EasyCGTree

An Easy Tool for Constructing Core-Gene Tree

Version 4.2

Dao-Feng Zhang

Updated on January 22, 2024

Any suggestions or questions, please contact: *zdf1987@163.com*

Contents

1. What is EasyCGTree? 1
2. How to install EasyCGTree? 1
3. Input data 2
   3.1 Genomic data (FASTA format) 2
   3.2 profile HMM 2
4. Run EasyCGTree 3
   4.1 Preparations 3
   4.2 Run EasyCGTree 3
   4.3 Options 4
5. Output files 6
6. Usefull scripts 7
   6.1 BuildHMM.pl 7
   6.2 Gene_Prevelence.pl 8
7. Hardware Requirement 8
8. FAQ 9
9. Citation 9
10. References 9

1. What is EasyCGTree?

EasyCGTree is a Perl script, developed to construct **genome-based Maximum-Likehood (ML) phylogenetic tree,** by taking microbial genomic data (Amino acid and/or DNA sequences) in FASTA format as input data. Profile hidden Markov models (HMMs) of core gene sets prepared in advance and enclosed in the package are used for homologs search by HMMER (http://hmmer.org/) [1], and customized gene sets (prepared as gene clusters, which means homologous gene sequences in one file) can also be used to build profile HMMs by EasyCGTree for homologs search. EasyCGTree includes two approaches to infer phylogeny: **supermatrix** and **supertree** [2]. It has integrated all the steps required between the input genomic data and the resulted tree file into one Perl script, which would make it easier to infer a core-gene tree. Furthermore, the intermediate data of an EasyCGTree run can be directly used as input data of many other applications.

2. How to install EasyCGTree?

After the downloaded package (https://github.com/zdf1987/EasyCGTree4; https://gitee.com/zdf1987/EasyCGTree4/releases) is decompressed and renamed (*e.g. '.../Downloads/EasyCGTree' for Linux; '/' should be '\' for Windows*), no installation is required for **EasyCGTree**, beside installing Perl language environment. Other extra programs invoked by **EasyCGTree** had been included within the package. The following is a list of the extra programs:

1) Perl (>v5.0): https://www.perl.org/

2*) HMMER[1] (v3.0): http://hmmer.org/

3*) FastTree[3] (>v2.0): http://www.microbesonline.org/fasttree/#Usage

4*) muscle[4] (v5): http://www.drive5.com/muscle/

5*) consense[5,6] (v3.698): https://evolution.genetics.washington.edu/phylip.html

6*) IQ-TREE[7] (>v2.0): http://www.iqtree.org/

7*) trimAl[8] (v1.2): http://trimal.cgenomics.org/trimal?do=backlink

8*) astral-weighted[9] (v1.15.23): https://github.com/chaoszhang/ASTER/

9*) prodigal (v2.6.3): https://github.com/hyattpd/Prodigal/wiki/introduction

*: a version of these software had been included in the **EasyCGTree** package. For Linux version.

Users may need to use "chmod +x filename" within folder 'bin' to make them executable.

Build-in Profile HMM need to be downloaded separately, and then decompress them and put .hmm files into 'HMM' folder.

Experienced users can replace the included tools with precompiled up-to-date versions easily to update EasyCGTree, and we will update these tools and the main scripts aperiodically to ensure the

longevity. We will also try to develop a version on MAC OS that is as portable as those on Windows and Linux.

EasyCGTree *DO NOT* support MAC OS currently.

3. Input data

Genomic data (**Protein sequences of CDS and/or genomic DNA sequences**) in FASTA format are enough to run **EasyCGTree**, unless users want to use customized gene set for homologs search.

3.1 Genomic data (FASTA format)

The FASTA files are required to be uncompressed and with a specialized name (unique, no spacer, no "_", and ending with ".fas"). ***Don't worry about that. The file names will be formatted automatically by EasyCGTree.*** The user just needs to gather the proteome of each genome into a folder.

3.2 profile HMM

The set of genes used to build the profile HMM determines that what genes will be extracted from the proteomes to infer a core-gene tree. How to define the core genes of the data set of interest? The best way is to define the pan-genome first. Nevertheless, if a researcher just wants a core-gene tree, pan-genome analysis is not the optimal way. Maybe, retrieving a core gene set from previous studies or public databases is a better choice.

