Phylogenetic systematics turns over a new leaf

Paul O. Lewis

Long restricted to the domain of molecular systematics and studies of molecular evolution, likelihood methods are now being used in analyses of discrete morphological data, specifically to estimate ancestral character states and for tests of character correlation. Biologists are beginning to apply likelihood models within a Bayesian statistical framework, which promises not only to provide answers that evolutionary biologists desire, but also to make practical the application of more realistic evolutionary models.

As an OPTIMALITY CRITERION (see Glossary) in phylogenetics, MAXIMUM LIKELIHOOD has been in use almost as long as parsimony, which remains the dominant method for phylogenetic analyses of discrete (morphological and molecular) character data. Unlike the other two major optimality criteria, both parsimony and likelihood operate directly on discrete character data rather than on a matrix of pairwise distances. This allows parsimony and likelihood methods to be used to estimate ancestral character states, in addition to estimating the tree topology. Felsenstein's[1] 'pruning algorithm' made it feasible to apply the likelihood criterion to nucleic acid sequence data. His freely distributed package of computer programs for phylogeny inference, called PHYLIP (Ref. 2), introduced the likelihood criterion into molecular systematics and molecular evolution.

Felsenstein's 1981 model[1] was the first in a series of improvements made to the pioneering model of Jukes and Cantor[3] and was the first to allow nucleotide frequencies to differ (the Jukes and Cantor model[3] assumes that the four types of nucleotide are present in equal proportions). Further developments led to models that accommodate both transition–transversion substitution bias[4] and unequal nucleotide frequencies[5,6]. This trend culminated in the general time-reversible (GTR) model, which not only allows unequal nucleotide frequencies, but also allows each of the six possible transitions between nucleotide states (i.e. A↔C, A↔G, A↔T, C↔G, C↔T and G↔T) to occur at different rates[7].

A further improvement, used in conjunction with the abovementioned models, is the ability to accommodate variation in substitution rates across sites, using several methods[8–10]. One of the most commonly applied methods for allowing rates to vary among sites is the discrete gamma method, which assumes that relative rates are distributed according to a gamma distribution with a mean of 1 and a variance of $1/\alpha$, where α defines the shape of the distribution. Large values of α translate to a tiny variance in relative rates (i.e. the rates at all sites are almost equivalent), whereas small values (e.g. 0.01) result in an L-shaped distribution of relative rates (most rates are low, but some rates are high). Yang[11] reviewed in depth this approach to accommodating site-to-site substitution rate heterogeneity, and other reviews of this and other topics include Swofford *et al.*[12], Huelsenbeck and Crandall[13] and Lewis[14].

Some computer programs that are used to determine phylogenies allow other options with respect to modeling rate heterogeneity, including: (1) invariable sites models, in which a proportion of sites is assumed incapable of undergoing substitution[15]; (2) site-specific rates models, which allow different rates for arbitrary subsets of sites that correspond to different genes, introns versus exons or different codon positions[15]; (3) hidden Markov models which allow for correlation in the rates at adjacent sites[2]; and (4) discrete gamma models in which α can differ between subsets of sites[16]. It is also possible to combine an invariable sites model with a discrete gamma model, letting the

Glossary

Bootstrapping: a statistical technique, first applied to phylogenetics by Felsenstein[39], in which new data sets are created by sampling randomly (and with replacement) from the original characters. These new data sets (called bootstrap data sets) are of the same size as the original. A desired quantity is computed for each bootstrap data set and the resulting distribution is used to estimate the dispersion that would be expected if the same number of new independent data sets had been collected. Bootstrapping assumes that the original characters were sampled independently.
Likelihood: a quantity that is proportional to the probability of the data (or probability density, if the data are continuous-valued), given specific values for all parameters in the model.The likelihood function provides a means to estimate the parameters of the model. Parameter values that are associated with the global maximum of the likelihood function are termed maximum likelihood estimates (MLEs).
Optimality criterion: a rule used to decide which of two trees is best. Four optimality criteria are currently widely used:
Maximum parsimony – the tree requiring the fewest character state changes is the better of the two trees.
Maximum likelihood – the tree maximizing the likelihood under the assumed evolutionary model is better.
Minimum evolution – the tree having the smallest sum of branch lengths (estimated using ordinary least squares) is better.
Least squares – the tree showing the best fit between estimated pairwise distances and the corresponding pairwise distances obtained by summing paths through the tree is better.

