Phylogenetic Inference using `RevBayes`

Model section using Bayes factors

Sebastian Höhna

1 Overview

This tutorial demonstrates how to set up and perform an analysis that computes the marginal likelihood of a given model. We will show how to calculates Bayes factors to select among different substitution model and different partition configurations of aligned DNA sequences.

1.1 Requirements

We assume that you have read and hopefully completed the following tutorials:

- RB_Getting_Started
- RB_Data_Tutorial
- RB_CTMC_Tutorial

This means, we expect that you are able to start `RevBayes`, know some basic commands, how to load data into `RevBayes` and run a single gene phylogenetic analysis (assuming an unconstrained/unrooted tree).

2 Data and files

We provide several data files which we will use in this tutorial. You may want to use your own data instead. In the **`data`** folder, you will find the following files

- **`primates_cytb.nex`**: Alignment of the *cytochrome b* subunit from 23 primates representing 14 of the 16 families (*Indriidae* and *Callitrichidae* are missing).
- **`primates_16s.nex`**: Alignment of the *16s ribosomal RNA* gene from the same 23 primates species.
- **`primates_cox2.nex`**: Alignment of the *COX-2* gene from the same 23 primates species.

3 Introduction

For most sequence alignments, several (possibly many) substitution models and partition schemes of varying complexity are plausible *a priori*, which therefore requires a way to objectively identify the model that balances estimation bias and error variance associated with under- and over-parameterized models, respectively. Increasingly, model selection is based on *Bayes factors* (*e.g.*, Suchard et al. 2001; Lartillot and Philippe 2006; Xie et al. 2011; Baele et al. 2012; 2013), which involves first calculating the marginal

likelihood under each candidate model and then comparing the ratio of the marginal likelihoods for the set of candidate model.

Given two models, M_0 and M_1, the Bayes factor comparison assessing the relative plausibility of each model as an explanation of the data, $BF(M_0, M_1)$, is:

$$BF(M_0, M_1) = \frac{\text{posterior odds}}{\text{prior odds}}.$$

The posterior odds is the posterior probability of M_0 given the data, $\mathbf{X}$, divided by the posterior odds of M_1 given the data:

$$\text{posterior odds} = \frac{\mathbb{P}(M_0 \mid \mathbf{X})}{\mathbb{P}(M_1 \mid \mathbf{X})},$$

and the prior odds is the prior probability of M_0 divided by the prior probability of M_1:

$$\text{prior odds} = \frac{\mathbb{P}(M_0)}{\mathbb{P}(M_1)}.$$

Thus, the Bayes factor measures the degree to which the data alter our belief regarding the support for M_0 relative to M_1 (Lavine and Schervish 1999):

$$BF(M_0, M_1) = \frac{\mathbb{P}(M_0 \mid \mathbf{X}, \theta_0)}{\mathbb{P}(M_1 \mid \mathbf{X}, \theta_1)} \div \frac{\mathbb{P}(M_0)}{\mathbb{P}(M_1)}. \tag{1}$$

This, somewhat vague, definition does not lead to clear-cut identification of the "best" model. Instead, *you* must decide the degree of your belief in M_0 relative to M_1. Despite the absence of any strict "rule-of-thumb", you can refer to the scale (outlined by Jeffreys 1961) for interpreting these measures (Table 1).

Table 1: The scale for interpreting Bayes factors by Harold Jeffreys (1961).

Strength of evidence	$BF(M_0, M_1)$	$\mathbf{log}(BF(M_0, M_1))$	$\mathbf{log_{10}}(BF(M_0, M_1))$
Negative (supports M_1)	< 1	< 0	< 0
Barely worth mentioning	1 to 3.2	0 to 1.16	0 to 0.5
Substantial	3.2 to 10	1.16 to 2.3	0.5 to 1
Strong	10 to 100	2.3 to 4.6	1 to 2
Decisive	> 100	> 4.6	> 2

For a detailed description of Bayes factors see Kass and Raftery (1995)

Unfortunately, direct calculation of the posterior odds to prior odds ratio is unfeasible for most phylogenetic models. However, we can further define the posterior odds ratio as:

$$\frac{\mathbb{P}(M_0 \mid \mathbf{X})}{\mathbb{P}(M_1 \mid \mathbf{X})} = \frac{\mathbb{P}(M_0)}{\mathbb{P}(M_1)} \frac{\mathbb{P}(\mathbf{X} \mid M_0)}{\mathbb{P}(\mathbf{X} \mid M_1)},$$

where $\mathbb{P}(\mathbf{X} \mid M_i)$ is the *marginal likelihood* of the data marginalized over all parameters for M_i; it is also referred to as the *model evidence* or *integrated likelihood*. More explicitly, the marginal likelihood is the probability of the set of observed data ($\mathbf{X}$) under a given model (M_i), while averaging over all possible values of the parameters of the model (θ_i) with respect to the prior density on θ_i

$$\mathbb{P}(\mathbf{X} \mid M_i) = \int \mathbb{P}(\mathbf{X} \mid \theta_i)\mathbb{P}(\theta_i)dt. \tag{2}$$

