Contents

1. About

PhyloToL is a phylogenomic pipeline composed for 4 major components: 1) Gene family assessment per taxon (adding taxa to the database), 2) Refinement of homologs and gene tree reconstruction, 3) Tree-based contamination removal and 4) building of supermatrix for species tree reconstruction. These components can be executed independently. PhyloToL is written primarily in Python 2.7 programming language but it also incorporates Perl, Ruby and Bash custom scripts. PhyloToL only runs through the command line (there is not GUI), therefore a minimum knowledge of UNIX is required.ttddsr PhyloToL can run in porwerful computers with multiple threads also in a normal computer in one thread.

2. Download and dependencies

Distribution: https://github.com/Katzlab/PhyloTOL
Dependencies:

1. Biopyton
2. Dendropy (https://dendropy.org/)
3. P4 (http://p4.nhm.ac.uk/)
4. bioperl
5. Mafft (any version; https://mafft.cbrc.jp/alignment/software/)
6. Usearch (any version; https://www.drive5.com/usearch/)
7. Guidance (v2.02; http://guidance.tau.ac.il/overview.html)
8. trimAl (v1.3; http://trimal.cgenomics.org/)
9. raxml

3. Databases

The databases should follow the next folder structure:

DataFiles/
    allOG5Files/
    ncbiFiles/
    BlastFiles/

The folder allOG5Files should contain the initial gene family dataset. A text file named with a unique code should represent each gene family. In our laboratory we used orthoMCL data for building this database. Then, we use codes like OG5_126595 (actin). The user can pick the name for these files but we recommend sticking with the prefix “OG5” or modifying scripts accordingly.

The folders ncbiFiles and BlastFiles represent the new taxa that are being added to the databases (see adding taxa section). While the folder ncbiFiles contains the actual sequences (e.g., transcriptome, genome or protein sequences), the folder BlastFiles contains the results of Blasting the sequences of the new taxa against the gene family database. There should be one file per taxon in both folders.

4. Taxa names

We use a 10 digits code for naming the taxa. This code is intended to represent the taxonomy. For instance, for the *Plasmodium falciparum*, the code is Sr_ap_Pfal. Here, the Sr_ represents the eukaryotic major clade SAR and Sr_ap_ represents the "minor" clade Apicomplexa. See annexed at the end of this document a list of minor clades that we recommend.

5. Running PhyloToL – first component (adding taxa)

In order to add new taxa to the database, the user can run the first of the four major component of phyloToL. This component would take High Throughput Sequencing data and conduct some steps such as Identify and remove sample bleeding in an illumina lane, removing prokaryotic and rDNA sequences and translating sequences using informed genetic codes. Finally, every sequence is classified into a gene family and this is represented in two files, a fasta file and a Blast report. The former one should be placed in the ncbiFiles folder and the last one should be placed in the BlastFiles folder.

Documentation for running this component of PhyoToL is available in the folder "AddTaxa" which came with this distribution.

6. Running PhyloToL – second and third components (Homology assessment, alignment and tree building and tree based contamination removal)

Quick start: Make sure you have this folders/files structure (Bold: input files)

PhyloToL/
- DataFiles/
  - BlastFiles/
  - allOG5Files/
  - ncbiFiles/
  - taxaDBpipeline3
  - **GFs_test**
  - **Taxa_test**
- Scripts/
  - PhyloToL scripts
  - pipeline_parameter_file.txt

GFs_test: In this file you put the list of GFs that you want to run through PhyloToL. For instance:

OG5_133844
OG5_133879
OG5_128106

Taxa_test: In this file you put the list of taxa that you want to run through PhyloToL. For instance:

EE_is_Tmar
EE_ka_Rtru
EE_ap_Asig
Ex_ma_Mjak
Am_ar_Enut
Am_ar_Mbal
Am_di_Acas

Set parameters in the parameters file, go to PhyloToL/Scripts/ and type “python phylotol.py”

Running with contamination removal: Make sure you have this folders/files structure (Bold: input files)

PhyloToL/
- DataFiles/
  - BlastFiles/
  - allOG5Files/
  - ncbiFiles/
  - taxaDBpipeline3
  - **GFs_test**
  - **Taxa_test**
  - **rules**
- Scripts/
  - PhyloToL scripts
  - pipeline_parameter_file.txt

Rules: Set of rules for contamination removal that PhyloToL will use for categorizing a sequence as contamination or not. This rules are set by the user by either manual inspection of a sample of trees, literature or any other method. A rule can be expressed as:

Sr_rh_Lvor Op_me Pl_

This will tell PhyloToL to consider as contamination a case in which the taxon Sr_rh_Lvor is nested among either Op_me or Pl_.

