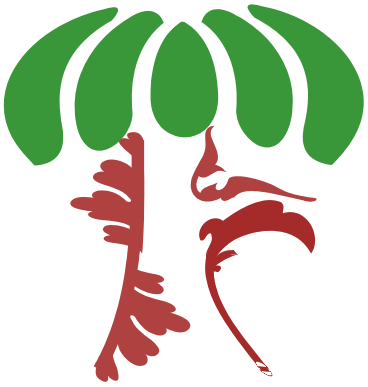
One-click Fungal Phylogenetic Tool

(OFPT)



Manual version 1.9

by **Xiang-Yu Zeng & Ting-Jun Tan**

2024/01/17

Supported OS: Windows 10 +

Contributors:

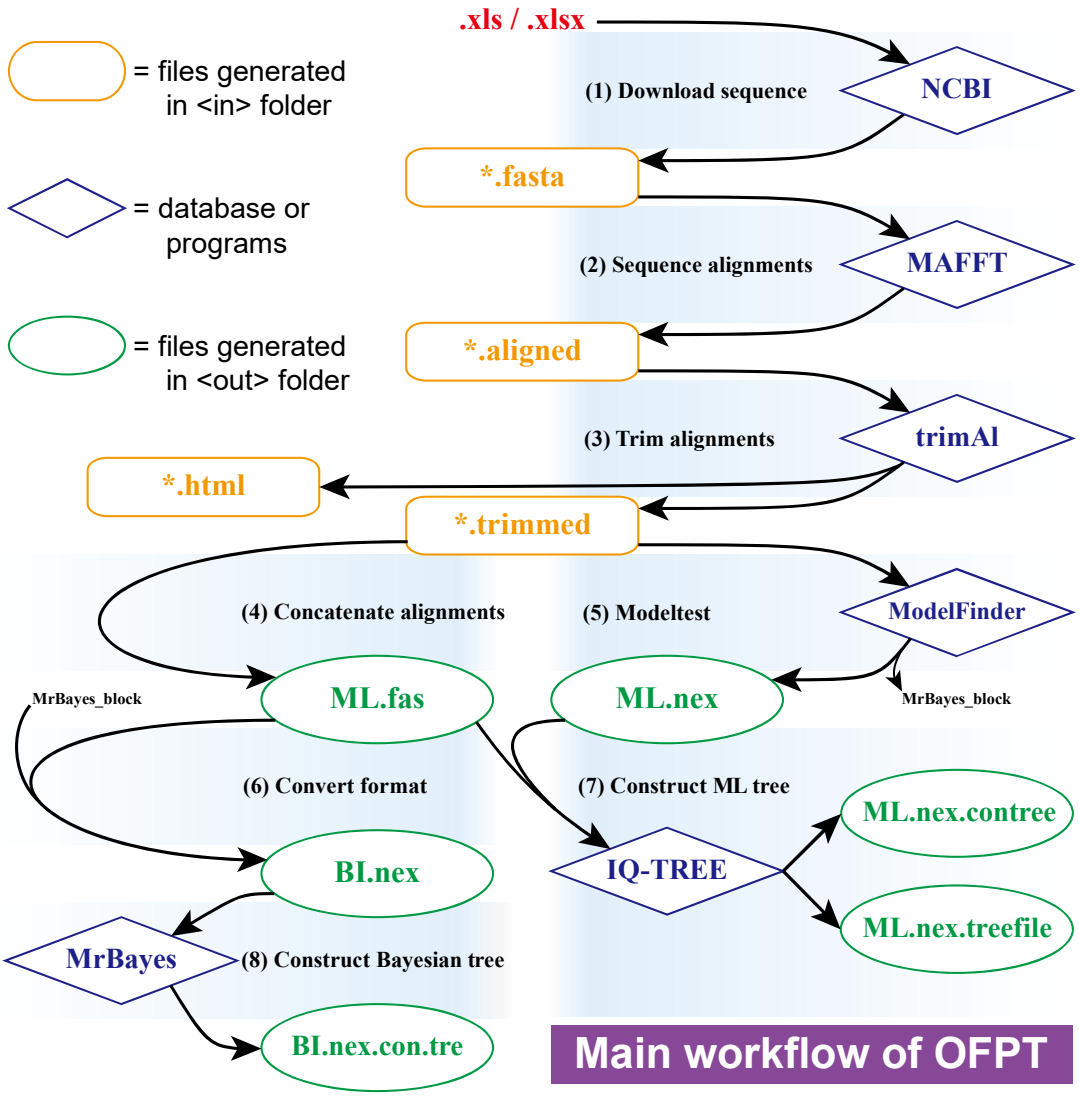
XC Peng, L Lu, D Gomdola, J Ma, HD Yang QF Meng, A Armand

