BAYESIAN STATISTICS IN PHYLOGENETIC INFERENCE

J.F. Scheepens

Sept-Oct 2006

Marine Biology

Supervised by Prof. Dr. W.T. Stam

Reference of front cover picture: http://facweb.cti.depaul.edu/bioinformatics/

Abstract

The development of traditional approaches to phylogeny inference has been governed by the trade-off between available time and accuracy of the method. The Neighbor-Joining method is the simplest and fastest of the discussed methods, but its lack of an optimality criterion (i.e. a numerical formula which calculates the optimal phylogeny according to the criterion) severely affects the probability to find the true phylogenetic tree. The Maximum Parsimony method includes an optimality criterion when it searches for the tree with the least nucleotide substitutions. The main problem with this method is that it thereby can underestimate the number of nucleotide changes that took place in reality such depending on the evolutionary time involved. The Maximum Likelihood method searches the tree that fits the data best by calculating likelihood values for the data regarding each possible tree. Although this approach is very accurate, it is computationally heavy and therefore restricted with regard to the number of taxa and/or number of informative nucleotide positions used in the phylogenetic inference.

Bayesian inference of phylogeny, like Maximum Likelihood, includes a likelihood value, but transforms it into a posterior probability which indicates the degree of belief for a certain tree regarding the data. The procedure allows the incorporation of prior knowledge, which is subsequently updated in the posterior probability in light of the new data. Like the Maximum Likelihood method, Bayesian inference is very accurate but slow. The development of the Markov chain Monte Carlo algorithm, which estimates posterior probabilities by regarding a subset of all trees slowly converging to the optimal tree, made the use of Bayesian inference of phylogeny feasible. Because Bayesian inference with MCMC is both accurate and fast, it has gained much attention and is increasingly used in answering phylogenetic questions.

Bayesian inference of phylogeny deals with some difficulties. First, the use of prior knowledge is controversial because it is thought to introduce subjectivity in the calculation which is harmful when the prior exerts much influence on the outcome of the calculation. Second, it is hard to guess when the MCMC has run long enough in order to having converged to the optimal tree. Third, Bayesian inference is said to be too liberal, i.e. presenting posterior probabilities which are too high. Tricks exist that overcome the first two points of critic, the use of prior knowledge and the length of the MCMC run. The third point, concerning the liberality of the calculation, is a more difficult problem, which will hopefully be solved in the future. Despite these difficulties, Bayesian inference is still the best method available at the moment since it is both fast and accurate.

Contents

Introduction

The theory of evolution states that species are related to each other by descent (Darwin 1859). Phylogenetics is the branch of biology that tries to reconstruct these relationships between taxa based on shared characters. In classical phylogeny, morphological, physiological or behavioural characters are regarded. Since the rise of molecular biology, phylogenetic inference methods are increasingly applied to DNA sequence data (Hillis *et al.* 1994). Because of the vast amount of characters obtained (each nucleotide is a character), manual inference of phylogenies has become unfeasible and computer algorithms (i.e. mathematical step-by-step procedures to infer phylogenies) are needed.

Phylogenies can be computationally inferred from DNA sequence data in many different ways. During the last decades, numerous methods have been designed and improved to approximate real molecular evolution better and better, thereby obtaining more confident phylogenic reconstructions (Lewis 2001). However, limits are set to the complexity of algorithms due to the exponential (or faster) increase in computing time with increasing number of taxa although the increasing power of desktop computers allows the use of larger datasets; there is a trade-off between available time and the complexity of the method used (Felsenstein 2004). For instance, the Maximum Likelihood method is one of the best performing methods to date, but it is very slow, so it can only be used when the number of taxa and/or number of informative nucleotide positions is small. The Maximum Parsimony method, though less precise, is a much faster method which is preferred when dealing with large datasets.

During the last several years, Bayesian inference of phylogeny has proven to be very fruitful since it overcomes some of the problems with older inference methods, especially the processing speed. It is both fast and accurate (Huelsenbeck *et al.* 2001). However, there is also critic on Bayesian methods, for instance concerning the use of prior knowledge (Shoemaker *et al.* 1999; Lewis 2001; Holder & Lewis 2003). The question central to this review is whether the Bayesian approach is better compared to traditional methods of phylogeny inference.

In this review, the three most popular traditional approaches to phylogeny inference, i.e. Neighbor-Joining, Maximum Parsimony and Maximum Likelihood, will be explained and advantages and disadvantages of these methods will be highlighted. Subsequently, Bayesian statistics, Bayesian inference of phylogeny and the Markov chain Monte Carlo (MCMC) algorithm will be explained. Then, the advantages and disadvantages of Bayesian over traditional phylogeny inference will be discussed. Finally, a conclusion will be drawn regarding the usability of the Bayesian approach based on the discussed advantages and disadvantages.

Traditional approaches to phylogeny inference

First of all, it should be made clear that "traditional approaches" are defined as approaches which predate the Bayesian method of phylogeny inference. It bears no notion of being outdated whatsoever. The Bayesian approach is relatively new, still controversial and quite different from the other methods that it is convenient to group all other (i.e. older) approaches together and call them traditional. The traditional methods that will be discussed are the Distance-Matrix method Neighbor-Joining and the two discrete data based methods Maximum Parsimony and Maximum Likelihood.

