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

Version 1.0

May 28, 2026

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

Introduction

The main program `gfmixf` fits the model described in McCarthy et al. (2026). The model is a profile mixture model that allows subsets of amino acid to show variation over edges. Given that a site has profile class c and rate r, evolution along edge e of length t_e, is according to a Markov process from time 0 to rt_e where the process has rate matrix

$$Q_{ij}^{(ce)} \propto \begin{cases} \gamma_G^{(e)} S_{ij} \pi_j^{(c)} & \text{if } j \in G \\ \gamma_F^{(e)} S_{ij} \pi_j^{(c)} & \text{if } j \in F \\ S_{ij} \pi_j^{(c)} & \text{otherwise,} \end{cases}$$

S_{ij} is the LG exchangeability matrix and $\pi_j^{(c)}$ is the stationary frequency for class c in the mixture.

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

```
$ gfmixf -c nclass -f freqfile -t treefile -r rootfile [-s seqfile] [-l seqfile]
       [-g gpfile] [-w in_wtfile] [-a alpha] [-e exchangeability]
       [-o deriv] [-p iprint] [-y paramfile] [-b numthreads]
```

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

`-c nclass`: the number of classes in the profile mixture

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

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

`-r rootfile`: A file with the names of taxa on one side of the root split.

`-s seqfile`: PHYLIP format sequence file. The format must comply with PHYLIP conventions which assumes taxon names are 10 characters and padded with blanks. See below for additional details.

`-l seqfile`: Alternative single line format sequence file that allows long names. Exactly one of -s or -l options must be present.

`-g gpfile`: Optional. A file with the 20 integer indicators of the G (value=0), F (value=1) and other (value=2) groups. Default: G=GARP, F=FYMINK.

`-w wtfile`: Optional. The (starting) weights for the classes. Default: $w_c \propto 1$.

`-a alpha`: Optional. The (starting) alpha parameter for a G4 rate variation model. Default: No rate variation.

`-e exchangeability`: Optional. The exchangeability matrix. Default: LG model exchangeabilities.

- `exchangeability=0`: $S_{ij} \propto 1$. Proportional model.
- `exchangeability=5`: JTT model
- `exchangeability=6`: WAG model
- `exchangeability=9`: LG model

`-o deriv`: Optional. Optimize parameters implied by `deriv`. Default: No optimization

- `deriv=1`: Edge-lengths
- `deriv=2`: Edge-lengths and alpha
- `deriv=3`: Edge-lengths, alpha and weights
- `deriv=4`: $\gamma_G^{(e)}$ and $\gamma_F^{(e)}$
- `deriv=5`: Edge-lengths, $\gamma_G^{(e)}$ and $\gamma_F^{(e)}$
- `deriv=6`: Edge-lengths, $\gamma_G^{(e)}$ and $\gamma_F^{(e)}$ alpha and weights

`-p iprint`: Optional. Additional output during optimization within the LBFGS routine is given per iteration with `iprint=1` and `iprint=2`. Default: No output during optimization.

`-y paramfile`: Optional. A file to which the parameters will be output. See below for more information. Default: Don't print parameters.

`-b numthreads`: Optional. The number of threads to use. Default: One thread.

The output to the screen is the log likelihood for the model.

An Illustrative Example

We illustrate usage with the Amborella data of Leebens-Mack et al. (2005). All of the input filles are available with the software. For the example we optimize only the γ_G and γ_F parameters because it illustrates the additional input that would not have been required with the usual option where all parameters are optimized. You can run that version by changing the argument -o 4 to -o 6 below. It will take much longer to run. The program was run at the commmand line with

```
$ gfmixf -c 21 -f amborella-c20-f.fr \
         -t amborella.treefile -r amborella.rootfile \
         -s amborella-interleave.phy \
         -w amborella.wtc -a 0.3396 \
         -g GARP -k FYMINK \
         -o 4 -p 1 \
         -b 4 -y amborella.yparam
```

The arguments `-c 21` and `-f amborella-c20-f.fr` are required. They indicate the number of classes and corresponding frequency file for the profile mixture model. For the example we used the C20+F frequencies. They were obtained by concatenating the `C20.aafreq.dat` frequency file with the observed frequencies for the data which were obtained separately. There are 21 lines in the file each with 20 entries for the frequencies. The first line gives the first of the C20 frequencies, the twentieth line gives the twentieth of the C20 frequencies and the final line gives the observed frequencies for the data.

