

原核生物核心基因进化树构建软件

**EasyCGTree v2**

使用说明

开发：张道锋

更新日期：2021 年 2 月 18 日

任何有关使用的问题和建议，欢迎联系：[zdf1987@163.com](mailto:zdf1987@163.com)

# 目录

1. EasyCGTree 简介 .....	1
2. EasyCGTree 的安装.....	1
3. 输入数据格式.....	3
4. 运行 EasyCGTree .....	4
4.1 前期准备 .....	4
4.1.1 Genomes.....	4
4.1.2 核心基因参比序列集.....	5
4.1.3 其他核心基因集.....	7
4.2 正式运行 EasyCGTree.....	8
5 输出文件 .....	10
6 参数设定 (可选项).....	12
7 硬件和系统要求 .....	13
8 运行时间 .....	14
9 FAQ.....	14
10 参考文献.....	15

## 1. EasyCGTree 简介

**EasyCGTree** 是一款使用 Perl 语言编写，用于构建原核生物核心基因进化树的脚本程序。用户只需要提供 fasta 格式的基因组或蛋白质组文件和一套核心基因的参比序列（蛋白序列），**EasyCGTree** 即可快速构建出一个基于指定核心基因集的最大似然法（maximum likelihood, ML）进化树。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。

## 2. EasyCGTree 的安装

**EasyCGTree** 软件包解压之后，无需安装，只要配置好需要调用的第三方软件，即可直接运行。根据不同的操作系统，用户需要自行安装的软件有以下几个。

### Windows 操作系统:

1) blast+ (不是 blast) 软件包（注意：最新的版本 2.10.x 在 Windows 10 操作系统上经常会出现一些问题）：下载地址 <https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/>。安装完成后，建议将 blast+ 相关子程序的路径加入系统环境变量。如果没有正确地加入环境变量，则需要在配置文件 parameters.txt 中设置 blast+ 的完整路径，并在运行 EasyCGTree 时输入 parameters.txt 参数（可选参数）（参见第 6 部分）。

2) ActivePerl: 下载地址 <https://www.perl.org/get.html>

3\*) FastTree: 下载地址 <http://www.microbesonline.org/fasttree/#Usage>

4\*) muscle: 下载地址 <http://www.drive5.com/muscle/>

\*: **EasyCGTree** 软件包中已经包括了最新版本的 muscle 和 FastTree，均为单线程版本。用户可直接使用，无需下载安装。

### Linux 操作系统:

1) blast+ (不是 blast) 软件包：下载地址 <https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/>。安装完成后，建议用户将 blast+ 相关子程序的路径加入系统环境变量。如果没有正确地加入环境变量，需要在配置文件 parameters.txt 中设置 blast+ 的完整路径，并在运行 EasyCGTree 时输入 parameters.txt 参数（可选参数）（参见第 6 部分）。

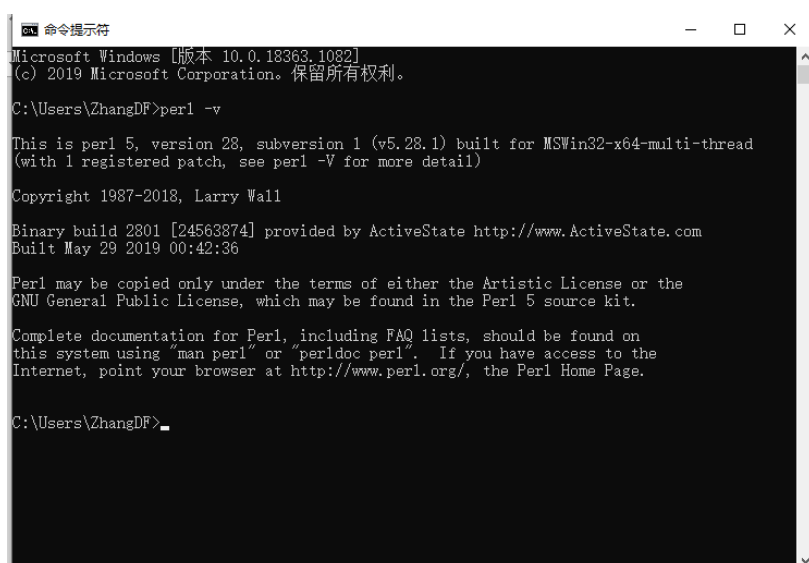
2) Perl: 下载地址 <https://www.perl.org/get.html>

3\*) FastTree: 下载地址 <http://www.microbesonline.org/fasttree/#Usage>

4\*) Clustal Omega: 下载地址 <http://www.clustal.org/>

\*: **EasyCGTree** 软件包中已经包括了最新版本的 Clustal Omega (clustalo) 和 FastTreeMP, 均支持多线程。用户可直接使用, 无需再次安装。但是用户可能需要在 bin 文件夹下使用 “chmod +x filename”命令来更改这两个软件的可执行权限。

