CLTree 1.0
User's Manual

Guanghong Zuo
ghzuo@ucas.ac.cn

March 24, 2022

Contents

1 Introduction

CLTree is a tool to annotate the phylogenetic tree by lineage and measure their differences in topology by Shannon entropy.

This manual is for the CLTree Standalone Version.

2 The Installation and Testing

CLTree is distributed by source code. It can be downloaded from Internet (https://github.com/ghzuo/Collapse). There are two ways to compile the source codes of CLTree: compile the source code by CMake; or use the docker image.

2.1 Normal Unix-like Mode

The program is implemented in C++. Some compile tools and libraries are required.

2.1.1 Preparation

- cmake $\geq$ 3.0
- g++ $\geq$ 7.0 or other compiler supporting C++11 standard
- require library: libz, nlohmann-json

2.1.2 Compile by CMake

1. unzip the package file and change into it
2. mkdir build and change into it
3. cmake .. and some options you wanted
4. make
5. make install (*option*)

2.1.3 Testing with Example

If this is the first time you use Collapse package, please go to the "example" folder. Run the collapse command to get an annotated phylogenetic tree and monophyly status by:

```
../build/bin/cltree
```

More detail of the command usage can be obtain by ‘-h’ option or read the follow sections.

2.2 Run Collapse in Container

The containers allows users run programs on both Windows and Linux/MacOS, and transfer the programs easily. To employ the container with Collapse, you should install `docker` at first. You can download docker free and reference from `https://docs.docker.com/install/` to how to install it. After install docker, basic usages for Collapse in container are shown blow:

1. Obtain image: You can build the Collapse docker image based on `Dockerfile` in the source code by command

   ```
   docker build -t="cltree-img" .
   ```

   Here option ‘-t’ set the image name. After build image, you can delete the dangling images for build by `docker image prune`. This will save much hard disk space. You can also download prebuilt Collapse image from internet by command:

   ```
   docker pull ghzuo/cltree .
   ```

   In this step, an image with Collapse programs will obtained.

2. Start container from image: run the follow command in the Collapse directory, i.e. the directory which include the ‘example’ directory of the Collapse

   ```
   docker run --rm -it -v $PWD/example:/root/data cltree-img
   ```

   In this step, you will enter the Collapse container, and the “example” folder of this project will be find in the “data” folder. Change path to the data folder, and run

   ```
   cltree
   ```

   You will get the result for eight genomes in the `list` file. You can change the path `'$PWD/example'` to your own data directory.

3. Exit and stop container: `exit` in docker terminal.

4. Run Collapse in a temporary container by one command without enter the container:

   ```
   cd <example> or <other data folder>
   docker run --rm -v $PWD:/root/data/ cltree-img cltree
   ```

5. More usage for `docker` can reference `https://docs.docker.com/`.

3 Workflow and Lineage Information Preparation

3.1 Scheme of CLTree and Basic Option

The scheme of program 'cltree' is shown in Figure 1. The main command is "cltree". Users can select different tokens (the read block in Figure 1) for different tasks. Before running the program, users must prepare two objects:

- a phylogenic tree (in newick format, e.g. Tree.nwk).
- the lineage information for leafs of the phylogenic tree.

The most easy way for prepare lineage information is utilize the NCBI Taxonomy database dump file package. It can be download from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz. With NCBI taxonomy dump file package. Users can obtain result by the command:

```
cltree -i Tree.nwk -d taxdump.tar.gz
```

