EasyCGTree

An Easy Tool for Constructing Core-Gene Tree

(Draft manual)

Dao-Feng Zhang

Update on December 20th，2020

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

Contents

1. What is EasyCGTree?

EasyCGTree is a Perl script, developed to construct phylogenetic tree by taking microbial genomes or proteomes in fasta format and reference sequences (either nucleotide or protein) of a set of core genes as input data. It has integrated all the steps needed between the input data and the resulted tree file into one Perl script, which would make it easier to infer a core-gene tree.

2. How to install EasyCGTree?

There is no necessary to compile **EasyCGTree** after uncompressing the downloaded package. But several extra programs would be invoked in **EasyCGTree**. Please install and configure the extra programs prior to running **EasyCGTree**. Here is a list of the extra programs:

Windows OS users:

1) blast+ (not blast) package (The latest 2.10.x may have problems when running on Windows 10): https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/. After installation, you can add it to the environment variable of the OS, or set the full path in "parameters.txt" (see ***part 6***).

2) ActivePerl: https://www.perl.org/get.html

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

4*) muscle: http://www.drive5.com/muscle/

*: a version of muscle and FastTree had been included in the **EasyCGTree** package.

Linux OS users (Tested under Ubuntu):

1) blast+ package: https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/. After installation, you can add it to the environment variable, or set the full path in "parameters.txt" (see ***part 6***).

2) Perl: https://www.perl.org/get.html

3*) FastTree: vailable from http://www.microbesonline.org/fasttree/#Usage

4*) Clustal Omega. Available from http://www.clustal.org/

*: a version of FastTree and clustalo had been included in the **EasyCGTree** package for Linux version. ***Users need to use "chmod +x filename" within folder 'bin' to make them executable.***

EasyCGTree *DO NOT* support MAC OS currently.

3. Input data

As mentioned above, genome/proteome data in fasta format and reference sequences (protein) of a set of core genes are required to run **EasyCGTree**.

3.1 Genomes/proteomes in fasta format (named Genome data)

The fasta files are required to be uncompressed and a specialized name (unique, no spacer, no "_", and ending with ".fasta"). ***Don't worry when you have dozens of genomes. There is a Perl script "formatGenomes.pl" to do this laborious job.*** The user just needs to gather the genome sequences.

3.2 Core-gene sets for gene calling (Query data)

This set of genes determines that which genes will be extracted from the genomes/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 its 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.

An optional way to retrieve a core gene set is to download from the Genome Taxonomy Database (GTDB) server (https://gtdb.ecogenomic.org/). GTDB defined a gene set named bac120 that includes 120 ubiquitous single-copy genes across the domain *Bacteria*, and a gene set named ar122 that includes 122 ubiquitous single-copy genes across the domain *Archaea*. Users just need to download all the marker genes of *Bacteria*, *Archaea*, or both as wishes, and related taxonomy lists. ***There is a Perl script "GetReferencFromGTDB.pl" in our package to help extract reference sequences (See 4.1.2.1).***

These gene sets can be selected from those mentioned in the following ***part 3.3*** of this manual. Gene sets mentioned in this part and **part 3.3** should ***have the same gene numbers and contain the same set of homologous genes.***

**Because of lower accuracy in gene calling, EasyCGTree requests protein sequence used for searching homologous gene.

3.3 Pre-prepared core gene sets (Reference data) *(optional)*

Some users may already have a pre-defined core-gene set of some genomes and probably a tree based on these genomes and this gene set, and now, they want to include some new genomes into this core-gene tree. In this case, it will save much time to

prepare such a data set and command a type "A3" run (see **part 4.2**). These gene sets should have the same gene numbers and contain the same set of homologous genes as those mentioned in **part 3.2**. ***Notably, these gene sets should be included in a folder named "Reference" (coercive).***

**Only gene sets of protein sequence is allowed.

4. Run EasyCGTree

4.1 preparations

4.1.1 Genome data (Genomes/proteomes)

1) gather the genomes/proteomes of fasta format into a folder, name them in English style. ***Copy the folder into the directory of EasyCGTree (e.g. D:/EasyCGTree).***

Name the folder whatever you like, it will be referred as "MyGenomes" as an example in the following steps. The file extension is inessential if users ensure the data is in fasta format.

2) format the genomes.

From now on, users need to run "cmd.exe" for Windows or "Terminal" for Linux before execute the Perl scripts in **EasyCGTree**.

Change the working directory to the parent folder of the folder including the genomes/proteomes (MyGenomes).

For Windows users:

```
d:     #####press the key "Enter" when finishing a line
cd ./EasyCGTree
perl formatGenomes.pl MyGenomes
```

4.1.2 Query data (used for gene calling)

***** Notably, protein sequence is requested.***

4.1.2.1 bac120/ar122 core-gene set

1) Create a folder (e.g. GTDBdata) in directory wherever you like, and name it.
2) Download data from GTDB (https://gtdb.ecogenomic.org/) and put them into the folder "GTDBdata".

