RevBayes

SEBASTIAN HÖHNA[1], JOHN P. HUELSENBECK[2], FREDRIK RONQUIST[3]

[1]*Department of Mathematics, Stockholm University, Stockholm, Sweden*
[2]*Department of Integrative Biology, University of California, Berkeley*
[3]*Swedish Museum of Natural History, Stockholm, Sweden.*

RevBayes is a program for the Bayesian estimation of phylogeny. The program takes as input character matrices, such as alignments of DNA or amino acid sequences. RevBayes uses a numerical method called Markov chain Monte Carlo (MCMC; Metropolis et al., 1953; Hastings, 1970) to approximate the posterior probabilities of phylogenetic trees; the program's output consists of files containing the MCMC samples. The MCMC samples can then be summarized in a variety of ways to make inferences about the phylogeny of the group of interest as well as other parameters of the phylogenetic model.

RevBayes grew out of developments in the MrBayes program (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Although the MrBayes program was quite popular, F.R. and J.P.H. made design decisions when writing MrBayes that made it difficult to accommodate new developments in the field. For example, MrBayes — like all phylogenetic programs — considers the phylogenetic model to be fixed. However, new developments allow the phylogenetic model itself to be considered an object of inference (*e.g.*, Huelsenbeck et al., 2004). Similarly, although MrBayes allows the user to partition data and model the evolutionary process independently in each data subset, it does not allow the partitioning scheme itself to be an object of inference. Instead, the user must specify the partition, which then remains an assumption of the analysis. New developments in the field, however, allow the partition itself to be considered a parameter to be estimated (Lartillot and Philippe, 2004; Huelsenbeck et al., 2006; Huelsenbeck and Suchard, 2007) or allow the evolution at an alignment site to be considered a mixture from several different models (Pagel and Meade, 2004, 2005). Implementing these new developments would require extensive rewriting of the MrBayes code.

One other limitation of the MrBayes program strongly guided the development of RevBayes: MrBayes uses a method for specifying evolutionary models that is quite limited. MrBayes, in an exercise of Müllerian mimicry[1], uses a command line interface and method for specifying models that closely resembles that used by PAUP* (Swofford, 1998), arguably the phylogeny program that has been the most influential in the field. However, we needed a language that was well-suited to specifying probability models which, after all, are what

[1] Müllerian mimicry, for the unfamiliar reader, is a phenomenon where two species that are both distasteful to predators evolve the same warning coloration. This is by contrast to Batesian mimicry, where a palatable species evolves a coloration similar to a distasteful species. We thought it unkind to describe PAUP* as the exclusively distasteful species when using the mimicry metaphore; we therefore call MrBayes a Müllerian mimic.

a Bayesian analysis assumes, and that could be easily extended; the MrBayes model-specification method placed constraints on how easily complex models could be described. Moreover, we wanted a language that could, in principle at least, allow the user to specify evolutionary models that have never been considered before. After careful consideration of the limitations of the MrBayes design and language, the decision was made to discontinue development of MrBayes and completely rethink how a phylogenetic program — specifically a Bayesian phylogenetic program — should be structured. The result is the RevBayes program.

This manual is intended to provide a background to the ideas implemented in the program and guidance on how to use the program. RevBayes is a purely Bayesian program. This simplifies matters from the perspective of the RevBayes development team because they do not need to consider ideas and methods that are not Bayesian. However, Bayesian statistical analysis can be quite complicated. One goal of this manual is to supply the user the necessary background information on Bayesian analysis to perform an adept Bayesian phylogenetic analysis. Similarly, phylogenetic analysis has become an incredibly baroque field with an extensive literature. Another goal of this manual is to provide the user background on phylogenetic models. Finally, RevBayes uses a new language for specifying phylogenetic models that is quite powerful and similar to the R language (Ihaka and Gentleman, 1996). (The program has switched models and now is a Müllerian mimic of R.) This manual provides background information necessary to specify complex phylogenetic models using RevBayes.

Bayesian Inference

Statistics is a set of methods for making inferences about the world in the face of uncertainty. In general, the statistical approach considers the data as potentially variable and assumes a probability model to describe this variability. For example, on tossing a coin the uncertainty in the outcome can be described by the Bernoulli model in which the probability of observing a head is θ and the probability of observing a tail is $1 - \theta$ (where $0 \leq \theta \leq 1$). The Bernoulli model expresses the uncertainty of the outcome of a coin toss by a single parameter, θ. One of the main goals of statistical inference is to assign values to parameters of a model based on observations, a process called 'estimation'. For coin tossing, the goal may be to estimate the value of θ (the parameter of the Bernoulli model) based on the results of tossing a coin repeatedly (the data)[2].