Paul O. Lewis
Dept of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road Unit 3043, Storrs, CT 06269-3043, USA.
e-mail: paul.lewis@uconn.edu

http://tree.trends.com 0169–5347/01/$ – see front matter PII: S0169-5347(00)02025-5

Box 1. The anatomy of a codon model

A small portion of the full (61 × 61) INSTANTANEOUS RATE MATRIX (see Glossary) for the Muse and Gaut[a] codon model is illustrated (Table I). The quantities π_A, π_C, π_G and π_T are the equilibrium nucleotide frequencies, only three of which represent free parameters of the model, because the fourth can be obtained by subtraction. The remaining two parameters in the model, α (synonymous substitutions) and β (nonsynonymous substitutions), are relative rate parameters. The diagonal elements are not shown, but can be calculated by subtraction. All changes between the codons that involve more than one nucleotide substitution have an instantaneous rate of zero. The remaining elements represent single nucleotide substitutions that result in a change from one codon to another. The rate of any one of this type of change depends on the nucleotide frequencies (e.g. the rate is lower for substitutions to a relatively rare base) and the nature of the change (synonymous versus nonsynonymous).

Reference

a Muse, S.V. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11, 715–724

Table I. Part of Muse and Gaut's 61 × 61 instantaneous rate matrix[a]

Codon before substitution (the 'from' state)	Codon after substitution (the 'to' state)							
	TTT (Phe)	TTC (Phe)	TTA (Leu)	TTG (Leu)	CTT (Leu)	CTC (Leu)	…	GGG (Gly)
TTT (Phe)	– – –	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$	$\beta\pi_C$	0	…	0
TTC (Phe)	$\alpha\pi_T$	– – –	$\beta\pi_A$	$\beta\pi_G$	0	$\beta\pi_C$	…	0
TTA (Leu)	$\beta\pi_T$	$\beta\pi_C$	– – –	$\alpha\pi_G$	0	0	…	0
TTG (Leu)	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	– – –	0	0	…	0
CTT (Leu)	$\beta\pi_T$	0	0	0	– – –	$\alpha\pi_C$	…	0
CTC (Leu)	0	$\beta\pi_T$	0	0	$\alpha\pi_T$	– – –	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
GGG (Gly)	0	0	0	0	0	0	…	– – –

Box Glossary

Instantaneous rate matrix: a square matrix of substitution rate parameters that serves to define a particular substitution model. The probability of observing state *j*, given starting state *i*, changes with time in a nonlinear fashion. The rate parameter represents the slope of this curve evaluated at $t = 0$. The elements on the diagonal are negative and equal to the sum of the other rates in the same row. This quantity is negative, because the probability of observing state *i*, given that *i* was the starting state, decreases with time.

gamma shape parameter (α) apply only to the variable fraction of sites[15].

This review focuses on new nucleotide sequence-based models (codon and secondary structure models), new applications of likelihood models (discrete morphological characters) and a new framework within which likelihood models can be applied to phylogenetics (Bayesian inference). First, models known as 'codon models' and 'secondary structure models' allow evolutionary dependence among nucleotide sites within the same codon and between sites that are opposite one another in a stem region of an rRNA gene, respectively. By contrast, parsimony and previous likelihood models assume that all nucleotide sites are evolutionarily independent. Therefore, codon and secondary structure models represent a significant step forward by allowing recognized evolutionary correlations between sites to be directly incorporated into a phylogeny method. Second, likelihood is being applied to discrete morphological characters, which have been limited previously to the realm of parsimony analysis. Applications that use likelihood models include the estimation of ancestral character states and tests for character correlation. Third, in addition to new models, the very framework in which likelihood models are used is changing. Likelihood models form the basis of Bayesian statistical methods, and it is only natural that Bayesian methods be applied to phylogenetic problems already being addressed using likelihood models.

Likelihood models that allow nonindependence

In 1994, Muse and Gaut[17] and Goldman and Yang[18] independently introduced likelihood models that were designed to account for evolutionary dependency among sites within codons. This is in contrast to parsimony and previous likelihood models, which were forced to assume independence among sites in a gene sequence, even if the independence assumption had been violated. For instance, in genes that encode proteins, the three nucleotide sites that form a single codon cannot evolve independently of one another if there is selection for a particular amino acid at the corresponding site in the polypeptide.