Contents

1. Introduction

DNA-based phylogeny is indispensable in current fungal taxonomic studies, especially for groups with limited morphological characteristics. However, conducting phylogenetic analysis comprises a complicated and time-consuming procedure, including retrieving sequence data from GenBank, sequence alignment and trimming, selecting the best-fit evolutionary model, sequence concatenation, converting sequence format with specific commands where necessary, and tree editing This whole process is very time-consuming and each individual who does phylogenetic analyses only focuses on a limit number of groups, where parameters for relevant analyses are mostly the same. Each individual who does phylogenetic analysis only focuses on a limit number of fungal groups, where parameters for relevant analysis are mostly the same. Therefore, many researchers have been spending a lot of time performing a large number of repetitive operations, which greatly reduces the efficiency of scientific researches. On the other hand, such analytical operations require researchers to have certain background knowledge of phylogeny, and be familiar with relevant websites and softwares.

Based on the above drawbacks, a click-and-run software named One-click Fungal Phylogenetic Tool (OFPT) is developed herewith, specifically for those who will conduct fungal phylogenetic analyses. OFPT is a streamlined software for conducting fungal phylogenetic analyses, with built-in programs of MAFFT, trimAl, IQ-TREE and MrBayes. The software automates and standardizes the repetitive process of constructing phylogenetic trees, which will make relevant studies much easier and faster, and its resulting trees are identical to the manually generated ones. OFPT is simple and straightforward, offering users a step-by-step process, and its workflow is illustrated below.


.xls / .xlsx
= files generated in <in> folder
= database or programs
= files generated in <out> folder
(1) Download sequence
NCBI
*.fasta
(2) Sequence alignments
MAFFT
*.aligned
(3) Trim alignments
trimAl
*.html
*.trimmed
(4) Concatenate alignments
(5) Modeltest
ModelFinder
MrBayes_block
ML.fas
ML.nex
MrBayes_block
(6) Convert format
(7) Construct ML tree
ML.nex.contree
BI.nex
IQ-TREE
ML.nex.treefile
MrBayes
(8) Construct Bayesian tree
BI.nex.con.tre
Main workflow of OFPT

2. How to use

2.1 Download and run

To use the software, you will first need to go to the url: https://ofpt.guhongxin.com/, click 'Download the newest version' button in the center, and a file named 'OFPT_v*.rar' (the asterisk represents the version number) will be downloaded to your computer. Then, you will need to extract the file, using 'WinRAR' or other compressing software, to a path without space or other illegal characters. Finally, you can double click the file 'OFPT.exe' to run the software.

2.2 Using the software for the first time

When first used, the user will be asked to enter their email address,and determine the working directory. Afterwards, a file named 'Config.ini' will be generated to record the setting information. The working directory can be changed afterwards by pressing 's' in the main interface.

