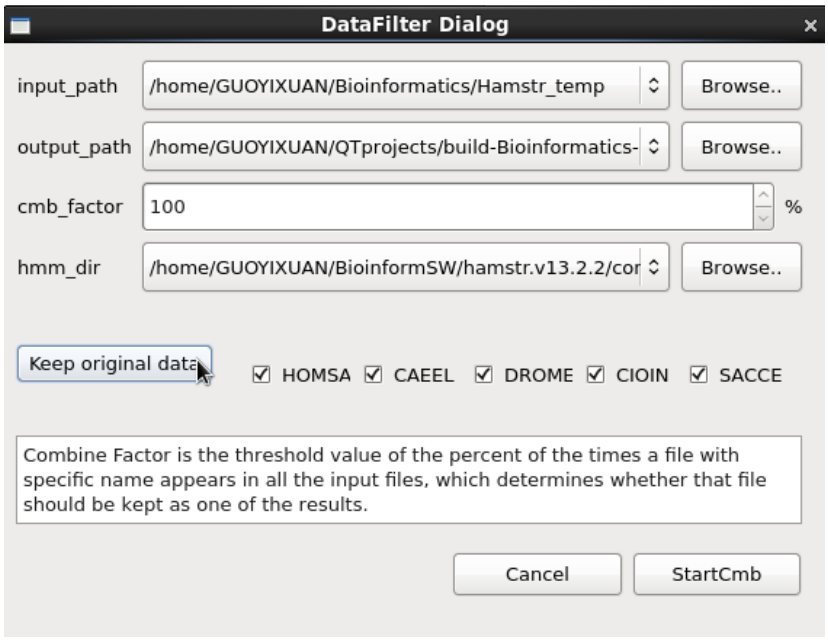
From all sequences that fulfill the reciprocity criterion the one showing the highest similarity to the core ortholog sequence in the reference species is identified and selected as representative.

3.2 Data Filter

DataFilter Dialog:



After choosing the input_path, the original data would show, you are allowed to keep some or all of the original data from the hmm_dir.



Data Filter is a way to filte the unnecessary output data after HaMStR. The parameters used in Data Filter are as follows.

Combine Factor

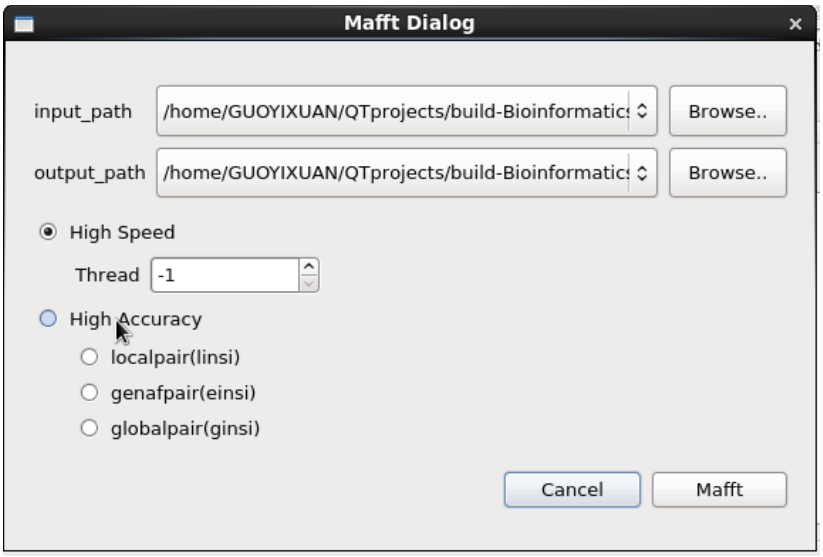
set this factor to determine the threshold value of the percent of times a file with specific name appears in all the input files, which decides whether that file should be kept as one of the results.

hmm_dir

you need to specify a specific hmm_dir to indicate the input files' hmm species.

3.3 Mafft

Mafft Dialog:



MAFFT offers various multiple alignment strategies. They are classified into three types, (**a**) the progressive method, (**b**) the iterative refinement method with the WSP score, and (**c**) the iterative refinment method using both the WSP and consistency scores. In general, there is a tradeoff between speed and accuracy. The order of speed is **a** > **b** > **c**, whereas the order of accuracy is **a** < **b** < **c**. Thus, there are two options for Mafft.