Codon models

Codon models represent an important advance in terms of the biological realism of substitution models. Codon models take the genetic code explicitly into account when computing the probability of a change at a site across a branch. Thus, rather than four possible states (A, C, G and T) there are 61 (64 possible codons, less the three stop codons), which means that codon models require a great deal more computational effort than classic DNA and RNA substitution models. In particular, likelihood methods consider each state in turn for every interior node of a tree. Unlike parsimony, the score used for comparing trees does not depend on any particular combination of (unobserved) ancestral states. The overall likelihood is a sum over all state

Box 2. Likelihood ratio test for character correlation

In Pagel's[a,b] likelihood ratio test, two models are considered, of which one is a more general version of the other. Normally, the constrained model represents the null hypothesis and is nested within (i.e. is a special case of) the more general model.

A likelihood ratio test returns a significant result if the data are better explained by the general model (for which the maximum of the likelihood function is L_1) than by the constrained model that represents the null hypothesis (for which the likelihood maximum is L_0). A ratio L_1/L_0 much greater than 1 indicates significance and the likelihood ratio test statistic, defined as $LR = -2(\ln L_0 - \ln L_1)$, is large and positive. LR is distributed as a chi-squared random variable with degrees of freedom equal to the difference in the number of free (estimated) parameters between the two models if the null model is perfectly nested within the unconstrained model.

Table I. Instantaneous rate matrix for the null hypothesis assuming evolutionary independence

Changing from:	Changing to:	
	0	**1**
0	$-\alpha$	α
1	β	$-\beta$

The null hypothesis model for a test of correlated evolution is based on the following instantaneous rate matrix (Table I).

The same matrix is used for each character, but α and β are estimated separately for each character. The number of free parameters in the null model is thus equal to 4. In practice, the TRANSITION MATRIX (see Glossary) that corresponds to this rate matrix is computed for each character (let the transition matrix for character 1 be P and let that for character 2 be R), which allows probabilities for specific events to be computed. For example, the joint probability that character 1 remains in state 0 and character 2 changes from state 0 to 1 across a specific segment of the tree, is obtained as follows:

$$\Pr(0\rightarrow0,0\rightarrow1) = P(0\rightarrow0)R(0\rightarrow1) \quad [1]$$

The fact that the joint transition probability is the product of the separate transition probabilities for each character makes explicit the assumption of independence in this model.

The general model enables (but does not force) the evolution of the two characters to be correlated and can account for simultaneous changes in the characters by using a single matrix (Table II).

Table II. Instantaneous rate matrix for the general hypothesis allowing evolutionary correlation

'After' state	'Before' state			
	0, 0	**0, 1**	**1, 0**	**1, 1**
0, 0	$-a-b$	a	b	0
0, 1	c	$-c-d$	0	d
1, 0	e	0	$-e-f$	f
1, 1	0	g	h	$-g-h$

This model has eight free parameters. The maximum-likelihood value under this model approaches that of the null model if the characters are evolving independently of one another. As in the codon models, a rate of 0 is specified for events that require two changes. The difference in the number of free parameters is 8–4 = 4, so the likelihood ratio test statistic is nominally distributed as a χ^2 random variable with four degrees of freedom. LR might not follow a χ^2 distribution in this case, because the models are not strictly nested (owing to the four rates with values of 0 in the general model, but see Ref. b). Pagel provides (in his computer program, DISCRETE) the means for using parametric bootstrapping to determine the significance level without making this distributional assumption.

References

a Pagel, M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. London B Biol. Sci.* 255, 37–45

b Pagel, M. (1997) Inferring evolutionary processes from phylogenies. *Zoologica Scripta* 26, 331–348

Box Glossary

Transition matrix: a matrix that shows the conditional probability of observing state j given starting state i after some arbitrary amount of time (t) for each pair of states i and j. This matrix can be derived from the instantaneous rate matrix, which describes fully the process at any given instant. (In this context, 'transition' should not be confused with the same term that is used in conjunction with 'transversions'.)

combinations, in much the same way that the probability of rolling a seven with two dice is a sum over the six possible ways in which the numbers on each die can add up to seven.

Although it would be possible to allow a different rate for every possible transition between codons, the estimation of so many parameters (3660 in total) would require an unreasonable amount of data. To keep the number of free (estimated) parameters manageable, the two codon models that have been proposed make some concessions. For instance, neither model permits more than a single change in any given instant (e.g. the rate of the change AAA→AGC is 0). However, this is a reasonable restriction, because two changes are allowed to occur in two consecutive infinitesimal periods of time.