EasyCGTree includes profile HMMs prepared in advance of several core gene sets ubiquitously present in Prokaryote or some taxa, which is referred to as profile HMM database (PHD). At present, PHD is available from https://github.com/zdf1987/EasyCGTree4 as HMM.zip (Decompress it and put .hmm files into 'HMM' folder), and includes:

a. bac120, 120 ubiquitous genes in the domain *Bacteria*[10].

b. ar122, 122 ubiquitous genes in the domain *Archaea*[10].

c. rp1, 16 ubiquitous ribosomal protein genes in *Prokaryote*[11].

d. rp2, 23 ubiquitous ribosomal protein genes in *Prokaryote*[12].

e. ery288, 288 core genes of the family *Erythrobacteraceae*[13].

f. rhodo268, 268 cores of the family *Rhodobacteraceae*[14].

PHD is extensible and can be customized. Gene sets (prepared as gene clusters) supplied by users can be used to build profile HMMs by EasyCGTree (see Section 6.1).

4. Run EasyCGTree

4.1 Preparations

Genomic data sets

Gather the genomic data sets of FASTA format into a folder, name them in English style. *Copy the folder into the home directory of EasyCGTree (e.g. '...Downloads/EasyCGTree' for Linux; '/' should be '\' for Windows).* The genomic data sets could be sequences of either proteins of CDS (i.e. proteome) or genomic DNA. Two types of sequences are also OK. EasyCGTree will judge the type of each genomic data set and **predict CDS for those with DNA sequences by using program prodigal.**

Name the folder whatever you like, and ensure each data set in FASTA format.

**Names excluding extension of each data set will be the labels present on the tree tips.

Select a profile HMM

User can select one in the PHD according to their input data (Section 3.2).

Otherwise, prepare gene clusters (homologous gene sequences in one file) of interest, and use the script "BuildHMM.pl" to build a customized profile HMM (Section 6.1). Then, a new profile HMM was added into PHD.

4.2 Run EasyCGTree

Note: EasyCGTree is a command line software, and can be run in **cmd.exe** (Windows) or **Terminal** (Linux).

With the input data prepared correctly, it is very easy to run **EasyCGTree**. **Ensure that the folder of proteomes (e.g. myGenome) is in the home directory** of **EasyCGTree** (e.g. '…/Downloads/EasyCGTree' *for Linux; '/' should be '\' for Windows*), change the working directory to that of **EasyCGTree** and run.

(i) Assume that a Windows user download the software into a folder "Downloads" under "D:" drive. There is an example of command lines for running (press "Enter" to end each line):

```
>d:
> cd .\Downloads\EasyCGTree
> perl EasyCGTree.pl -input myGenomes
```

Note: Windows uses '\' to separate hierarchy, while Linux uses '/'. However, '/' is usually acceptable by Windows for most of the time, via transferring it to '\' automatically. We only use '\' to separate hierarchy in the part below of this manual for convenience.

(ii) Linux users can run EasyCGTree in the same way after change the working directory to that of **EasyCGTree** using the command 'cd'.

Then, you will get a tree (named after the name of the folder containing the genomes, e.g. myGenome.bac120.119.supermatrix.fasttree.tree) in Newick format; and many files/folders generated during running the script.

4.3 Options (do not need to be ordered)

-input <String> (The sole essential option)

Input directory including all genomic data sets

-task <String, 'all', 'hmmsearch', 'refine', 'alignment', 'tree_infer'>

Set run mode (Fig. 1). [default: all]