   Download the marker genes of *Bacteria*, *Archaea*, or both as you wish. Taking release95.0 as an example: ar122_taxonomy_r95.tsv

(https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/ar122_taxonomy_r95.tsv) and ar122_marker_genes_reps_r95.tar.gz (https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_reps/ar122_marker_genes_reps_r95.tar.gz) for *Archaea*; or bac120_taxonomy_r95.tsv (https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/bac120_taxonomy_r95.tsv) and bac120_marker_genes_reps_r95.tar.gz (https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_reps/bac120_marker_genes_reps_r95.tar.gz) for *Bacteria*.

Marker genes of all genomes included in this release is available at: https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_all/

Uncompress the files and move the "faa" folder into "GTDBdata".

3) Copy GetReferencFromGTDB.pl into the folder "GTDBdata".
4) Assume that the bac120 gene sets of the genus *Bacillus* was wanted.

   Change the working directory to the parent folder "GTDBdata" and type:

   perl GetReferencFromGTDB.pl bac120_taxonomy_r95.tsv faa g__Bacillus

   The bac120 genes of each genomes, belonging to *Bacillus* and included in GTDB, will be extracted and write into a file named with the species label and in the format as descript in **part 4.1.2.2**. A log file (.csv) named with the taxon specified will be created to report the included genomes and related details (not all the genes of bac120/ar122 could be found in the genomes included in GTDB). The "g" means genus, and the users can specify any taxa ranging from species to phylum with a label of "s, g, f, o, c, p", respectively. Please note, there are two "_", not one.

   If you specify a higher taxon or a big genus, you will get a lot of bac120 gene sets. In most instances, one gene set for a genus is feasible, because the divergence within a genus is limited for bac120/ar122 genes (https://gtdb.ecogenomic.org/). However, more gene sets used in **EasyCGTree** will increase the accuracy of the gene-calling step. ***In contract, too much gene sets will slow down the running speed. We recommend to use gene sets that meets the following criteria:*** **a)** each gene set contain all the genes (120 or 122) of bac120/ar122; **b)** the number of gene sets between M/60 and M/15 be used in following analysis (M means the number of members of a genus); **c)** and the species of these gene sets should be widespread in the genus tree based on 16S rRNA gene.
5) Gather the selected gene sets in a folder (e.g. query), and ***move it into the directory of EasyCGTree.***

4.1.2.2 Personal core-gene set

The users can use personal core-gene sets, but ensure that: ***a) they contain no***

paralogs; b) each gene is labeled in a format of XXX_XXX_XXX_geneSymbol and is unique within a gene set and among gene sets; c) they have the same gene numbers; they contain the same set of homologous genes. The latter two are recommended but not imperative if the user was well-known about how **EasyCGTree** works. For gene labels, we recommend to use "generic name"_"specific epithet"_"strain number"_"geneSymbol". The separator "_" is not allowed in each of the four divisions. For running **EasyCGTree** correctly, ***the position of gene symbol is coercive, and gene symbol should be kept consistent among homologs in all gene set.*** The "generic name", "specific epithet" and "strain number" are not coercive, and can be set waywardly (e.g. 1_1_1_ geneSymbol).

4.1.3 Reference data (used for tree inference) (*optional*)

***** Notably, protein sequence is requested.***

This data set should be prepared for a type "A3" run (see **part 4.2**). These gene sets should have the same gene numbers and contain the same set of homologous genes as those mentioned in **part 3.2 and 4.1.2**. ***Notably, these gene sets should be included in a folder named "Reference" (coercive and no need to be specified in command line) and be prepared in the format as descript in part 4.1.2.2.***

4.2 Ready to Run EasyCGTree

With the input data prepared correctly, it is very easy to run **EasyCGTree**. Ensure that the two folders are in the directory of **EasyCGTree**, change the working directory to that of **EasyCGTree** and type:

Perl EasyCGTree.pl myGenomes query A/A1/A2/A3 nucl/prot parameters.txt(optional).

Table 1 Allowed combinations of command line settings and input data

Command line settings		Input data		
A/A1/A2/A3	nucl/prot	Genome	Query	Reference#
A	nucl	nucl*	prot	not requested
	prot	prot		
A1	nucl	nucl*		
	prot	prot		
A2	nucl	nucl*		
	prot	prot		
A3	prot	prot		prot

*, nucl means either genome sequence or proteome in nucleotide sequence. #, the folder name of Reference data is coercive to be "Reference", and is not needed to specified in the command line.

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.

