Bayesian Phylogenetic Methods

Z Yang, University College London, London, UK

Introduction

Bayesian statistics has the unique feature that uncertainties in all unknowns (such as the unknown parameters in a model or the competing hypotheses for explaining the same data) are described using statistical distributions. In classical statistics (such as the maximum likelihood method), parameters and hypotheses cannot be assigned distributions. Suppose one wants to analyze the data (x) to estimate the unknown parameter θ under a model. In a Bayesian analysis, one assigns a distribution on θ before the analysis of the data. This is called the 'prior distribution' and reflects one's knowledge or belief about the possible values of θ. The Bayesian analysis of the data then produces the distribution of θ given the data, $f(\theta|x)$, called the 'posterior distribution.' The two are related through the Bayes theorem

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} \propto f(\theta)f(x|\theta) \qquad [1]$$

Here the probability of the data given the parameter θ, $f(x|\theta)$, is the likelihood, and represents the information about the parameter θ in the data x. The marginal probability of the data, $f(x) = \int f(\theta)f(x|\theta)\mathrm{d}\theta$, is a normalizing constant, and its role is to ensure that $f(\theta|x)$ is a proper statistical distribution and integrates to 1. Equation [1] thus says that the posterior is proportional to the prior times the likelihood, or equivalently, the posterior combines information in the prior and in the data sample.

Note that the likelihood function is the basis for classical statistical methods, especially the maximum likelihood method. Thus all models developed for the maximum likelihood method can be implemented in the Bayesian framework. In analysis of large datasets, the two methods often produce numerically very similar results even though the interpretations differ. However, different results may be obtained by the two methodologies if the data are not informative, and in particular, if the focus of the analysis is on model selection.

In molecular phylogenetics, the data x is an alignment (or alignments) of sequences of nucleotides, codons, or amino acids from several species. Here, we assume that the sequences are already aligned and we ignore alignment errors. Our focus is the phylogenetic tree, which consists of the tree topology (τ) and the lengths of branches (denoted collectively as $\boldsymbol{b}$). The branch length is measured by the expected number of substitutions per site, and quantifies the amount of evolution along the branch. Given the tree, the sequence data at the tips of the tree (for extant species) are the product of the process of sequence evolution along the branches. This process is typically described by a continuous time Markov chain (Felsenstein, 1981). The model of substitution may include additional parameters, denoted ϕ, such as the relative substitution rates between nucleotides and the equilibrium frequencies of the nucleotides. More complex models may include parameters to describe the rate variation across sites in the sequence or the nonsynonymous/synonymous substitution rate ratio in comparisons of protein-coding gene sequences (Yang, 1993; Goldman and Yang, 1994). For more details on various models used in molecular phylogenetics.

The posterior distribution of the tree topology, branch lengths, and substitution parameters is then given by eqn [1] with parameter θ replaced by τ, $\boldsymbol{b}$, and ϕ:

$$f(\tau,\boldsymbol{b},\phi|x) \propto f(\phi)f(\tau,\boldsymbol{b})f(x|\tau,\boldsymbol{b},\phi) \qquad [2]$$

Here $f(\phi)$ is the prior distribution on substitution parameters, $f(\tau, \boldsymbol{b})$ is the prior on tree topology and branch lengths, while $f(x|\tau, \boldsymbol{b}, \phi)$ is the likelihood or probability of the sequence data given the tree topology and branch lengths, given by the model of sequence evolution (Felsenstein, 1981).

The Bayesian approach to molecular phylogenetics was introduced by Rannala and Yang (1996), Yang and Rannala (1997), Mau and Newton (1997), and Li *et al.* (2000). The early studies used simple models of sequence evolution and assumed a constant rate of evolution (the molecular clock). Nowadays, we have several Bayesian phylogenetic programs that implement a wide range of complex models that account for various aspects of the sequence data. General Bayesian programs for phylogeny reconstruction include MrBayes (Ronquist *et al.*, 2012), BEAST (Drummond and Rambaut, 2007), and PhyloBayes (Lartillot *et al.*, 2009). A number of Bayesian programs are also available for estimating species divergence times incorporating information in both fossils and molecules, such as MCMCTREE (Yang, 2007) and DPPDIV (Heath *et al.*, 2012).

For an extensive discussion of Markov chain Monte Carlo (MCMC) algorithms used in Bayesian phylogenetics, see Chapters 7 and 8 of Yang (2014). The edited book by Chen *et al.* (2014) summarizes recent developments, especially concerning model selection in Bayesian phylogenetics.

