原核生物核心基因进化树构建软件

EasyCGTree v4.2

使用说明

编写：张道锋

更新日期：2024 年 01 月 22 日

任何有关使用的问题和建议，欢迎联系：*zdf1987@163.com*

目录

1. EasyCGTree 简介

EasyCGTree 是一款使用 Perl 语言编写，用于构建**基于指定核心基因集的最大似然法（maximum likelihood，ML）进化树**的脚本程序。以 fasta 格式的**蛋白质组序列、基因组序列或者二者混合数据**作为输入数据，内置了常用基因集的 HMMs 数据用于同源基因检索，用户提供的基因序列也可以被 EasyCGTree 构建成 HMMs 并用于同源基因检索。，**EasyCGTree** 即可快速构建出一个。EasyCGTree 包括两种推演系统发育的方法：**supermatrix** 和 **supertree**。该程序整合了从基因组数据到进化树的所有数据处理步骤，使用第三方软件进行同源基因检索、序列比对和进化树构建，使核心基因进化树的构建变得非常方便、快捷。另外，EasyCGTree 产生的中间数据也可以直接用于其他软件进行分析。

2. EasyCGTree 的安装

EasyCGTree 软件包 (https://github.com/zdf1987/EasyCGTree4; https://gitee.com/zdf1987/EasyCGTree4/releases)解压之后，无需编译，只需安装 Perl 语言环境，并配置好需要调用的第三方软件，即可直接运行。**EasyCGTree** 软件包已经包含了第三方程序，如果用户需要使用这些程序的最新版本，则需要进行额外的安装和配置。以下是所需程序的列表：

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/

*: 这些软件已包含在 **EasyCGTree** 软件包里。对于 Linux 版本，可能需要在 bin 文件夹下使用**“chmod +x filename”命令让这些软件变为可执行文件。**

内置的 HMM 文件需要单独下载，解压后放在 HMM 文件夹下。

注意：EasyCGTree 暂*不支持*MAC OS。

3. 输入数据

只需要 FASTA 格式的基因组序列（DNA、蛋白质或者二者混合）数据即可

运行 EasyCGTree。每个基因组序列文件内序列类型需保持一致。

3.1 基因组数据（FASTA 格式）

FASTA 格式的基因组文件必须是解压后的形式，不同基因组文件不可重名（文件名内部不能包含空格和“_”，并且以“.fas”作为扩展名）。***如果使用的基因组非常多的话，也不用担心，EasyCGTree 会自动对所有的基因组文件名格式化。***用户只需要整理每个基因组的序列文件并放入一个文件夹即可。

3.2 profile HMM

用于构建 profile HMM 的基因集决定了将从蛋白质组中提取哪些基因来推断核心基因树。如何定义感兴趣基因组数据集的核心基因?最好的方法是先定义泛基因组（Pan-genome）。然而，如果研究人员只想要一个核心基因树，泛基因组分析并不是最佳方法。也许，从以前的研究或公共数据库中检索核心基因集是更好的选择。

EasyCGTree 包含了原核生物或某些分类单元中普遍存在的几个核心基因集，并构建了相应的 profile HMM，称为 profile HMM Database (PHD)。目前，PHD 以 HMM.zip(解压后将 .hmm 文件放入 'HMM' 文件夹中) 的形式从 https://github.com/zdf1987/EasyCGTree4 获得，包括:

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 是可扩展和个性化的。用户可以提供基因序列然后使用 EasyCGTree 中的脚本程序构建 profile HMM (see Section 6.1).