Neighbor-Joining

The oldest approaches to phylogeny inference are based on distance matrices, such as Neighbor-Joining (Saitou & Nei 1987), which is still in use as it is a very fast method. In Distance-Matrix methods a matrix (taxa x characters) is created which values indicate the distance between pairs as the fraction of overall similarity between the data. These similarity values are used to construct the tree. Various Distance-Matrix methods, including Neighbor-Joining, search for the tree which minimizes all branch length distances in relation to the similarity values between pairs by means of clustering (Felsenstein 2004).

For the calculation of the similarities between sequences, a model of DNA evolution can be used. For example, transitions and transversions of nucleotides can be given equal weight with the Jukes-Cantor model or different weights with Kimura's two-parameter model. Another example of a nucleotide substitution model is the Sankoff algorithm, which calculates the minimum total cost of a certain tree with different costs for different types of nucleotide changes. These models are reversible, meaning that a change from A to G is as likely as a change from G to A. The most flexible model of nucleotide substitution rates is the general time-reversible (GTR) model, which allows different rates for each type of nucleotide change at different sites and through time (Felsenstein 2004).

The best nucleotide substitution model can be determined with the computer programme Modeltest (Posada & Crandall 1998), which tests which model explains the data best. In Modeltest a Likelihood Ratio Test algorithm can be performed on the data comparing models in a hierarchical fashion starting with the most general model (Jukes-Cantor) and increasing in model complexity. The algorithm compares the fit of different models to the data by iteratively comparing a simpler model with a more complex model and returning a likelihood value for each comparison. Finally, the programme interprets the results and presents the best-fitting model (Huelsenbeck & Rannala 1997; Felsenstein 2004).

After the similarities have been calculated based on the chosen nucleotide substitution rates, the construction of the tree is a simple and fast calculation. It is also fast compared to other methods because there is no optimality criterion involved, like in Maximum Parsimony, Maximum Likelihood and Bayesian inference. An optimality criterion is a numerical formula which calculates the optimal phylogeny according to a criterion, e.g. the least number of nucleotide changes. The use of an optimality criterion makes the calculation a slow process because it theoretically compares all possible phylogenies, after which it presents the optimal phylogeny (Felsenstein 2004).

The lack of an optimality criterion is also the main drawback of Distance-Matrix methods compared to other methods, because it does not optimise phylogenies according to a certain criterion but it calculates the phylogeny from the distance values. In fact, the mere result of the computation is a tree without anything known about ancestral states of sequences. However, the phylogenies of recently diverged sequences can often be inferred quite well with Neighbor-Joining, but older relationships may obscure multiple nucleotide transitions in one character, which this method does not take into account (Baldauf 2003; Felsenstein 2004).

Maximum Parsimony

The Maximum Parsimony method was developed to be a better alternative to distance methods for phylogeny inference by taking into account an optimality criterion. The Maximum Parsimony method searches for the tree with the least total number of nucleotide changes; it is therefore based on the principle of William of Ockham that the simplest hypothesis that explains the data is the one to be preferred (Felsenstein 1988). The main advantage of this method is that, contrary to the Neighbor-Joining method, Maximum Parsimony takes into account the evolution of the DNA sequence. Similar to Neighbor-Joining, the nucleotide substitution rates can be adjusted to some extent in the Maximum Parsimony method (Felsenstein 2004).

Besides using an optimality criterion to infer phylogeny, another advantage of the Maximum Parsimony method is that gaps, which describe insertions and deletions (indels) of nucleotides after sequence alignment, can be simply coded as a fifth type of 'nucleotide' with a different nucleotide substitution rate. Each insertion or deletion could encompass more than one nucleotide, but should be counted as one evolutionary event. These events should therefore be encoded by one additional nucleotide at the end of the sequence with a different nucleotide substitution rate (Meusnier *et al.* 2004).

A disadvantage of the Maximum Parsimony method is that it can show long branch attraction (Hendy & Penny 1989; Felsenstein 2004). In long branches leading to two different species, the probability that part of the branches show a similar route is higher than the probability of a single change in an interior branch that split the two species in

the far past (Fig. 1). This is particularly a problem in datasets with distantly related species so that long branches are likely to occur.

The underlying cause for long branch attraction is that Maximum Parsimony methods are based on the assumption that the tree with the minimum changes is the best tree. This is a major drawback of Maximum Parsimony since in this way the number of changes is structurally underestimated. In reality multiple changes in one character in a branch are possible, but these are neglected by Maximum Parsimony. Therefore, chance is high that a slightly less parsimonious tree is the real tree (Huelsenbeck *et al.* 2002).