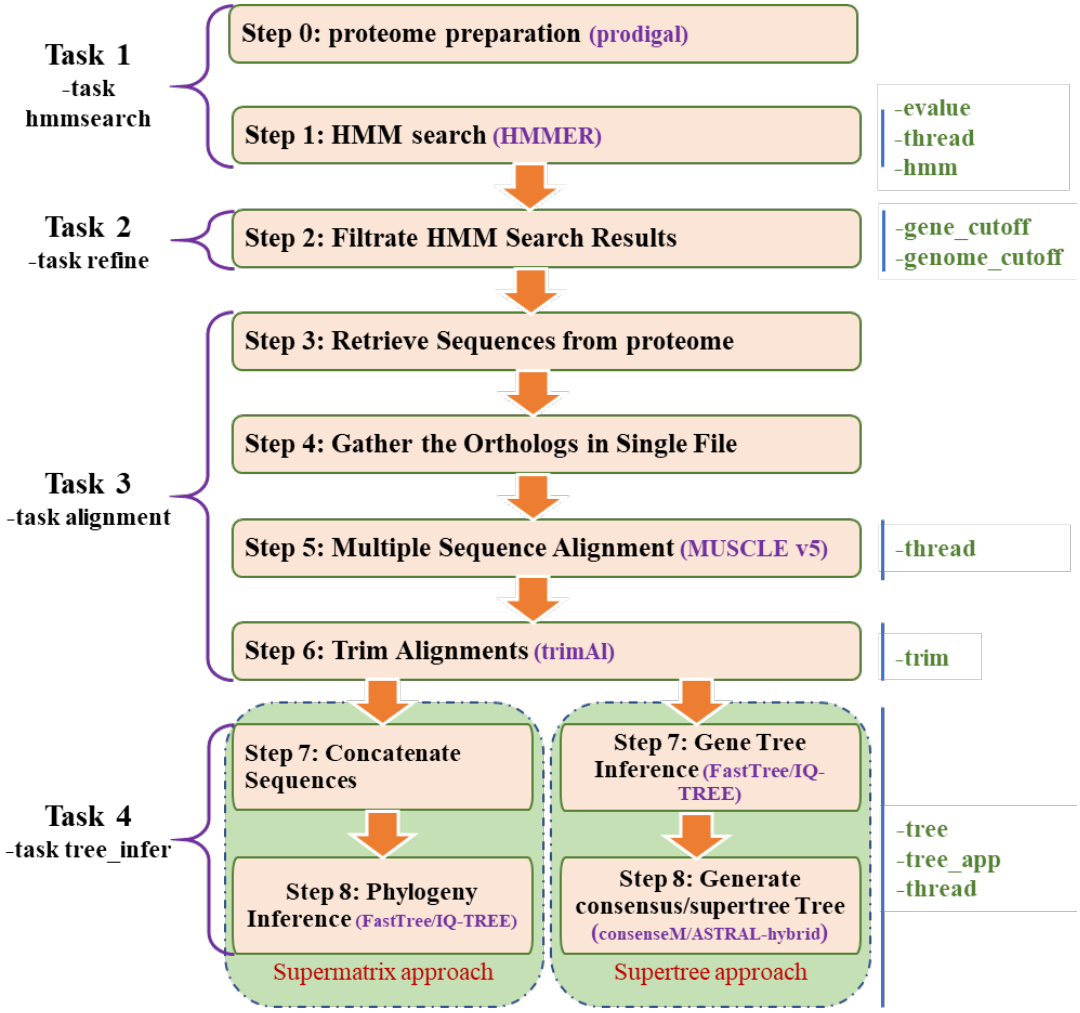

Fig. 1 Program flowchart of EasyCGTree for a complete run

-tree <String, 'sm', 'st', 'cs'>

Approach used for tree inference. [default: sm]

sm, supermatrices; st, supertree; cs, consensus tree

-tree_app <String, 'fasttree', 'iqtree'>

Applications used for tree inference. [default: fasttree]

-hmm <String, 'bac120', 'rp1', et al.>

Profile HMM used for gene calling. [default: bac120]

Available HMM could be found in '../EasyCGTree/bin/'.

-thread <Int>

Number of threads to be used by hmmer, clustalo (Linux) and iqtree. [default: 2]

-trim <String, 'gappyout','strict', 'strictplus'>

Standard for trimming alignment used by trimAl. [default: strict]

-evalue <Real, 10..0>

Expect value for saving hits during HMM search. [default: 1e-10]

-gene_cutoff <Decimal, 0.5..1>

Cutoff for omitting low-prevalence gene. [default: 0.8]

-genome_cutoff <Decimal, 0.5..1>

Cutoff for omitting low-quality genomes. [default: 0.8]

-help

Display help message.

Options of FastTree or IQ-TREE can be changed in 'bin/tree_app-options.txt'.

#More comments on '-genome_cutoff' and '-gene_cutoff'

The two parameters may be confusing to some users. In most situations, leave them alone and the default settings will be OK for a run involving limited number of proteome/genomes (<30), especially when the data are all from complete genomes. However, if much more proteomes are used, it is inevitable that some genes are not detected in some proteomes, and the number of common genes decrease. In fact, many genomes collected in the representative database of GTDB (Genome Taxonomy Database, https://gtdb.ecogenomic.org) contain an incomplete bac120/ar122 gene set. It is competing between common gene volume and genome data size.