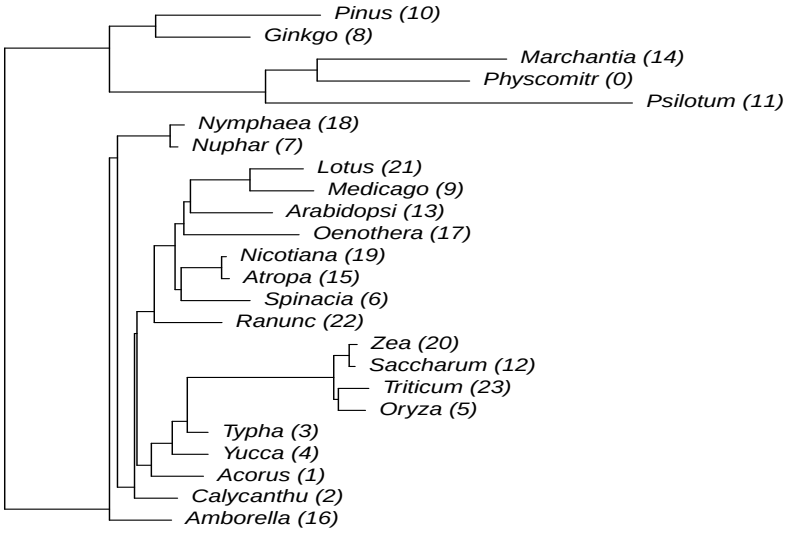

Figure 1: The tree for the example.

```
$ cat amborella-c20-f.fr
3.5098836193887414e-02 .... 1.1722486380046611e-02
...
6.7572740635108958e-02 .... 6.1906612625617675e-02
```

The arguments `-t amborella.treefile -r amborella.rootfile` are required. They indicate the treefile in newick format and corresponding rootfile. The tree is given in Figure 1 and rooted at the correct location. In the example we will only be optimizing the γ_G and γ_F parameters, so the tree should have edge-lengths; edge-lengths need not be present in the tree file. The tree used was the output tree from the LG+C20+F+Gamma model. To indicate where the root is we indicate the names of the taxa on one side of the root

```
$ cat amborella.rootfile
Pinus Ginkgo Marchantia Physcomitr Psilotum
```

The argument `-s amborella-interleave.phy` indicates the sequence data.

```
$ head amborella-interleave.phy
24 15688
Physcomitr MVKI--RPDE ISSIIRKQIE DYSQEIKVVN VGTVLQVGDG IARIYGLDKV MAGELVEFED
Acorus     MATL--RADE ISNIIRERIE QYTREVKVVN TGTVLQVGDG IARIHGLDEV MAGELVEFEE
Calycanthu MVTI--RADE ISNIIRERIE QYNREVKIVN TGTVLQVGDG IARIHGLDEV MAGELVEFEE
Typha      MVTL--RADE ISNIIRERIE QYSREVKIVN TGTVLQVGDG IARIHGLDEV MAGELVEFEE
Yucca      MVTL--RADE ISNIIRERIE QYNREVKVVN TGTVLQVGDG IARIHGLDEV MAGELVEFEE
Oryza      MATL--RVDE IHKILRERIE QYNRKVGIEN IGRVVQVGDG IARIIGLGEI MSGELVEFAE
```

```
Spinacia   MATI--RADE ISKIIRERIE GYNREVKVVN TGTVLQVGDG IARIHGLDEV MAGELVEFEE
Nuphar     MVTI--RAEE ISNIIRERIE QYNREVKIVN TGTVLQVGDG IARIHGLDEV MAGELVEFEE
Ginkgo     MVTI--RPDE ISSIIRKQIE QYNQEVEVAN IGTVLRVGDG IARIHGLDEV MAGELVEFVD
```

The taxon names here are all 10 characters or less which is required if interleaved form is used. The alternative is to use single line format and would be necessary with longer names. For instance, the single line format of the sequence data is

```
$ head amborella.phy
 24 15688
Physcomitr MVKI--RPDEISSIIRKQIEDYSQEIKVVNVGTVLQVGDGIARIYGLDKVMAGELVEFEDN...
Acorus     MATL--R...
...
```

At least one of `-s amborella-interleave.phy` or `-l amborella.phy` was required.

