Heterotachy models in *BayesPhylogenies*

BayesPhylogenies is a general software package for inferring phylogenetic trees using Bayesian Markov Chain Monte Carlo (MCMC) methods. The program allows a range of models of gene sequence evolution, models for morphological traits, models for rooted trees, and gamma and beta distributed rate-heterogeneity. In addition, the program implements a model to detect heterogeneity in the pattern of evolution (Pagel and Meade, 2004) that allows the user to fit more than one model of sequence evolution, without partitioning the data. The program also the implements a mixture model for heterotachy described by Meade and Pagel (2008) and Pagel and Meade (2008).

This document provides an introduction and examples to using the heterotachy models implemented in BayesPhylogneies. Heterotachy is defined as variation in the rates of evolution, as might be expected in adaptively evolving genes. The model we implement allows this variation to be distributed non-parametrically in contrast to the covarion model (see Pagel and Meade 2008 for further discussion).

We introduce the heterotachy model here, it is described in more detail in (Meade and Pagel, 2008; Pagel and Meade, 2008). The heterotachy model for phylogenetic inference allows the likelihood of the model to be calculated over 'mixtures' of branch lengths on the tree. Each component of a mixture is a set of branch-lengths for all *2s-3* branches (for an un-rooted tree), where *s* is the number of species. The mixture model approach calculates the likelihood over more than one set of lengths for these branches. A site or set of sites that accelerated in one part of the tree could be characterised by longer branches there, while other sites could be characterised by a shorter branch. It is because the position or number of longer or shorter branches is not specified in advance that model is non-parametric.

Finding 'mixtures' on branches is costly because each mixture requires *2s-3* new parameters in the model, and these extra parameters need to be 'paid for' by improvement in the likelihood. This limits the power of mixture models to detect heterotachy. As such, a 'reversible-jump' version of the mixture model method is also implemented. This approach allows the investigation and parameter estimation for only those branches of a mixture that contribute significantly to the inference of the tree. The RJ method typically reduces the number of parameters by 50-90% and this method can detect far smaller amounts of heterotachy at any given place in the tree (see Pagel and Meade, 2008).

Running *BayesPhylogenies*

The program can be run interactively from a console, command line interface or terminal, or its commands can be set in the input file of aligned sequences or other traits (see the BayesPhylogenies Manual (www.evolution.rdg.ac.uk/Files/bayesphylogenies.pdf).

To run the program, first change into the directory that the program was unpacked into. Start the program by typing the binary name followed by the nexus file name.

for OS X and Linux machines type

```
./BayesPhylogenies filename.nex
```

for Windows

```
BayesPhylogenies.exe filename.nex
```

The input file must be in nexus format. The program expects as input a standard nexus file such as generated from PAUP or Clustal. We provide and example Nexus file for the worked example below.

Once started the following will appear:

```
BayesPhylogenies
Mark Pagel and Andrew Meade
www.evolution.reading.ac.uk


BayesPhylogenies:
```

Now the program is ready to accept commands from the user. It is also possible to set up and run BayesPhylogenies using a Nexus input file at the end of the alignment (see the BayesPhylogenies Manual for further details).

Setting up and Running Heterotachy models

This section provides an example of how to run the heterotachy models implemented in BayesPhylogneies. There are two categories of heterotachy models implemented; we provide here an example of how to run each approach. The first approach is the model where each component of a mixture is a set of branch-lengths for all *2s-3* branches in an un-rooted tree (where *s* is the number of species), hereafter this will be called *the whole branch length set mixture model.* The second approach is the 'reversible-jump' version of the mixture model, from here on this will be called *the reversible-jump branch length set mixture model.*

The phylogenetic mixture models for heterotachy implemented in BayesPhylogenies are described in the following papers:

The whole branch length set mixture model is described in **Meade, A and Pagel, M. 2008. A phylogenetic mixture model for heterotachy. In Evolutionary biology from concept to application (ed. P. Pontarotti), pp. 29–41. Heidelberg, Germany: Springer Verlag.**

The reversible-jump branch length set mixture model is described in **Pagel, M. and Meade, A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos Trans R Soc Lond B Biol Sci.* 363, 3955-3964.**

Running the whole branch length set mixture model

An example input file called 'example1.nex' is provided with the program download. This is an alignment of simulated data generated from a random topology of 70 taxa in Nexus format. To simulate heterotachy in a sequence alignment two random sets of branch lengths were generated from a uniform distribution ranging between 0 and 0.2 for the simulated topology. One thousand five hundred simulated nucleotides were generated for each branch-length set. These two simulated datasets were concatenated in to a single alignment of 3000 simulated nucleotides. (See Pagel and Meade, 2008 for full details of how this simulation was generated). We would expect the heterotachy model, which can sum the likelihood over two sets of branches to fit these data significantly better than a model with a single set of branch lengths.