High speed:

thread: Number of threads (if unsure, --thread -1)

High accuracy (for <~200 sequences x <~2,000 aa/nt):

localpair(linsi)

genafpair(einsi)

globalpair(ginsi)

In order to obtain more accurate alignments in extremely difficult cases, three new options, L-INS-i, G-INS-i and E-INS-i, have been added to recent versions (v.≥5) of MAFFT. These options use a new objective function combining the WSP score (Gotoh) explained above and the COFFEE-like score (Notredame et al.), which evaluates the consistency between a multiple alignment and pairwise alignments (Katoh et al. 2005).

For pairwise alignment, three different types of algorithms are implemented, global alignment (Needleman-Wunsch), local alignment (Smith-Waterman) with affine gap costs (Gotoh) and local alignment with generalized affine gap costs (Altschul). The differences in the accuracy values among these methods are small for the currently available benchmarks, as shown here. However, each of them has different characteristics, according to the algorithm in the pairwise alignment stage:

E-INS-i (genafpair(einsi)) is suitable for alignments like this:

```
ooooooooXXX------XXXX-------------------------------XXXXXXXXXX-XXXXXXXXXXXXXXXoooooooooooo
---------XXXXXXXXXXXXooo-------------------------------XXXXXXXXXXXXXXXXX-XXXXXXXX-----------
-----oooXXXXXX---XXXXoooooooooo---------------------XXXXX----XXXXXXXXXXXXXXXXXoooooooooooo
---------XXXXX----XXXXoooooooooooooooooooooooooooooooXXXXX-XXXXXXXXXXX--XXXXXXX-----------
---------XXXXX----XXXX-------------------------------XXXXX---XXXXXXXXXX--XXXXXXXooooo--------
```

where 'X's indicate alignable residues, 'o's indicate unalignable residues and '-'s indicate gaps. Unalignable residues are left unaligned at the pairwise alignment

stage, because of the use of the generalized affine gap cost. Therefore E-INS-i is applicable to a difficult problem such as RNA polymerase, which has several conserved motifs embedded in long unalignable regions. As E-INS-i has the minimum assumption of the three methods, this is recommended if the nature of sequences to be aligned is not clear. Note that E-INS-i assumes that the arrangement of the conserved motifs is shared by all sequences.

L-INS-i (localpair(linsi)) is suitable to:

```
ooooooooooooooooooooooooooooooXXXXXXXXXXX-XXXXXXXXXXXXXXX------------------
--------------------------------XX-XXXXXXXXXXXXXXX-XXXXXXXXoooooooooo-------
------------------oooooooooooooXXXXX----XXXXXXXX---XXXXXXXoooooooooo-------
--------ooooooooooooooooooooooooXXXXX-XXXXXXXXX----XXXXXXXoooooooooooooooooo
--------------------------------XXXXXXXXXXXXXXX----XXXXXXX------------------
```

L-INS-i can align a set of sequences containing sequences flanking around one alignable domain. Flanking sequences are ignored in the pairwise alignment by the Smith-Waterman algorithm. Note that the input sequences are assumed to have only one alignable domain. In benchmark tests, the ref4 of BAliBASE corresponds to this. The other categories of BAliBASE also correspond to similar situations, because they have flanking sequences. L-INS-i also shows higher accuracy values for a part of SABmark and HOMSTRAD than G-INS-i, but we have not identified the reason for this.

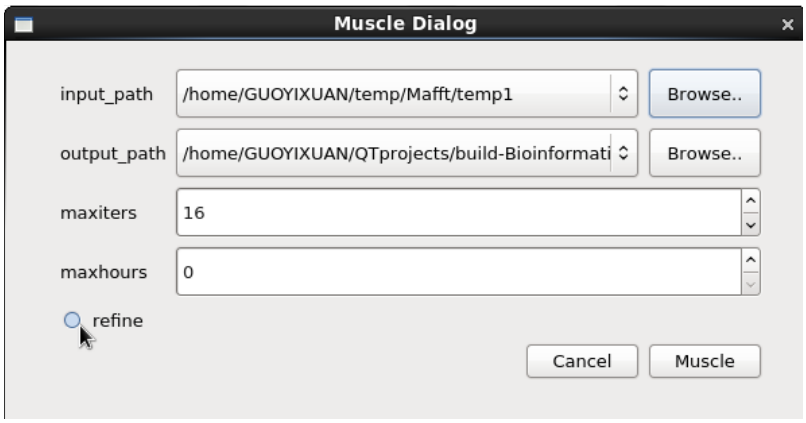
G-INS-i (globalpair(ginsi)) is suitable to:

```
XXXXXXXXXX-XXXXXXXXXXXXXX
XX-XXXXXXXXXXXXX-XXXXXXXX
XXXXX----XXXXXXX---XXXXXXX
XXXXX-XXXXXXXXX----XXXXXXX
XXXXXXXXXXXXXXX----XXXXXXX
```

G-INS-i assumes that entire region can be aligned and tries to align them globally using the Needleman-Wunsch algorithm; that is, a set of sequences of one domain must be extracted by truncating flanking sequences. In benchmark tests, SABmark and HOMSTRAD correspond to this.

