gfmix: Phylogenetic analyses using the site-and-branch-heterogeneous GFmix model

Version 1.2

November 22, 2024

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

Introduction

The main program `gfmix` fits the models orginally described in Muñoz Gómez et al. (2021) but also used in Baker et al. (2024). It takes as input a sequence file, a treefile with edge-lengths, a file giving the split of taxa at the root, a frequency mixture file and the output file from a run of the phylogenetic software IQ-TREE (Nguyen et al. 2015). Installation information is available towards the end of this document.

The program `gfmix` can be run at the command line with

```
$ gfmix -s seqfile -t treefile -i iqtreefile -f frfile -r rootfile [-fclass FYMINK] [-gclass GARP]
```

A brief description of the options and output is given below. Additional information is available in subsequent sections.

`-s seqfile`: The input sequence file. NOTE: The format must comply with PHYLIP conventions. See below for additional details.

`-t treefile`: A Newick tree file with edge-lengths.

`-i iqtreefile`: The output file from IQ-TREE with extension `.iqtree`.

`-f frfile`: A file with the frequencies for each frequency class as rows.

`-r rootfile`: A file with the integer labels $0, 1 \ldots$ of taxa on one side of the root split. The integer labels should match the ordering of sequences in `seqfile`.

`-fclass string`: Optional. A string giving the amino acids in the F class. Default is 'FYMINK'

`-gclass string`: Optional. A string giving the amino acids in the G class. Default is 'GARP'

The output is the log likelihood for the model.

Additional Information about Program Usage

Input

`-s seqfile`: The file should conform to the requirements of the PHYLIP package (Felsenstein, 1989, 2004). Sequence names should be 10 characters long and padded by blanks. The names should match the names used in the input treefile. Input can be either interleaved or sequential with one caveat: The lines 2 through $m + 2$, where m is the number of taxa, must contain the name of taxa followed by sequence data. For instance the start of a sequence file might be

```
6 3414
Homsa      ANLLLLIVPI LI...
Phovi      INIISLIIPI LL...
...
```

but not

```
6 3414
Homsa
ANLLLLIVPI LI...
Phovi
INIISLIIPI LL...
...
```

which would be allowed under the sequential format by PHYLIP.

Additional information is available at

`http://evolution.genetics.washington.edu/phylip/doc/sequence.html`

R functions have been included in a file, `convert2integer.q`, that can be used to convert names in some sequence files and corresponding tree files to integer labels.

`-t treefile`: The tree should conform to the Newick standard. A discussion of this standard as implemented in PHYLIP is given at

`http://evolution.genetics.washington.edu/phylip/newicktree.html`

and a more formal description is available at

`http://evolution.genetics.washington.edu/phylip/newick_doc.html`

`-i iqtreefile`: `gfmix` assumes that IQ-TREE has been run with a class-frequency mixture with `+F` option and gamma rate variation. The software currently fits with a `LG` exchangeability matrix, so it is best to run IQ-TREE with this option as well.

If IQ-TREE is run with sequence file `seqfile`, it will create a file `seqfile.iqtree` in the directory where `seqfile` is located. `gfmix` uses this file to get the weights of the class-frequency mixture and α parameter for the gamma rate variation model.

`-f frfile`: A file with the frequencies for each frequency class as rows, not including the `+F` frequencies which are determined within `gfmix`. These frequencies should match the frequency classes in the call to IQ-TREE.

In Muñoz Gómez et al. (2021), the MAM60 model was the main frequency class model. MAM60 frequency classes are determined from the sequence data at hand using the methods of Susko, Lincker and Roger (2018). The program `mammal` can be used to get these. It is available at

`https://www.mathstat.dal.ca/~tsusko/`

Files with the frequencies for the C-series models are included in the packaged software. So if IQ-TREE is called with option `-m LG+C20+F+G`, `C20.aafreq.dat` can be used as `frfile`; include the full pathname.

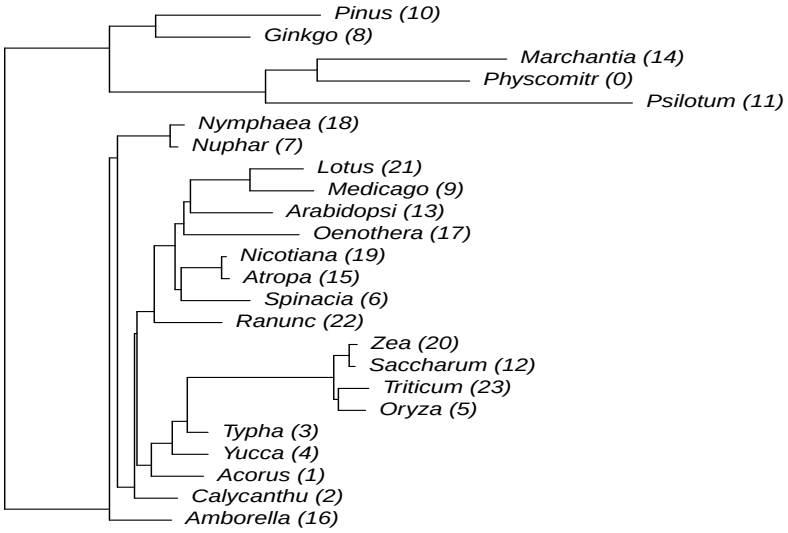

Figure 1: An example tree to illustrate rooting.