Priors

The prior distribution is supposed to summarize one's objective information (according to 'Objective Bayesian') or personal beliefs (according to 'Subjective Bayesian') about the likely values of the model parameters. In Bayesian phylogenetics, the tree topologies (τ) represent discrete statistical models, the branch lengths ($\boldsymbol{b}$) are continuous parameters that are defined only on specific trees, while the substitution parameters (ϕ) are often defined for all possible trees. The parameter space of the inference problem is high-dimensional and also complex. Specification of the prior is thus a nontrivial task. Indeed, a few cases have been identified in which innocent-looking priors adopted in common Bayesian programs lead to unreasonable extreme results.

Here we describe a few commonly used prior distributions in Bayesian phylogenetics. First we consider the prior on the tree topology. Most phylogenetic analyses are conducted

doi:10.1016/B978-0-12-800049-6.00208-0

without assuming the molecular clock and use unrooted trees. The total number of unrooted trees T_n for n species is

$$T_n = (2n–5)(2n–7)\cdots 1 \qquad [3]$$

It is common to assign a uniform prior on all possible trees, with each assigned the probability $1/T_n$.

If the species are closely related, the evolutionary rate may be roughly constant among species. One can then use the molecular clock to infer rooted trees. Rooted trees are also used to infer species divergence times in the so-called molecular clock or relaxed-clock dating analysis. A prior distribution over the rooted trees and node ages (branching times) can be generated using a model of cladogenesis. For example, a birth–death process conditioned on the number of observed or sampled species can be used to describe the biological process of speciation and extinction, and to generate a prior on the rooted tree topologies and node ages. The birth–death process includes the Yule pure-birth process as a special case. Parameters for the birth–death model include the birth rate, the death rate, and the sampling fraction (the proportion of extant species that are actually included in the data). Those parameters in the prior can be changed to assess the impact of the prior on the posterior inference, or they may be estimated from the data by assigned prior distributions on them (called 'hyper-priors').

For DNA sequences sampled from the same species, Kingman's (1982) coalescent process provides a prior distribution for the gene genealogies. However, this is not a suitable prior model for inferring species phylogenies.

Next, we consider the prior for branch lengths. A binary unrooted tree for n species has $2n - 3$ branches. Given each unrooted tree topology, the $2n - 3$ branch lengths can be assigned independent and identical distributions (i.i.d.) such as the uniform or exponential. In the case of the uniform, an upper bound (such as 100) is specified by the user. However, those i.i.d. priors on branch lengths have been found to be problematic, as they may be very informative and unreasonable about the tree length (sum of branch lengths) (Rannala *et al.*, 2012). For example, a tree of 100 species has 197 branch lengths. If each is assigned the uniform prior U(0, 100), the tree length will have the prior mean 9850 and the 99% prior interval (8806, 10 894), with ~10 000 substitutions at an average site. When the data are not very informative (as is the case when the sequences are highly similar), this unreasonable prior can overwhelm the Bayesian analysis and leads to unreasonably long trees with large tree lengths (Brown *et al.*, 2010). An alternative has been suggested to fix this problem (Rannala *et al.*, 2012; Zhang *et al.*, 2012), in which a gamma prior is assigned to the tree length and then the sum is partitioned into branch lengths according to a uniform Dirichlet distribution (a multivariate extension of the uniform distribution).

Markov Chain Monte Carlo

Note that the normalizing constant $f(x)$ in eqn [1] involves an integral. When there are many parameters in the model, this integral will be multidimensional and may be very hard to compute. Modern Bayesian inference is often achieved through a computational algorithm called MCMC. This is an iterative simulation algorithm that generates a sample from the posterior distribution $f(\theta|x)$.

