

# biphy 1.0

Walker Pett

November 5, 2014

## 1 Introduction

### 1.1 Markovian Substitution Process

The process of gain and loss of binary characters is modeled as a continuous time Poisson process along each branch, where the rates of gain and loss on branch  $j$  of duration  $t_j$  are indicated by  $\lambda_j$  and  $\mu_j$  respectively. The stochastic kernel of this substitution process on branch  $j$ , denoted  $\mathbf{Q}_j$ , can be expressed as follows:

$$\mathbf{Q}_j = \begin{bmatrix} -\lambda_j & \lambda_j \\ \mu_j & -\mu_j \end{bmatrix}$$

When not constrained to ultrametric trees, it is more natural to use the normalized rates,  $\rho_j = \lambda_j + \mu_j$  and  $\pi_j = \frac{\lambda_j}{\lambda_j + \mu_j}$ , such that

$$\mathbf{Q}_j / \rho_j = \begin{bmatrix} -\pi_j & \pi_j \\ (1 - \pi_j) & -(1 - \pi_j) \end{bmatrix}$$

The parameters  $\pi_j$  and  $\rho_j$  can be interpreted as the stationary gene frequency and total rate of character replacement on branch  $j$ , respectively. Using this reparameterization, the branch lengths can then be re-expressed in units of the expected number of substitutions per site,  $\nu_j = \rho_j t_j$

Note that because, in general, the entire substitution process along  $\tau$  is not time-reversible, the placement of the root is not trivial. Furthermore, because the process is not stationary, we specify an independent parameter  $\phi$  representing the stationary frequency in the ancestral sequence at the root of  $\tau$ .

## 2 Basic usage

### 2.1 Starting a run

a run is initiated by providing the `biphy` command with a data file and a run name

```
biphy [options...] -d <data file> <run name>
```

### 2.2 Summarizing the output

`biphy` produces a trace file (`<run name>.trace`) and a treelist file (`<run name>.treelist`) suitable as inputs respectively to the `tracecomp` and `bpcomp` programs in the PhyloBayes package [2].

## 2.3 Restarting a run

In the event of a power failure or other error, a **biphy** run can be restarted using the run name without any other options:

```
biphy <run name>
```

## 2.4 Stream-reading mode

Stream-reading mode is activated when some options are specified, but a data file is not specified via the **-d** option. this mode is used only for performing posterior predictive simulations and computing cross-validation scores via the **-ppred** and **-cv** options:

```
biphy [-ppred -cv <test file>] <run name>
```

note that a run for which the **-s** option was not specified produces a stream that contains only one sample (the most recent sample). thus, activating stream-reading mode on such a run will give an error.

# 3 Detailed Options

## General options

**-d <datafile>**

option for specifying the multiple sequence alignment to be analysed (see: Input Data Format section)

**-x <every> [<until>]**

specifies the saving frequency and (optional) the number of points after which the chain should stop. If this number is not specified, the chain runs forever. By definition, **-x 1** corresponds to the default saving frequency.

**-t <tree file>**

constrains the analysis to be conducted on the specified fixed tree

**-o <outgroup file>**

constrains trees to be rooted using the specified outgroup

**-f**

forces the program to overwrite an already existing run with same name.

## Model options

**-dollo**

the stochastic Dollo model of Nicholls and Gray [3]. under this model, branch lengths are expressed in expected numbers of losses per site. this option automatically activates the **-u 1** option

**-h**

time-homogeneous model. a single stationary frequency  $\phi$  is assumed across the entire tree.

**-nh**

time-heterogeneous, hierarchical beta model. branch-wise stationary frequencies are drawn iid from a beta distribution with hyperparameters  $\beta_1$  and  $\beta_2$

**-m <int>**

time-heterogeneous, beta mixture model. branch-wise stationary frequencies are drawn iid from a mixture of <int> Beta distributions with hyperparameters  $\beta_1$  and  $\beta_2$

**-dpp**

time-heterogeneous, infinite mixture model. branch-wise stationary frequencies are drawn from a Dirichlet process with Exponential concentration parameter of mean 1, and with a Beta base distribution with hyperparameters  $\beta_1$  and  $\beta_2$

## Additional priors

**-dgam <int>**

this option specifies that site-specific rates should be drawn from a  $\text{Gamma}(\alpha, \alpha)$  distribution, discretized on <int> quantiles. the options **-dgam 0** or **-dgam 1** specify a constant rates model. the default is 4 rate categories.