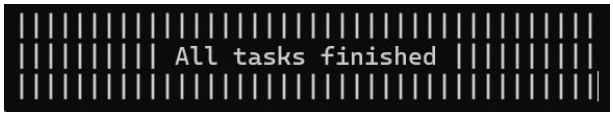
```
***** Config file does not exist, proceeding for settings *****


(Please enter your email address): xyzeng3@gzu.edu.cn

(Please enter the path of the working directory): D:\test
```

2.3 Main interface

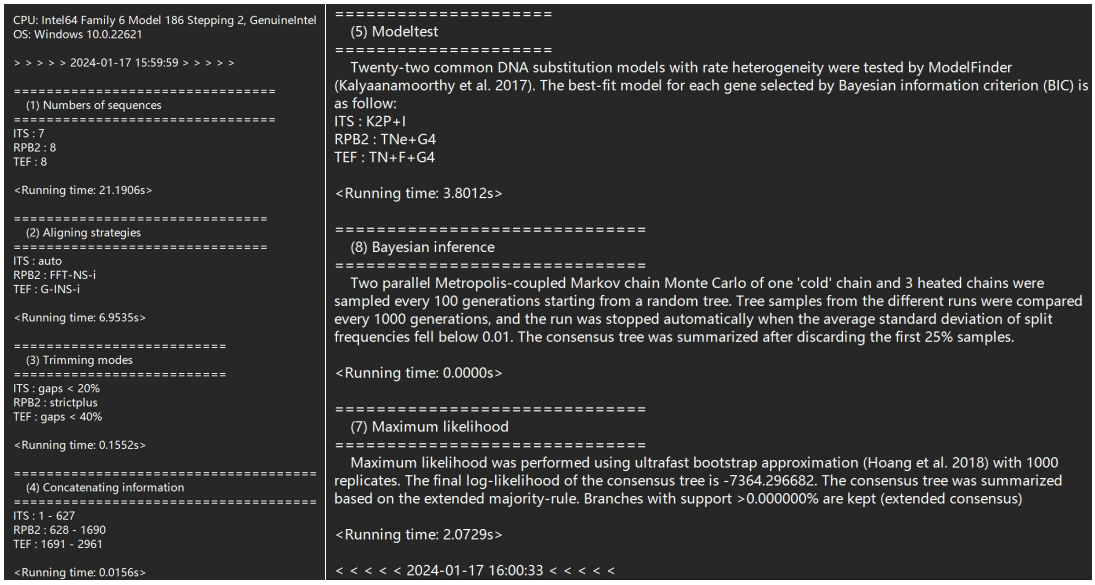
If the 'Config.ini' file exists, the main interface of the software will appear. Users can simply type '0' to get a ML tree with a standard procedure, or run the other 8 different modules individually or sequentially according to their demands. This can be either a single number, or a series of numbers, such as '1234' or '1234567'. After press 'ENTER', the software will run the selected modules in sequence without human interventions. If the program reruns with a different Excel file, please delete all generated folder in advance.

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* (H) elp                 /|||\   |||||  /|||\  ||||||            ____    *
* (S) ettings             ||  ||  ||||   ||||/    ||    \  /  ||  ||__||  *
*                         \|||/   ||     ||       ||     \/   || . ___||  *
* (0)          One-click mode - Run from step (1) to (7). Requires a single *
*                               Excel file in the working directory         *
* (1)       Download sequence - Requires a single '*.xls' or '*.xlsx' file  *
*                               in the working directory                    *
* (2)     Sequence alignments - Requires '*.fasta'   file(s) in <in>        *
* (3)         Trim alignments - Requires '*.aligned' file(s) in <in>        *
* (4)  Concatenate alignments - Requires '*.trimmed' file(s) in <in>        *
* (5)               Modeltest - Requires '*.trimmed' file(s) in <in>        *
* (6)          Convert format - Requires 'ML.fas' file in <out> and will    *
*                               read 'MrBayes_block' file in <temp>         *
* (7)       Construct ML tree - Requires 'ML.fas' & 'ML.nex' file in <out>  *
* (8) Construct Bayesian tree - Requires 'BI.nex' file in <out>             *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
===============================================================================
 <Current working directory>: D:\test
===============================================================================
Please select the function(s) you would like to proceed (e.g. 1234567 or 1234):
```

When all selected tasks are completed, the main interface will appear again with the message ‘All tasks finished’ showing above.



Information of the each run will be recorded in the ‘Information.txt’ file in the <out> folder once all tasks are finished.



2.4 Data preparation

Before starting to use the software, users only need to put a single Excel file under the working directory, with the 1st sheet providing names and accession numbers of the datasets and the 2nd sheet with your own sequence data in FASTA format. In the 1st sheet, species name and strain code of each entry are required in the first two columns, which will then be treated as the sequence ID by the program. The rest columns should include accession numbers of each strain, with their corresponding gene name in the header of each column. If you would like to add reference after each entry, the corresponding column should be named as ‘Reference(s)’. In the 2nd sheet, the header of each columns should be identical to the gene name included in the 1st sheet, otherwise your own data will not be included to the corresponding dataset. The FASTA format of users’ sequence data should be placed into two cells, where sequence ID followed by the cell including DNA sequences. The template of the input Excel file is as follow, as well as in the ‘example.xlsx’ file under the directory <examples>.

Species	Strain	Gene1	Gene2	Gene3
sp.1	st.1	ac. no. 1	ac. no. 2	
sp.2	st.2	ac. no. 3		ac. no. 4
...	...	...	...	...

Sheet1

Gene2	Gene3
>id1	>id2
seq1	seq2

Sheet2

2.5 Aligning strategies and trimming methods

The software offers the following parameter for aligning and trimming.