用户可通过下图方式确认 Perl 语言环境否正确安装: 在 Windows 系统下打开 cmd.exe (在开始菜单搜索即可), 或者在 Linux 系统下打开终端(Terminal); 输入: **perl -v**。如果显示类似以下信息的界面, 则表明 Perl 语言环境正确安装。



```
命令提示符
Microsoft Windows [版本 10.0.18363.1082]
(c) 2019 Microsoft Corporation. 保留所有权利。

C:\Users\ZhangDF>perl -v

This is perl 5, version 28, subversion 1 (v5.28.1) built for MSWin32-x64-multi-thread
(with 1 registered patch, see perl -V for more detail)

Copyright 1987-2018, Larry Wall

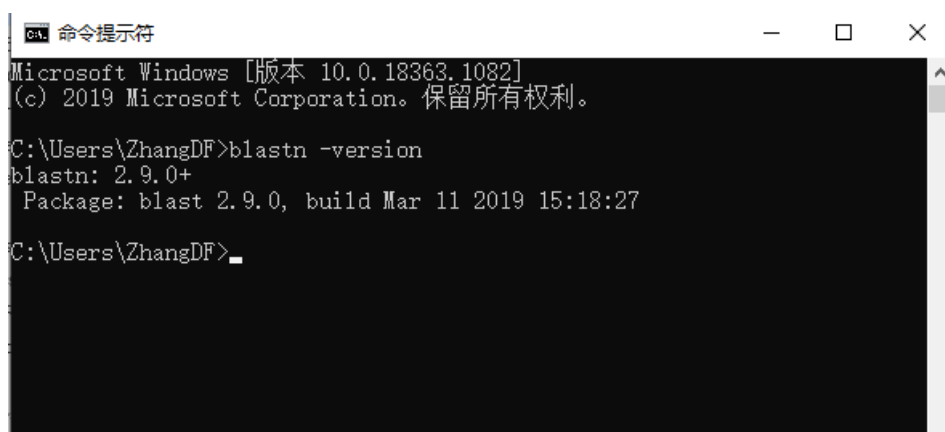
Binary build 2801 [24563874] provided by ActiveState http://www.ActiveState.com
Built May 29 2019 00:42:36

Perl may be copied only under the terms of either the Artistic License or the
GNU General Public License, which may be found in the Perl 5 source kit.

Complete documentation for Perl, including FAQ lists, should be found on
this system using "man perl" or "perldoc perl".  If you have access to the
Internet, point your browser at http://www.perl.org/, the Perl Home Page.

C:\Users\ZhangDF>
```

用户可通过以下方式确认 blast+是否正确安装并已加入系统环境变量: 在 Windows 系统下打开 cmd.exe (在开始菜单搜索即可), 或者在 Linux 系统下打开终端 (Terminal); 输入: **blastn -version**。如果显示类似于以下信息的界面, 则表明 blast+正确安装, 且已加入系统环境变量。



```
命令提示符
Microsoft Windows [版本 10.0.18363.1082]
(c) 2019 Microsoft Corporation. 保留所有权利。

C:\Users\ZhangDF>blastn -version
blastn: 2.9.0+
Package: blast 2.9.0, build Mar 11 2019 15:18:27

C:\Users\ZhangDF>
```

如果已安装却没有加入环境变量,则需要在配置文件 parameters.txt 中设置 blast+的完整路径,并在运行 EasyCGTree 时输入 parameters.txt 参数(可选参数)(参见第 6 部分)。

**注意: EasyCGTree 暂不支持 MAC OS。**

### 3. 输入数据格式

EasyCGTree 需要 fasta 格式的基因组/蛋白质组数据(Genome data)和核心基因参比序列(Query data/Reference data)作为输入数据。EasyCGTree 使用 Query data 对 Genome data 作 BLAST 检索,根据 BLAST 结果提取相应的同源序列形成每个基因组的同源基因集,然后和 Reference data(如果提供)中的基因集一起进行序列比对、修剪和串联,最终使用串联序列构建进化树。

#### 3.1 fasta 格式的基因组/蛋白质组数据 (Genome data)

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,可以使用 EasyCGTree 中名为“formatGenomes.pl”的脚本程序来对所有的基因组进行文件名格式化。用户只需要保证文件内的数据是 fasta 格式即可(参见 4.1.1 部分)。

#### 3.2 核心基因数据集 (Query data)

核心基因数据集决定了 EasyCGTree 需要从基因组上提取哪些基因序列来构建进化树。那么如何确定所需基因组的核心基因呢?最好的方法是先进行泛基因组(Pan-genome)分析。然而,对于只需要构建一个核心基因进化树的用户来讲,泛基因组分析很不划算。这时从已有的研究中获得核心基因来构建进化树就更加方便、高效。

对于研究较少的微生物类群来讲,可能从来没有人开展过相关的泛基因组分析,那么从别人的研究中获得核心基因集,是不现实的。对于这种情况,用户可以考虑从 GTDB 数据库(Genome Taxonomy Database, (<https://gtdb.ecogenomic.org/>))获取 bac120 或者 ar122 核心基因集。GTDB 在整个细菌域定义了一个称为 bac120 的核心基因集,包含 120 个在细菌域普遍存在的单拷贝基因;在古菌域定义了一个称为 ar122 的核心基因集,包含 122 个在古菌域中普遍存在的单拷贝基因。用户可根据具体情况下载细菌域或古菌域的 bac120/ar122 数据库,以及相应的分类表。然后使用 EasyCGTree 中的“GetReferenceFromGTDB.pl”脚本程序来提取感

