Molecular Microbiology 2024

Bioinformatics Protocol 9

Phylogenomics: GTDBtk for making a concatenated protein alignment tree, FastANI/MiGA for classification of genomes, and AutoMLST for classification and tree in one online service

This protocol covers the use of FastANI for rapid estimation of average nucleotide identity, GTDBtk (Quick Concatenated Ortholog Alignment Tree) and IQ-TREE for building whole-genome phylogenetic trees, and the use of the AutoMLST. To view your finished tree you will need to install Figtree (http://tree.bio.ed.ac.uk/software/figtree/) on your own computer.

1. Log into your account on the `dnaseq1a.bio.au.dk` server.
2. Obtain `fasta` files (files ending in `.fna`) for the scaffolds of the genomes you wish to include in your phylogenetic tree and fastANI analysis. For the most part, you can follow the instructions in protocol 4 to obtain these genomes and place them into a directory in your home directory called `relative_genomes_fna`. However, this time you need to retrieve the fasta scaffold files (ending in `.fna` or `.fna.gz`) rather than the Genbank files. You can also include `.fna` files from other sources (such as IMG) if you want to. If you haven't already done so, you can BLAST the 16S rRNA gene sequence using the NCBI online `blastn` service to find the best related genomes to your genome. Include an outgroup genome - something from a neighbouring genus that you can use to find the root of your phylogenetic tree.
3. Once the genome files are all in your `relative_genomes_fna` directory, unzip any gzipped files that are amongst them. Make sure all files end with the file extension `.fna` - this is how `GTDBtk` will know which files to use as input files. Return to your home directory and run the following script (custom written for this class) to rename your files to something easier to understand than the cryptic IDs that NCBI has given the files. This script takes the species name for the genome out of the first line of the fasta file and uses it as the file name.

   ```
   cd ~
   rename_fna_files.py relative_genomes_fna
   ```

4. Copy the fasta decontaminated scaffolds file for your genome of interest (the one you sequenced) into the `relative_genomes` directory. If it doesn't already have a `.fna` file extension, change its filename now using `mv`.
5. Change back into the `relative_genomes` directory and make a text file of all the `.fna` filenames, one per line.

   ```
   cd relative_genomes
   ls -1 *.fna > relative_genome_filenames.txt
   ```

6. Run `fastANI` to quickly determine the average nucleotide identity between your genome of interest and your relative genomes.

```
fastANI --ql relative_genome_filenames.txt \
--rl relative_genome_filenames.txt \
-o fastANI_results.txt \
--matrix
```

   - `fastANI` - the name of the program
   - `-ql relative_genome_filenames.txt` - the filename for the text file you made in the previous step. A list of `fna` files, one on each line - your "query" genomes.
   - `--rl relative_genome_filenames.txt` - the filename for the text file you made in the previous step. A list of `fna` files, one on each line - your "reference" genomes.
   - `-o fastANI_results.txt` - the output file name.
   - `--matrix` - This generates a full matrix output file for pairwise comparisons of all genomes.

7. Open the `.matrix` output file in a text editor - you will see a table of pairwise comparisons. It might be easiest to view if copied into Excel and the list of names pasted laterally so you can see which ANI value (in %) refers to which set of two genomes. Does your genome have >95% identity with any others? How about other species within the genus you're investigating - do multiple strains from within a species have >95% ANI values? If you were to rearrange the classification of this genus based purely on genomic information, would you make any changes?

8. Change back into your home directory and then run `GTDBtk` to identify single-copy orthologues (phylogenetic marker genes) common to all the genomes in the `relative_genomes` directory.

```
cd ~
gtdbtk identify \
--extension fna \
--genome_dir genomes \
--out_dir gtdb_identify \
--cpus 3
```

   - `gtdbtk identify` - This is the name of the program we are using (`gtdbtk`) and the chosen workflow (`identify`, to identify phylogenetic marker genes).
   - `--extension fna` - This specifies that all the `fasta` files in the directory end with the `.fna` file extension. If you made sure all the files are named as specified in step 3, you shouldn't have to change this.
   - `--genome_dir relative_genomes` - This specifies the name of the directory containing all the genomes you plan to include in your phylogenetic tree. If you are trying to make several different phylogenies with different sets of genomes, you may need to change this name each time.
   - `--out_dir gtdb_identify` - This specifies the name of the output directory to be used as input in the next step.
   - `--cpus 3` - This specifies the number of cpus allocated to this job.