A/A1/A2/A3 means running mode:

A, command complete run without input data mentioned in **part 3.3 and 4.1.3**, and yield all output files (See **part 5**);

A1, first part run (yield TEM1-3 folders, query sequences and a log file);

A2, second part run (yield TEM4-9 folders, a log file, concatenated sequences and a tree);

A3, complete run with input data mentioned in **part 3.3 and 4.1.3,** conflicting to parameter nucl.

The division of A1 and A2 mode is aimed to save running time when users need to optimize parameters (see **part 6**), because BLAST search (step 2) is time-consuming when the gene sets used in gene calling and/or genomes increase (see **part 4.1.2.1**).

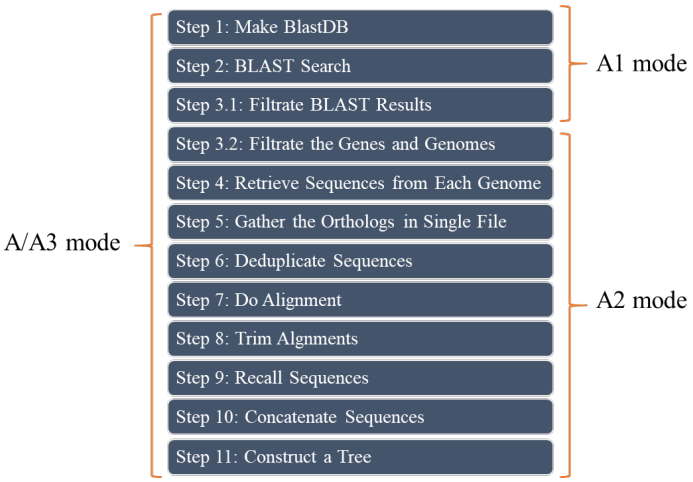

Figure 1 Program flowchart of EasyCGTree for a complete run

nucl/prot specialized the sequence type (protein or nucleotide) used for tree inference. Setting nucl means the input data mentioned in ***part 3.1and 4.1.1*** should be nucleotide sequence, while setting prot means the input data should be protein sequence.

5. Output files

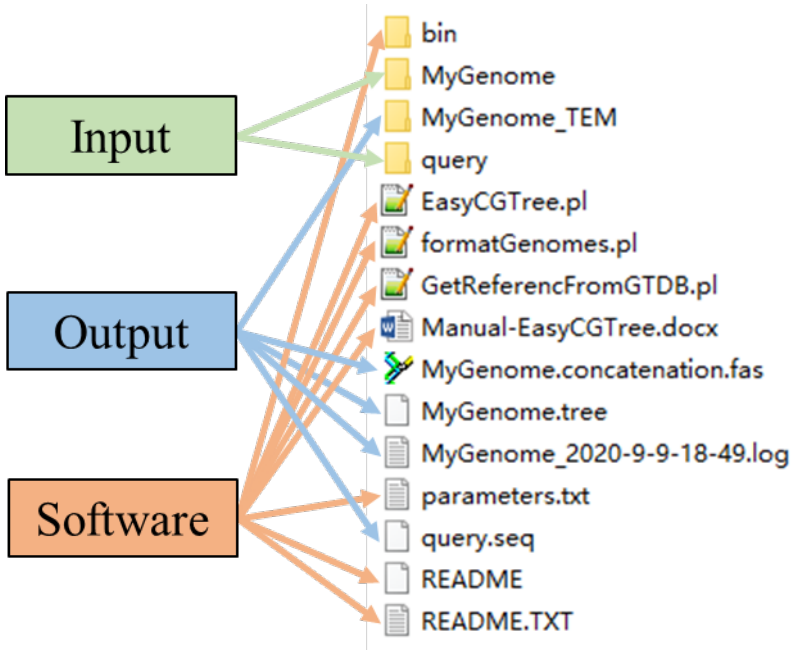

Figure 2 Content of the EasyCGTree home directory after a run

1) A file in Newick format of a ML tree. Users could display it by using FigTree, MEGA, iTOL or other software.

2) A fasta-format file contains concatenated sequences of wanted genes extracted from each genome, which is used by FastTree software 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 TEM9_ReCallAln (see below).

3) A log file 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. This information was also displayed on the screen during the run.

4) A file contains the query sequences used for BLAST search, which is named after the folder containing reference gene sets and ends with ".seq".