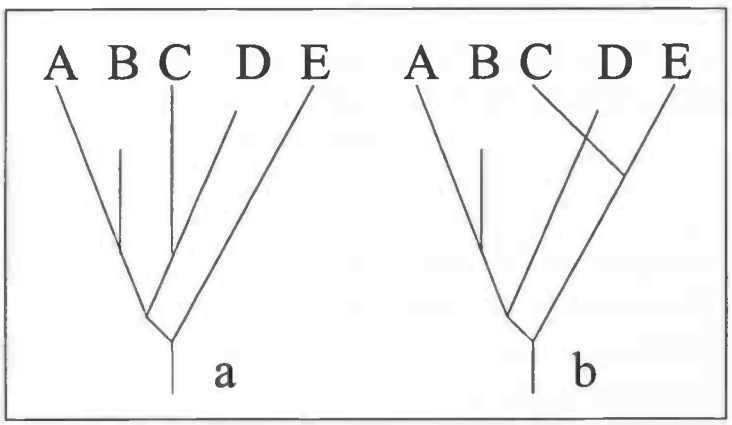

Figure 1. (a) True tree; (b) Long branch attraction resulted in the node between C and E. Based on Felsenstein (2004).

Maximum Likelihood

The Maximum Likelihood method is among the best methods for phylogeny inference (Huelsenbeck & Rannala 1997). It is both accurate (more likely to predict the real tree) and robust (less sensitive to incorrect models and assumptions) (Huelsenbeck & Rannala 1997). In short, the Maximum Likelihood method determines how well each possible tree predicts the data. While calculating likelihood values, the parameter values (e.g. for nucleotide substitution rates) in the model are estimated. The tree with the highest likelihood to predict the data is the resulting tree. However, there are many trees possible for a data set and the search through so-called tree space is multi-dimensional, which explains why this approach is one of the slowest of all (Holder & Lewis 2003).

For the likelihood calculations, a model for nucleotide substitution rates is needed for which the computer programme Modeltest (Posada & Crandall 1998) can be used (explained above). The best performing model can then be implemented in the Maximum Likelihood calculations. (Huelsenbeck & Rannala 1997; Felsenstein 2004).

When a certain model is chosen for the Maximum Likelihood calculation, the actual rates of nucleotide substitutions are still unknown. These parameters are often not of major importance to the question, so they are regarded as 'nuisance' parameters and,

as explained above, they are usually estimated in accordance with the model during the calculations (Holder & Lewis 2003). However, by estimating the nuisance parameters, a large range of possible trees is not regarded, which influences the robustness of the method (Holder & Lewis 2003).

An advantage of the Maximum Likelihood method is that estimation of branch lengths, although often regarded as a nuisance parameter, is an important component of the method. Changes are more likely along long than short branches and branch lengths are therefore important for the outcome of the likelihood calculations. Maximum Parsimony, in contrast, does not take the relationship between substitution chances and branch lengths into consideration; it just minimizes the total length of the branches in the tree. Maximum Likelihood tries to estimate the actual amount of change based on the chosen evolutionary model. Maximum Likelihood therefore results in more accurate trees than Maximum Parsimony (Hillis *et al.* 1996).

Bootstrapping

Neighbor-Joining, Maximum Parsimony and Maximum Likelihood produce phylogenies as a point estimate from which it is impossible to deduce confidence levels for relationships within the tree. Some nodes may be weakly supported by the data while others have strong support. For calculation of confidence levels bootstrapping is a commonly used method. Bootstrapping is a semi-statistical approach because the calculation of confidence levels is based on multiple (often 1000) pseudo-replicate datasets generated from the original data by performing resampling with replacement. Replacement is necessary because otherwise the same dataset would be obtained every resampling. In the case of species x characters matrices, characters are resampled since species are not independent samples (characters might not be independent either, but that problem is not considered here). Some characters are drawn more than once and some not at all; the number of times a character is drawn is a sample from a multinomial distribution. The idea behind resampling is that the variation obtained between pseudo-replicate datasets is typical for variation found when new data would be gathered. From these pseudo-replicate datasets new trees are calculated. Each node in the original tree is given a support value indicating its frequency of occurrence in the bootstrap trees (often as a percentage). Bootstrap values higher than 50% therefore indicate that in more than half of the phylogenies the node was present, which is a weak but positive support. Bootstrap values of 70% or higher indicate strong support. A drawback of this method is that bootstrapping extends the processing time linearly with the number of resamplings. This may especially hamper the use of bootstrapping in the already computationally heavy Maximum Likelihood method (Holder & Lewis 2003).

NP-complete problem

The NP-complete (Non-Deterministic Polynomial-Time Complete) problem is the major obstacle towards accurate phylogeny inference. In short, the more accurate a model needs to be, the more complex it will be, the more processing time it needs to reach that accuracy. Illustrating this, the Neighbor-Joining method carries out a number of operations proportional to the polynomial function n^3, where n is the number of species (Felsenstein 2004). Regarding this function, the number of operations can increase very fast if more species are added. However, the relation between the number of species and the number of operations in methods like Maximum Parsimony and Maximum Likelihood can be described by an exponential function, denoted as e^n, which will always overtake n^x functions if n is large enough. In fact, exponential functions increase so fast that it quickly becomes infeasible to work with them. As the number of operations is related to the processing speed, exponential functions are a general problem in algorithms for phylogeny inference since the time needed for computations sharply rises to immense amounts with increasing number of species. This is called the NP-complete problem. It is assumed that this problem cannot be solved, i.e. the formula cannot be rewritten so that the number of computations needed do not increase exponentially with the number of species. A way to solve this problem is to improve existing or new algorithms so that they allow a higher n before they start to increase into too high number of computations, e.g. $0.0000001e^n$ or $e^{0.001n}$ (Felsenstein 2004).

