Manual for tip dating in bpp

Anna Nagel, Tomáš Flouri, Ziheng Yang, and Bruce Rannala

August 24, 2023

1 New Control File options in bpp

Options in the control file that are either new or modified when using tip dating are described below. For options that are unchanged, we refer readers to the full bpp manual for a full description of all of the options.

1.1 BPP control file variables

1

`datefile = s`

DESCRIPTION

Sets the path/name of the date file to be the string s

VALUES

s, a string of characters specifying the directory path and/or name of a file that contains sequence sample date information.

DEPENDENCIES

Model A00 must be used with tip-dating (`speciesdelimitation = 0`, `speciestree = 0`). Checkpointing cannot be used. The tip-dating `locusrate` option must be used (`locusrate = 3 d d`). A global clock must be used. This is the default or can be set with `clock = 0`.

COMMENTS

See section 2.1 for a description of the file format.

EXAMPLES

```
datefile = dates.txt
```

```
datefile = /home/foo/seqDates.txt
```

2

`locusrate = +d[(f f f s, s)]`

DESCRIPTION

Specifies the model of substitution rate variation among loci, as well as the parameters, and the prior on the parameters of the model.

VALUES

`d+[(f f f s, s)]`, specifies the model and parameters. The variable d+ takes one of 4 possible values with different numbers of additional variables depending on d+. The permissible combinations are:

d+	[args]	Description
0		Loci have same rate
1	f f f s $\alpha_{\bar{\mu}}$ $\beta_{\bar{\mu}}$ α_{μ_i} prior	Locus rates variable and estimated
2	s Filename with locus rates	Locus rates variable and specified
3	α β	Loci have same rate which is estimated using tip dating

where $\alpha_{\bar{\mu}}$ and $\beta_{\bar{\mu}}$ are the parameters of the Gamma distribution prior on the average rate among loci $\bar{\mu}$ and α_{μ_i} is the parameter of the prior on the rate at the ith locus given $\bar{\mu}$. The prior variable must be either `iid` for a conditional iid (hierarchical) prior on μ_i or `dir` for a Gamma-Dirichlet prior. The prior variable is optional, if it is not specified `iid` is used.

DEFAULT

`0`

DEPENDENCIES

If both the `prior` option for the variable `locusrate` and the `prior` option for the variable `clock` are set, they should take the same value. If the `locusrate = 3 d d`, the `datefile` option must be specified.

COMMENTS

The setting `locusrate = 0` (default) means that all loci have the same mutation rate. Setting `locusrate = 1 f f f s` specifies a model with rate variation among loci in which rates are estimated from the sequence data. Including the first integer (1) which species the model there are 5 arguments, the last one (prior) being optional. Parameters $\alpha_{\bar{\mu}}$ and $\beta_{\bar{\mu}}$ specify the shape and rate parameters of the gamma distribution for the mean rate across loci ($\bar{\mu}$), while α_{μ_i} and prior are used to specify the locus rates (μ_i) given the mean rate ($\bar{\mu}$). The option prior can take two values dir (for Gamma-Dirichlet rates for loci), and iid (for conditional i.i.d. or hierarchical prior). `locusrate = 2 LocusRateFileName` specifies the fixed-rates model of locus-rate variation (Burgess and Yang, 2008). This is the strategy used by Yang (2002), with the relative rates estimated by the distance to an outgroup species. The relative locus rates are listed in the file: there should be as many numbers in the file, separately by spaces or line returns, as the number of loci (nloci). The program re-scales those rates so that the average across all loci is 1 and then use those relative rates as fixed constants. Specifically the mean rate across loci ($\bar{\mu}$) is assigned a gamma prior:

$$\bar{\mu} = \mathrm{G}(\alpha_{\bar{\mu}}, \beta_{\bar{\mu}}),$$

with mean and variance

$$\begin{aligned} \text{Mean}(\bar{\mu}) &= \frac{\alpha_{\bar{\mu}}}{\beta_{\bar{\mu}}}, \\ \text{Var}(\bar{\mu}) &= \frac{\alpha_{\bar{\mu}}}{\beta_{\bar{\mu}}^2}. \end{aligned}$$

When there are no fossil calibrations in the species tree, the rates should all be relative. In this case we suggest fixing $\alpha_{\bar{\mu}} = \beta_{\bar{\mu}} = 0$ (with 0 causing the program to fix both parameters at ∞), and the program will fix the mean rate across loci at $\bar{\mu} = 1$. Otherwise one can use equal and large values such as $\alpha_{\bar{\mu}} = 100$ and $\beta_{\bar{\mu}} = 100$, so that $\bar{\mu}$ is nearly fixed at 1. Given the mean rate $\bar{\mu}$, two priors are available to specify the locus rates μ_i, with $i = 1, 2, \cdots, L$, where L is the number of loci (`nloci`). If `prior = dir` (for Gamma-Dirichlet distribution of locus rates), the total rate $L\bar{\mu}$, given the mean rate ($\bar{\mu}$), is partitioned into locus rates (μ_i), using the concentration parameter α_{μ_i} The model and notation follow dos Reis *et al.* (2014, eqs. 3-5). If `prior = iid` (for conditional-i.i.d. or hierarchical prior of locus rates), the locus rates (μ_i) are i.i.d. given the mean rate ($\bar{\mu}$):