兴趣类群的 bac120 或者 ar122 核心基因集作为参比序列（参见 4.1.2.1）。

这些基因集也可以从接下来 3.3 部分筛选得到。

**\*\*由于使用核酸序列进行同源基因检索时精准度较低，EasyCGTree 只支持使用蛋白序列进行同源基因检索。**

### 3.3 已有的核心基因数据集（Reference data）（可选项）

有些时候，或者是前期分析得到，或者是从其他的研究报道中获得，用户已经拥有一套根据一个基因组数据集定义的核心基因集，或许也已经有一个进化树。现在，用户想增添新的基因组到进化树中。在这种情况下，使用 EasyCGTree 的 A3 运行模式（见 4.2）可以很方便地完成分析。这些基因集应该是和 3.2 部分描述的基因集在基因数量上是相对应的，并且同源基因使用统一的基因名。

**\*\*这些基因集必须是 fasta 格式的蛋白序列，且需要放在名为“Reference”的文件夹中。**

## 4. 运行 EasyCGTree

### 4.1 前期准备

#### 4.1.1 Genomes

1) 将所有基因组文件放在一个文件夹中，用英文字符进行命名（例如 MyGenome）。将该文件夹复制到 **EasyCGTree** 的主文件夹中，比如：  
D:/EasyCGTree。

包含基因组的文件夹可以任意命名，在后续的描述中，均以“MyGenomes”为例。文件夹内文件的扩展名没有特殊要求，只要用户保证文件内的数据是以 fasta 格式存放即可。

#### 2) 基因组文件名格式化

从这一步开始，用户就需要使用 “cmd.exe”（Windows 系统）或者“Terminal”（Linux 系统）来执行 **EasyCGTree** 软件包中的脚本程序。

首先改变当前的工作目录为 **EasyCGTree** 的主文件夹，确保基因组文件夹 MyGenomes 在当前目录下。

以 Windows 系统为例，用户依次输入：

d: (每输入完一行，敲击 “Enter” 键)

cd ./EasyCGTree

perl formatGenomes.pl MyGenomes

格式化之后的基因组文件名(除去扩展名),即是最终进化树上现实的内容。

#### 4.1.2 核心基因参比序列集

\*\*注意: 这部分数据要求是蛋白序列。

##### 4.1.2.1 以 bac120/ar122 作为核心基因集

- 1) 根据用户的喜好创建一个文件夹, 比如: GTDBdata。
- 2) 从 GTDB (<https://gtdb.ecogenomic.org/>) 数据库下载数据并放在 GTDBdata 文件夹中。

用户可以根据实际情况选择下载细菌域或古菌域的相关基因数据。以 95.0 版本为例: 古菌需要下载的文件有 ar122\_taxonomy\_r95.tsv ([https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/ar122\\_taxonomy\\_r95.tsv](https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/ar122_taxonomy_r95.tsv)) 和 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](https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_reps/ar122_marker_genes_reps_r95.tar.gz)); 细菌需要下载的文件有 bac120\_taxonomy\_r95.tsv ([https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/bac120\\_taxonomy\\_r95.tsv](https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/bac120_taxonomy_r95.tsv)) 和 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](https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_reps/bac120_marker_genes_reps_r95.tar.gz))。

以上下载的是数据库中 representatives 基因组的相关数据, 几乎覆盖目前所有已获得纯培养的主要类群, 质量较高。如果用户希望下载所有收录的数据, 可以从以下地址下载: [https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic\\_files\\_all/](https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/genomic_files_all/)。

下载完成后, 将其解压并放在 GTDBdata 文件夹中, 为了保证特异性 EasyCGTree 使用“faa”文件夹中的氨基酸序列进行分析, fna 文件夹可以删除。

- 3) 从 EasyCGTree 主文件夹中将 GetReferencFromGTDB.pl 脚本程序复制到 GTDBdata 文件夹中。
- 4) 假设我们想构建属于 *Bacillus* 的一些基因组的核心基因进化树。首先切

换工作目录为 GTDBdata，并输入：

```
perl GetReferencFromGTDB.pl bac120_taxonomy_r95.tsv faa g__Bacillus
```

命令行中“g”表示“属”，用户也可以指定从“种”到“门”的任意分类单元，只需要在命令行中分类单元的前面添加“s, g, f, o, c, p”（各分类级别首字母）等标记即可。需要注意的是字母代码和分类单元中间有两个“\_”，不是一个。然后下载的 GTDB 数据库中所有属于 *Bacillus* 属的基因组的 bac120 基因都将被提取出来，每个基因组生成一个独立的文件，并以种属名命名（如下图）。可使用文本编辑软件打开查看。

名称	修改日期
Alteromonas_A-lipolytica.fas	2020/9/15 14:46
Alteromonas_A-sp002335925.fas	2020/9/15 14:46
Alteromonas_A-sp002993365.fas	2020/9/15 14:46
Alteromonas_B-confluentis.fas	2020/9/15 14:46
Alteromonas_B-sp002729795.fas	2020/9/15 14:46
Alteromonas_C-pelagimontana.fas	2020/9/15 14:46
Alteromonas-abrolhosensis.fas	2020/9/15 14:46
Alteromonas-australica.fas	2020/9/15 14:46
Alteromonas-gracilis.fas	2020/9/15 14:46
Alteromonas-macleodii.fas	2020/9/15 14:46
Alteromonas-marina.fas	2020/9/15 14:46
Alteromonas-mediterranea.fas	2020/9/15 14:46
Alteromonas-nauphtalenivorans.fas	2020/9/15 14:46