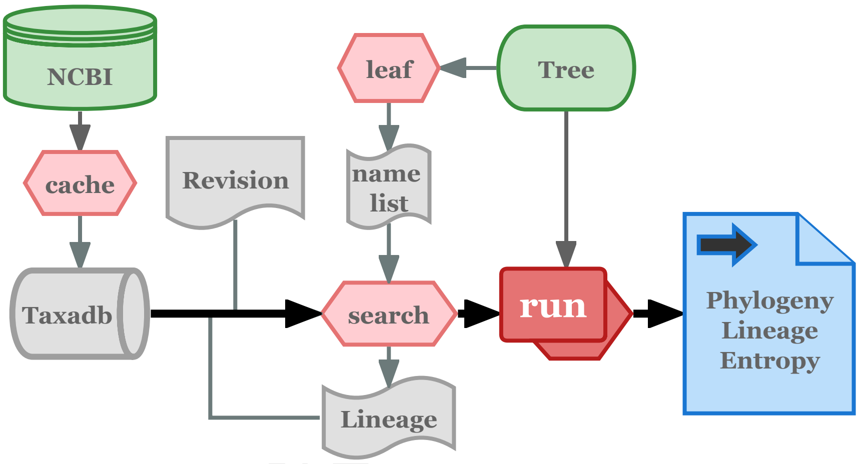

Figure 1: The Workflow of CLTree commands. The red blocks are the tasks of CLTree, and other blocks indicate the input/output files.

3.2 Lineage Information Preparation

Users can also edit the lineage manually or make batch modifying. We will describe them in the sections.The manually lineage file is shown in follow. There are two column in this file. The first column is the name of the species, and the lineage is shown in the second column. Here

3.3 Lineage String

- Example of lineage file:

```
Chlamydia_pecorum_...RefSeq <D>Bacteria<P>Chlamydiae...<S>Chlamydia_pecorum
Chlamydia_trachomatis...RefSeq <D>Bacteria<P>Chlamydiae...<S>Chlamydia_trachomatis
Chlamydia_pecorum...RefSeq  <D>Bacteria<P>Chlamydiae...<S>Chlamydia_pecorum
Chlamydia_trachomatis...RefSeq  <D>Bacteria<P>Chlamydiae...<S>Chlamydia_trachomati
Corynebacterium_jeikeium...  <D>Bacteria<P>Actinobacteria...<S>Corynebacterium_jei
Leptospira_interrogans...RefSeq <D>Bacteria<P>Spirochaetes...<S>Leptospira_interro
```

- Example for abbreviation of taxon ranks

A Example for batch revision

```
# Phylum B13 Firmicutes
<F>Erysipelotrichaceae<G>Eubacterium <F>Erysipelotrichaceae<G>Erysipelothrix   # CVTree
<P>Firmicutes<C>Erysipelotrichia <P>Tenericutes<C>Erysipelotrichia  # CVTree
<F>Unclassified<G>Exiguobacterium <F>Bacillaceae<G>Exiguobacterium   # CVTree
<F>Planococcaceae<G>Solibacillus <F>Bacillaceae<G>Solibacillus  #   NCBI taxonomy -> LE
```

4 Programs and Command-Line Options

4.1 Basic Usages

To obtain a basic usage of the program and tasks, you can used the “help” task, as “cltree help” or “cltree -h”. The output of the help information is shown in follow. And Users can type ”cltree help Task” to get more help information about other tasks.

```
cltree Task [options]
 Available Task:
   run       Annotate phylogenetic tree with taxonomy system
   cache     Make NCBI database cache from taxdump.tar.gz
   query     Query lineage from local NCBI taxonomy database
   search    Search lineage from lineage files and NCBI taxonomy
             database, and revised by the revision file
   leaf      Obtain the species name list of phylogenetic tree
   rank      Output the default rank names and abbreviations
   help      Provide the help information for <Task>
 [ -h ]      Display this information
```

4.2 Default Task

- run – Annotate phylogenetic Tree by Lineage and measure their Differences by Shannon Entropy. The "run" task is the default task of the command, so this task can be omitted except the "-h" option.

```
cltree run
 [ -D ./ ]               The work directory, default: ./
 [ -i Tree.nwk ]         Input newick tree, default: Tree.nwk
 [ -o collapsed ]        Output prefix name: default: collapsed
 [ -m <Revision.txt> ]   Lineage revision file for batch edit,
                         default: None
 [ -l Lineage.txt ]      Lineage file for leafs of tree,
                         default: Lineage.txt or Lineage.csv
 [ -d taxadb.gz ]        Taxa database file or directory,
                         default: taxadb.gz or taxdump.tar.gz
 [ -R <None> ]           List of rank names and abbreviations,
                         default: use the setting of program
 [ -r DKPCOFGS ]         Abbreviation of output taxon rank,
                         default: according to source
 [ -O <Outgroup> ]       Set the outgroup for the unroot tree.
                         default: None, rearranged by taxonomy
 [ -P ]                  Output prediction for undefined leafs
 [ -q ]                  Run command in quiet mode
 [ -h ]                  Display this information
```