Bayesian statistics

The first work on Bayesian statistics, formulating Bayes' theorem, was published in 1763, two years after the death of the author Reverend Thomas Bayes (1702-1761). For a long time after the publication of Bayes' work, the proposed Bayesian formula has been forgotten or ignored and the frequentist approach dominated statistical thought. Only during the last decades has Bayesian inference increasingly been applied in most branches of science (Beaumont & Rannala 2004).

Bayesian statistics is based on quite different principles and assumptions than frequency statistics. Frequentists describe the probability of an event as the relative frequency of that event in a hypothetically unlimited number of trials. Contrasting, Bayesian analysis defines probability as a degree of belief in the likelihood of an event. This makes Bayesian inference deal with likelihood values instead of probabilities when regarding the data. Another distinction between the two approaches is that, in the Bayesian approach, the probabilities are based only on observed data, whereas the frequentist approach does involve predictions on data that have not been gathered (Dennis 1996; Quinn & Keough 2002; Ellison 2004).

Bayes' theorem can be stated mathematically as follows (Bayes 1763):

$$\text{Prob}(H \mid X) = \text{Prob}(X \mid H)\ \text{Prob}(H) / \text{Prob}(X)$$

Here, Prob($H \mid X$) means the posterior probability distribution, or the probability that a certain hypothesis is true given the data and the prior probabilities. Prob ($X \mid H$) is the likelihood that the data are true given a certain hypothesis, which is the basis of the Maximum Likelihood method discussed earlier. Prob (H) is the prior probability distribution. It is *a priori* chosen by the experimenter and reflects the pregiven probability that a certain hypothesis is true. The denominator, Prob (X), is a normalizing factor, calculated as the sum of all possible numerators, so that all probabilities add up to one.

Bayes' theorem has two interesting properties. First, it changes prior probabilities to posterior probabilities based on likelihood values and therefore updates degrees of belief that a certain event will occur. Second, it allows estimation of probabilities of one or more hypotheses based on the data. This is in sharp contrast with likelihood functions that give probabilities for the data considering different hypotheses (Prob ($X \mid H$)). It may not be directly clear why hypotheses should be treated as random variables. It is accepted that a probability can be calculated for observed data given a model, like in the Maximum Likelihood method, but there are difficulties to accept that a probability can be calculated for a model given the data. How can a probability be assigned to a one-time event which is either true or false? The reasoning behind this is that different histories (models) are observed to which extent they result in the given dataset. In this way it would actually be better to speak of degree of belief instead of probability to discern it from frequentist nomenclature (Eddy 2004).

To get more insight on Bayes' formula and its power, the following example is given. Consider two species A and B that have an identical morphology and therefore cannot be distinguished by eye (cryptic species). We would like to know the chance that a random species is A given the outcome of a new chemical test. It is already known that A and B occur in a ratio 1:99, so 1% of the cases is species A. This is the prior probability (Prob (H)). To determine a random sample being A or B, the chemical test indicates the presence of species A in 80% of the cases of A. Unfortunately, this test gives a false positive in 9.6% of the cases (indicating A whereas in fact it is species B). Both probabilities indicate the likelihood of either hypothesis (Prob ($X \mid H$)). In order to calculate the probability that a species is A given that the chemical test showed a positive result, which is the posterior probability (Prob($H \mid X$), we need all three pieces of information. According to Bayes' theorem, this probability needs to be calculated as $(0.8*0.01) / (0.8*0.01 + 0.096*0.99) = 0.078$. The probability that the chemical test gives the correct answer is therefore 7.8%, which is much higher than the prior probability of 1% that we deal with species A. So from this example, it becomes clear that

incorporation of the prior probability can have much influence on the outcome. In this particular case, the prior is even necessary to solve the problem. In addition, Bayes' theorem calculates probabilities for different hypotheses based on the data (chance of species A given a positive result). It derives these from likelihood values for the data given certain hypotheses (chance of positive result if species A & chance of positive result if species B). So the important point here is that the probability that a species A gives a positive result (Prob ($X \mid H$)) is not the same as the probability that a positive result implies that the species is A (Prob($H \mid X$)) (Example based on Yudkowsky 2006).

Bayesian phylogeny inference

Both Bayesian analysis and Maximum Likelihood are likelihood-based methods to infer phylogenies. Different tests have shown that both Bayesian analysis and Maximum Likelihood outcompete other methods (Huelsenbeck *et al.* 2002). Bayesian statistics is used to estimate probabilities of different hypotheses based on the data, whereas Maximum Likelihood calculates likelihood values for the data given the hypotheses.