After moving to the directory where the binary and data file have been placed, typing the following will start the program with the example described above:

For OS X and Linux machines type

```
./BayesPhylogenies example1.nex
```

for Windows

```
BayesPhylogenies.exe example1.nex
```

You should see something like the following:

```
BayesPhylogenies
Mark Pagel and Andrew Meade
www.evolution.reading.ac.uk


BayesPhylogenies:
```

To implement the whole branch length set model use the **BLS =** command. For example, typing '**BLS = 2**' causes two independent branch length sets of the chosen model of evolution to be calculated at each site. This command can be used in conjunction with the mixture model for pattern heterogeneity (see Pagel and Meade, 2004) as well as gamma. If no other commands are specified the default model setting will be used (GTR).

After the program has started, typing the following will run the two whole-branch-lengths-sets model using the simulated data:

```
BLS = 2
Seed = 1635
Printfreq = 10000
Run
```

This will cause the program to begin a MCMC analysis of the input data (example1.nex). (The **PrintFreq** command simply sets the frequency with which the Markov chain is sampled.) It will print output to the screen showing the iteration number of the chain and the log-likelihood of the tree at that iteration. Using the above commands and seed for the dataset example1.nex the output will look as follows (One would not normally specify the random number seed manually as one is automatically generated if not specified, but including it here will ensure that the example run included is exactly the same as the one produced):

```
BayesPhylogenies
Mark Pagel and Andrew Meade
www.evolution.reading.ac.uk


BayesPhylogenies: bls = 2
BayesPhylogenies: seed = 1635
BayesPhylogenies: pf =10000
```

```
BayesPhylogenies: run


Current settings for BayesPhylogenies.

Number of chains:                   1
Number of iterations:               + Infinite
Taxa:                               71
Current iterations:                 0
Print frequency:                    10000
Random seed:                        1635
Output File Base:                   example1.nex
Input file:                         example1.nex
Current partition:                  Default
Auto run:                           False
Burn in period:                     None
No Of Topologies:                   1
No Of Brach Length Sets:            2
Using an Un-Rooted tree
Bayes Trees outputPartition name:                       Default
      No:                           0
      Start:                        0
      End:                          2999
      Model:                        GTR
      Base Frequencies:             Estimated
      No of patterns                1
      No of rates                   1
      No of Covarion categories     0
      Invariant sites:              False
29/07/2009 11:41:51
0           -303568.455993
10000   -138983.685908
20000   -133361.004955
30000   -132441.129641
40000   -131994.002608
50000   -131781.666804
60000   -131672.786603
```

The program will also write two output files to the directory. One is a parameters file – a tab-delimited file containing the tree likelihoods and the estimated parameters for each iteration of the Markov chain including the weights associated with the branch length sets. The other is a trees file – this contains the trees and branch lengths sets. As there are two or more sets of branches associated with each tree these will **not** be able to be viewed in programs such as TreeView, FigTree or PAUP. The program BayesTrees which is available from our software pages (www.evolution.reading.ac.uk/SoftwareMain.html) can be freely downloaded and can be used to view the trees.

For reference a parameters file (Wholeexample1.Parameters) and Trees file (Wholeexample1.trees) for the analysis above are included with the download. These are the results of running the MCMC chain for 3000000 iterations – this chain has reached apparent convergence. Using the random seed provided it is possible to duplicate the results from this chain (that is, produce exactly the same output files).

It is possible to estimate more than two whole branch length sets, however estimating a large number is extremely computationally intense and may often be unnecessary. It is possible to determine the optimal number of branch lengths sets using the Akaike information criterion (AIC) or the more stringent Bayesian information criterion (BIC) (See Pagel and Meade, 2008 for further discussion). For example, the mean likelihood from the converged chain for a model with a single set of branches returns a log-likelihood score of -131951. The heterotachy model with two whole branch lengths sets improves this by 923 log-likelihood units to -131028 (this is significant by both AIC and BIC).

Running the reversible-jump branch length set mixture model

Using the same example file (example1.nex) and starting BayesPhylogenies in the same way as above, the reversible-jump branch length set mixture model can be set up as follows.