4.3 Lineage Tasks

You can also obtain the leaf lineages of the phylogenetic tree by other task step by step and review them manually, i.e. `leaf, cache, rank, query, search`. The function and usage of these three programs are show blow:

- leaf – Obtain the name list of phylogenetic tree

```
cltree leaf
 [ -i Tree.nwk ]       Input tree file, default: Tree.nwk
 [ -o namelist.txt ]   Output name list, default: name.list
 [ -q ]                Run command in quiet mode
```

```
  [ -h ]                Display this information
```

- cache – Package the dump files of NCBI taxonomy database as cache to speed up query lineage from database.

```
cltree cache
 [ -d taxdump.tar.gz ]  NCBI taxon dumpfile directory, default: taxd
 [ -o taxadb.gz ]       Packaged taxon database, default: taxadb.gz
 [ -q ]                 Run command in quiet mode
 [ -h ]                 Display this information
```

- search – Search the lineage of the genome

```
cltree search
 [ -i namelist.txt ]    Input name list, ':N' after the file name
                        select the N column of the file
                        default: first column of namelist.txt
 [ -o Lineage.csv ]     Output lineage file, default: lineage.csv
 [ -m <Revision.txt> ]  Lineage revise file for batch edit,
                        default: None
 [ -l Lineage.txt ]     Lineage file for leafs of tree,
                        default: Lineage.txt
 [ -d taxadb.gz ]       Taxa database file or directory,
                        default: taxadb.gz or taxdump.tar.gz
 [ -R <None> ]          List file for rank names and abbreviations,
                        default: use the setting of program
 [ -r <DKPCOFGS> ]      Set output taxon rank by abbreviations,
                        default: according to source

 [ -q ]                 Run command in quiet mode
 [ -h ]                 Display this information
```

4.4 NCBI Taxonomy Database Tasks

superkingdom	D
kingdom	K
subkingdom	k
superphylum	Q
phylum	P
subphylum	p

```
superclass      L
class           C
subclass        c
superorder      W
order           O
suborder        o
family          F
subfamily       f
genus           G
subgenus        g
species         S
subspecies      s
```

- rank – Output an example of example of the list of taxon rank names and abbreviations.

  ```
  cltree rank
   [ -o ranklist.txt ]  Output name list, default: ranklist.txt
   [ -q ]               Run command in quiet mode
   [ -h ]               Display this information
  ```

- query – Query the lineage of the genome from NCBI Taxonomy database dump files or the database cache

  ```
  cltree query
   [ -I <Taxon ID> ]     Query a taxon id
   [ -N <Taxon Name> ]   Query a taxon name
   [ -i namelist.txt ]   The query list file default: name.list
   [ -d taxadb.gz ]      The dump of NCBI taxonomy database
   [ -o Lineage.txt ]    Output file, default: Lineage.txt
   [ -R <None> ]         List file for rank names and abbreviations,
                         default: use the setting of program
   [ -r <DKPCOFGS> ]     Set output taxon rank by abbreviations,
                         default: same to the source
   [ -H ]                Don't output missing items
   [ -q ]                Run command in quiet mode
   [ -h ]                Display this information
  ```

5 Algorithm

5.1 Annotate the Phylogeny Tree

5.2 Measure Difference by Shannon Entropy

6 Citing Collapse in a Publication

Please cite:

1. Guanghong Zuo (2021) Collapse: Annotate Phylogenetic Tree by Lineage and Measure their Differences by Shannon Entropy. `in preparation`.

References

Zuo, G. (2021) Collapse: Annotate Phylogenetic Tree by Lineage and Measure their Differences by Shannon Entropy. *Journal*, `in submission`.