Here we illustrate the main features of the MCMC algorithm by applying it to the simple phylogenetic problem of estimating the distance θ between two sequences under the JC69 model (Jukes and Cantor, 1969). The data consist of the human and orangutan mitochondrial 12S rRNA genes, with $x=90$ differences at $n=948$ sites. The parameter θ is the expected number of nucleotide substitutions per site between the two sequences. Given θ, the likelihood or the probability of observing the data is given by the binomial probability

$$f(x|\theta) = p^x(1-p)^{n-x} = \left(\tfrac{3}{4} - \tfrac{3}{4}e^{-4\theta/3}\right)^x \left(\tfrac{1}{4} + \tfrac{3}{4}e^{-4\theta/3}\right)^{n-x}, \qquad [4]$$

where $p = \frac{3}{4} - \frac{3}{4}e^{-4\theta/3}$ is the probability that a site is occupied by two different nucleotides in the two sequences separated by a distance θ. We assign a uniform prior on θ in the range (0, 1) so that $f(\theta)=1$ for $0<\theta<1$. The posterior is then given by eqn [1] as

$$f(\theta|x) = \frac{1}{f(x)} f(\theta) f(x|\theta) = \frac{1}{f(x)} \left(\frac{3}{4} - \frac{3}{4}e^{-4\theta/3}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-4\theta/3}\right)^{n-x} \qquad [5]$$

The following algorithm generates a sample from this posterior distribution.

1. Initialize: $n=948$, $x=90$, $w=0.25$. Set initial state: $\theta=0.1$, say.
2. Loop
   a. (Propose a new value θ^*.) Generate $u \sim U(0, 1)$ and set $\theta^* = \theta + w(\frac{1}{2} - u)$. Note that θ^* is a uniform random variable over the interval $U(\theta - \frac{w}{2}, \theta + \frac{w}{2})$. If $\theta^* < 0$, set $\theta^* = -\theta^*$.
   b. (Accept or reject the proposed value.) Compute the posterior density ratio $\alpha = \frac{f(\theta^*|x)}{f(\theta|x)} = \frac{f(\theta^*)f(x|\theta^*)}{f(\theta)f(x|\theta)}$. If $\alpha > 1$, accept θ^*. Otherwise accept θ^* with probability α. This can be achieved by drawing another random number $v \sim U(0, 1)$, and accepting θ^* if and only if $v < \alpha$. If θ^* is accepted set $\theta = \theta^*$. Otherwise set $\theta = \theta$.
   c. Print out θ.

It is easy to see that the algorithm simulates a Markov chain; the next θ value the algorithm will visit depends on the current θ only, but not the θ values visited in the past. Second, the algorithm tends to visit θ values with high posterior more often than θ values with low posterior. Indeed, the probability that the visited θ value is in the interval $(\theta, \theta + \Delta\theta)$ is $f(x|\theta)\Delta\theta$. In other words, the θ values generated by the algorithm constitute a sample from the posterior distribution $f(x|\theta)$. Lastly, there is no need to compute the normalizing constant $f(x)$ of eqn [5] since it cancels in the calculation of the posterior ratio α in Step 2b. This is the feature that allows us to avoid the calculation of the high-dimensional integrals, making it possible to implement sophisticated parameter-rich models that may not be feasible for maximum likelihood implementation.

Figure 1(a) shows the paths of two Markov chains from two runs of the algorithm, using different starting positions. **Figure 1(b)** shows the histogram and smoothed density estimate of posterior using a large sample from a long chain.

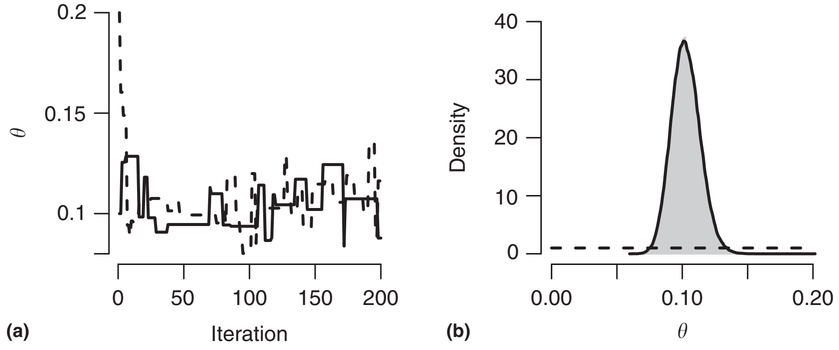

Figure 1 Markov chain Monte Carlo algorithm to sample from the posterior for the JC69 distance θ between two sequences. (a) Trace plot of two chains which started from different positions (0.1 and 0.2), each run over 200 iterations. (b) Histogram (shaded area) and smoothed density (solid curve) of the posterior sample obtained by running the algorithm over 10^6 iterations. The prior (dashed line) is shown as well for comparison.

The posterior mean is 0.1027, standard deviation is 0.0110, and the central 95% posterior credibility interval is (0.0824, 0.1253). In comparison, the famous JC69 distance formula gives the maximum likelihood estimate to be $\hat{\theta} = 0.1015$.