3.4 Muscle

Muscle Dialog:



The parameters used in Muscle are as follows:

inputpath

path of the input files in FASTA format (default stdin)

outputpath

path of the output alignment in FASTA format (default stdout)

maxiters

aximum number of iterations (integer, default 16)

maxhours

maximum time to iterate in hours (default no limit)

refine

select this flag if you want to run Muscle on a prealigned matrix. This is very fast, average accuracy similar to T-Coffee

3.5 MatrixMasking

MatrixMasking Dialog:

This function cut the random similarity of matrix concerning the matrixmasking operation.

inputpath

path of the input files.

outputpath

path of the output files.

Options

You are free to set these two options or leave it alone.

-r if -r is used without an argument 4*N random pairs are compared, checking for replications (which are avoided). If -r is used with an argument, this number of randomly selected pairs is analysed and used to infer the consensus profile, if -r used with an argument which is beyond the maximal number of possible non-overlapping pairs, only the maximal number of pairs is compared. If the -r option and the -t option are not used, random pairs are compared as default, with 4*N selection of pairs.

-e for nt sequences disables N replacement for fuzzy ends of sequences.

3.6 Concatenate

Concatenate Dialog:

This function allows the user to concatenate the matrix and results in a supermatrix used for phylogenomic analysis, and in the meantime, generate a partition file contains the partition information of the supermatrix. (Note: the partition file contains no information about the best model of each partition, you need assistance from other software such as partitionFinder to get this information.)

inputpath

path of the input files.

matrix_name

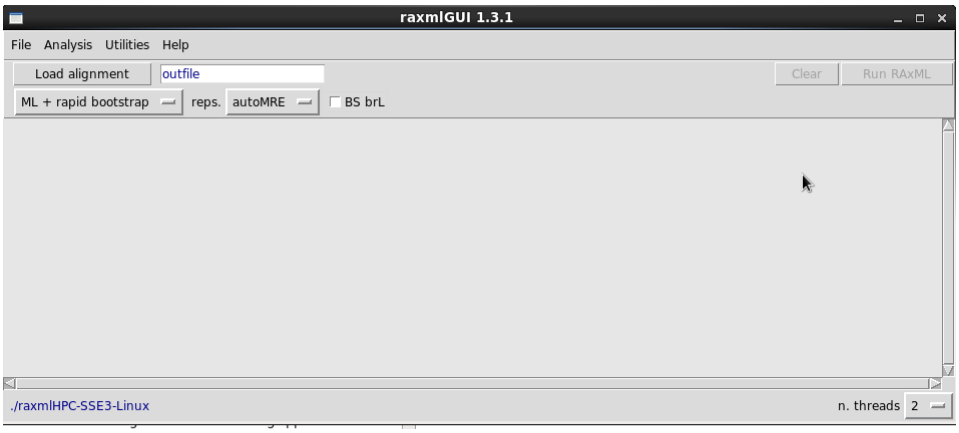
name of the matrix file.

charset_list

name of the partitionfile.

3.7 RaxMLGUI

RaxMLGUI dialog:



Three different ML analyses can be set up through RaxMLGUI:

(1) Maximum likelihood reconstruction using the rapid hill-climbing algorithm (RAxML option " -f d " ; " ML search " option in the GUI)

(2) Rapid bootstrap analysis and search for a best-scoring ML tree (R AxML option " -f

a ”).
(3) Thorough bootstrap analysis (RAxM L option “ -b ”), followed by a ML search. Subsequently, BS support values are drawn on the best-scoring ML tree (RAxML option “-f b”). The number of independent ML searches and BS replicates can be set using the program’s toolbar. As an alternative to the predefined number of BS replicates, different “bootstopping” options are available, which automatically stop the bootstrap r un after the necessary number of replicates. RaxmlGUI further incorporates RAxM L’s options to generate consensus trees , compute persite log-likelihoods with output compatible with the software CONSEL (Shimodaira 2001), and to compute R obinson-Foulds pairwise distances (Robinson and Foulds 1981) between trees.

I hope that you enjoy your journey with PhyloPipe!