9. Now use the identified phylogenetic marker genes to produce a concatenated alignment of protein sequences from those genes using `GTDBtk`.

```
gtdbtk align \
--identify_dir gtdb_identify \
--out_dir gtdb_align \
--cpus 3
```

   - `gtdbtk align` - This is the name of the program we are using (`gtdbtk`) and the chosen workflow (`align`, to align the phylogenetic marker genes identified in the last step).
   - `--identify_dir gtdb_identify` - This specifies the output directory from the previous step (identification), which becomes the input for this step (alignment).
   - `--out_dir gtdb_align` - This specifies the output directory for the alignment.
   - `--cpus 3` - This specifies the number of cpus allocated to this job.

10. Now use the output from `GTDBtk` (the concatenated ortholog protein alignment) to build a rough, preliminary phylogenetic tree using `FastTree`. The file needs to be `gunzip`ped first.

```
gunzip gtdb_align/align/gtdbtk.bac120.user_msa.fasta.gz
fasttree gtdb_align/align/gtdbtk.bac120.user_msa.fasta > gtdbtk_fasttree_phylogeny.tr
```

    - `gtdb_align/align/gtdbtk.bac120.user_msa.fasta` - This is the input file - the concatenated protein alignment from `GTDBtk`.
    - `gtdbtk_fasttree_phylogeny.tree` - This is the output file name. If you build multiple trees you should change this for each tree.

11. Download the file `gtdbtk_fasttree_phylogeny.tree` to your own computer and open it in FigTree. It may ask you what the branch labels are when you open it - ignore this. How does the tree look? Try rooting it on a genome you designate as an outgroup. Do you want to add or remove some genomes? Repeat steps 2 and 8–10 until you have a tree you are happy with.

12. Now use the output from `GTDBtk` (the concatenated ortholog protein alignment) to build a more accurate phylogenetic tree using `IQ-TREE`. `IQ-TREE` tests many different evolutionary models and finds the best fit, then uses bootstrapping to determine the statistical likelihood of the branching. However, expect that it will take several days to run (thus the use of the `screen` prefix).

```
screen -L iqtree \
-s gtdb_align/align/gtdbtk.bac120.user_msa.fasta \
-m TESTNEW \
-b 100 \
-nt 3
```

    - `iqtree` - The name of the program for building a phylogenetic tree based on a multiple sequence alignment.
    - `-s gtdb_align/align/gtdbtk.bac120.user_msa.fasta` - The fasta file containing the multiple sequence alignment of concatenated

phylogenetic marker gene protein sequences generated by `gtdbtk`.

- `-m TESTNEW` - A parameter to tell `iqtree` to test various phylogenetic models and figure out the best one - for more information see *L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, and B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol., 32:268-274. DOI: 10.1093/molbev/msu300.*
- `-b 1000` - The number of bootstrap replicates to test the likelihood of the reconstructed phylogeny. 1000 is typically used in most publications, but 100 is enough here - a smaller number will run faster, and a higher number will give a more precise probabilities.
- `-nt 3` - The number of threads for IQ-TREE to use (3 in this case).

13. Download the output file from `IQ-TREE`, `gtdbtk.bac120.user_msa.fasta.treefile` (in the `gtdb_align` directory), to your own computer and open it in FigTree. It may ask you what the branch labels are when you open it - these are bootstrap numbers.

14. **AutoMLST**: https://automlst.ziemertlab.com/. Go to this website and click *Start Analysis*. Click *Browse...* and upload your genome `.fna` file, then click *Upload* to the right. Enter your email address and give the analysis job a name that makes sense to you, then click *Submit Job*. **AutoMLST** will run MASH ANI analysis (similar to FastANI) to find closely related genomes then build a concatenated protein alignment phylogeny based on those relative genomes (basically an automated version of this protocol).