In phylogeny inference, many trees can be produced with a given dataset, so many hypotheses need to be compared. It depends on the chosen model of nucleotide evolution which tree is the most probable. Some models fit a certain dataset better than other models. The Maximum Likelihood approach searches through all possible trees while estimating the nucleotide substitution rates for a certain model of evolution. In this way, an optimal tree will be found which is a point estimate based on the data. As explained above, this approach may take unreasonable amounts of time.

When Bayes' theorem is applied to phylogeny inference, the data (X) are the sequences obtained from the different individuals; the hypotheses (H) are all possible trees. Bayesian analysis does not estimate the nucleotide substitution rate parameters like Maximum Likelihood, but searches through all possible combinations of parameters (tree topology, branch lengths and/or nucleotide substitution rates) and comes up with a likelihood for the data for each possible configuration of parameters. The obtained likelihood value is then used to calculate the posterior probability of that particular configuration of parameters based on the data and in relation to the prior probabilities. A consensus tree can then be created by observing all trees and their posterior probabilities. Based on the frequency of occurrence of nodes in the possible trees, probability values are given to each node. Another possibility is that different tree topologies or particular (groups of) configurations of parameters can be given a posterior probability indicating their chance of occurrence based on the data. In this way, different hypotheses can be compared according to their probability of occurrence given the data. Compared to frequentist approaches, Bayesian inference allows addressing new questions. For example, it allows comparisons between multiple hypotheses like the case of monophyly

against non-monophyly in *Ipomoea* species discussed later on (Huelsenbeck *et al.* 2002). The basis of questioning is not traditionally statistical with a null and alternative hypothesis, but a broad range of outcomes can be compared directly by interpreting their posterior probabilities. Traditional approaches check how unlikely the data would be if a certain situation is assumed. The Bayesian approach questions what the probability is of one or more hypotheses given the data (Shoemaker *et al.* 1999).

In theory, the Bayesian approach has to search in every possible combination of parameter values, which makes the search space multi-dimensional. Theoretical Bayesian inference of phylogeny therefore deals with NP-completeness. It may be clear to the reader that in theory the Bayesian method is not fast at all. In contrast, it involves far more complex calculations than, e.g., the Maximum Likelihood method which time-consuming nature is already a major obstacle to phylogeny inference. The reason why Bayesian analysis has become feasible in practice is because a sophisticated algorithm, called the Markov chain Monte Carlo algorithm, drastically lowers the processing time while still giving accurate results (Felsenstein 2004).

Markov chain Monte Carlo methods

Markov chain Monte Carlo methods (MCMC methods) have been designed to overcome the problem that all hypotheses need to be tested for probability analysis which makes analyses dauntingly slow. MCMC methods calculate optimal trees and their accompanying posterior probabilities with randomly chosen parameter values. A calculated tree is compared with the preceding tree and then either accepted or rejected as the new tree based on whether it improves the posterior probability or not. In this way, the MCMC methods can produce a reliable estimation of the posterior distribution and the accompanied parameters in a very quick way.

The Metropolis algorithm is the most widely used MCMC method and uses the following procedure. It calculates the optimal tree T_i and its posterior probability with randomly chosen parameter values. Then it calculates a neighbouring (i.e. closely related; with slightly changed parameter values) tree T_j and computes the ratio of the posterior probabilities of these trees. If the ratio is >1, the new tree T_j is accepted and the next neighbouring tree T_k is calculated. If the ratio is <1, a random number between 0 and 1 is drawn. If the drawn number is less than the ratio, the new tree T_j is accepted and the next tree T_k calculated. Otherwise, the new tree T_j is rejected and another neighbouring tree T_j is calculated.

In this way, a long chain of trees is calculated that over time converge towards trees with high probabilities. The danger of getting stuck in local optima is lowered by the chance that a calculated tree, although worse than its preceding tree, can be accepted if the ratio is higher than a randomly drawn value. From the long chain of trees, only a subset (e.g. every 1000^{th} tree) is recorded including the accompanying posterior

probabilities. The log-likelihood of these trees is plotted over time, showing a function that increases towards an optimal value (Fig. 2). The initial increasing phase is called the burn-in phase. In this phase, the posterior distribution increases rapidly from the initial tree with random parameter settings towards the region with high posterior probabilities. Only the trees from this seemingly asymptotic phase are used as these contain the trees with the highest posterior probabilities. Based on this subset of high probability trees, a consensus tree can be constructed by observing all recorded trees. The subset of trees can also be used to calculate posterior probabilities for different groups of hypotheses. These probabilities indicate the chance of occurrence of these hypotheses based on the data just like in the full Bayesian approach without using the MCMC algorithm. Concluding, the MCMC approach allows much faster inference of phylogenies and their probabilities by using only a small subset of possible trees. It does not estimate parameters like Maximum Likelihood, but calculates probabilities over the whole spectrum of parameter configurations that make up the multidimensional tree space.