The model by Muse and Gaut[17] assigns rates to all other possible changes (i.e. the changes that do not require more than two nucleotide point mutations) on the basis of two parameters that represent the rate of synonymous and nonsynonymous substitutions,

Box 3. Prior and posterior probabilities

Bayesian inference involves interplay between the likelihood function and the prior and posterior distributions. To illustrate this concept, consider the following simplistic problem. Suppose a black marble has come from one of two urns that each contain millions of marbles. The challenge is to decide from which urn the marble was removed. Suppose further that it is known that 40% of the marbles in urn A and 80% of the marbles in urn B are black. Choosing a marble from urn B would be clearly more likely to yield a black marble. Nevertheless, it is instructive to calculate the posterior probability that the marble came from urn B and to compare this to the posterior probability that it came from urn A.

Bayes' Rule, which is used to obtain the posterior probability from the likelihood and prior probability, is based on the definition of conditional probability. Accordingly, for two events, A and B, the joint probability of A and B equals the conditional probability of one, given the other (i.e. 'A given B' or 'B given A') multiplied by the probability of the condition (Eqn 1):

$$\Pr(A,B) = \Pr(A)\Pr(B|A) = \Pr(B)\Pr(A|B) \quad [1]$$

Focusing only on the equation on the right, dividing both sides by Pr(A) yields Bayes' Rule (Eqn 2):

$$\Pr(B|A) = \frac{\Pr(B)\ \Pr(A|B)}{\Pr(A)} \quad [2]$$

In the Bayesian framework, A represents data and B a hypothesis (or parameter):

Pr($B|A$) is termed the posterior probability and is the probability of the hypothesis (or parameter value), given the data.

Pr($A|B$) is termed the likelihood and is the probability of the data, given the hypothesis (or parameter value).

Pr(B) is termed the prior probability and is the unconditional probability of the hypothesis (or parameter value). This must be specified by the investigator without reference to the data.

Pr(A) is the unconditional probability of the data, which can be obtained, using the law of total probability, by calculating the sum of the product Pr(B)Pr(A/B) for all possible values of B. This serves as a normalizing constant, which ensures that the sum of the posterior probabilities is 1.

In this example, there is only one datum (the black marble). Therefore, computing the likelihood (probability of the data, given the hypothesis) is straightforward and is simply the probability that a single marble is black, given a particular urn hypothesis. The likelihood is 0.4 for urn A and 0.8 for urn B. If maximum likelihood were being used to decide between the two urns, urn B would be selected, because the likelihood of drawing a black marble is greater for urn B than for urn A (0.8 versus 0.4). However, to choose a prior distribution that reflects complete prior ignorance of which urn was used when drawing the black marble, we specify that the prior probability of each urn is 0.5. The normalizing constant is therefore (Eqn 3):

$$\Pr(\text{black marble chosen}) = (0.5)(0.4) + (0.5)(0.8) = 0.6 \quad [3]$$

The posterior probability can now be computed for each urn as follows (Eqns 4,5):

$$\Pr(\text{urn A/black marble chosen}) = \frac{(0.5)(0.4)}{0.6} = \frac{1}{3} \quad [4]$$

$$\Pr(\text{urn A/black marble chosen}) = \frac{(0.5)(0.8)}{0.6} = \frac{2}{3} \quad [5]$$

Thus, the probability that the black marble came from urn B, given the datum, is two thirds. The posterior distribution is often described as an updated version of the prior distribution. In this case, the posterior distribution (0.33, 0.67) represents an updated version of the prior distribution (0.5, 0.5), the evidence used for the update being the fact that the single marble drawn was black. A major benefit of taking the Bayesian perspective lies in the fact that it produces probabilities for hypotheses of interest, which is exactly what investigators desire. Likelihoods are useful, but not easily interpreted, because they represent the probability of the data given the hypothesis rather than the probability of the hypothesis given the data.

A criticism of Bayesian approaches is the subjectivity of the prior. Note that in the example above (with only a single datum), it would be necessary to stipulate a prior probability for urn A that was greater than two thirds to 'rig' the analysis (that is, to ensure that the conclusion is in favor of urn A). While the posterior distribution always changes if the prior is changed, the conclusions are usually not overly sensitive to the prior and any effect of the prior decreases as the amount of data increases. A counterargument to the subjectivity criticism is that subjectivity that is inherent in the prior is explicit and must be defensible.