These two parameters are expected to compromise this contradiction by excluding low-prevalence genes and low-quality genomes. Users should be informed that these two parameters determine the floor level on the quality of input genomes (the selected gene set is determined by the selected genomes). Lower cutoffs make no sense if the genomes are all of high quality. For example: users will always get 120/122 genes to infer a tree if all the genomes contain a complete bac120/ar122 gene set, no matter what is specified (0.5-1).

5. Output files

Assume the input proteome folder is 'myGenome'.

1) Phylogenetic trees of Newick format. Users could display it by using FigTree, MEGA, iTOL or other tree viewers.

Final tree: the tree file name looks like this.

myGenome.bac120.119.supermatrix.fasttree.tree

(1) (2) (3) (4) (5)

(1) indicates name of the folder that includes input proteome (-proteome); (2) gene set in './HMM' used for tree inference (-hmm); (3) how many genes of the gene set are used for tree inference; (4) approach used for tree inference (-tree: cs, sm, or st); (5) tool used for tree inference (-tree_app; iqtree or fasttree).

2) A log file (myGenome…._2021-8-2-13-21.log) **named after the genome folder, HMM, tree-making approach, tree-inference tool and the starting time.** It records the detailed information about the run, such as: how many/which genes was included or excluded; how many/which genomes was included or excluded.

3) A folder (myGenome_TEM) **named after the genome folder and ending with "_TEM".** Subfolders' names in this folder are universal to all the runs of EasyCGTree. It contains the following folders and files.

a. TEM0_Proteome: proteomes either copied from input directory "myGenome" or outputted via program prodigal.

b. TEM1_HMMsearch_out: HMM search results against each proteome.

c. TEM2_HMMsearch_outS:. filtrated HMM search results of 'TEM1_HMMsearch_out'. If two or more hits were identified, the best result for each gene will be selected based on E-value. Results below the E-value cutoff (default 1e-10) will be discarded.

d. TEM3_GeneSeqs: gene sequences retrieved from each proteome based on the results of 'TEM2_HMMsearch_outS'.

e. TEM4_GeneCluster: containing files gathering homologs from different genomes. One gene (cluster) was gathered in a single file. **Files of each gene cluster could be directly used to build a profile HMM (see Section 6.1)**.

f. TEM5_Alignment: FASTA-format files of alignments created from data in 'TEM4_GeneCluster'.

g. TEM6_AlignTrimmed: FASTA-format files of trimmed alignments created from data in 'TEM5_Alignment' with trimAI.

h. TEM7_geneTree: tree files (and related information generated by IQ-TREE) constructed from the aligned gene clusters in TEM6_AlignTrimmed.

i. FASTA-format concatenated sequences (myGenome……concatenation.fas) **of selected genes, which is used to infer the ML tree.** This file is named after the genome folder, HMM, gene number, tree-making approach, and tree-inference tool with addition of ".concatenation.fas". This file is generated from the folder

'TEM6_AlnTrimmed' (see below). This file is also accessible to other phylogeny-inferring software, e.g. FastME and RaxML.

j. GenomeGeneScreened.txt: gene and proteome list after filtration.

k. HMMinfo.txt: gene families' list and the length of each domain.

l. intree: includes the trees constructed from each selected gene. It was generated from the tree files in 'myGenome/TEM7_geneTrees' (see below). The taxa names are not identical to those users provided, and the correspondence can be found in 'myGenome_TEM/intreeInfo.txt'.

m. intreeInfo.txt: correspondence relationship of taxa names between those users give and those used by program consense and ASTRAL-Hybrid.

n. other files generated by IQ-TREE: IQ-TREE will write several files during a run, including '.bionj', '.cpk.gz', '.contree', '.iqtree', '.log', '.mldist', '.model.gz', '.splits.nex', and '.treefile'**.** Content in '.contree' is identical to that in 'myGenome.supermatrix.iqtree.tree', if supermatrix approach was used. These files will be written in home directory of EasyCGTree or the folder '../TEM7_geneTree'.

o. outtree/outfile: generated by consense program in phylip software package (set '-tree st'). It records the details to generate '**myGenome.XXX.consensus.XXX.tree**' from' **intree**' by consense program. The taxa names are not identical to those users provided, and the correspondence can be found in 'myGenome_TEM/intreeInfo.txt'.

p. trimAl_Details.txt: A file records the alignment lengths of each gene before and after trimming by program trimAl.