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

a. TEM1_blastDB: containing the databases created by makeblastdb program for each genome.

b. TEM2_blastOUT: containing BLAST results of each genome in format 6 by tblastn program.

c. TEM3_blastOUT_S: containing screened BLAST results of TEM2_blastOUT. If two or more references gene sets were used, the best result for each gene will be selected based on the bitscore. Results below the identity cutoff will be dropped and the cutoff (default 50%)

could be set in parameters.txt (part 6).

d. TEM4_GeneSeqs: containing gene sequences retrieve from each genome based on the results of TEM3_blastOUT_S.

e. TEM5_GeneCluster: containing files gathering homologs from different genomes. One gene (cluster) was gathered in a single file.

f. TEM6_DedupGeneCluster: containing files including non-redundant sequences. These files are descendant of those in TEM5_GeneCluster. Genes with the same sequence are deduplicated, and the correspondence between representatives and analogues are recorded in files of TEM60_DedupList.

g. TEM60_DedupList: see f. TEM6_DedupGeneCluster.

h. TEM7_Alignment: containing fasta-format files of alignments created based on files in TEM6_DedupGeneCluster.

i. TEM8_AlignTrimmed: containing fasta-format files of trimmed alignments created based on files in TEM7_Alignment.

j. TEM9_ReCallAln: containing fasta-format files of trimmed alignments of all the genomes used to infer ML tree. These files are generated based on the files in TEM8_AlignTrimmed and TEM60_DedupList.

6. Parameter setting (optional)

Five parameters are optional to be set in 'parameters.txt' (the name can be changed, but the format in it should be following the instructions in the model file parameters.txt). ***If users want to set them, please remember to specify the filename including specific parameters in the command line (see part 4.2).***

a. tblastn: specify the location of the program tblastn, e.g. /share/bin/. If you have added blast path to the environment variable of the OS, ignore it.

b. num_threads: specify the number of threads used by the program tblastn. It depends on your computer and should be integer).

c. blastIdentitycutoff: specify the cutoff for filtering the tblastn results (0-100). The bac120 and ar122 consist of highly conserved house-keeping genes. So lower cutoff might introduce wrong signals when inferring ML trees, if some genes of bac120/ar122 were not well assembled or sequenced. We recommend to use ⩾50 if reference gene sets can be prepared for all the genera of interest.

d. geneCutoff: specify the cutoff for omitting low-prevalence gene (0-1). It means that only genes present in more than 80% (default, geneCutoff=0.8) of the genomes will be used in following analysis. Lower cutoff will keep more genes used to infer ML tree, and missing genes in some genomes will be treat as gaps.

e. genomeCutoff: specify the cutoff for omitting low-quality genomes (0-1). It means that only genomes harboring more than 80% (default, genomeCutoff=0.8) of the genes determined by setting geneCutoff (see previous paragraph) will be used to infer a ML tree. Lower cutoff will keep more genomes used to infer ML tree, but the genes used will be fewer.

The latter two parameters may be confusing to some users. In most situations, leave them alone and the default setting is OK for a run involving limited number of genomes (<30), especially when the genomes are all from type strains. However, if much more genomes are used, it is inevitable that some genes are not detected in some genomes, and the number of common genes decrease. In fact, many genomes collected in the representative database of GTDB contain an incomplete bac120/ar122 gene set, and users will find it when gather reference gene sets of bac120/ar122 (see ***part 3.2***). It is competing between increasing common genes and increasing genomes of interest when inferring a core-gene tree.

The latter 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 lower limit 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 (0-1).

7. Hardware Requirement

A normal PC is good enough to run EasyCGTree, because clustalo, FastTree, and muscle are fast and approachable. However, the speed depends on the size of the input data. When Genome data <100, Query data < 10, and gene number in Query data < 200, a PC made in last 3 years will finish the analysis within several hours. 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).

8. Performance

Table 2 Running performance with tested data

Memory	16G	16G
CPU	i7-9700	i7-9700
OS	Windows 10	Windows 10
Threads used	2 (only for BLAST)	2 (only for BLAST)
Run mode	A3	A1
Genomes/proteomes	1	450
Taxonomy	Erythrobacteraceae	Enterobacteriaceae
Query	2	3
Genes	288	120
Reference	75	/
Running time	20 min 27 sec	1 hour 56 min 5 sec

9. FAQ

Waiting for questions and suggestions……

10. References

Zhang, D.F.; Cui, X.W.; Zhao, Z.; Zhang, A.H.; Huang, J.K.; Li, W.J. *Sphingomonas hominis* sp. nov., isolated from hair of a 21-year-old girl. Antonie Van Leeuwenhoek 2020, doi:10.1007/s10482-020-01460-z. (**If you use EasyCGTree, please cite this paper befor the paper describing EasyCGTree is published.**)

Parks, D.H.; Chuvochina, M.; Waite, D.W.; Rinke, C.; Skarshewski, A.; Chaumeil, P.A.; Hugenholtz, P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **2018**, *36*, 996-1004, doi:10.1038/nbt.4229.

Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**, *25*, 3389-3402, doi:10.1093/nar/25.17.3389.

Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.Z.; Lopez, R.; McWilliam, H.; Remmert, M.; Soding, J., 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.

Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **2009**, *26*, 1641-1650, doi:10.1093/molbev/msp077.

Edgar, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **2004**, 32(5), 1792-1797, doi: 10.1186/1471-2105-5-113.