Software Manual: dusti

Version: 0.0.1

Megan L. Smith

January 1, 2025

Contents

1 Installation

Install the package with:

```
pip install git+https://github.com/SmithLabBio/dusti.git
```

2 Input files

2.1 Alignments

The primary input is the path to a folder with a set of alignments from gene families. These can be either in fasta or phylip format. Fasta files should be named with the suffix '.fa' or '.fasta', while phylip files should be named with the suffic '.phy'.

2.2 Mapping File

A file mapping gene names to species is also required. This file should be a text file and delimited either by tabs or spaces. The first column contains the species name, and the second column contains information regarding gene names.

There are three options for providing mapping information:

- **Exact mapping:** For each gene name in the input gene trees, include the full gene name and the corresponding species name.

  Example: A A_0_0

  Result: Maps A_0_0 to species A.

- **Prefix mapping:** Map gene names to species names based on a prefix. The first understore in the gene names will be used to distinguish the prefix from the remainder of the name.

  Example: A A*

  Result: Maps A_* to species A, where * can be any set of characters.

- **Suffix mapping:** Map gene names to species names based on a prefix. The final underscore in the gene names will be used to distinguish the suffix from the remainder of the name.

  Example: A *A

  Result: Maps *_A to species A, where * can be any set of characters.

An example mapping file is available in ./example/s_map.txt

3 Running dusti

The user must supply several commands to run dusti:

- -i (–input): a path to a directory of alignments

- -a (–map): a path to the mapping file
- -o (–output_directory): a path to an output directory
- –qfm: a path to wQFM program
- -q (–max_quartets): an integer specifying the maximum number of quartets to use (default: all)
- -s (–seed): an integer specifying a random number seed to use when sampling quartets (default: None)
- –force: used to overwrite an existing output directory (default: False)
- –svd: perform inference based on SVD (default: False)
- –parsimony: perform inference based on parsimony (default: False)

To run dusti on the example files:

```
dusti --input example/alignments --map example/s_map.txt -o example_results/
--qfm  /Documents/programs/wQFM-2020/wQFM-v1.4.jar --force
```

4 Output

Dusti will output two files per analysis:

- SVD analysis:
  - svdquartets.trees: quartets inferred using the SVD method
  - svd.tre: species tree inferred using SVD method and quartet puzzling
- Parsimony analysis:
  - parsimonyquartets.trees: quartets inferred using the parsimony method
  - parsimony.tre: species tree inferred using parsimony method and quartet puzzling