同时也将生成一个以指定的分类单元命名的日志文件（例如：g\_\_Alteromonas\_log.csv），该文件记录了每个基因组提取到的基因数目，缺失的基因数目及列表等信息（对于基因组草图来讲，一些基因组会缺失 bac120/ar122 的部分基因，主要是测序和组装环节导致的）。该文件以逗号分隔不同信息，可以直接使用 Excel 打开。结果如下图所示。

Accession	Assembly	seq. no.	Taxonomy	Present	absent	absent list	PF00380.14	PF00410.14	PF00466.15	PF01025.14
teromonas-macleodii	GCF_000172635.2	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002335925	120	--	--	+	+	+	+
teromonas-sp002335925	GCF_000172635.2	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002335925	119	1	TIGR00065	+	+	+	+
teromonas-gracilis	GCF_0002993325.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002993365	120	--	--	+	+	+	+
teromonas-sp002993365	GCF_0002993325.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002993365	120	--	--	+	+	+	+
teromonas-marina	GCF_0002993365.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002993365	120	缺失的基	缺失的基	+	+	+	+
teromonas-mediterranea	GCF_0002993365.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002993365	120	--	--	+	+	+	+
teromonas-nauphtalenivorans	GCF_0002993365.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002993365	120	--	--	+	+	+	+
teromonas-australica	GCF_0002993365.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002993365	120	--	--	+	+	+	+
teromonas-sp002729795	GCF_0002729795.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002729795	111	9	PF00410.14 TIGR000967 TIGR01017 TIGR01021 TIGR0107	+	+	+	+
teromonas-sp001885075	GCF_001885075.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp001885075	120	--	--	+	+	+	+
teromonas_A-sp002335925	GCF_0002335925.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002335925	103	17	PF00410.14 PF00466.15 TIGR00059 TIGR00967 TIGR0101	+	+	+	+
teromonas_C-pelagimontana	GCF_000499975.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp00499975	119	1	TIGR00065	+	+	+	+
teromonas-australica	GCF_000730385.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp00730385	119	1	TIGR00065	+	+	+	+
teromonas-nauphtalenivorans	GCF_000730385.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp00730385	119	1	TIGR00065	+	+	+	+
teromonas-mediterranea	GCF_00020585.3	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp0020585	117	3	TIGR00065 TIGR01011 TIGR01063	+	+	+	+
teromonas_B-confluentis	GCF_001757105.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp001757105	120	--	--	+	+	+	+
teromonas-abrolhosensis	GCF_001953635.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp001953635	120	--	--	+	+	+	+
teromonas-trifoliorum	GCF_001561115.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp001561115	120	--	--	+	+	+	+
teromonas-sp002691125	GCF_002691125.1	1	d_Bacteria.p_Proteobacteria.c_Gammaproteobacteria.o_Alteromonadales.f_Alteromonas.g_Alteromonas_ssp002691125	91	29	PF00410.14 TIGR00066 TIGR00065 TIGR00088 TIGR0013	+	+	+	+

如果用户指定了一个较高级别的分类单元或者一个较大的属，可能会得到非常多的核心基因集，太多的核心基因集会拖慢 EasyCGTree 的运行速度，但是却无法明显提高分析的质量。一般来说，一个属包含的成员不太多的话，选取一个核心基因集即可，因为 bac120 和 ar122 在属内的变异是非常有限的。反过来讲，对于较大的



类群，在用户要分析的基因组彼此之间的分歧程度未知的情况下，选用更多的核心基因集可以进一步保证同源基因检索的准确性。基因集的多少和分析准确性成正相关（非线性），但是与运行时间成负相关。由于 EasyCGTree 运行速度非常快（比如，15 个基因集和 15 个基因组在大多数 PC 上均可以在 10-20 分钟内完成分析），用户可以把所涉及的属在 GTDB 数据库中所有的基因集都用来进行同源基因检索。如果基因集或基因组较多（>100），我们提出以下几条建议帮助用户选择基因集：选择的每一个基因集要尽可能保证包含 bac120/ar122 所有的 120/122 个基因；每个属至少有一个参比基因集；选择的基因集数量尽量在 M/60 和 M/15 之间（M 表示用户要分析的类群的基因组数量）；选中的基因集归属的物种在相应的属或者用户所使用的基因组的 16S rRNA 基因进化树上要尽量分散。

- 5) 选好参比基因集后，将它们放入同一个文件夹（例如 query），并移动到 EasyCGTree 主目录。

#### 4.1.2.2 其他核心基因集

用户可以准备自己的本地核心基因集，但是需要确保：这些基因中不包含旁系同源基因；每一条基因序列按照“XXX\_XXX\_XXX\_基因符号”的格式进行标注；每个基因集含有同样数目的基因；不同的基因集中的基因一一对应、互为直系同源基因。对于最后两条，希望用户能尽量保证，但是如果用户对 EasyCGTree 的原理和分析流程非常熟悉，可以根据需求进行调整。对于每条基因序列的标签，我们建议使用“属名\_种加词\_菌株号\_基因符号”的方式命名。分隔符号“\_”不允许出现在种属名、菌株号和基因符号内部，基因符号必须放在末尾，并且要保证不同基因集之间同一个基因使用同一个基因符号。否则 EasyCGTree 就无法正常运行。属名、种加词和菌株号也可以使用其他内容替换，顺序也不做强制要求。例如，1\_1\_1\_基因符号的格式也是许可的。

#### 4.1.3 其他核心基因集

**\*\*注意：这部分数据要求是蛋白序列。**

该部分数据只需要在进行 A3 模式（见 4.2）分析时提供。这些数据需要和 3.2 和 4.1.2 部分描述的数据保持同样的基因数目和基因目录。注意，这些数据集必须放在名为“Reference”的文件集中（强制要求，所以不必再命令行进行文件夹名的指定），也需要满足 4.1.2.2 描述的格式要求。每个基因集的文件名（除去扩展名）即是最终进化树上显示的内容。

## 4.2 正式运行 EasyCGTree

基因组文件夹和核心基因集准备好之后，**EasyCGTree** 主程序的运行就变得非常简单。确保这三个文件夹都在 **EasyCGTree** 主目录，切换到 **EasyCGTree**，然后输入以下命令即可。

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	prot		
A1	nucl	nucl*		not
	prot	prot	prot	requested
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.