4. 运行 EasyCGTree

4.1 前期准备

输入数据（氨基酸序列文件、DNA 序列文件或者二者混合）

将 FASTA 格式的基因组数据整理到一个文件夹中，用英文命名。将文件夹

复制到 EasyCGTree 的主目录中(例如 Linux 的“…Downloads/EasyCGTree”)；' / '对于 Windows 应该是'\ ')。

将文件夹命名为任意名称，并确保数据为 FASTA 格式。

**去掉扩展名的蛋白质组名称将作为树上的标签。

选择一个 HMM

用户可以在 PHD (见 3.2 节)，也就是 HMM 文件夹中选择一个。

或者，准备感兴趣的基因簇(同源基因序列在一个文件中)，并使用脚本“BuildHMM.pl”来构建个性化的 HMM(见第 6.1 节)，同时会自动添加新的 profile HMM 到 PHD 中。

4.2 运行 EasyCGTree

注意：EasyCGTree 是一个命令行软件，需要在 cmd.exe (Windows)或 Terminal (Linux)中运行。

正确准备好输入数据后，就可以很容易地运行 EasyCGTree 了。确保基因组组文件夹(例如 myGenome)在 EasyCGTree 的主目录下(例如 Linux 的“…/Downloads/EasyCGTree”)；“/”对于 Windows 应该是“\”)，将工作目录更改为 EasyCGTree 的工作目录并运行。

(i)假设 Windows 用户将软件下载到“D:”驱动器下的“Downloads”文件夹中。命令行运行的示例如下(在完成每一行时按“Enter”)：

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

(ii) Linux 用户使用“cd”命令将 EasyCGTree 的工作目录改为 EasyCGTree 的工作目录后，可以按照相同的方式运行 EasyCGTree。

然后，将获得 Newick 格式的进化树（以包含基因组的文件夹的名称命名，例如 myGenome.bac120.supermatrix.fasttree.tree）和运行脚本期间生成的许多文件/文件夹。

4.3 基本设置（无需按顺序输入）

-input <String> (The only essential option)

Input directory including all genomic data sets

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

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

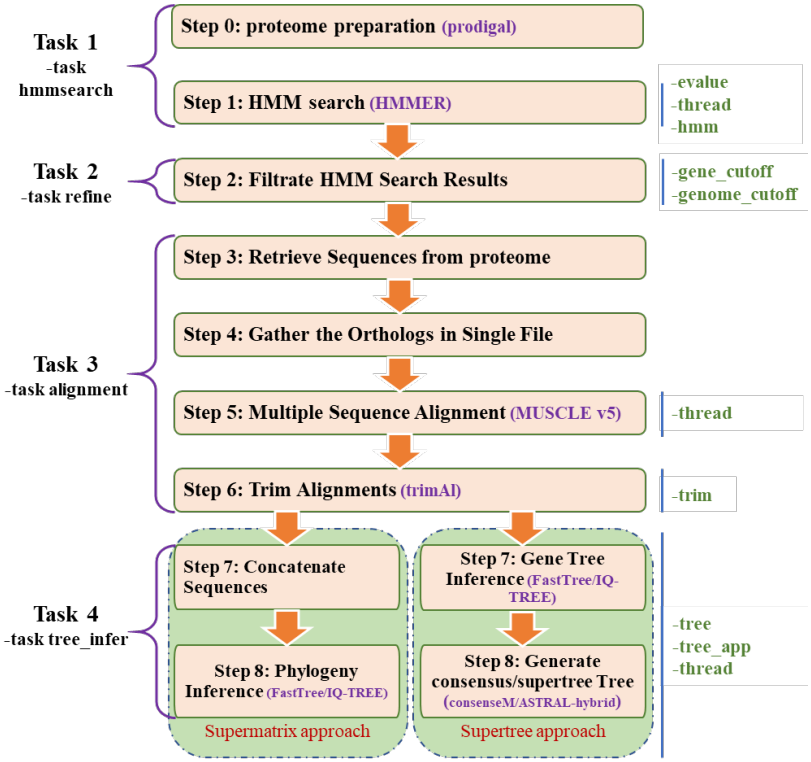

Fig. 1 EasyCGTree 工作流程图

-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.