**-lexp**

hierarchical exponential prior on branch lengths. branch lengths are drawn from an exponential distribution of mean  $\mu$ , which is itself drawn from an exponential distribution of mean 1/10 .

**-ldir**

compound Dirichlet prior on branch lengths [4]. branch lengths are drawn from a Dirichlet distribution of concentration parameter 1, scaled by the tree length which is drawn from an exponential distribution of mean 1.

**-rp <min> <max>**

the root frequency  $\phi$  is truncated on the (unnormalized) interval (**min,max**). applies only to the time homogeneous model, and the hierarchical beta model.

**-rigid**

the root frequency  $\phi$  and the stationary frequencies on the branches subtending the root node of the tree are constrained to be equal. this can improve mixing due to the non-identifiability of these parameters in the unconstrained model [?]. applies only to time-heterogeneous models.

## Correction for unobservable site-patterns

**-u <int>**

use this option to specify site-patterns that cannot be observed or which have been omitted from the dataset. this results in a correction to the likelihood that accounts for the absence of these site-patterns [1].

the option `<int>` is one of:

- 0 = no site-patterns have been omitted (default)
- 1 = constant absence site-patterns have been omitted
- 2 = constant presence site-patterns have been omitted
- 4 = singleton gains have been omitted
- 8 = singleton losses have been omitted

combinations of the above site-patterns are achieved by adding the indicated values. for example, `-u 3` indicates that constant sites have been omitted, and `-u 15` indicates that parsimony uninformative sites have been omitted.

## MCMCMC options

`-n <int>`

this option specifies the number of Metropolis-coupled chains to use, with one chain per available thread. the default is 1 chain.

`-delta <float>`

`-sigma <float>`

these options specify the  $\delta$  and  $\sigma$  tuning parameters for the calculation of chain heats during a parallel Metropolis-coupled run. the heat of chain  $k$  is computed as follows

$$\beta_k = (1 + \delta)^{-\sigma^k}$$

the default values are  $\delta = 0.05$  and  $\sigma = 1$ . these values should be appropriate for small values of  $n$  (less than 10), but should be adjusted on a case by case basis to improve mixing, usually by increasing  $\sigma$  and/or decreasing  $\delta$  as the number of chains is increased.

`-si <int>`

this option specifies the number of generations between proposals to swap Metropolis-coupled chains. the default value is 1.

## Output options

`-e`

in addition to the Newick treelist output, a Nexus-formatted tree file with extra branch parameters is produced. This is useful for mapping branch-wise stationary frequencies to branches in tree visualisation software.

`-s`

the entire output of all chains is saved to the file `<run name>.stream`. This option is required if subsequent posterior predictive analysis or cross-validation checks are desired. note that the saving frequency of this stream is the single-chain saving frequency specified by the `-x` option, multiplied by the swap interval of the Metropolis-coupled chain specified by the `-si` option.

## Model-checking options

the following options are to be used in stream-reading mode, which is activated by running **biphy** without specifying a data file via the **-d** option. in stream-reading mode, the parameter configuration of each sample is read sequentially from the stream of the indicated run, and computations are performed by conditioning on this parameter configuration

### **-ppred**

using this option, an alignment is simulated at each sample point, then the frequency of “1” states is calculated for each species and saved to the file **<run name>.ppred**

### **-cv <test file>**

using this option, the likelihood of the specified test alignment is calculated at each sample point, and saved to the file **<run name>.cv**

## References

- [1] Joseph Felsenstein. Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution*, 46(1):159–173, 1992.
- [2] Nicolas Lartillot and Hervé Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–109, June 2004.
- [3] Geoff K. Nicholls and Russell D. Gray. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):545–566, July 2008.
- [4] Bruce Rannala, Tianqi Zhu, and Ziheng Yang. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Molecular biology and evolution*, 29(1):325–35, January 2012.