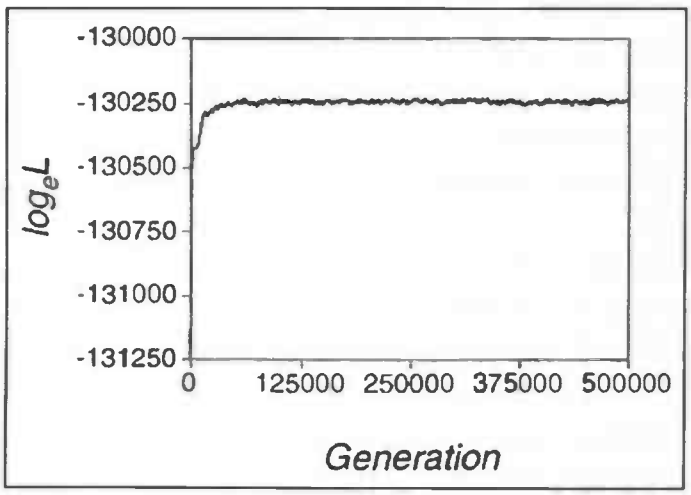

Figure 2. An example of a log-likelihood plot of posterior probabilities (from Huelsenbeck *et al.* 2000).

Critic on and problems with Bayesian phylogeny inference

Prior probability

The incorporation of a prior probability distribution can be seen as an advantage, because previously gathered informative data is not discarded but used in a statistically sound way. These previous probabilities, or beliefs, can then be updated in the light of the newly gathered data. However, if no information on probabilities for certain hypotheses exists beforehand, the prior probability distribution should be an equal distribution

(Huelsenbeck *et al.* 2002). Such non-informative priors are regarded as being objective as they give equal weight to different hypotheses (Ellison 2004).

Huelsenbeck *et al.* (2002) give an example of phylogeny inference of *Ipomoea*, where researchers could give their own belief in monophyly of *Ipomoea* species based on up-to-date morphology-based trees. Most researchers gave a probability of 0.5 reasoning that it was fair to give equal weight to monophyly and non-monophyly. All possible trees can be divided in monophyletic and non-monophyletic trees with a ratio of 296,366 against one. In this way, a prior distribution of 0.5 gives a much higher weight to monophyly. Despite this, a higher posterior probability was found for non-monophyletic trees and most researchers changed their belief accordingly. Even if a high prior probability (e.g. 0.9) for monophyly was assumed, posterior distributions would still indicate non-monophyly. This is an example where the prior distribution does not affect the outcome very much.

If the posterior distribution is very sensitive to the choice of prior, different conclusions can be drawn from the same data but based on different priors. This is the main critic on Bayesian phylogeny inference, or Bayesian statistics in general; the choice of the prior distribution is too subjective. Unfortunately, the prior probability is a necessary evil if the hypothesis-based likelihood needs to be transformed into a data-based probability. Another reason why the prior is indispensable is that it enables integration over all possible parameter values by weighting the parameters according to their posterior probability (Holder & Lewis 2003).

To estimate the effect of the prior on the outcome, multiple analyses can be performed with different priors. The relative contribution of prior and data on the posterior distribution can also be determined by calculating the odds of prior against posterior distribution in favor of a certain hypothesis. This ratio is called the Bayes factor. If the prior has much power, the reasoning behind choosing a particular prior should be explained very well in order to be accepted by the scientific community. Generally, the larger the dataset, the less influence the priors have on the outcome. (Huelsenbeck *et al.* 2002).

The critic that the use of priors in Bayesian inference is subjective can be countered by stating that the frequentist method also involves subjectivity when setting the critical level for rejection of a hypothesis (i.e. α-value). In addition, frequentist approaches are often burdened with prior information, although it is often obscured. For instance, knowledge on how large the sample size should be or what type of analysis to use is not discarded for the sake of objectivity but used to optimize an experiment and the data analysis.

As explained above, a flat (or uniform) prior is often used which gives equal *a priori* probability to all possible trees. However, this could result in problems if, for instance, a uniform prior is put on unbounded quantities, creating a zero probability for

every possible tree. Therefore, the prior needs to be truncated somewhere, but this results in slanted probabilities. For example, if the prior gives equal weight to all branch lengths (t), than the substitution rates (p) accompanied with branch lengths shows a much higher probability in the range of ¾ as this is the highest rate of change (Fig. 3). Vice versa, a flat prior probability for substitution rate creates distribution of branch lengths slanted to values close to zero (Fig 4.; Felsenstein 2004).

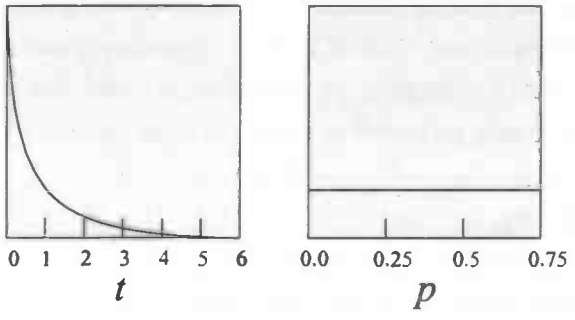

Figure 3. Prior probability distributions for branch length (t) and nucleotide substitution rate (p). A flat prior is used for substitution rate.