The argument `-w amborella.wtc` indicates the weights of the frequency mixture. These were obtained from a fitted LG+C20+F+Gamma model.

```
$ cat amborella.wtc
0.0093
0.0237
...
0.6791
```

There are 21 entries in the file, one for each of the frequency classes. The weights for the frequency mixture are generally not required but in this case we are only optimizing the γ_G and γ_F parameters, so reasonable values are needed. Another reason for input values would be as starting points for the optimizer. The default values are all set to `1/nclass`. These are not appropriate unless you are optimizing weights

The argument `-a 0.3996` indicates an alpha value to use with the discretized gamma rates-across-sites model that has four rate classes. As with other parameters for this example, these were obtained from a fitted LG+C20+F+Gamma model. The default is no rate variation, so if you want rate variation, even if you are optimizing parameters, you should include a starting value.

The arguments `-g GARP` and `-k FYMINK` indicate the amino acids in the G and F classes. Strings with the single letter amino acid codes giving the amino acids in the two groups are required. In the present case, these were included for illustration. The default G and F groups are GARP and FYMINK, so they would have been used without argument

The argument `-o 4` indicates that we want to optimize only the γ_G and γ_F parameters. There are a number of optimization options. The other main one would be to optimize all parameters. This would be specified with `-o 6`. If you don't include a `-o` option, the routine will return the likelihood for the input parameters and γ_G=1, γ_F=1, so with no gfmix element.

The argument `-p 1` indicates that we want to see the progress of the algorithm as it is running.

```
$ gfmixf -c 21 -f amborella-c20-f.fr \
         -t amborella.treefile -r amborella.rootfile \
         -s amborella-interleave.phy \
         -w amborella.wtc -a 0.3396 \
```

```
         -g GARP -k FYMINK \
         -o 4 -p 1 \
         -b 4 -y amborella.yparam

           * * *
        RUNNING THE L-BFGS-B CODE
           * * *
Machine precision = 2.22e-16
 N =         94
 M =          5
At X0, 0 variables are exactly at the bounds
At iterate     0, f(x)= 1.76e+05, ||proj grad||_infty = 2.60e+02
At iterate     1, f(x)= 1.76e+05, ||proj grad||_infty = 9.87e+01
....
....
-1.7511257895973272e+05
```

The last line gives the optimized log likelihood. If you don't include a `-p 1` option, only the final line will be indicated.

The `-b 4` option indicates the number of threads to use. The default is to use 1. We have found that performance can become poor with many threads, particularly if there are competing jobs. We recommend using `-b 2` or `-b 4`.

The final argument `-y amborella.yparam` indicates an output parameter file. The default is not to output parameters.

```
$ cat amborella.yparam
alpha = 0.340

Estimated Tree
((Amborella:0.069577,((Nuphar:0.008663,Nymphaea:0.015854):0.059088,...:0.117527);

Weights
  1 9.30093e-03
  2 2.37024e-02
  ...
 21 6.79168e-01
```

The first three entries in the output file are the alpha parameter, the estimated tree and the weights in the mixture. In the example, because we are not optimizing parameters these are the same as the input values but more generally, they would give the estimated values.

The next set of entries give the γ_G, γ_F parameters for each of the edges (equivalently child nodes). The first column indicates the node, the next two columns give the γ_G and γ_F parameters. The fourth columns gives the model-based ratio of the cumulative frequency of G to the cumulative frequency of F amino acids at the child node of the edges. Information about nodes is provided later in the parameter file and will be discussed presently but nodes are always ordered 0 through $2m-2$ (the root node=46 here) where m is the number of taxa. The first m nodes the terminal nodes ordered according to the appearance of

the corresponding taxa in the sequence file. For each of these nodes, the observed ratio of the cumulative frequency of G to the cumulative frequency of F amino acids can be calculated. This is what is given in the final column and only for the terminal nodes (equivalently taxa). In the present case is appears that the model systematically chose small facultes than for the observed data. It is not clear why that is the case.

```
$ cat amborella.yparam
alpha = 0.340
...
gamma(G), gamma(F), GF ratio each node and emprical GF for taxa
  0 5.20008e-01 1.63165e+00 6.88019e-01 7.66761e-01
  1 1.44134e+00 8.75285e-01 8.51635e-01 9.44573e-01
  ....
 23 1.10444e+00 9.50097e-01 8.34921e-01 9.00763e-01
 24 8.67560e-01 1.50511e+00 8.03236e-01
 25 9.31700e-01 9.44982e-01 8.30503e-01
 ....
 46 1.05170e+00 8.04918e-01 8.52235e-01
```