If you refer back to equation 1, you can see that, with very little algebra, the ratio of marginal likelihoods is equal to the Bayes factor:

$$BF(M_0, M_1) = \frac{\mathbb{P}(\mathbf{X} \mid M_0)}{\mathbb{P}(\mathbf{X} \mid M_1)} = \frac{\mathbb{P}(M_0 \mid \mathbf{X}, \theta_0)}{\mathbb{P}(M_1 \mid \mathbf{X}, \theta_1)} \div \frac{\mathbb{P}(M_0)}{\mathbb{P}(M_1)}. \tag{3}$$

Therefore, we can perform a Bayes factor comparison of two models by calculating the marginal likelihood for each one. Alas, exact solutions for calculating marginal likelihoods are not known for phylogenetic models (see equation 2), thus we must resort to numerical integration methods to estimate or approximate these values. In this exercise, we will estimate the marginal likelihood for each partition scheme using both the stepping-stone (Xie et al. 2011; Fan et al. 2011) and path sampling estimators (Lartillot and Philippe 2006; Baele et al. 2012).

3.1 Phylogenetic Models

The models we use here are equivalent to the models described in the previous exercise on substitution models (continuous time Markov models). To specify the model please consult the previous exercise. Specifically, you need to specify a

- Jukes-Cantor substitution model
- Hasegawa-Kishino-Yano substitution model
- General-Time-Reversible substitution model
- Gamma distributed rate variation among sites
- Invariant sites model

3.2 Estimating the Marginal Likelihood

With a fully specified model, we can set up the **`powerPosterior()`** analysis to create a file of 'powers' and likelihoods from which we can estimate the marginal likelihood using stepping-stone or path sampling. This method computes a vector of powers from a beta distribution, then executes an MCMC run for each power step while raising the likelihood to that power. In this implementation, the vector of powers starts with 1, sampling the likelihood close to the posterior and incrementally sampling closer and closer to the prior as the power decreases.

Just to be safe, it is better to clear the workspace (if you did not just restart `RevBayes`)

```
clear()
```

Now set up the model as in the previous exercise. You should start with the simple Jukes-Cantor substitution model. Setting up the model will involve

1. Loading the data and retrieving useful variables about the data (e.g., number of sequences and taxon names).

2. Specify the rate matrix of the substitution model.
3. Specify the tree model including branch length variables.
4. Create a random variable for the sequences that evolved under the **PhyloCTMC**.
5. Clamp the data.
6. Create a model object.
7. Don't forget the moves!

The following description of how-to specify a marginal likelihood computation is valid for any model in RevBayes. You need to repeat this later for other models. First, we create the variable containing the power posterior analysis. This requires us to provide a model and vector of moves, as well as an output file name. The **cats** argument sets the number of power steps.

```
pow_p = powerPosterior(mymodel, moves, "model1.out", cats=50)
```

We can start the power posterior analysis by first burning in the chain and and discarding the first 10000 states. This will help the analysis to start from the posterior distribution instead of any random parameter state.

```
pow_p.burnin(generations=10000,tuningInterval=1000)
```

Now execute the run with the **.run()** function:

```
pow_p.run(generations=1000)
```

Once the power posteriors have been saved to file, create a stepping stone sampler. This function can read any file of power posteriors and compute the marginal likelihood using stepping-stone sampling.

```
ss <- steppingStoneSampler(file="model1.out", powerColumnName="power", likelihoodColumnName="
    likelihood")
```

Compute the marginal likelihood under stepping-stone sampling using the member function **marginal()** of the **ss** variable and record the value in Table **??**.

```
ss.marginal()
```

Path sampling is an alternative to stepping-stone sampling and also takes the same power posteriors as input.

```
ps = pathSampler(file="model1.out", powerColumnName="power", likelihoodColumnName="likelihood")
```

Compute the marginal likelihood under stepping-stone sampling using the member function **marginal()** of the **ps** variable and record the value in Table **??**.

```
ps.marginal()
```

→ As an example we provide the file **RevBayes_scripts/marginalLikelihood_JukesCantor.Rev**.