FastTree 和 IQ-TREE 的相关设置可以在**'tree_app-options.txt'**文件中进行。

#关于'-genome_cutoff'和'-gene_cutoff'

这两个参数对于一些用户可能较难理解。在大多数情况下，如果只涉及数目较少的基因组（<30），使用默认参数就能得到很好地结果，尤其是所有基因组都来自于典型菌株时。然而，当使用的基因组较多时，无法在所有基因组中检索到的基因数目就会显著增加，而所有基因组共同含有的基因就会显著下降。事实上，GTDB 数据库(Genome Taxonomy Database, https://gtdb.ecogenomic.org)中很多基因组都只收录了不完整的 bac120 或 ar122 基因集。这一点，用户在提取参比基因集的时候都会或多或少看到一些例子。共有基因的数目和基因组数目是此消彼长的矛盾关系。

这两个参数便是用来缓解这种矛盾，主要是通过排除缺失次数较多的基因和含有基因数量较少的基因组的方式来实现。需要注意的是，这两个参数决定的是基因组质量的下限（所选的基因集由所选的基因组决定），如果所有的基因组均具有较高的质量，那么这两个参数设置为多少都不影响分析结果。例如：如果所有的基因组都含有 bac120 或 ar122 的所有基因，那么不管这两个参数如何设置，我们总能得到所有的 120 或 122 个基因（0.5-1），因为所有的基因组都能满足这两个参数设定的要求。

5 输出文件

假设包含基因组数据文件夹的名称是'myGenome'。

1) Newick 格式的系统发育树。用户可以通过使用 FigTree, MEGA, iTOL 或其他进化树查看器来显示。

最终进化树：树文件名看起来像这样：

myGenome.bac120.119.supermatrix.fasttree.tree
(1) (2) (3) (4) (5)

(1)表示输入蛋白质组(-proteome)所在文件夹的名称；(2) /HMM 中用于进化树推演的基因集（-hmm）；(3)基因集中有多少个基因被用于构建进化树进化树建树方法（-tree: cs, sm 或 st）；(4)进化树推演工具（-tree_app：Iqtree 或 fasttree）。

2) 根据基因组文件夹、HMM、建树方式、建树软件和运行起始时间命名的日志文件(myGenome…._2021-8-2-13-21.log)。该文件记录了 **EasyCGTree** 运行过程中的一些细节，例如：多少以及哪些基因和基因组最终被选择出来用于后续分析。在程序运行过程中，该文件中绝大多数信息也同步在运行窗口界面显示。

3) 以基因组文件夹(myGenome_TEM)**命名、以"_TEM"为后缀的文件夹。**该文件夹包含多个子文件夹，子文件名在不同的分析中保持一致，且对所有 EasyCGTree 的运行都是通用的。它包含以下文件夹。

a. **TEM0_Proteome:** 包含用于后续分析的蛋白质组，或者直接从输入文件夹拷贝过来，或者是通过 prodigal 软件获得。

b. **TEM1_HMMsearch_out：**HMMER 对每个基因组的检索结果。

c. **TEM2_HMMsearch_outS：**TEM1_HMMsearch_out'中检索结果过滤后的结果。如果确定了两个或两个以上的匹配结果，则根据 e 值选择每个基因的最佳结果。低于 e 值的阈值(默认为 1e-10)的结果将被丢弃。

d. **TEM3_GeneSeqs：**根据 **TEM2_HMMsearch_outS** 文件夹中的结果从每个基因组中提取的基因序列。

e. **TEM4_GeneCluster：**包含从不同基因组中整理得到的同源基因序列文件。每一个同源基因被保存在同一个文件中。

f. **TEM5_Alignment：**包含根据 TEM4_GeneCluster 中的文件构建的 fasta 格式序列比对文件。

g. **TEM6_AlignTrimmed：**包含 TEM5_Alignment 中的文件使用 trimAI 修剪后的 fasta 格式序列比对文件。

h. **TEM7_geneTree：**包含根据 TEM6_AlignTrimmed 文件夹中每个基因簇文件构建的进化树（以及其他 IQ-TREE 产生的文件）。

i. **所有核心基因串联序列文件(myGenome…..concatenation.fas)。**该文件以 FASTA 格式存放，也是最终用来构建进化树的输入文件。该文件以基因组文件夹、HMM、基因数量、建树方式和建树软件命名，以".concatenation.fas"为后缀，根据 TEM6_AlnTrimmed 文件夹中的数据生成（见下文）。IQ-Tree、RaxML 等进化树构建软件也可以直接使用