Information is then provided about the nodes. As mentioned the first m nodes are terminal, labeled $0, \ldots m-1$ and are ordered according their appearance in the sequence file. The remaining entries indicate the taxa descending from the labeled node.

```
$ cat amborella.yparam
alpha = 0.340
...
Edge/node labels. Taxa descending from edge/node are indicated
0 Physcomitr
1 Acorus
...
23 Triticum
24 Medicago Lotus
25 Saccharum Zea
...
46 Amborella ... Marchantia
```

As a final option for visualizaing how labels correspond to nodes we include a Newick file representation. The labels of the internal nodes indicate what they are. For the terminal nodes, `name` is replaced by `node_number|name` in the Newick file.

```
$ cat amborella.yparam
alpha = 0.340
...
Newick treefile with node labels
((16|Amborella:1,((7|Nuphar:1,18|Nymphaea:1)38:1,....)46;
```

Additional Information about Program Usage

Input

`-f frfile`: A file with the frequencies for each frequency class as rows. If the fit of a `+F` model is desired, these should include the `+F` frequencies as a row.

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 the base mdoel `-m LG+C20+G` is of interest, `C20.aafreq.dat` can be used as `frfile`; include the full pathname.

`-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`

`-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`

`-l seqfile`: An alternative sequence file format for long taxon names. As with PHYLIP format, the first two entries should be integers giving the number of taxa and the number of sites. This is followed by lines of the form

```
taxon_name sequence_data
```

The `sequence_data` string should be contiguous without blanks and follow the taxon name. For instance the following format would work.

```
6 3414
donald_j_trump
ANLLLLIVPILI...
Phovi        INIISLIIPILL...
...
```

Taxon names can be of any length but should not have spaces.

`gpfile`: There should be 20 entries in the file, each giving the group G (value=0), F (value=1) and other (value=2) that the corresponding amino acid has. Amino acids have the standard ordering illustrated below with the labels for G=GARP and F=FYMINK.

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0	0	1	2	2	2	2	0	2	1	2	1	1	1	0	2	2	2	1	2

The corresponding `gpfile` would be

```
0 0 1 2 2 2 2 0 2 1 2 1 1 1 0 2 2 2 1 2
```

System Requirements and Installation

Requirements and Installation of External Packages

The main program `gfmixf` 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 programming environment. 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. The program utilizes a C version of the L-BFGS-B algorithm of Zhu et al. (1997) written by Stephen Becker `http://amath.colorado.edu/faculty/becker/`.

Installation

1. Download and unpack the software

   ```
   $ tar zxf gfmixf.tar.gz
   ```

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

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

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

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 gfmix_lnl_align charfeq-taxa gfmixf an_freq
   ```

   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 <- ""` at the top of the file `gfmixf` 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 using the C20 model.

5. The source code and directory can be removed:

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

References

Leebens-Mack J., Raubeson L.A., Cui L., Kuehl J.V., Fourcade M.H., Chumley T.W., Boore J.L., Jansen R.K., depamphilis C.W. (2005). Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol. 22:1948-1963.

McCarthy, C.G.P., Susko, E., Harada, R. and Roger, A.J. (2026). Modeling site-and-branch-heterogeneity with GFmix. Systematic Biology. syag040. https://doi.org/10.1093/sysbio/syag040

Muñoz Gómez, S., Susko, E., Williamson, K., Eme, L., Slamovitz, C.H., Moreira, D. Purificación, L. and Roger, A.J. (2021). A site-and-branch-heterogeneous model on an expanded dataset favor mitochondria as sister to known Alphaproteobacteria. DOI:10.21203/rs.3.rs-557223/v1

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

Zhu, C., Byrd, R.H., Lu, P. and Nocedal, J. (1997). L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. ACM Trans. Math. Soft. 23:550–560.