6. Usefull scripts

6.1 BuildHMM.pl

Used to build profile HMM from customized gene clusters (protein sequences). The profile HMM will be written into the PHD (directory './HMM') with a label the same to the folder including customized gene clusters.

Usage:

```
> perl BuildHMM.pl -gc Dir_name [options]
```

Options:

-gc <String> (Essential)

Input directory of gene clusters (gc) in fasta format.

-aln (Optional)

Gene clusters in the input directory are regarded as Alignment. BuildHMM.pl will not align them.

-thread <Int> (Optional)

Number of threads to be used by clustalo (Linux). [default: 8]

-help (Optional)

Display this message.

6.2 Gene_Prevelence.pl

Used for statistics of the prevalence of genes in a gene set among genomes/proteome, based on the homologues searching against related HMM in local PHD. The homologues searching will be performed using EasyCGTree.pl if related '../TEM1_HMMsearch_out' was not detected.

Usage:

```
> perl Gene_Prevelence.pl Dir_name
```

Options:

-proteome <String> (Essential)

Input data (proteomes) directory.

-hmm <String, 'bac120', 'rp1', et al.> (Optional)

Profile HMM (in folder 'HMM') used for gene calling. [default: bac120]

It will be ignored when a '~_TEM/TEM1_HMMsearch_out' folder is present.

-thread <Int> (Optional)

Number of threads to be used by hmmsearch. [default: 4]

It will be ignored when a '~_TEM/TEM1_HMMsearch_out' folder is present.

-evalue <Real, 10..0> (Optional)

Expect value for screening hmmsearch hits. [default: 1e-10]

-help (Optional)

Display this message.

7. Hardware Requirement

A normal PC is good enough to run EasyCGTree with limited input data, because clustalo, FastTree, and muscle are fast and approachable. However, the speed depends on the size of the input data. When proteomes <100, a PC will finish the analysis within several hours based on bac120. If bigger size input data was used, the version of Linux OS and powerful PC/server are recommended, because FastTree and clustalo under Linux support multi-threads (muscle and FastTree under Windows only support single thread). Please keep in mind that IQ-TREE will take much more time and larger memory than FastTree. If users fail with IQ-TREE (-tree_app iqtree) on a PC, try default settings with FastTree, or use more powerful PC or server.

8. FAQ

1. Why not support Mac OS?

A: Taking “easy to install” and “easy to use” as the philosophy, EasyCGTree is arbitrary in selecting third-party applications that can be used directly. Unfortunately, we failed to find valid applications to implement its function under Mac OS.

9. Citation

Zhang Dao-Feng, He Wei, Shao Zongze, Ahmed Iftikhar, Zhang Yuqin, Li Wen-Jun & Zhao Zhe. EasyCGTree: a pipeline for prokaryotic phylogenomic analysis based on core gene sets. *BMC Bioinformatics* 2023 24:390. https://doi.org/10.1186/s12859-023-05527-2

10. References

1. Jaina M, Robert DF, Sean RE, Alex B, Marco P, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 2013 41(12):e121.
2. Snel B, Huynen MA, Dutilh BE. Genome trees and the nature of genome evolution. *Annual Rev Microbiol* 2005 59:191-209.
3. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009 26: 1641-1650.
4. Edgar, RC (2021), MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping, bioRxiv 2021.06.20.449169. https://doi.org/10.1101/2021.06.20.449169.
5. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.7a. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. 2009.
6. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989 5: 164-166.
7. Bui QM, Heiko AS, Olga Cet al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era, *Mol Biol Evol* 2020 37(5):1530–1534.
8. Salvador CG, José MSM, Toni G. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, *Bioinformatics* 2009 25(15): 1972–1973.
9. Zhang C, Mirarab S. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. *Mol Biol Evol* 2022 39(12): msac215.
10. Parks DH et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiol* 2017 2, 1533–1542.
11. Brown CT et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 2015 523, 208–211.
12. Rinke, C et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 2013 499, 431–437.
13. Xu L et al. Genomic-based taxonomic classification of the family *Erythrobacteraceae*. *Int J Syst Evol Microbiol* 2020 70:4470–4495.
14. Zhang DF et al. Phylotaxonomic assessment based on four core gene sets and proposal of a genus definition among the families *Paracoccaceae* and *Roseobacteraceae*. Int J Syst Evol Microbiol. 2023 73(11): 006156.