**EasyCGTree** 的工作流程可以大致划分为以下 11 个步骤。运行结束之后，用户即可得到一个 Newick 格式的进化树文件，该文件根据基因组文件夹命名，例如 myGenome.tree。同时，程序运行过程中也会产生一些其他的文件和文件夹（见第 5 部分）。

A/A1/A2 设置的是不同的运行模式：

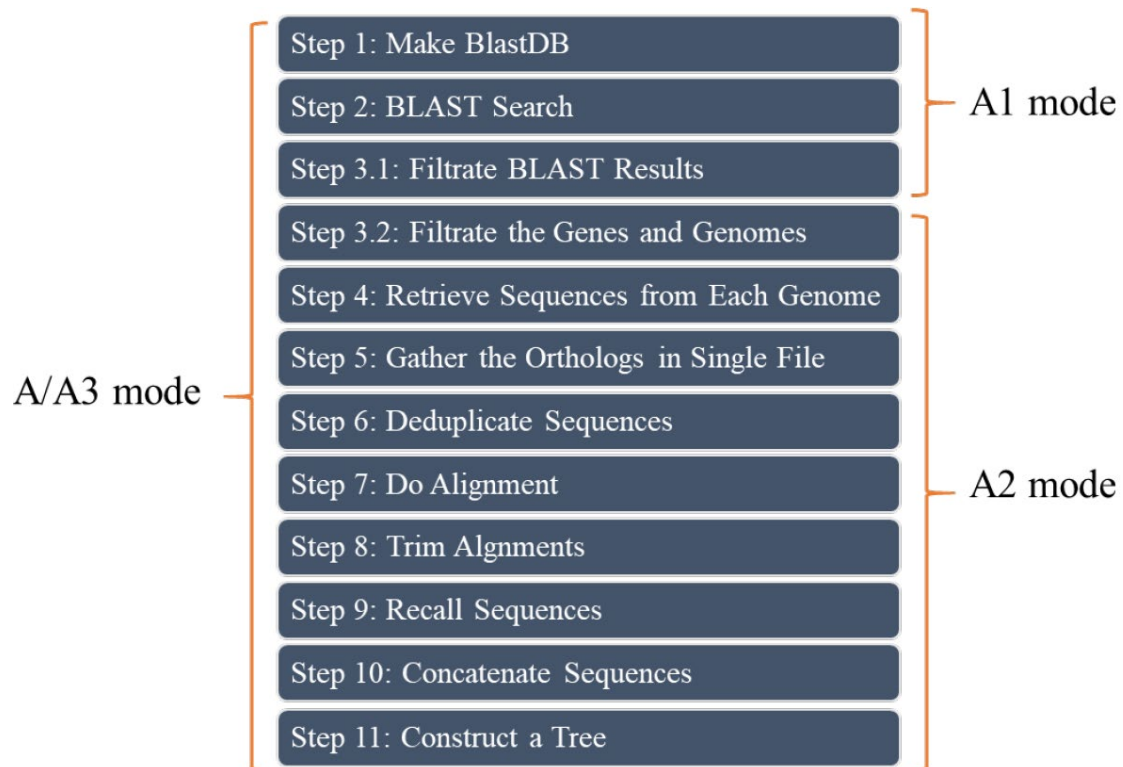
A 表示执行完整的分析流程，不使用 3.3 和 4.1.3 描述的数据（Reference 文件夹），产生所有的输出文件（参见 5）；

A1 表示进行前半部分分析，生成 TEM1-3 文件夹，query 序列文件和一个运行日志文件；

A2 表示只进行后半部分分析，生成 TEM4-9 文件夹、一个运行日志文件、各个基因组所有可用的核心基因的串联序列文件和一个进化树文件；

A3 表示执行完整的分析流程，需要使用 3.3 和 4.1.3 描述的数据（Reference 文件夹），产生所有的输出文件（参见 5），要求基因组数据为蛋白质组序列。

**\*\*A1 和 A2 类型的分析主要是便于用户根据 Step4 的结果及时调整参数 geneCutoff 和 genomeCutoff（见第 6 部分），因为 BLAST 检索分析在核心基因集较多的情况下需要花费大量的时间（参见 4.1.2.1）。**



Nucl/prot 用于指定最终用于构建进化树的序列类型（蛋白或核酸）。设定为 **nucl** 即要求 3.1 和 4.1.1 部分的基因组数据应该为核酸序列或所有 CDS 的氨基酸序列；设定为 **prot** 则要求所有的输入数据为蛋白序列。