In the example above, discrete hypotheses were used. However, continuous parameters are more often the targets of Bayesian analyses. In such cases, probability density functions replace the probabilities of discrete hypotheses, but Bayes' rule is still applicable. Figure I illustrates the prior and posterior density curves for the parameter p (probability of heads) in a simple coin-flipping experiment in which six heads were observed for ten flips. Two different prior distributions were used, namely a flat [Beta(1,1), Fig. Ia] and a more informative [Beta(2,2), Fig. Ib] prior distribution. As an example of how such curves could be used, the posterior probability that p is between 0.45 and 0.55 could be obtained by integrating the posterior density curve between 0.45 and 0.55.

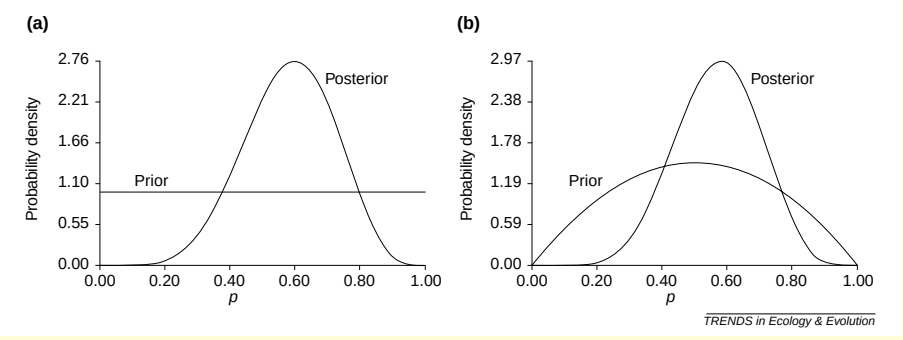
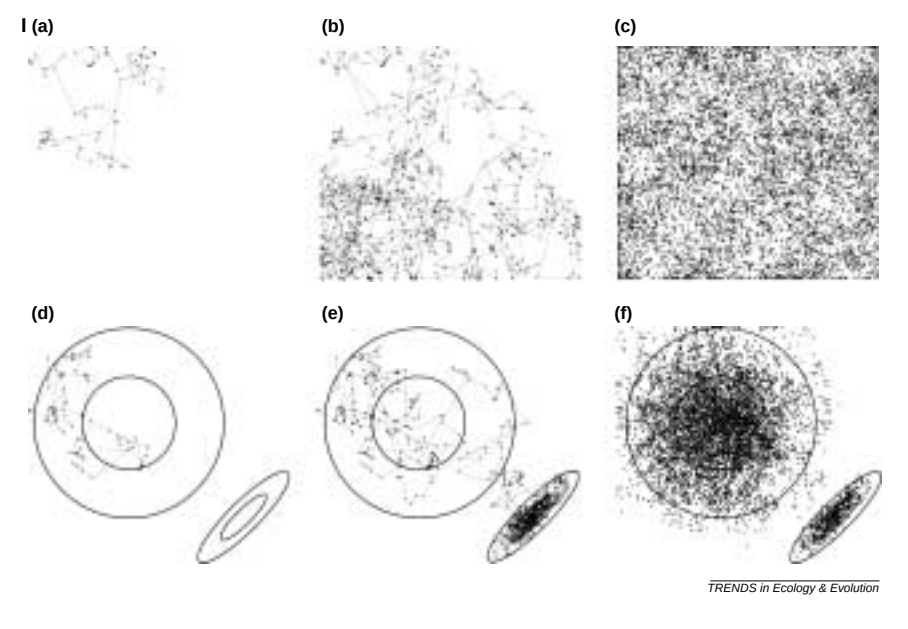


Fig. I. Prior and posterior density curves for p (probability of heads) in a coin-flipping experiment, illustrating a flat [Beta(1,1)] (a) and an informative [Beta(2,2)] (b) prior distribution.

Box 4. Markov Chain Monte Carlo methods



The principles behind Markov Chain Monte Carlo (MCMC) methods can be illustrated with a simple analogy. A robot is allowed to 'walk' in a square field. The robot takes steps that vary in length and randomly selects a direction for each step. The robot never leaves the field, because if a step is about to take it outside the boundary, it is reflected back into the field. A representative path followed by such a robot is illustrated in Fig. Ia–c for 100 (Fig. Ia), 1000 (Fig. Ib) and 10 000 steps (Fig. Ic). (The path lines have been removed from Fig. Ic to clarify the distribution of steps.) Note that, even though the robot is walking randomly, eventually every portion of the field is visited if it is allowed to walk for long enough.

