# MAMMaL: (M)ultinomial (A)pproximate (M)ixture (Ma)ximum (L)ikelihood
# Accelerated Estimation of Frequency Classes in Site-heterogeneous Profile Mixture Models
## Version 1.1
## May 6, 2019

**Edward Susko**

*Department of Mathematics and Statistics, Dalhousie University*

## Introduction

The main program `mammal` takes as input a number of classes, a sequence file and a tree and outputs estimated frequencies for classes using the methods described in Susko, Lincker and Roger (2018). Installation information is available towards the end of this document. As a convention for referring to models that use the estimated frequencies, we recommend, for instance, `LG+MAM20+G` for a model that has an LG exchangeability matrix, gamma rate variation and 20 frequency classes constructed using the MAMMaL software.

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

```
$ mammal -s seqfile -t treefile -c number_of_classes [OPTIONS]
```

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. Used to determine high-rate sites.

`-c number_of_classes`: The number of frequency classes.

`-h`: Use hierarchical clustering to get starting frequencies with the base R clustering routines. Default: The C-series frequencies of Le et. al. (2008) are used if the number of classes are in $\{10, 20, \ldots, 60\}$.

`-l`: Don't use likelihood weighting. Default: Use likelihood weights.

`-q quantile`: Use sites with rates $> q \times 100$th percentile of mean DGPE rates. Default: 0.75.

`-C penalty`: The penalty parameter, $\eta$ in Susko, Lincker and Roger (2018). Default: $\eta = 5$.

The output consists of two text files:

`estimated-frequencies`: Each row gives the amino acid frequencies for a class. An additional row is included at the end that gives the overall frequencies for the alignment.

`esmodel.nex`: A nexus file that can be used with the options `-mdef esmodel.nex`, `-m LG+ESmodel+G` to define a mixture model for ML estimation using IQ-TREE (Nguyen et al. 2015)

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 `mammal` is run, they should be renamed or moved.

## 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`

`-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`

`-h`: The default hierarchical clustering routine in R is used. This routine can require more memory than is available if the number of sites in the alignment is large. The default starting frequencies when the number of classes are in $\{10, 20, \ldots, 60\}$ are the C-series classes, so using hierarchical starting frequencies is not strictly necessary.

Susko, Lincker and Roger (2018) references the verb-Rclusterpp- package of Linderman and Bruggner (2013), which was available in version 1.0 of the MAMMaL software. We have since removed this option as the package has been archived on the CRAN repository and the last version in that repository had a bug.

### Output

`estimated-frequencies`: Each row gives the amino acid frequencies for a class. An additional row is included at the end that gives the overall frequencies for the alignment. Thus the returned frequencies are actually suitable for a mixture with one additional class corresponding to the overall frequencies in the data set. You can delete the last row to fit a model without the +F component.

On any given row, the ordering of frequencies is the conventional ordering expected by most packages, alphabetical on their three-letter codes:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| A | R | N | D | C | Q | E | G | H | I |
| Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile |
| L | K | M | F | P | S | T | W | Y | V |
| Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |

`esmodel.nex`: A nexus file that can be used to fit a mixture model with IQ-TREE at the command line through

```
$ iqtree -s seqfile -m LG+ESmodel+G -mdef esmodel.nex [OPTIONS]
```

With no options, it will fit the mixture and search for the best tree. The term `ESmodel` should always be present in the model statement. The example above fits a model that has the LG exchangeability matrix and a discretized gamma rates-across-sites model on top of the mixture. The elements of the model can be changed. Note that because the overall frequencies are included as the last set of frequencies in `esmodel.nex`, the model being fit is effectively mixture+F model. You should not include a +F in the model specification. See

```
http://www.iqtree.org/doc/Command-Reference#general-options
```

for additional information about options and specifying substitution models (search 'Specifying substitution models' on that page).

## System Requirements and Installation

**Requirements and Installation of External Packages** The main program `mammal` 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, a working installation of the R statistical package and the R package `quadprog` of Turlach and Weingessel (2013). The statistical package R is freely available at `https://cran.r-project.org`. Once R is installed, the `quadprog` R package can be installed at the R command line with

```
> install.packages("quadprog")
```

The source code has been compiled and tested using gcc computers running Linux and Mac OS X. While the program has not been tested on another platform, it should compile under any Linux distribution as well as Mac OS X. On Mac OS, to install gcc, bring up a terminal and type

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

### Installation

1. Download and unpack the software

   ```
   $ tar zxf mammal.tgz
   ```

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

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

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

   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 and C-series frequencies

   ```
   keep-sites mammal-sigma mult-data mult-mix-lwt charfreq
   C10.dat, ..., C60.dat
   ```

   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 `mammal` 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 file `mammal` to a location in your `PATH`.

5. The source code and directory can be removed:

   ```
   $ cd ../
   $ rm -rf mammal.tgz mammal/
   ```

## References

Felsenstein, J. (2004). PHYLIP Phylogeny Inference Package (version 3.6). Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (version 3.2). Cladistics 5: 164-166.

Le S.Q., Gascuel O., Lartillot N. (2008). Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24:23172323.

Linderman, M. and Bruggner, R. (2013). Rclusterpp: Linkable C++ clustering. R package version 0.2.3.

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.. Mol. Biol. Evol., 32:268-274.

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.

Turlach, B.A. and Weingessel, A. (2013). quadprog: Functions to solve quadratic programming problems. R package version 1.5-5.