软件运行起始界面如下：

```

C:\Windows\system32\cmd.exe
D:\EasyCGTree-Win>perl EasyCGTree.pl MyGenome query A
Smartmatch is experimental at EasyCGTree.pl line 308.
Smartmatch is experimental at EasyCGTree.pl line 340.
Smartmatch is experimental at EasyCGTree.pl line 513.
Smartmatch is experimental at EasyCGTree.pl line 809.
===== EasyCGTree =====
Version 0.0
by Dao-Feng Zhang (Hohai University)

#####Starting Informations #####
The user ordered a complete analysis of the pipeline!
Command: perl EasyCGTree.pl MyGenome query A
Job Started at: 13:22:21,12-9-2020

##### Parameters (Default) #####
tblastn= specify the location of the program tblastn, e.g. /share/bin
/:
num_threads=2 specify the number of threads used by the program tblastn;
blastIdentitycutoff=70 specify the cutoff for filterring the tblastn results;
geneCutoff=0.8 specify the cutoff for omitting low-prevalence gene;
genomeCutoff=0.8 specify the cutoff for omitting low-quality genomes;

##### Step 1: Make BlastDB #####
  
```

软件运行结束界面如下：

```

C:\Windows\system32\cmd.exe
9 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92
93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 11
9 120.
Step 10: I got 12 concatenated sequences!

##### Step 11: Construct a Tree #####

Step 11: I am making the tree.....FastTree Version 2.1.11 No SSE3
Alignment: MyGenome.concatenation.fas
Nucleotide distances: Jukes-Cantor Joins: balanced Support: SH-like 1000
Search: Normal +NNI +SPR (2 rounds range 10) +ML-NNI opt-each=1
TopHits: 1.00*sqrtN close=default refresh=0.80
ML Model: Generalized Time-Reversible, CAT approximation with 20 rate categories
Ignored unknown character X (seen 1 times)
Initial topology in 0.43 seconds
Refining topology: 14 rounds ME-NNIs, 2 rounds ME-SPRs, 7 rounds ML-NNIs
Total branch-length 1.174 after 7.04 sec, 1 of 10 splits
ML-NNI round 1: LogLk = -912176.241 NNIs 0 max delta 0.00 Time 11.98
GTR Frequencies: 0.1611 0.3447 0.3499 0.1443ep 11 of 12
GTR rates(ac ag at cg ct gt) 1.9906 2.7699 1.6742 4.4399 3.4209 1.0000
Switched to using 20 rate categories (CAT approximation)20 of 20
Rate categories were divided by 0.736 so that average rate = 1.0
CAT-based log-likelihoods may not be comparable across runs
Use -gamma for approximate but comparable Gamma(20) log-likelihoods
ML-NNI round 2: LogLk = -800062.914 NNIs 0 max delta 0.00 Time 33.24
Turning off heuristics for final round of ML NNIs (converged)
ML-NNI round 3: LogLk = -799874.012 NNIs 0 max delta 0.00 Time 38.99 (final)
Optimize all lengths: LogLk = -799873.543 Time 40.97
Total time: 62.01 seconds Unique: 12/12 Bad splits: 0/9

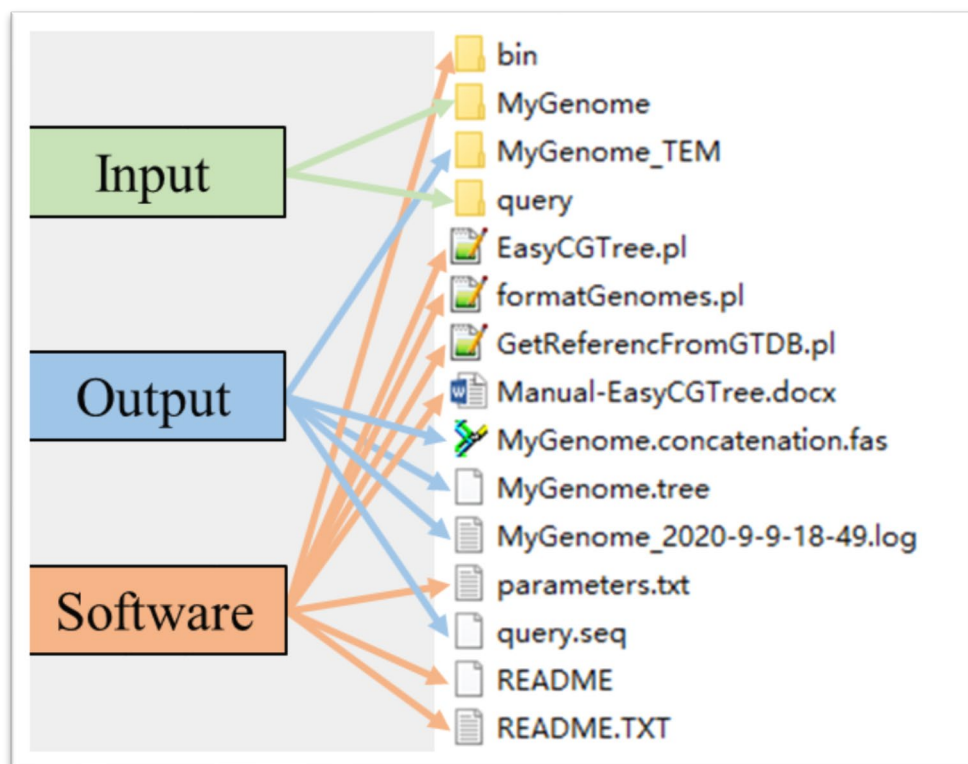
##### Ending Information #####

Job was finished at: 32:7:22,12-9-2020
Running time: 0-0-8-53 (day-hour-min-sec)

D:\EasyCGTree-Min>

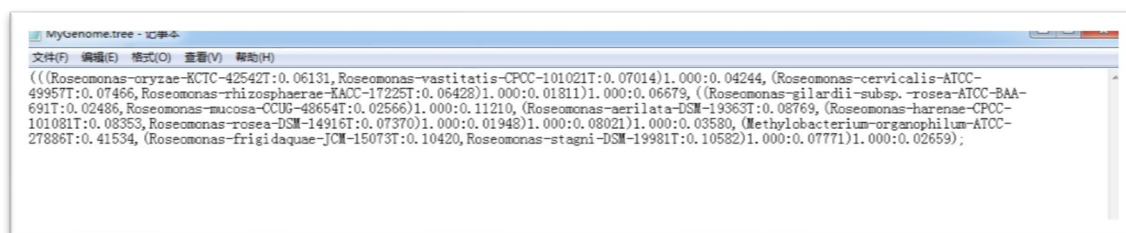
```