To implement the reversible-jump branch length set mixture model use **RJBLS =** command. For example, typing '**RJBLS = 2**' causes the program to accommodate heterotachy by allowing some or all of the branches in the phylogenetic tree to have up to two distinct branch lengths. The reversible-jump procedure allows multiple branches in part of the tree where it significantly improves the likelihood of the model. This method has the potential to account for heterotachy with fewer parameters than whole branch length sets model. (The reversible-jump branch length set mixture model is fully described in Pagel and Meade, 2008.) Again, this command can be used in conjunction with the mixture model for pattern heterogeneity (see Pagel and Meade, 2004) as well as gamma. If no other commands are specified the default model setting will be used (GTR).

After the program has started, typing the following will run the reversible-jump branch length set mixture model with up to two branches using the simulated data:

```
RJBLS = 2
Seed = 1248381896
Printfreq = 10000
Run
```

This will cause the program to begin a MCMC analysis of the input data (example1.nex). (The PrintFreq command simply sets the frequency with which the Markov chain is sampled.) It will print output to the screen showing the iteration number of the chain and the log-likelihood of the tree at that iteration. Using the above commands and seed for the dataset example1.nex the first part of output will look as follows (as before, one would not normally specify the random number seed manually as one is automatically generated in not specified, but including it here will ensure that the example run included is exactly the same as the one produced):

```
BayesPhylogenies
Mark Pagel and Andrew Meade
www.evolution.reading.ac.uk


BayesPhylogenies: seed 1248381896
BayesPhylogenies: pf = 10000
BayesPhylogenies: rjbls = 2
BayesPhylogenies: run
```

```
Current settings for BayesPhylogenies.

Number of chains:                   1
Number of iterations:               + Infinite
Taxa:                               71
Current iterations:                 0
Print frequency:                    10000
Random seed:                        1248381896
Output File Base:                   example1.nex
Input file:                         example1.nex
Current partition:                  Default
Auto run:                           False
Burn in period:                     None
No Of Topologies:                   1
RJ Branch Length Sets:              2
Using an Un-Rooted tree
Bayes Trees outputPartition name:   Default
            No:                     0
            Start:                  0
            End:                    2999
            Model:                  GTR
            Base Frequencies:       Estimated
            No of patterns          1
            No of rates             1
            Invariant sites:        False
28/07/2009 13:37:45
0       -314959.595582
10000   -140898.507600
20000   -132504.934055
30000   -132092.648795
40000   -131969.317491
50000   -131973.374846
60000   -131966.999271
```

The program will also write two output files. One is a parameters file – a tab-delimited file containing the tree likelihoods and the estimated parameters from each iteration of the Markov chain including the weights associated with the branch length sets and the number of branches for which there is signal for two branch lengths. The other is a trees file – this contains the trees and branch lengths sets. As there is two or more sets of branches associated with each tree these will **not** be able to be viewed in programs such as TreeView, FigTree or PAUP. The program BayesTrees which is available from our software pages (www.evolution.reading.ac.uk/SoftwareMain.html) can be freely downloaded as can be used to view the trees.

For reference a parameters file (RJexample1.Parameters) and Trees file (RJexample1.trees) for the analysis above are included with the download. These are the results of running the MCMC chain for 3 million iterations – this chain has reached apparent convergence. Using the random seed and commands provided it is possible to exactly duplicate this chain (and produce exactly the same output files).

One advantage of the reversible-jump procedure is that no post analysis model testing has to be carried out. But for comparison the reversible-jump model returned a mean log-likelihood score from the converged chain of -131098. This is only 70 log-likelihood units worse that the model with two whole branch lengths, but the RJ model set uses many fewer parameters.

More than two branch-length sets can be explored using the reversible jump procedure, but we recommend that additional branch-length sets are justified first using the whole-branch-length-sets procedure. Then the reversible-jump approach can be used to investigate which branches are most affected by heterotachy.