```
[Aligning strategies]
=============================================================================================
0.     auto - FFT-NS-1, FFT-NS-2, FFT-NS-i or L-INS-i depends on data size (default).
1. FFT-NS-i - The iterative refinement is repeated until no more improvement in the WSP
              score is made or the number of cycles reaches 1,000.
2.  E-INS-i - is suitable for alignments like this:
              oooooooXXX------XXXXoooooooo----------ooooXXXXXoooooooooooooo--oooooooo
              --------XXXXX----XXXXoooooooooooooooooooooXXXXX-ooooooooooooooooo------
              ooooooo-XXXXX----XXXX-----------------------XXXXX---ooooooo--oooooooooo---
              --------XXXXXX---XXXX-----------------------XXXXX-------------------------
              --------XXXXXXXXXXXX-----------------------XXXXX-------------------------
              --------XX-------XXXX-----------------------XXXXX-------------------------
3.  L-INS-i - is suitable for alignments like this:
              ooooooooooooooooooooooooooooooXXXXXXXXXX-XXXXXXXXXXXXXX-----------------
              --------------------------------XX-XXXXXXXXXXXXXX-XXXXXXXXooooooooooo------
              ----------------oooooooooooooXXXXX----XXXXXXXX---XXXXXXXooooooooooo------
              -------oooooooooooooooooooooooXXXXX-XXXXXXXXXX----XXXXXXXoooooooooooooooo
              --------------------------------XXXXXXXXXXXXXXX----XXXXXXX-----------------
4.  G-INS-i - is suitable for alignments like this:
              XXXXXXXXXX-XXXXXXXXXXXXXX
              XX-XXXXXXXXXXXXXX-XXXXXXXX
              XXXXX----XXXXXXXX---XXXXXXX
              XXXXX-XXXXXXXXX----XXXXXXX
              XXXXXXXXXXXXXXX----XXXXXXX
```