## 5 输出文件



Content of the **EasyCGTree** home directory after a run

1) **Newick** 格式的 **ML** 进化树文件。用户可以使用 FigTree, MEGA, iTOL 等软件来显示和美化进化树。使用“记事本”打开进化树文件，内容如下图所示。



2) 用于构建进化树的基因组的核心基因串联序列文件。该文件以 fasta 格式存放，也是最终使用 FastTree 构建进化树的输入文件。该文件根据基因组文件夹命名，以“.concatenation.fas”为后缀，根据 TEM9\_ReCallAln 文件夹中的数据生成（见 5.4j）。IQ-Tree、RaxML 等进化树构建软件也可以直接使用该文件。

3) 根据基因组文件夹和运行起始时间命名的日志文件。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。

4) 用于 BLAST 检索的序列文件。该文件根据包含核心基因集的文件夹命名，以“.seq”为后缀，由所有的核心基因集文件合并而来。

5) 以基因组文件夹命名、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。这些文件夹如下图所示。

名称	修改日期	类型
TEM1_blastDB	2020/9/12 21:58	文件夹
TEM2_blastOUT	2020/9/12 21:26	文件夹
TEM3_blastOUT_S	2020/9/12 21:26	文件夹
TEM4_GeneSeqs	2020/9/12 21:26	文件夹
TEM5_GeneCluster	2020/9/12 21:26	文件夹
TEM6_DedupGeneCluster	2020/9/12 21:26	文件夹
TEM7_Alignment	2020/9/12 22:06	文件夹
TEM8_AlnTrimmed	2020/9/12 22:06	文件夹
TEM9_ReCallAln	2020/9/12 22:06	文件夹
TEM60_DedupList	2020/9/12 21:26	文件夹

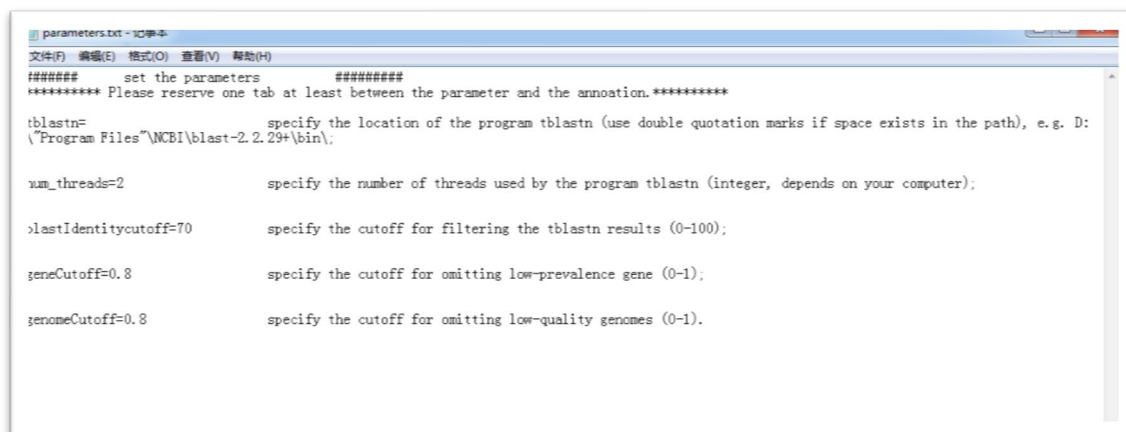
- TEM1\_blastDB:** 包含 makeblastdb 软件根据每个基因组构建的数据库文件。
- TEM2\_blastOUT:** 包含每个基因组的 BLAST 检索结果，以格式 6 存放。
- TEM3\_blastOUT\_S:** 从 TEM2\_blastOUT 筛选后得到的 BLAST 结果。如果使用多个核心基因集，EasyCGTree 将根据 bitscore 为每个基因筛选出最佳结果。同源性低于

parameters.txt 中设定的阈值（默认 50%）的话，将被舍弃（见第 6 部分）。

- d. **TEM4\_GeneSeqs**: 包含根据 TEM3\_blastOUT\_S 文件夹中的结果从每个基因组中提取的基因序列。
- e. **TEM5\_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. **TEM6\_DedupGeneCluster**: 包含去重复后的同源基因序列。这些文件是根据 TEM5\_GeneCluster 中的文件去重复后生成。代表基因序列和相应的基因名被保存在 TEM60\_DedupList 文件夹中。
- g. **TEM60\_DedupList**: 见 f. TEM6\_DedupGeneCluster 文件夹相关说明。
- h. **TEM7\_Alignment**: 包含根据 TEM6\_DedupGeneCluster 中的文件构建的 fasta 格式序列比对文件。
- i. **TEM8\_AlignTrimmed**: 包含根据 TEM7\_Alignment 中的文件修剪后的 fasta 格式序列比对文件。
- j. **TEM9\_ReCallAln**: 包含所有用于构建进化树基因组的修剪后的 fasta 格式比对文件。这些文件是根据 TEM8\_AlignTrimmed 和 TEM60\_DedupList 中对应的文件生成。