There are many methods for estimating the parameters of a statistical model. Here, we will consider only two: the method of maximum likelihood, first described by the great population geneticist and statistician R. A. Fisher, and

[2] Throughout this manual, we use the convention of denoting the observations as X and the parameters using greek letters, such as θ. We also use the convention of describing a conditional probability as $f(\cdot \mid \cdot)$. Because all probability functions are described in this manner, there is the potential for confusion. However, it should be clear from the context and the parameters of the function what its identity is and whether the probability distribution is discrete or continuous.

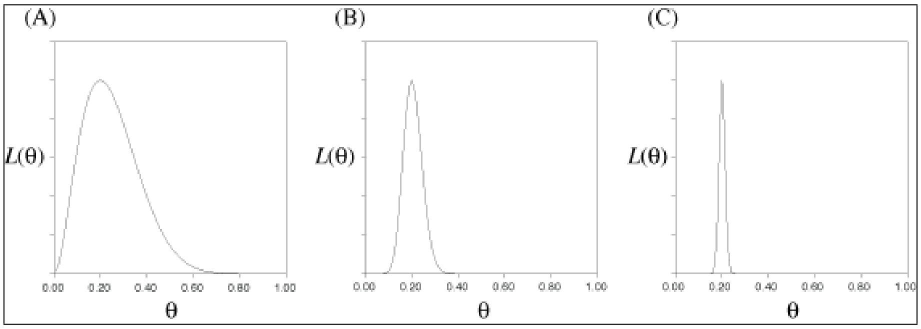

Figure 1: The likelihood function for three different sets of observations that all have the property of having the same maximum likelihood estimate of $\hat{\theta} = 0.2$. A, $n = 10$ and $x = 2$; B, $n = 100$ and $x = 20$; C, $n = 1000$ and $x = 200$.

Bayesian estimation. Maximum likelihood calls the best estimate of a parameter that parameter value that maximizes the probability of the observations. This estimate is called the 'maximum likelihood estimate' (or 'MLE'). The probability of observing the data is called the likelihood function:

$$L(\text{Parameter}) = C \times \Pr(\text{Observations} \mid \text{Parameter})$$

where the constant C is arbitrary, but allows both continuous and discrete probability distributions to be evaluated.

A simple example illustrates the method of maximum likelihood. Consider the problem of estimating the probability that heads appear face up on a single toss of a coin. For a fair coin, the probability that heads lands face-up on a single toss is $\theta = 1/2$. However, one can also estimate the probability of heads for any particular coin; perhaps one is interested in testing whether the coin is, indeed, a fair one. In this case, the parameter θ is considered a parameter of the statistical model, and is allowed to vary. As a scientist, the natural way to determine whether some coin is in fact fair is to toss the coin many times (*i.e.*, the scientist performs an experiment). One way to summarize the results of a coin-tossing experiment is to simply note the number of times heads land face up on n tosses of the coin. We will denote the number of heads observed on n tosses x. The likelihood for the coin tossing experiment is

$$L(\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

which is the binomial probability distribution. (The binomial coefficient, $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, is the number of ways to choose x objects from n.)

Figure 1 plots the parameter θ against the likelihood for several different experimental outcomes. Note that the likelihood appears to be maximized when θ is equal to the proportion of heads that were observed. With a modest amount of calculus, one can show that the likelihood is in fact maximized at $\theta = x/n$.

Bayesian Inference of Phylogeny

Phylogenetic Models

Graphical Representation of Probability Models

Acknowledgments

Sebastian Höhna thanks his mütterchen. The development of RevBayes was partially supported by grants from the NSF (DEB-0918791) and NIH ([NUMBER]) awarded to J.P.H.

References

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Huelsenbeck, J. P., S. Jain, S. W. D. Frost, and S. L. K. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proceedings of the National Academy of Science, U.S.A. 103:6263–6268.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Molecular Biology and Evolution 21:1123–1133.

Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Huelsenbeck, J. P. and M. Suchard. 2007. A nonparametric method for accommodating and testing across-site rate variation. Systematic Biology 56:975–987.

Ihaka, R. and R. Gentleman. 1996. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5:299–314.

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution 21:1095–1109.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21:1087–1092.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology 53:571–581.

Pagel, M. and A. Meade. 2005. Mixture models in phylogenetic inference. Pages 121–142 *in* Mathematics of Evolution and Phylogeny (O. Gascuel, ed.) Oxford University Press.

Ronquist, F. and J. P. Huelsenbeck. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Swofford, D. L. 1998. PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods. Sinauer Associates, Inc., Sunderland, Massachusetts.