该文件。

j. genomeGeneScreened.txt：过滤后的基因和蛋白质组列表。

k. HMMinfo.txt：基因家族列表和每个结构域的长度。

l. Intree：包括由每个选定基因构建的树。从'myGenome/TEM7_geneTrees'中的基因树文件生成的。树上的标签名称与用户提供的不一致，对应关系可在'myGenome_TEM/intreeInfo.txt '中找到。

m. intreeInfo.txt：用户提供的基因组数据文件名和 consense 与 ASTRAL-Hybrid 使用的名称的对应关系。

n. IQ-TREE 生成的其他文件：IQ-TREE 在运行期间会生成几个文件，包括'.bionj '，'.cpk.gz'，'.contree'，'.iqtree'，'.log'，'.mldist'，'.model.gz'，'.split.next '和'.treefile '。 “.contree”中的内容与上文提到的最终进化树（如果使用 IQ-TREE）相同。如果使用 supermatrix 方法。这些文件将写入 EasyCGTree 的主目录或文件夹“../TEM7_geneTree”。

o. outtree/outfile： 由 philips 软件包中的 consense 程序计算一致性树时生成(设置'-tree cs')。它记录了生成“myGenome.XXX.consensus.XXX.tree”的详细信息。通过 consense 程序从 intree 中获取 Tree。树上的标签名称与用户提供的不一致，对应关系可在'myGenome_TEM/intreeInfo.txt '中找到。

p. trimAl_Details.txt: 记录了每一个基因在使用 trimAl 修剪前后的比对序列长度。

6 有用的脚本

6.1 BuildHMM.pl

用于从个性化的的基因簇(蛋白质序列)中构建 profile HMM。配置文件 HMM 将被写入 PHD(目录' ./HMM ')，其标签与包含自定义基因簇的文件夹相同。

使用方法：

```
> 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

用于统计基因组/蛋白质组中某一基因集中各个基因的普遍性，基于本地 PHD 中相关 HMM 的同源基因检索结果。如果文件夹 '../TEM1_HMMsearch_out '未被检测到，将使用 EasyCGTree.pl 执行同源基因搜索。

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 硬件和系统要求

一台普通的电脑足以运行 EasyCGTree，因为 clustalo、FastTree 和 muscle 是非常高效、简便的软件。但是，速度取决于输入数据的大小。当基因组数据<100，Query data <10，其中基因数<200 时，一台个人电脑将在数小时内完成分析。如果

使用比较大的输入数据，则建议使用 Linux 操作系统版本和性能强劲的 PC 或服务器，因为 Linux 下的 FastTree 和 clustalo 支持多线程(Windows 下的 muscle 和 FastTree 只支持单个线程)。请注意，IQ-TREE 将占用比 FastTree 更多的时间和更大的内存。如果用户在 PC 上使用 IQ-TREE (-tree_app iqtree)运行失败，请尝试使用 FastTree（默认设置），或使用功能更强大的 PC 或服务器。

8 FAQ

1. 为什么不支持 Mac OS 操作系统呢？

A: EasyCGTree 以“易于安装”和“易于使用”为理念，在选择可以直接使用的第三方应用程序时是以简单为先决条件。不幸的是，我们没有找到比较简单的应用程序来在 Mac 系统上实现 EasyCGTree 的功能。

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 参考文献

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.