## 6 参数设定 (可选项)

**EasyCGTree** 中有 5 个可选的参数供用户设置，参数修改需要在 parameters.txt 文件中进行。该文本文件名可以任意修改，但是文件内的格式不能随意修改。如果用户通过该文件设定了参数，那么在运行 **EasyCGTree** 时就需要在命令行指定相应的参数文件。否则 **EasyCGTree** 将会按照默认参数进行执行，具体方式请参考 4.2 部分内容。Parameter.txt 文件内容如下图所示。



```
parameters.txt - 记事本
文件(F)  编辑(E)  格式(O)  查看(V)  帮助(H)
##### set the parameters #####
***** Please reserve one tab at least between the parameter and the annoation.*****

tblastn=          specify the location of the program tblastn (use double quotation marks if space exists in the path), e.g. D:
                  \Program Files\NCBI\blast-2.2.29\bin\;

num_threads=2     specify the number of threads used by the program tblastn (integer, depends on your computer);

slastIdentitycutoff=70    specify the cutoff for filtering the tblastn results (0-100);

geneCutoff=0.8      specify the cutoff for omitting low-prevalence gene (0-1);

genomeCutoff=0.8    specify the cutoff for omitting low-quality genomes (0-1).
```

- a. **tblastn**: 用于指定 blast+软件包的安装位置，例如: ./share/bin/。如果用户已经将 blast+ 加入了系统环境变量，那么也可以不设置该参数。
- b. **num\_threads**: 指定使用 tblastn 软件进行序列检索时使用的 CPU 线程数量。该参数要

求必须是整数，设置范围取决于用户的电脑配置。理论上越大越好。

**c. blastIdentitycutoff:** 指定筛选 BLAST 结果时使用的同源性阈值，设定范围 0-100。因为 bac120 和 ar122 均是由高度保守的看家基因组成，因此较低的阈值可能会匹配到非特异片段（如果基因组上相应的基因没有被很好地测序或组装），从而在构建进化树时引入错误的信号。如果用户为每一个用于构建进化树的基因组所在的属都准备了参比基因集，建议将阈值设置为 $\geq 50$ 。

**d. geneCutoff:** 设定舍弃缺失较高频次基因的阈值，理论设定范围 0-1。该参数默认值为 0.8，是指一个基因如果在 80%以上的基因组中被检索到，那么该基因将被用于后续分析。较低的阈值会保留更多的基因用于后续分析，构建进化树时缺失的基因将被补充为 gap 形式。

**e. genomeCutoff:** 指定舍弃含有较少基因的基因组的阈值，理论设定范围 0-1。默认参数为 0.8，是指一个基因组如果含有 80%以上最终使用的核心基因（见上一段），那么该基因组将被保留用于后续分析。较低的阈值将保留更多的基因组来构建进化树，但是最终保留的遗传信号将减少。

最后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，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（参见 3.2）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一种此消彼长的矛盾关系。例如，我们曾使用大约 500 个 *Alteromonadaceae* 基因组来构建 bac120 进化树，结果发现 bac120 中没有任何一个基因是存在于所有基因组中的。

后面的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，如果所有的基因组均具有较高的质量（比如都是完成图），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，因为所有的基因组都能满足这两个参数设定的要求。

## 7 硬件和系统要求

EasyCGTree 依赖的 Perl 语言环境、blast+、fastTree 等软件在 Windows 和 Linux 操作系统上均有相应的版本可以使用，因此 EasyCGTree 是一个跨平台、能在主流操作系统上完美运行的软件包。在 Mac OS 下，FastTree 安装较为繁琐



（本软件的核心思想是简化），因此 **EasyCGTree** 暂不提供相应的版本。

在硬件方面的要求，主要取决于基因组数据量和核心基因集的大小。如果用于分析的基因组数目小于 100 个，任何一台个人电脑即可在数小时以内完成分析。对于普通数据量，建议的配置为：双核以上 CPU；4G 以上内存；5G 以上空闲磁盘空间。对于基因组数量介于 100-1000 个，建议使用 Linux 操作系统和性能较强的个人电脑或服务器，因为 Windows 下 **EasyCGTree** 调用的第三方软件 muscle 和 FastTree 都只有单线程版本可用，运行时间会明显延长。对于基因组数量大于 1000 个，建议使用服务器运行 **EasyCGTree**。

## 8 运行时间

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

## 9 FAQ

收集提问中.....



## 10 参考文献

- 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**, *113*(10), 1523-1530. doi:10.1007/s10482-020-01460-z. (如果 EasyCGTree 为您提供了便利, 在描述软件的文章发表之前, 请您引用该文章)
- 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.