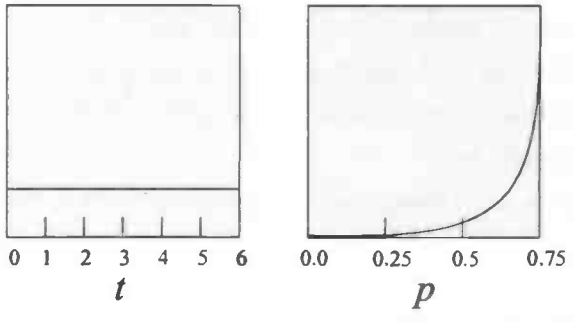

Figure 4. Prior probability distributions for branch length (t) and nucleotide substitution rate (p). A flat prior is used for branch length.

Convergence and mixing

Another point of discussion is how long the burn-in phase should be before convergence to a stable likelihood value is achieved. The use of log-likelihood plots, which show the change in likelihood of the trees during the calculation, has been shown to be unreliable to assess convergence (Huelsenbeck *et al.* 2002). Sometimes there appears to be a stable likelihood value, which then suddenly starts climbing again to higher values. A way to assess whether the algorithm gets stuck in local optima is to observe the (apparent) final states of multiple analyses starting with different trees (Huelsenbeck *et al.* 2002). If different stable states have been reached, the lower probability values represent local optima. If all trials converge to the same value, this is less likely to be a local optimum.

An additional problem is that the degree of mixing of the chain, which is the variability in accepted trees, cannot be detected with log-likelihood plots (Huelsenbeck *et al.* 2002). The higher the variability, the more chance that the optimal tree will be found. However, a relatively constant likelihood value says nothing about the amount of mixing. Again, if multiple random starting trees converge to the same tree, this may indicate that mixing is sufficient (Huelsenbeck *et al.* 2002).

Metropolis coupling is a method that enhances mixing and increases the chance that full convergence is found (Huelsenbeck *et al.* 2002). Although the Metropolis algorithm in the MCMC method can leave local optima by accepting a tree with a lower likelihood, there is a chance that the same local optimum is found time and again. This is because all proposed trees are neighbouring trees which look like each other converging towards the same optimum. The problem that the Metropolis algorithm gets stuck in local optima can be overcome by Metropolis coupling (Huelsenbeck *et al.* 2002). In a Metropolis-coupled procedure, a number of parallel chains are run. One chain samples from the original posterior distribution, referred to as the cold distribution. The other chains sample from heated distributions, which means that the posterior distribution is raised by a randomly chosen but fixed factor between 0 and 1. Two chains are randomly chosen at a regular interval and the present trees are proposed to each other's chain like a normal Metropolis step. Heated chains are more likely to leave local optima. Several times a tree from a heated chain is proposed to the cold chain, which may help the cold chain to leave a local optimum. Metropolis-coupling in this way proposes intelligent alternative trees to the cold chain. Metropolis-coupled MCMC, or $(MC)^3$ has improved the potential of phylogeny inference considerably, although convergence and mixing remain problematic.

Probabilities and likelihood values

The Bayesian method is said to be too liberal in its assignment of probabilities to overall trees or separate nodes (occurrence of Type I error is higher than expected), whereas Maximum Likelihood and Neighbor-Joining bootstrap calculations are too conservative (occurrence of Type I error is lower than expected), according to a study by Suzuki *et al.* (2002). They calculated the number of false-positives in a comparison between three concatenated genes. The genes, when analysed separately, would result in three different tree topologies. The three sequences differed equally from each other which, when concatenated, would theoretically result in low support values for each tree topology. However, in more than 40% of the Bayesian analyses, the result was a false-positive. This is more than the 5% false-positives generally accepted as a Type I error. Contrasting, the Neighbor-Joining method and the Maximum Likelihood approach were found to be too conservative, revealing less than 5% Type I errors.

In line with these results, bootstrap values of around 50% are often regarded as enough evidence for moderate support for a bifurcation (Baldauf 2003, Holder & Lewis

2003), whereas 95% is usually the minimum allowed in Bayesian analyses. There are several potential explanations why bootstrap support values are often lower than Bayesian posterior probabilities. A plausible explanation is the following. Bootstrap support values, calculated for Neighbor-Joining, Maximum Parsimony or Maximum Likelihood methods, are measures of uncertainty based on the outcomes of resampled data matrices. As explained above, a value below 50% means that the node in question was found in less than 50% of the bootstrap replicates. The main difference with posterior probabilities is that bootstrap values have no relationship with evolutionary uncertainty whereas posterior probabilities are in fact measures of evolutionary uncertainty. This is because posterior probabilities are based on integration over all possible parameter values calculated from the likelihood values and the prior distribution (Huelsenbeck *et al.* 2002; Felsenstein 2004). This methodological difference may be the reason for the discrepancy between bootstrap values and posterior probabilities as found by Suzuki *et al.* (2002).

Conclusion

An overview of the different methods of phylogeny inference and their computing speed, advantages and disadvantages is given in table 1.

Table 1. Computing speed, advantages and disadvantages of different phylogeny inference methods.

