EasyCGTree

An Easy Tool for Constructing Core-Gene Tree

Version 4

Dao-Feng Zhang

Update on August 18rd，2022

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

Contents

1. What is EasyCGTree?

EasyCGTree is a Perl script, developed to construct **genome-based Maximum-Likehood (ML) phylogenetic tree,** by taking microbial genomic data (proteome) 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/EasyCGTree) is decompressed, 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: https://www.perl.org/

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

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

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

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

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

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

8*) Clustal Omega[9]: http://www.clustal.org/ (Linux only)

*: 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.***

EasyCGTree *DO NOT* support MAC OS currently.

3. Input data

Proteome data (Protein sequences of genomes) in fasta format are enough to run **EasyCGTree**, unless users want to use customized gene set for homologs search.

3.1 Proteome 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 in 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 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. spi269, 269 core genes of the family *Spirosomaceae* [14].

g. rhodo268, 268 cores of the family *Rhodobacteraceae*.

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

Proteome data

Gather the proteomes of fasta format into a folder, name them in English style.

Copy the folder into the home directory of EasyCGTree (e.g. D:/EasyCGTree).

Name the folder whatever you like, and ensure the data in fasta format.

**Names excluding extension of each proteome 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. D:/EasyCGTree), change the working directory to that of **EasyCGTree** and type (press "Enter" when finishing a line):

```
> d:
> cd ./EasyCGTree
> perl EasyCGTree.pl -proteome myGenomes
```

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

4.3 Options (do not need to be ordered)

-proteome <String> (The only essential option)

Input proteome data directory

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

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

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

Approach (sm, supermatrices; st, supertree) used for tree inference. [default: sm]

-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]

-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.

Settings for FastTree or IQ-TREE can be made in 'tree_app-options.txt'.

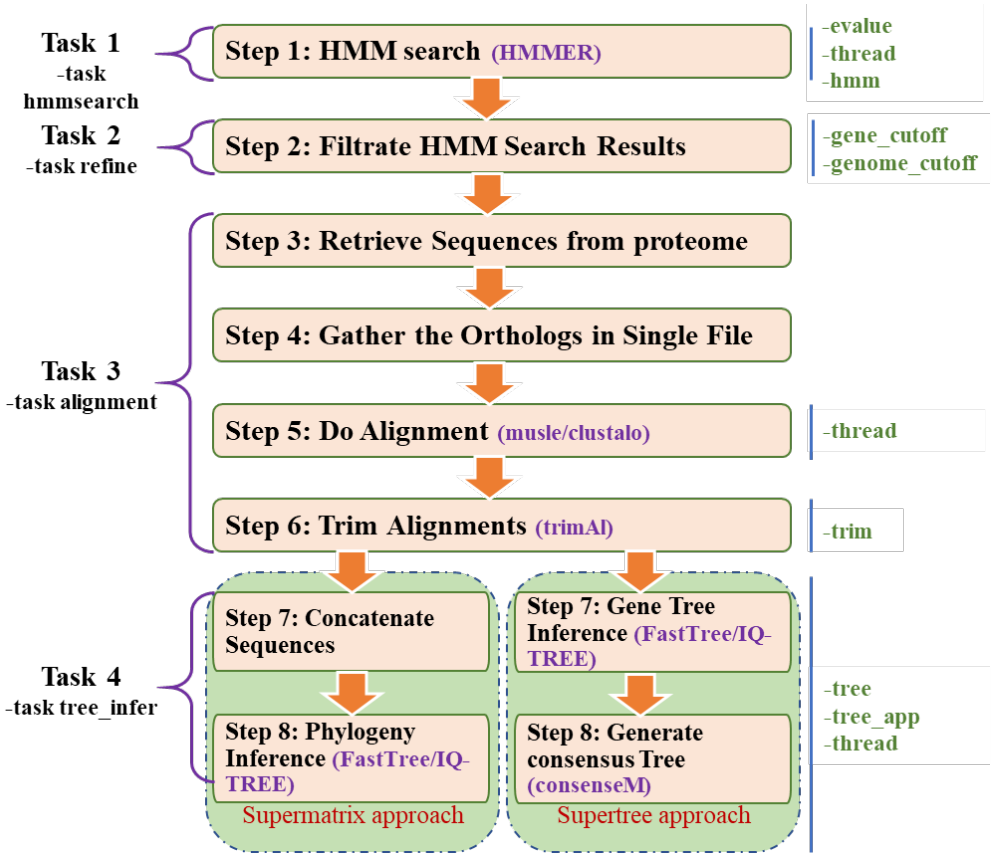

Fig. 1 Program flowchart of EasyCGTree for a complete run

#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.

myGenome.supermatrix.fasttree.tree/myGenome.supermatrix.iqtree.tree: generated by supermatrix approach using FastTree or IQ-TREE (set '-tree sm ' or default; -tree_app fasttree (default) or -tree_app iqtree).

myGenome.supertree.fasttree.tree/myGenome.supertree.iqtree.tree: generated by supertree approach using consense program in phylip software package (set '-tree st'; -tree_app fasttree (default) or -tree_app iqtree).

outtree/outfile: generated by supermatrix approach using consense program in phylip software package (set '-tree st'). It records the details to generate '**myGenome.supertree.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'.

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'.

2) A fasta-format file (myGenome.concatenation.fas) **contains concatenated sequences of wanted genes extracted from each genome, which is used to infer a ML tree.** This file is named after the genome folder 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.

3) A log file (myGenome_2021-8-2-13-21.log) **named after the genome folder 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.

4) 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. TEM1_HMMsearch_out: HMM search results against each proteome.

b. 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 50%) will be discarded.

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

d. 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)**.

e. TEM5_Alignment: fasta-format files of alignments created from data in 'TEM4_GeneCluster'.

f. TEM6_AlignTrimmed: fasta-format files of trimmed alignments created from data in 'TEM5_Alignment'.

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

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

i. intreeInfo.txt: correspondence relationship of taxa names between those users give and those used by program consense.

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

4) 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'.

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 Dir_name
```

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. Q: I have a genome with no proteome/CDS sequence released. How to use EasyCGTree to infer a tree from amino sequences?

A: You must get the proteome/CDS sequence before starting. There is an easy way to get CDS quickly, that is using the online tool CDSeasy (https://bioinfo-mml.sjtu.edu.cn/STEP/STEP_CDSeasy.php). You will get the CDS in a few minutes.

2. 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.

9. Citation

Xue HP, Zhang DF, Xu L et al. *Actirhodobacter atriluteus* gen. nov., sp. nov., isolated from the surface water of the Yellow Sea. Antonie van Leeuwenhoek 114, 1059–1068 (2021).. (**If you use EasyCGTree, please cite this paper before the paper describing EasyCGTree is published.**)

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. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004 32(5): 1792-1797.
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. Sievers F, Wilm A, Dineen D et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **2011** *7*, doi:10.1038/msb.2011.75.
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 (2017).
11. Brown CT et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 2015 523, 208–211 (2015).
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. Characterization of *Marinilongibacter aquaticus* gen. nov., sp. nov., a unique marine bacterium harboring four CRISPR-Cas systems in the phylum *Bacteroidota*. *J Microbiol*. 2022: https://doi.org/10.1007/s12275-022-2102-3.