Using Heterotachy models with real data sets and other models of evolution

So far the heterotachy models have been discussed it terms of simulated datasets. The datasets that researchers use to infer phylogenetic trees often contain far more complex signal. For example, it is well established that multiple gene alignments contain significant heterogeneity in the rate and pattern of evolution across sites. *BayesPhylogenies* allows the user to combine models which account for rate and pattern heterogeneity with heterotachy models. This combination has been shown to improve the log-likelihood of the data (see Meade and Pagel 2008; Pagel and Meade 2008). To illustrate how to combine these models start the program with the data set called kiontke.nex. Type the following:

for OS X and Linux machines type

```
./BayesPhylogenies kiontke.nex
```

for Windows

```
BayesPhylogenies.exe kiontke.nex
```

This dataset is a multiple gene dataset used in a paper by Kiontke *et al.* 2004 to study the phylogenetic relationship of *Caenorhabditis elegans* to its closest relatives. It has been shown to contain rate and pattern heterogeneity as well as heterotacy (see Meade and Pagel, 2008; Pagel and Meade, 2008).

After the program is started using this dataset it can be run using the following commands:

```
RJBLS = 2
RJPA
Seed = 946432
Printfreq = 10000
Gamma = 4
Run
```

The **RJBLS**, **Seed** and **Printfreq** commands are as before. The **RJPA** command is a reversible-jump implementation of the Pagel and Meade (2004) mixture model for pattern heterogeneity to determine how many different rate matrices were required to explain the data. This reversible-jump procedure automatically moves among Markov chains with different numbers of rate matrices, and at convergence estimates the posterior support for these different chains (there is no a priori limit to the number of matrices that can be estimated). Rate heterogeneity is included in the model using Yang's (1994) discrete-gamma rate heterogeneity model (the **Gamma** command). Using the above commands and seed for the dataset rydin.nex the first part of output will look as follows:

```
BayesPhylogenies
Mark Pagel and Andrew Meade
www.evolution.reading.ac.uk
```

```
BayesPhylogenies:       RJBLS = 2
BayesPhylogenies:       RJPA
BayesPhylogenies:       Seed = 938153
BayesPhylogenies:       Printfreq = 10000
BayesPhylogenies:       Gamma = 4
BayesPhylogenies:       Run


Current settings for BayesPhylogenies.

Number of chains:                   1
Number of iterations:               + 19000000
Taxa:                               14
Current iterations:                 0
Print frequency:                    10000
Random seed:                        946432
Output File Base:                   kiontke.nex
Input file:                         kiontke.nex
Current partition:                  Default
Auto run:                           True
Debuging Mode:                      True
Burn in period:                     None
No Of Topologies:                   1
RJ Branch Length Sets:              2
Using an Un-Rooted tree
Bayes Trees outputPartition name:   Default
      No:                           0
      Start:                        0
      End:                          2316
      Model:                        GTR
      Base Frequencies:             Estimated
      No of patterns                RJ MCMC
      No of rates                   4
      No of Covarion categories     0
      Rate distribution:            Gamma
      Invariant sites:              False
31/07/2009 12:51:07
0       -55200.980964
10000   -43550.810518
```

The program will produce the usual output file from which it is possible to determine how many branches show support of multiple branch lengths and associated weights. It also shows how many patterns of evolution are optimal for this dataset.

For reference a parameters file (kiontke.Parameters) and Trees file (kiontke.trees) for the analysis above are included with the download. These are the results of running the MCMC chain for 19 million iterations – this chain has reached apparent convergence. Using the random seed and commands provided it is possible to exactly duplicate this chain (and produce exactly the same output files). This dataset is large and will take a long time to run on a standard desktop computer and would require a large amount of memory. From inspection of the output file the best fitting model allows a second branch length for approximately 17 of the branches. The data set supports 4 different pattern matrices.

Acknowledgements and Disclaimer

The development and implementation of the Heterotachy models in *BayesPhylogenies* has been supported in by grant NE/C51992X/1 from Natural Environment Research Council. We make every effort to ensure the program works as described but we cannot guarantee that it is free of bugs. Please report any problems to Mark Pagel (m.pagel@reading.ac.uk) or Andrew Meade (a.meade@reading.ac.uk).

References

Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., and Fitch, D.H. 2004. Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci.* 101: 9003-9008

Meade, A and Pagel, M. 2008. A phylogenetic mixture model for heterotachy. In Evolutionary biology from concept to application (ed. P. Pontarotti), pp. 29–41. Heidelberg, Germany: Springer Verlag.

Pagel, M. and Meade, A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos Trans R Soc Lond B Biol Sci.* 363, 3955-3964.

Pagel, M. and Meade, A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol*. 53:571–81.

Rydin, C., Kallersjo, M. and Friis, E.M. 2002. Seed plant relationships and the systematic position of Gnetales based on nuclear and chloroplast DNA: conflicting data, rooting problems, and the monophyly of conifers. *Int. J. Plant Sci*. 163, 197-214.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–314.