3.3 Exercises

- Compute the marginal likelihoods of the *cytb* alignment for the following substitution models
  - Jukes-Cantor substitution model
  - Hasegawa-Kishino-Yano substitution model
  - General-Time-Reversible substitution model
  - General-Time-Reversible substitution model with gamma distributed rate variation among sites
  - General-Time-Reversible substitution model with invariant sites
  - General-Time-Reversible substitution model with gamma distributed rate variation among sites and the invariant sites model
- Fill the marginal likelihood estimates in Table 2.
- Repeat the marginal likelihood computation for the *MT-ND1 gene* and fill Table 3.
- Which is the best fitting substitution model?

Table 2: Estimated marginal likelihoods for different substitution models for the cytb alignment*.

	Marginal lnL estimates	
Substitution Model	*Stepping-stone*	*Path sampling*
JC (M_1)		
HKY (M_2)		
GTR (M_3)		
GTR+Γ (M_4)		
GTR+I (M_5)		
GTR+Γ+I (M_6)		
Any other model (M_7)		
Any other model (M_8)		
Any other model (M_9)		

*you can edit this table

Table 3: Estimated marginal likelihoods for different substitution models of the ND1 gene*.

	Marginal lnL estimates	
Substitution Model	*Stepping-stone*	*Path sampling*
JC (M_1)		
HKY (M_2)		
GTR (M_3)		
GTR+Γ (M_4)		
GTR+I (M_5)		
GTR+Γ+I (M_6)		
Any other model (M_7)		
Any other model (M_8)		
Any other model (M_9)		

*you can edit this table

4 Compute Bayes Factors and Select Model

Now that we have estimates of the marginal likelihood under each of our different models, we can evaluate their relative plausibility using Bayes factors. Phylogenetics software programs log-transform the likelihood to avoid underflow, because multiplying likelihoods results in numbers that are too small to be held in computer memory. Thus, we must use a different form of equation 3 to calculate the ln-Bayes factor (we will denote this value $\mathcal{K}$):

$$\mathcal{K} = \ln[BF(M_0, M_1)] = \ln[\mathbb{P}(\mathbf{X} \mid M_0)] - \ln[\mathbb{P}(\mathbf{X} \mid M_1)], \tag{4}$$

where $\ln[\mathbb{P}(\mathbf{X} \mid M_0)]$ is the *marginal lnL* estimate for model M_0. The value resulting from equation 4 can be converted to a raw Bayes factor by simply taking the exponent of $\mathcal{K}$

$$BF(M_0, M_1) = e^{\mathcal{K}}. \tag{5}$$

Alternatively, you can interpret the strength of evidence in favor of M_0 using the $\mathcal{K}$ and skip equation 5. In this case, we evaluate the $\mathcal{K}$ in favor of model M_0 against model M_1 so that:

if $\mathcal{K} > 1$, then model M_0 wins
if $\mathcal{K} < -1$, then model M_1 wins.

Thus, values of $\mathcal{K}$ around 0 indicate ambiguous support.

Using the values you entered in Table 2 and equation 4, calculate the ln-Bayes factors (using $\mathcal{K}$) for the different model comparisons. Enter your answers in Table 4 using the stepping-stone and the path-sampling estimates of the marginal log likelihoods.

Table 4: Bayes factor calculation*.

Model comparison	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
M_1	-								
M_2		-							
M_3			-						
M_4				-					
M_5					-				
M_6						-			
M_7							-		
M_8								-	
M_9									-

*you can edit this table

5 Partitioned Analysis

Now that you have identified the best substitution model for each of the two genes, we will run a joint analysis under both genes: a partitioned analysis.

Questions about this tutorial can be directed to:

- Tracy Heath (email: tracyh@berkeley.edu)
- Michael Landis (email: mlandis@berkeley.edu)
- Sebastian Höhna (email: sebastian.hoehna@gmail.com)

References

Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. Suchard, and A. Alekseyenko. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution 29:2157–2167.

Baele, G., W. Li, A. Drummond, M. Suchard, and P. Lemey. 2013. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Molecular Biology and Evolution 30:239–243.

Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis. 2011. Choosing among partition models in bayesian phylogenetics. Molecular Biology and Evolution 28:523–532.

Jeffreys, H. 1961. The theory of probability. Oxford University Press.

Kass, R. and A. Raftery. 1995. Bayes factors. Journal of the American Statistical Association 90:773–795.

Lartillot, N. and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. Systematic Biology 55:195.

Lavine, M. and M. J. Schervish. 1999. Bayes factors: what they are and what they are not. The American Statistician 53:119–122.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time markov chain evolutionary models. Molecular Biology and Evolution 18:1001–1013.

Xie, W., P. Lewis, Y. Fan, L. Kuo, and M. Chen. 2011. Improving marginal likelihood estimation for bayesian phylogenetic model selection. Systematic Biology 60:150–160.

Version dated: January 21, 2015