Method of Inference	Speed	Advantages	Disadvantages
Neighbor-Joining	++	-Conservative method.	-No optimality criterion.
Maximum Parsimony	+	-Robust if evolutionary change is small.	-Disregards influence of branch length on substitution chance. -Can show long branch attraction
Maximum Likelihood	-	-Results in the optimal tree based on estimated parameters of evolution. -Conservative method.	-Only one tree is given as a point estimate.
Bayesian Inference	+	-Creates posterior probabilities of different hypotheses given the data by integrating over all parameter values.	-May result in too high probabilities. -Use of prior distribution is subjective. -Hard to guess when MCMC has run long enough.

Recapitulating the advantages and disadvantages of Bayesian inference compared to traditional methods, it can be concluded that the appearance of the Bayesian method represents a step forward in the development of sophisticated inference of phylogeny.

Three main reasons can be given. First, Bayesian inference is based on the likelihood function, which already proved to be advantageous in the Maximum Likelihood method compared to other traditional methods. Second, prior information can be incorporated in the calculation, although this can also be seen as undermining the objectivity of the statistical method. Third, the MCMC method makes the calculations feasible by approximating the posterior probabilities. The main disadvantages of Bayesian inference are threefold. First, it tends overestimate posterior probabilities, thereby falsely giving support for trees or nodes. Second, as already mentioned the use of prior probabilities undermines the objectivity but in addition can also exert much influence on the outcome. Third, it is difficult to estimate when the MCMC method has run long enough to have converged to the global optimum (Table 1.). These disadvantages are no reason to abandon the Bayesian method but ask for caution when using it. Whereas the influence of the prior can be estimated by regarding Bayes factors and the results of multiple chains can be compared to estimate mixing and convergence, the problem of overestimation of posterior probabilities seems harder to tackle. However, the advantages of Bayesian inference compared to traditional methods have created a large active scientific community working on and with Bayesian inference. Bayesian inference therefore is not a finished project but is continuously improved, increasing the chance that solutions will be found for the still existing problems. Although, that is what I belief.

Acknowledgement

I would like to thank Wytze Stam for his supervision.

References

Baldauf, S.L. (2003) Phylogeny for the faint of heart: a tutorial. *Trends in Genetics*, **19**, 345-351.

Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, **53**, 370-418.

Beaumont, M.A. & Rannala, B. (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics*, **5**, 251-261.

Darwin, C. (1859) *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* Penguin Books, London, England, 480 pp.

Dennis, B. (1996) Discussion: should ecologists become Bayesians? *Ecological Applications*, **6**, 1095-1103.

Eddy, S.R. (2004) What is Bayesian statistics? *Nature Biotechnology*, **22**, 1177-1178.

Ellison, A.M. (2004) Bayesian inference in ecology. *Ecology Letters*, **7**, 509-520.

Felsenstein, J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annual Reviews of Genetics*, **22**, 521-565.

Felsenstein, J. (2004) *Inferring Phylogenies.* Sinauer Associates, Inc., Sunderland, Massachusetts, 668 pp.

Hendy, M.D. & Penny, D. (1989) A framework for the study of evolutionary trees. *Systematic Zoology*, **38**, 297-309.

Hillis, D.M., Huelsenbeck, J.P. & Cunningham, C.W. (1994) Application and accuracy of molecular phylogenies. *Science*, **264**, 671-677.

Hillis, D.M., Moritz, C. & Mable, B. (1996) *Molecular Systematics*, 2nd edition, Sinauer Associates, Sunderland, Massachusetts, 655 pp.

Holder, M. & Lewis, P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews – Genetics*, 4, 275-284.

Huelsenbeck, J.P., Larget, B. & Swofford, D. (2000) A compound poisson process for relaxing the molecular clock. *Genetics*, **154**, 1879-1892.

Huelsenbeck, J.P. & Rannala, B. (1997) Phylogenetic methods coming of age: testing hypotheses in an evolutionary context. *Science*, **276**, 227-232.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R. & Bollback, J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310-2314.

Huelsenbeck, J.P., Larget, B., Miller, R.E. & Ronquist, F. (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, **51**, 673-688.

Lewis, P.O. (2001) Phylogenetic systematics turns over a new leaf. *Trends in Ecology & Evolution*, **16**, 30-37

Meusnier, I., Valero, M., Olsen, J.L. & Stam, W.T. (2004) Analysis of rDNA ITS1 indels in *Caulerpa taxifolia* (Chlorophyta) supports a derived, incipient species status for the invasive strain. *European Journal of Phycology*, **39**, 83-92.

Posada, D. & Crandall, K.A. (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*, **14**, 817-818.

Saitou, M. & Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406-425.

Quinn, G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, UK, 542 pp.

Shoemaker, J.S., Painter, I.S. & Weir, B.S. (1999) Bayesian statistics in genetics – a guide for the uninitiated. *Trends in Genetics*, **15**, 354-358.

Suzuki, Y., Glazko, G.V. & Nei, M. (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, **99**, 16138-16143.

Yudkowsky, E.S. (2006) An Intuitive Explanation of Bayesian Reasoning. URL: http://sysopmind.com/bayes/bayes.html