Once the rules are set runPhyloToL typing …

python phylotol-concleaner.py ../DataFiles/rules

Running PhyloToL partially: Depending on what is the type of study, you might want to run PhyloToL just for collecting candidate gene families (For instance, if you want to apply a different tool to test for homology) or for collecting homologs but not trees (for instance, if you want to try another tree inference tool). The use can run PhyloToL in two different modes: “ng” and “nr”, respectively.

python phylotol.py ng (Runs phylotol until Guidance)
python phylotol.py nr (Runs phylotol until raxml)

Running PhyloToL partially and restart:

The user can run PhyloToL up to guidance (mode ng); the user can resume the run and either produce post-guidance files and trees or only post-guidance files (with option "nr" - no raxml)

python phylotol-resumer.py path_to_working_directory
python phylotol-resumer.py path_to_working_directory nr (no tree)

If the user produced post-guidance files but not tree, then the user can resume the run and produce trees in this way

python phylotol-resumer.py path_to_working_directory

In order to run partially and re-start, PhyloToL requires a working directory. If the user ran phylotol up to guidance and wants to resume, the working directory will be a folder (e.g., out) with all pre-guidance files inside. The folders/files structure will be like … (bold: working directory)

PhyloToL/
- DataFiles/
  - BlastFiles/
  - allOG5Files/
  - ncbiFiles/
  - taxaDBpipeline3
  - GFs_test
  - Taxa_test
- **out/**
  - **pre-guidance files**
- Scripts/
  - PhyloToL scripts
  - pipeline_parameter_file.txt

If the user ran phylotol up to raxml and wants to resume (to produce trees). Then, the user has to keep a structure like this… (bold: working directory)

PhyloToL/
- DataFiles/
  - BlastFiles/
  - allOG5Files/
  - ncbiFiles/
  - taxaDBpipeline3
  - GFs_test
  - Taxa_test
- **out/**
  - **out_resume**/

GFs_test_ results2keep/
Pre-guidance files
Post-guidance files

Scripts/
PhyloToL scripts
pipeline_parameter_file.txt

7. Running PhyloToL – fourth component (supermatrix)

The user run this component for choosing orthologous sequences and produce alignments for concatenation. In order to do this, the user should set the option "concatAlignment = y" in the parameters file. Then run PhyloToL as shown in "Quick start". Once PhyloToL has finished type …

python concatenateFastas.py path_to_alignments_for_concatenation

This will produce a supermatrix that can be used for species tree building

8. Annexed. Recommended minor clades:

Amoebozoa,Am_ar,Archamoebae
Amoebozoa,Am_di,Discosea
Amoebozoa,Am_my,Mycetozoa
Amoebozoa,Am_hi,Himatismenida
Amoebozoa,Am_is,incertaesedis
Amoebozoa,Am_th,Thecamoebida
Amoebozoa,Am_tu,Tubulinea
Amoebozoa,Am_va,Vannellidae
Orphans (Enything else),EE_ap,Apusozoa
Orphans (Enything else),EE_br,Breviatea
Orphans (Enything else),EE_cr,Cryptophyta
Orphans (Enything else),EE_ha,Haptophyceae
Orphans (Enything else),EE_is,incertaesedis
Orphans (Enything else),EE_ka,Katablepharidophyta
Excavata,Ex_eu,Euglenozoa
Excavata,Ex_fo,Fornicata
Excavata,Ex_he,Heterolobosea
Excavata,Ex_is,incertae sedis
Excavata,Ex_ja,Jakobida
Excavata,Ex_ma,Malawimonadidae
Excavata,Ex_ox,Oxymonadida
Excavata,Ex_pa,Parabasalia
Opisthokonta,Op_ch,Choanoflagellida
Opisthokonta,Op_fu,Fungi
Opisthokonta,Op_ic,Ichthyosporea
Opisthokonta,Op_is,incertae sedis
Opisthokonta,Op_me,Metazoa
Opisthokonta,Op_nu,Nucleariidae and Fonticula group
Plantae,Pl_gl,Glaucophytes
Plantae,Pl_gr,Green algae

Plantae,Pl_rh,Red algae
SAR,Sr_ap,Apicomplexa
SAR,Sr_ch,Chromerida
SAR,Sr_ci,Ciliates
SAR,Sr_di,Dinoflagellates
SAR,Sr_is,incertae sedis
SAR,Sr_pe,Perkinsea
SAR,Sr_rh,Rhizaria
SAR,Sr_st,Stramenopiles