```
[Trimming method]
=============================================================================================
0.   gappyout - trim based on gaps' distribution (default)
1.      strict - trim based on the fraction of gaps in a column and their similarity scores
2. strictplus - more stringent than strict
3.    -gt 0.2 - allow 20% fraction of sequences to have gaps
4.    -gt 0.3 - allow 30% fraction of sequences to have gaps
5.    -gt 0.4 - allow 40% fraction of sequences to have gaps
6.    -gt 0.5 - allow 50% fraction of sequences to have gaps
7.    -gt 0.6 - allow 60% fraction of sequences to have gaps
8.    -gt 0.7 - allow 70% fraction of sequences to have gaps
9.    -gt 0.8 - allow 80% fraction of sequences to have gaps
```

By default, the software will apply a global strategy to all gene regions, with 'auto' for aligning and 'gappyout' for trimming. If users would like to perform different strategies for aligning and trimming to different genes, two digits representing aligning strategy and trimming method together at the end of the headers (e.g. ITS26, TEF145) are required. This also requires users to run the program from step 1, as only the first step reads the information of the Excel file.

Species	Strain	Gene126	Gene2	Gene345
sp.1	st.1	ac. no. 1	ac. no. 2	
sp.2	st.2	ac. no. 3		ac. no. 4
...	...	...	...	...

Sheet1

Gene2	Gene345
>id1	>id2
seq1	seq2

Sheet2

3. Features

OFPT is written and packaged by Python 3.9. It integrates eight modules corresponding to the process of doing phylogenetic analysis into a pipelined process, viz. sequence downloading, sequence alignment, alignment trimming, alignment concatenating, model selection, format conversion, construction of ML tree, and construction of Bayesian tree. These modules can be executed separately or in sequence with the above orders.

3.1 Download sequence

The program will first detect if only a single Excel file (.xls or .xlsx) is included in the working directory and then read the accession numbers from the 1st sheet of the excel file. The 2nd sheet will be detected simultaneously. Request of downloading will then be sent to NCBI (https://www.ncbi.nlm.nih.gov/), and sequences in FASTA format will be downloaded and renamed according to the content in the first two columns. When finished, the program will verify if the number of entries in the Excel sheet is identical to the number of sequences downloaded from GenBank. An error message will appear if some of the listed accession numbers were not downloaded successfully. In that case, the user must correct the mistakes manually, or use the function 'check' to allow the program find the problematic entries. Afterwards, sequences (FASTA format) in the 2nd sheet, if any, will be matched based on their gene names (the header of each column) and added to the corresponding dataset accordingly. In addition, the program will detect the last two characters of the header, which represents the gene name, and check if they are digits matching the option of aligning strategy and trimming mode. If so, the program will treat these two digits as a customized option for aligning and trimming regarding each genomic region. Finally, '*.fasta' files for each gene region will be generated separately in the folder <in>.

3.2 Sequence alignments

Files with the '*.fasta' extension in the folder <in> will first be detected and proceeded for alignment using MAFFT (Katoh & Standley 2013). The global aligning strategy of commonly used 'auto', 'FFT-NS-i', 'E-INS-i', 'L-INS-i' and 'G-INS-i', can be selected when setting the configurations. If two digits at the end of the gene name have been detected in the excel file, the aligning strategy will be set according to the first digit. Aligned sequence datasets will finally be written in '*.aligned' file(s) separately in the folder <in>.

3.3 Trim alignments

Files with the '*.aligned' extension in the folder <in> will first be detected and proceeded for the alignment trimming using trimAl (Capella-Gutiérrez et al. 2009). The global trimming mode of commonly used '-gappyout', '-strict', '-strictplus', '-gt 0.2–0.8' can be selected when setting. If two digits at the end of the gene name have been detected in the excel file, the trimming mode will be set according to the second digit. Trimmed datasets will finally be written in '*.trimmed' files separately in the

folder <in>, and '*.html' files demonstrating the trimmed regions will be generated in the folder <in>.

3.4 Concatenate alignments

Files with the extension of '*.trimmed' in the folder <in> will first be detected and proceeded for concatenating. Sequences with the same ID will be concatenated in the order of file names. Sequences of each missing ID will be filled with question marks according to the length of corresponding gene regions. The concatenated dataset will finally be written into a single 'ML.fas' file in the folder <out>. The 'ML.fas' file can be directly used as the input file of IQ-TREE and RAxML (Stamatakis 2014).