$$\mu_i \sim \text{G}(\alpha_{\mu_i}, \alpha_{\mu_i}/\bar{\mu}).$$

This model is described in Zhu *et al.* (2015, eq. 8) and is also implemented in MCMCTREE. In both the Gamma-Dirichlet and the conditional i.i.d. models, parameter α_{μ_i} is inversely related to the extent of rate variation among loci, with a large α_{μ_i} meaning similar rates among loci. If all loci are noncoding, the rates are probably similar, so $\alpha_{\mu_i} = 10$ or 20 may be reasonable, while for coding loci or exons, $\alpha_{\mu_i} = 2$ or 1 may be appropriate. The α_{μ_i} parameter may affect the estimates of the population size parameter (θ) for the root node on the species tree (Burgess and Yang, 2008). The L locus rates (μ_i) are parameters in the model. If $\alpha_{\bar{\mu}} > 0$, the mean rate ($\bar{\mu}$) is a parameter as well.

The setting `locusrate = 3 d d` means that all loci have the same mutation rate, and the rate is estimated using tip dating. The rate prior is $\Gamma(\alpha, \beta)$. The units of the prior should match the units of in the date file. For example, if the dates are specified in years, the rate should be in expected number of substitutions per year.

EXAMPLES

```
locusrate = 0

locusrate = 2 rates.txt

locusrate = 1 0 0 2

locusrate = 1 2 3 2 dir

locusrate = 3 20 1000000

locusrate = 1 1 1 1 iid
```

2 Input file format

2.1 Date File

Each individual has an associated sample age. The **date file** assigns a date to the individual using the individual ID tag (described in the main bpp manual). The dates are in units before time

present, so larger numbers are older. Any desired time units (e.g. years, thousands of year, etc) can be used so long as the mutation rate prior is on the same time scale (e.g. expected substitutions per year, expected substitutions per thousand years). For example, the following date file assigns `specimen1` an age 500, `specimen2` an age of 10000. The population name can be included before the caret ^ symbol or it can be excluded.

```
A^specimen1 500
specimen2 10000
B^specimen3 45000
B^specimen4 35000
```

Each line should have one individual followed by the sample age. Each individual must be included in the date file. See the dates file `mammoth/dates.txt` in the `examples` subdirectory.

3 Simulator

Most of the options in the simulator remain the same. The only differences in the simulator options are related to specifying tip ages. We refer the reader to the main `bpp` manual for a more complete description of the control file options for simulation.

3.1 Control file

`datefile` specifies the name/path to the file containing the sequence ages. The ages are in units of expected number of substitutions, which is not the same as the inference program (see 1.1 for more detail). The dates are assigned to populations. The number of sample dates must match the number of individuals for each population. For example, the below date file has 3 individuals from population A with sample ages of 0.001, 0.0016, and 0.001 in units of expected number of substitutions. There are also two individuals from population B and two individuals from population C.

```
A 0.001
A 0.0016
A 0.001
B .0005
B .0002
C .0014
C .0012
```

`seqDates` specifies the name/path to the output file for the date file with individual names with their sample dates. The `bpp` simulator automatically names the indidivduals and sequences based on the population names. Since the individuals are not named in the input files for the simulator, the simulator must generate a file to match the dates to the individuals. Below is the `seqDates` file that corresponds to the `datefile` shown above.

```
A^a1 0.001000
A^a2 0.001000
A^a3 0.001600
```

```
B^b1 0.000200
B^b2 0.000500
C^c1 0.001200
C^c2 0.001400
```

3.2 Using the simulator

The `seqDates` and `datefile` options must be used together in the simulator. A strict clock model (the default, which is equivalent to `clock = 1`) is required for simulating with tip dates. The simulator uses dates in units of expected substitutions, but the inference program uses dates in calendar time such as years or thousand years. The dates in the `seqDates` file can be converted to a unit of years by dividing by the per year substition rate. This is required to use the simulator to generate data to use in the inference program. For example, if we assume a substitution rate of 10^{-8} per year and use the same example as above, the `datefile` to use for inference would be

```
A^a1 100000
A^a2 100000
A^a3 160000
B^b1  20000
B^b2  50000
C^c1 120000
C^c2 140000
```

`bpp` does not generate a file with the dates in years or similar units. The user must prepare this file after running the simulator.

References

Burgess, R. and Yang, Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.*, 25(9): 1979–1994.

dos Reis, M., Zhu, T., and Yang, Z. 2014. The impact of the rate prior on bayesian estimation of divergence times with multiple loci. *Syst. Biol.*, 63(4): 555–565.

Yang, Z. 1994. Estimating the pattern of nucleotide substitution. *J Mol Evol*, 39(1): 105–11.

Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4): 1811–1823.

Zhu, T., dos Reis, M., and Yang, Z. 2015. Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. *Syst. Biol.*, 64(2): 267–280.