In phylogenetic reconstruction, the parameter space consists of several components: the tree topology τ, the branch lengths $\boldsymbol{b}$, and the substitution parameters ϕ. In each iteration, the different components may be updated in turn. For example, variants of tree search algorithms such as nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) can be used to update the tree topology. The branch lengths and substitution parameters can be updated using sliding windows, as in the simple MCMC algorithm above. See Chapter 8 of Yang (2014) for a detailed discussion of MCMC proposal algorithms in phylogenetics. The phylogenetic MCMC algorithm generates a sample from the joint posterior distribution of the tree topologies (τ), the branch lengths ($\boldsymbol{b}$), and the substitution parameters (ϕ).

Output Analysis from Simulation

The MCMC sample from the posterior distribution can be summarized in different ways.

For scalar parameters such as branch lengths ($\boldsymbol{b}$) and substitution parameters (ϕ), the posterior means or medians are often used, together with the 95% posterior credibility intervals (CIs). Two types of intervals are commonly used. The 95% central (equal-tail) CI lies between the 2.5% and 97.5% quantiles of the posterior sample. The highest posterior density (HPD) CI includes values that make up 95% of the posterior probability and that have the highest posterior density. When the data are informative so that the posterior of the parameter is nearly symmetrical, the two intervals will be nearly identical. Otherwise they can be very different. The HPD interval is generally preferred over the equal-tail interval since it has the shortest length and includes only the most likely parameter values.

For the tree topology, a simple summary is the 'maximum a posteriori' (MAP) tree, which is the tree topology with the highest posterior probability (that is, the tree topology that is most visited during the MCMC algorithm). This gives a point estimate of the true tree. However, when the data are not very informative, the MAP tree may have a very low posterior probability, and is a poor summary. We also have an analogue of interval estimates for trees. The 95% credible set of trees contains those trees that have the highest posterior probabilities such that the total probability of the entire set is at least 95%. However, if this set contains a large number of trees, it will not be very useful.

The most commonly used summary is the so-called majority-rule consensus tree. Note that each internal branch defines a split (a bipartition) of the species. The majority-rule consensus tree includes splits that appear in at least half of the trees sampled, with the posterior probability of each split indicated on the internal branch of the tree. For more details on constructing consensus trees.

See also: Consensus Methods, Phylogenetic. Directed Evolution, History of. Maximum Likelihood Phylogenetic Inference. Molecular Evolution, Models of. Phylogenetic Invariants. Searching Tree Space, Methods for

References

Brown, J.M., Hedtke, S.M., Lemmon, A.R., Lemmon, E.M., 2010. When trees grow too long: Investigating the causes of highly inaccurate Bayesian branch-length estimates. Systematic Biology 59, 145–161.

Chen, M.-H., Kuo, L., Lewis, P., 2014. Bayesian Phylogenetics: Methods, Algorithms, and Applications. London: Chapman & Hall/CRC.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7, 214.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17, 368–376.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11, 725–736.

Heath, T.A., Holder, M.T., Huelsenbeck, J.P., 2012. A Dirichlet process prior for estimating lineage-specific substitution rates. Molecular Biology and Evolution 29, 939–955.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. New York, NY: Academic Press, pp. 21–123.

Kingman, J.F.C., 1982. The coalescent. Stochastic Processes and Their Applications 13, 235–248.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.

Li, S., Pearl, D., Doss, H., 2000. Phylogenetic tree reconstruction using Markov chain Monte Carlo. Journal of the American Statistical Association 95, 493–508.

Mau, B., Newton, M.A., 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. Journal of Computational and Graphical Statistic 6, 122–131.

Rannala, B., Yang, Z., 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Journal of Molecular Evolution 43, 304–311.

Rannala, B., Zhu, T., Yang, Z., 2012. Tail paradox, partial identifiability and influential priors in Bayesian branch length inference. Molecular Biology and Evolution 29, 325–335.

Ronquist, F., Teslenko, M., van der Mark, P., *et al.*, 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61, 539–542.

Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10, 1396–1401.

Yang, Z., 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24, 1586–1591.

Yang, Z., 2014. Molecular Evolution: A Statistical Approach. Oxford: Oxford University Press.

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo Method. Molecular Biology and Evolution 14, 717–724.

Zhang, C., Rannala, B., Yang, Z., 2012. Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. Systematic Biology 61, 779–784.