Now suppose that two hills are present (represented by bivariate normal densities – one uncorrelated and with a correlation coefficient of 0.9) and the robot is programmed to use the following simple rules (based on the local environment) when taking steps:

- If a proposed step will take the robot uphill, it automatically takes the step.
- If a proposed step will take the robot downhill, it divides the elevation at the proposed location by its current elevation and only takes the step if it draws a random number (uniform on the open interval 0,1) that is smaller than this quotient.
- The proposal distribution is symmetrical, which means that the probability of proposing a step from point A to B is the same as that of proposing a step from point B to A.

By following these rules, the path followed by the robot will take it to points in the field in proportion to their elevation, with higher points being visited more often than lower ones. Letting the robot begin its walk near the upper left-hand corner, its movements are illustrated in Fig. Id–f for 100 (Fig. Id), 1000 (Fig. Ie) and 10 000 steps (Fig. If). (As in Fig. Ic, path lines are not shown in Fig. If.)

The robot's movements can be used to estimate the volume under any specified portion of this landscape. In the example above, if the first few steps (called the 'burn-in' stage) are disregarded then the resulting distribution of points is a good approximation of the landscape, which is a mixture of two separate bivariate normal densities. The longer the robot is allowed to walk, the closer the approximation becomes.

This analogy also illustrates one of the pitfalls of approximating surfaces using MCMC methods. Were the robot to take only a few steps, it is clear that it would not be likely to encounter the second hill. Therefore, care should be taken to ensure that Markov chains are run for long enough to provide a good approximation of the surface over the entire parameter space. One simple way to investigate this is to run several chains, each starting from a different (randomly selected) starting point. If the resulting approximations are substantially different, this indicates that none of the chains were run for long enough.

respectively (Box 1). By estimating only five parameters (the two rate parameters and three nucleotide frequency parameters), the entire 61×61 matrix of transition probabilities required to compute the likelihood can be specified. The Muse and Gaut codon model is implemented in Muse and Kosakovsky's 'Hypothesis Testing Using Phylogenies' (HYPHY) computer package[19]. The commonly employed HKY85 model[5] also requires five parameters, but the two rate parameters in this case represent the rates of transitions and transversions, rather than nonsynonymous and synonymous substitutions.

The codon model proposed by Goldman and Yang[18] goes a step further in allowing both transition–transversion bias and synonymous-nonsynonymous substitution bias. This model also allows biochemical differences among amino acids to play a role in determining the rate at which changes among amino acids occur. For example, a change from one hydrophobic to another hydrophobic residue can occur at a higher rate than a change from a hydrophobic to a hydrophilic residue. Goldman and Yang's model is implemented in Yang's Phylogenetic Analysis By Maximum Likelihood (PAML) computer package[16].

Secondary structure models

Secondary structure models represent an attempt to add biological realism to analyses of genes with RNA products that have a secondary structure. Tillier and Collins[20] and Muse[21] independently introduced

models that can account for the evolutionary dependence between nucleotide sites that are opposite one another in the stem regions of an RNA gene product. Following substitutions in these sites, selection favors compensatory substitutions to preserve the stability of stem regions and hence also the secondary structure that is needed for the catalytic activities of these RNA molecules. Muse's[21] model requires only a single additional free parameter to account for the evolutionary correlation between paired stem sites and reduces to a standard nucleotide substitution model (the HKY85 model) if no such correlation exists in the stem regions.

Applications of likelihood to discrete morphological characters

The most common method for obtaining estimates of ancestral character states for discrete morphological characters involves parsimony optimization, although likelihood-based alternatives have been proposed[22–24]. Cunningham[25] reviewed ancestral character state estimation. I will therefore focus on a different application of likelihood models in inferences concerning morphological features, namely testing for evolutionary correlation.

Pagel[26] described a likelihood ratio test (Box 2) that allows an investigator to determine whether two discrete morphological traits are correlated to a greater extent than can be explained simply by the phylogeny. Two traits are evolutionarily correlated if a change in state of one of the characters predisposes the other character to change state soon after. For example, the evolution of feathers on the forelimbs of dinosaurs is correlated with the evolution of wings. In this case, the appearance of feathers for reasons other than flight[27] apparently encouraged a change from the state 'wings absent' to 'wings present', which might not have otherwise occurred. Phylogenies impose correlations on characters, even if those characters evolve independently. Pagel's test addresses whether an evolutionary correlation exists between two characters that acts to increase the observed correlation above and beyond the level imposed by the phylogeny.