The software does not currently deal well with exact 0 frequencies. To adjust for this, for each class, all frequencies that are less than `1.0e-8` are set to `1.0e-8` and then the frequencies for the class are rescaled to sum to 1.

`-r rootfile`: A file with the integer labels $0, 1 \ldots$ of taxa on one side of the root split. The integer labels should match the ordering of sequences in `seqfile`. So 0 corresponds to the first sequence in the file, 1 to the second sequence and so forth.

For example, the sequence file used to estimate the tree in Figure 1 was of the form

```
24 15688
Physcomitr MVKI--RPDE...
Acorus     MATL--RADE...
...
Triticum   MATL--RVDE...
```

So the inter label 0 corresponds to the sequence `Physcomitr`, the label 1 corresponds to `Acorus` and label 23 to `Triticum`; labels are indicated in the figure. A valid `rootfile` for this tree would be

```
10 8 14 0 11
```

because those taxa are on one side of root.

NOTE: The program creates a number of files prefixed with `tmp` which are removed upon conclusion. If you have files of the form `tmp.*` in the directory where `gfmix` is run, they should be renamed or moved.

System Requirements and Installation

Requirements and Installation of External Packages

The main program `gfmix` is an R language script file that effectively pastes together results from a number of smaller programs, some of which were written in R and some in C. To install the package you will need a C compiler and a working installation of the R statistical package

The source code has been compiled and tested using the `gcc` compiler on linux operating systems. While the program has not been tested on another platform, it should compile on other operating systems. On Mac OS, to install `gcc`, bring up a terminal and type

```
$ xcode-select --install
```

The program uses output files from IQ-TREE. It is possible that these files may have a different format in some past or future versions of IQ-TREE. Versions of IQ-TREE considered in testing were 2.1.4beta and 1.5.5.

Installation

1. Download and unpack the software

   ```
   $ tar zxf garp.tar.gz
   ```

   This will create a directory `garp` that contains the source code.

2. Change directories to `garp` and create the main program files with the make command

   ```
   $ cd
   $ make
   $ chmod a+x gfmix
   ```

   The default installation assumes the `gcc` compiler is available. To use a different compiler, change the variable `CC` in `Makefile`.

3. Copy the program files

   ```
   treecns rert alpha_est_mix_rt gfmix
   ```

   to a location in your `PATH` or to a known directory. If the directory that these files are copied to is not in your `PATH`, you should change the line `bindir <- ""` in the file `gfmix` to

   ```
   bindir <- "dir_with_files/"
   ```

   where `dir_with_files` is the name of the directory that the files above have been copied to.

4. Copy the C-series frequencies

   ```
   C10.aafreq..dat, ..., C60.asfreq.dat
   ```

   to a known directory. These can be used with the `-f` option to fit with a C-series model. For instance if they were stored in `dir_with_files`, `-f dir_with_files/C20.aaafreq.dat` would fit yusing the C20 model.

5. The source code and directory can be removed:

```
$ cd ../
$ rm -rf garp.tar.gz garp/
```

References

Baker, B.A, Gutiérrez-Preciado, A. and Rodiriguez del Rio, A., McCarthy, C.G.P., López-Garcia, P., Huerta-Cepas, J., Susko, E., Roger, A.J., Eme, L. and Moreira, D. (2024). Expanded phylogeny of extremely halophilic archaea shows multiple independent adaptations to hypersaline environments. *Nature Microbiology*, **9**:964–975.

Muñoz Gómez, S., Susko, E., Williamson, K., Eme, L., Slamovitz, C.H., Moreira, D. Purificación, L. and Roger, A.J. (2022). Site-and-branch-heterogeneous analyses of an expanded dataset favor mitochondria as sister to known Alphaproteobacteria. *Nature Ecology & Evolution.* **6**:253–262.

Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Molecular Biology and Evolution.* **32**:268—274.

Susko, E., Lincker, L. and Roger, A.J. (2018). Accelerated Estimation of Frequency Classes in Site-heterogeneous Profile Mixture Models. *Molecular Biology and Evolution.* **35**:1266–1283.