3.5 Modeltest

Files with the '*.trimmed' extension in the folder <in> will first be detected and proceeded for modeltest using ModelFinder (Kalyaanamoorthy et al. 2017). Then the program will delete and recreate the <temp> folder before running this module to avoid errors caused by previously generated files, especially when users changed the data in the Excel file and rerun the program with the same working directory. Five files will be generated for each gene dataset in the folder <temp> including a '*.trimmed.iqtree' file, which records the detailed results of the model test. The 'ML.nex' file will be generated in the folder <out> as one of the input file for IQ-TREE, which list the information of each gene partition and their corresponding model selected by the Bayesian information criterion (BIC). The command block for MrBayes will also be written in the 'MrBayes_block' file in the folder <temp>, including the commands for applying different models to different partitions. As some of the models are not supported in MrBayes, 'nst=Mixed' will be used when the best-fit model is not GTR/SYM, HKY/K80 or F81/JC.

3.6 Convert format

The 'ML.fas' file in the folder <out> will first be detected and converted from FASTA to NEXUS format, and add the content of the 'MrBayes_block' file in the folder <temp> to the end. The converted dataset will finally be written into a 'BI.nex' file in the folder <out>, which can be directly used as the input file for MrBayes. The 'BI.nex' file can be directly used as the input file of MrBayes.

3.7 Construct ML tree

The 'ML.fas' and 'ML.nex' file in the folder <out> will be detected and proceed for maximum likelihood (ML) tree construction using IQ-TREE (Nguyen et al. 2015), performing an ultra-fast bootstrap with 1000 replicates (Hoang et al. 2018). The program will read the partition and model information in the 'ML.nex' file to perform the analysis. The consensus ML tree with bootstrap supports will be generated in the folder <out> namely 'ML.nex.contree', and the ML tree with both bootstrap supports and SH-aLRT result will be generated in the folder <out> namely 'ML.nex.treefile' file.

3.8 Construct Bayesian tree

The ‘BI.nex’ file in the folder <out> will be detected and proceed for Bayesian inference (BI) using MrBayes (Ronquist et al. 2012). The default Markov chain Monte Carlo analysis will be performed with two parallel runs, each including four chains and 100 sampling frequencies with 1,000 diagnostic frequencies and 10,000 checkpoint frequencies. The runs will stop automatically when the standard deviation falls below 0.01 or the number of generation reaches 50,000,000 (can be changed in the MrBayes block manually). The first 25% sample fraction will be discarded when calculating convergence diagnostics. The consensus BI tree with posterior probabilities will be generated in the folder <out> namely ‘BI.nex.con.tre’. This function is not recommended to use on PC, as it takes a long time and will consume lots of CPU resources.

4. Notice

02. If the anti-virus software reports a virus, please add the software to the trust list or temporarily disable the anti-virus software.
03. Paths of both the software and the working directory MUST NOT include SPACE or any other illegal characters!!!
04. When an error occurs, the software will prompt an error message and quit the current process.
05. Current version only supports DNA data.
06. Only a single Excel file should be included in the working directory.
07. Content in the first two columns of the Excel file will be treated as sequence ID, information of accession numbers should be included in the other columns.
08. The missing accession number in the Excel file should be represented by an empty cell, a single hyphen, or a single en-dash.
09. Gene names (header of columns) in the Excel file should only include alphabets, numbers and underscores.
10. The column header representing the same gene in sheets 1 and 2 should always be identical.
11. The sequence ID of the same strain the users provided should be identical, in order to avoid errors when concatenating multi-genes.
12. Please ensure all sequences are forward before aligning, as the function of adjusting sequence direction are not available in the software.
13. Details of modeltest results can be found in '*.iqtree' files in <temp> folder.
14. Outgroups will not be assigned in the when constructing trees.
15. Running 'convert sequence format' alone will not generate the MrBayes command block, unless a file named 'MrBayes_block' already exists in <temp> folder.
16. If an unexpected error occurs, the software window may automatically shut down while running. In this case, users can run the software from the Windows command prompt to check the exact errors during execution. First, press 'win + R' and type 'cmd' to open the command prompt. Then, type 'cd <path of OFPT>' to enter the path of the software. Finally, type the file name 'OFPT.exe' to run the software.