Suppose that two characters each change state just once on a phylogeny and happen to change along the same segment of the phylogeny. At first sight, this coincidence might appear to be the result of an evolutionary correlation. However, evidence to support this interpretation would be weak, because the co-evolutionary event occurred only once. By contrast, the evidence of correlated evolution of the two characters would be stronger if changes in the two characters co-occur on many different parts of the phylogeny. Pagel and others have emphasized the importance of using explicit phylogeny-based methods for assessing the strength of evidence for such correlations, because the number of degrees of freedom available for the test can otherwise be greatly exaggerated[26].

In Pagel's test, the null hypothesis is that the two characters of interest have evolved independently. The general (unconstrained) model allows for some correlation in the evolution of the two characters. The null model is a constrained version of the general model, because the independence assumption constrains the correlation to equal zero. The maximum of the likelihood function under the unconstrained model will, thus, approach that of the constrained model as the correlation approaches zero. If the true correlation is not zero, then the maximum of the likelihood function under the general model will be significantly larger than that which can be attained by the null model, reflecting the better fit of the general model. Pagel argues that examination of the estimated values of the parameters of the general model enables a determination of which character is most likely to have changed state first, precipitating a change of state in the other character. However, such fine-scaled interpretations should be made with caution, because several parameters are estimated under this model using data from only two characters.

An interesting example of the application of Pagel's method involves testing whether switches to mutualism in fungi, either in the form of association with algae (i.e. lichenization) or liverworts, are associated with changes in the rate of molecular evolution[28]. Lutzoni and Pagel[28] began by classifying a group of closely related *Omphalina* mushroom species as either fast (F) or slow (S) evolving and as either mutualistic (M) or nonmutualistic (N). Estimates of the rate of evolution were based on nuclear ribosomal DNA (nrDNA) from the 25S and 5.8S genes and the internal transcribed spacer (ITS) region. They then obtained likelihoods for both the null model (assuming independent evolution of both mutualism and the rate of molecular evolution) and the general model (in which the rate of evolution and mutualism can be correlated). The three sets of data fit the correlation model better than the independence model. However, the improved fit was only significant for the 25S data. Therefore, Lutzoni and Pagel[28] concluded that there is a correlation between the rate of molecular evolution at the 25S nrDNA locus and the evolution of mutualism. These authors also looked at individual rate parameters that were estimated in the nonindependence model. For the 25S data, the estimated rate for the transition M,S→M,F was significantly greater than that for N,S→N,F, indicating that increases in the rate of molecular evolution are more likely to occur in lineages that are already mutualistic. Also (again for the 25S data), the rate of the transition N,S→M,S was significantly greater than that for N,S→N,F, indicating that the tendency for slowly evolving lineages to evolve mutualism is greater than the tendency for nonmutualistic lineages to evolve fast rates of molecular evolution. This ability to tease apart the correlations among characters sets Pagel's method apart from other methods, which simply assess whether a correlation exists.

A Bayesian future for model-based phylogenetics?

Likelihood methods in phylogenetics take longer than other methods. For large data sets, it is not uncommon for single heuristic searches to take days, if not months, for a computer to complete. Even when complete, such analyses only represent a point estimate of a phylogeny. A measure of support for individual clades requires BOOTSTRAPPING, which effectively turns an already time consuming analysis into one that requires a far greater amount of time. This situation might eventually be ameliorated by taking a Bayesian approach. Larget and Simon[29] have summarized the recent literature on Bayesian approaches to making inferences about phylogenies[29–34] (S. Li, PhD thesis, Ohio State University, 1996; B. Mau, PhD thesis, University of Wisconsin, 1996) and have presented novel algorithms to estimate simultaneoulsy the tree topology and provide a measure of nodal support. These algorithms have the potential to transform model-based phylogenetic inference.

Bayesian approaches to statistical problems involve making inferences using the posterior distribution (or simply 'the posterior') of hypotheses or parameters of interest (Box 3). This is feasible when problems are simple enough for an analytical formula for the posterior to exist. However, for most 'real' problems, models involve several parameters and, thus, a complicated joint posterior for which there is no simple formula. The posterior for an unrooted phylogenetic tree, for example, involves at least the topology (a discrete parameter) and $2N–3$ branch length parameters, where N is the number of tip nodes.

Fortunately, a technique exists for approximating complicated posteriors, such as those that are characteristic of problems that involve phylogenies. The technique is called Markov Chain Monte Carlo (MCMC) or the Metropolis–Hastings algorithm[35,36]. The idea underlying MCMC is that a Markov chain that takes the form of a correlated random 'walk' through the parameter space can be conducted in such a way that any probability distribution (however complicated) can be approximated by periodically sampling values (Box 4). The approximation can be made arbitrarily accurate by running the Markov chain for a sufficient number of steps.

