ezTree Tutorial v0.1

Yu-Wei Wu <yuwei.wu@tmu.edu.tw>

Graduate Institute of Biomedical Informatics, Taipei Medical University

In this tutorial I provide a brief guide on how to use ezTree generating a phylogenetic tree for a set of genomes. Please first setup ezTree and all related software following the instruction of the ezTree **README.txt** file. ezTree will attempt to download the Pfam database during the first use of ezTree.

When the ezTree script was executed without any parameters, the following screen will appear to indicate the parameter inputs.

```
root@cd866d9d8c2c:/home/yuwei.wu/ezTree# ./ezTree
Please input both a list file consisting of genomes and a output header.

ezTree - building phylogenetic trees for a set of genomes
Usage:
  ezTree
    -list (list file of genomes)
    -out (output header)
   (Either -list or -dir is required for running ezTree)

   (Other parameters)
    [-thread (thread num; default 4)]
    [-evalue (evalue for HMMER; default 1e-10)]

  Please read README file for more details.
```

Now let's run a very simple example. There are dozens of Proteobacteria genome that are downloaded along with ezTree and put in the "test_example" folder. Enter the folder and type "ls." You will see its content as follows.

```
root@cd866d9d8c2c:/home/yuwei.wu/ezTree# cd test_example/
root@cd866d9d8c2c:/home/yuwei.wu/ezTree/test_example# ls
Brevundimonas_abyssalis_TAR-001.fasta          Escherichia_coli_JJ1886.fasta              Shewanella_baltica_OS678.fasta
Brevundimonas_diminuta_ATCC_11568.fasta        Escherichia_coli_O157_H7_Sakai.fasta       Shewanella_putrefaciens_200.fasta
Brevundimonas_naejangsanensis_DSM_23858.fasta  Morganella_morganii_KT.fasta               Yersinia_aldovae_IP07632.fasta
Buchnera_aphidicola_Bp.fasta                   Morganella_psychrotolerans_GCSL-Mp3.fasta  Yersinia_aleksiciae_159.fasta
Buchnera_aphidicola_Sg.fasta                   README.txt                                 list_class
Erwinia_amylovora_CFBP1430.fasta               Salmonella_bongori_N268-08.fasta           list_family
Erwinia_billingiae_Eb661.fasta                 Salmonella_bongori_NCTC_12419.fasta        list_genus
Escherichia_albertii_EC06-170.fasta            Salmonella_enterica_serovar_Albany.fasta   list_order
Escherichia_albertii_KF1.fasta                 Salmonella_enterica_serovar_Dublin.fasta   list_phylum
Escherichia_coli_E24377A.fasta                 Shewanella_algae_MARS_14.fasta             list_species
root@cd866d9d8c2c:/home/yuwei.wu/ezTree/test_example#
```

There are totally 23 genomes in fasta format and six list files, namely "list_phylum," "list_class," …, "list_species." The content of each list is simply a list of genomes. For example, the file "list_species" consists of three *E. coli* genomes.

```
root@cd866d9d8c2c:/home/yuwei.wu/ezTree/test_example# head list_species
Escherichia_coli_E24377A.fasta
Escherichia_coli_JJ1886.fasta
Escherichia_coli_O157_H7_Sakai.fasta
root@cd866d9d8c2c:/home/yuwei.wu/ezTree/test_example#
```

Generating list file is very easy—simply put all genomes of interest in a folder and use “ls *.fasta > listfile” to get the list (assuming that all genomes end in .fasta). No need to type in all filenames manually. One can also input protein sequences instead of whole genome into ezTree; in this case ezTree will skip the protein-prediction step for the files with protein sequences.

Now we can run ezTree by indicating the list file and the output header. I use list_species as an example to showcase how ezTree runs.

```
# /home/yuwei.wu/ezTree/ezTree -list list_species -out list_species.out
```

The above command will read in the genomes defined in the list file, predict genes from the genomes, annotate the functional profiles for the genes, and get single-copy marker genes for building the phylogenetic tree. The screenshot below is the running result of ezTree for the list file “list_species.” Note that the download of PFAM database will only be performed once.

PFAM Download

```
Did not find Pfam database--possibly first-time use of ezTree.
Trying to download pfam data file from ftp://ftp.ebi.ac.uk/. Please be patient in this process...
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  245M  100  245M    0     0  4198k      0  0:00:59  0:00:59 --:--:-- 2423k
Predicting genes for Escherichia_coli_E24377A.fasta
Predicting genes for Escherichia_coli_JJ1886.fasta
Predicting genes for Escherichia_coli_O157_H7_Sakai.fasta
344 seconds spent on predicting genes.
Running hmmscan on Escherichia_coli_O157_H7_Sakai.fasta
Running hmmscan on Escherichia_coli_JJ1886.fasta
Running hmmscan on Escherichia_coli_E24377A.fasta
11281 seconds spent on mapping genes against PFAM.
Start Processing [Escherichia_coli_O157_H7_Sakai.fasta]
Start Processing [Escherichia_coli_JJ1886.fasta] remaining number of PFAMs: 1220
Start Processing [Escherichia_coli_E24377A.fasta] remaining number of PFAMs: 1161
Identified 1161 marker genes for the genomes.
60 seconds spent on processing mapping results.
25 seconds spent on making tree.
root@cd866d9d8c2c:/home/yuwei.wu/ezTree/test_example#
```

The ezTree pipeline will generate totally three files and one working directory for the input list file, as shown in the following screensot.

```
root@cd866d9d8c2c:/home/yuwei.wu/ezTree/test_example# ls -l list_species*
-rw-r--r-- 1 root root      98 Apr 13 07:11 list_species
-rw-r--r-- 1 root root 1101002 Apr 13 10:39 list_species.out.aln
-rw-r--r-- 1 root root     125 Apr 13 10:39 list_species.out.nwk
-rw-r--r-- 1 root root   35879 Apr 13 10:39 list_species.out.pfam
```

These files are:

1. .aln: the concatenated alignment file of all marker proteins.
2. .nwk: the Newick tree for the genomes defined in the list file
3. .pfam: the identified single copy marker genes in terms of PFAM
4. .work directory: this is the work directory of ezTree. If one needs to re-run ezTree, simply input the same “-out” parameter. ezTree will locate all temporary files in the

working directory and calculate the results in a whim. Below is a screenshot for re-running ezTree on the “list_species” list file.