In phylogenetic applications, each step in a Markov chain involves a random modification of the tree topology, a branch length or a parameter in the substitution model (e.g. substitution rate ratio). If the posterior that is computed for a proposed step is larger than that of the current tree topology and parameter values, the proposed step is taken. Proposed steps that result in downhill moves on the posterior surface are not automatic and depend on the magnitude of the decrease. The function that determines the probability of a downhill step in this case is based on the ratio of the new and current posteriors. Using these rules, the Markov chain visits regions of tree space (*sensu lato*, including the tree topology space and dimensions for other parameters, such as branch lengths) in proportion to their posterior.

For a practical example of how this method could be used to answer a question posed by a systematist, consider the question, 'What is the probability that group X is monophyletic?' To answer this question, we would run the Markov chain, sampling trees periodically. Suppose that in a sample of 100 000 trees, group X appeared as a monophyletic group in 74 695 trees. The probability (given the observed data) that group X is monophyletic is approximately 0.74695, because the Markov chain visits trees in proportion to their posterior probability. This value is a natural measure of nodal support and is easier to interpret than existing measures that are based on parsimony (decay index and bootstrap values) or likelihood (bootstrap values). Larget and Simon[29] make the point that only one run of the Markov chain is needed to gain such information, compared to the many bootstrap runs needed in a maximum likelihood analysis.

Another important application of MCMC in Bayesian phylogenetic inference involves estimating divergence times in a 'relaxed molecular clock' model, that is, a model in which substitution rates vary across the phylogeny, but in which rates in descendant lineages are correlated with the rate in their common ancestor[37,38]. MCMC methods provide Bayesian 'credibility intervals' for divergence dates without having to assume a perfect clock and also allow considerable flexibility (and even uncertainty) in how the clock is calibrated.

The ability of MCMC models to provide answers to virtually any question that might be imagined by the investigator is a thrilling prospect. The ability to provide answers much more efficiently than is possible with current likelihood methods should make the Bayesian approach to phylogenetics well-worth watching in the next few years.

Acknowledgements
I would like to thank Kent Holsinger, Louise A. Lewis, Mark Pagel, Chris Simon and ZihengYang for their careful reading of earlier drafts of this article. I also gratefully acknowledge the Alfred P. Sloan Foundation/ National Science Foundation for funding (grant 98-4-5 ME).

References

1 Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376
2 Felsenstein, J. (1999) *PHYLIP, Phylogeny Inference Package* (University of Washington, Seattle), Version 3.572
3 Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In *Mammalian Protein Metabolism* (Munro, H.N., ed.), pp. 21–132, Academic Press
4 Kimura, M. (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120
5 Hasegawa, M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 21, 160–174
6 Kishino, H. and Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* 29, 170–179
7 Rodríguez, F. *et al.* (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142, 485–501
8 Churchill, G.A. *et al.* (1992) Sample size for a phylogenetic inference. *Mol. Biol. Evol.* 9, 753–769
9 Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39, 306–314
10 Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10, 1396–1401
11 Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–372

12 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Hillis, D.M. *et al.*, eds), pp. 407–514, Sinauer Associates
13 Huelsenbeck, J.P. and Crandall, K.A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28, 437–466
14 Lewis, P.O. (1998) Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. In *Molecular Systematics of Plants II* (Soltis, D.E. *et al.*, eds), pp. 132–163, Kluwer
15 Swofford, D.L. (2000) *PAUP*, Phylogenetic Analysis Using Parsimony* (Sinauer, Sunderland, MA), Version 4.0b3,
16 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13, 555–556
17 Muse, S.V. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11, 715–724
18 Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736
19 Muse, S.V. and Kosakovsky Pond, S.L. (2000) *HYPHY, Hypothesis Testing Using Phylogenies*, (North Carolina State University, Raleigh) Version 1
20 Tillier, E.R.M. and Collins, R.A. (1995) Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Mol. Biol. Evol.* 12, 7–15
21 Muse, S.V. (1995) Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics* 139, 1429–1439
22 Schluter, D. *et al.* (1997) Likelihood of ancestor states in adaptive radiation. *Evolution* 51, 1699–1711
23 Mooers, A.Ø. and Schluter, D. (1999) Reconstructing ancestor states with maximum likelihood: support for one- and two-rate models. *Syst. Biol.* 48, 623–633
24 Pagel, M. (1999) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* 48, 612–622
25 Cunningham, C.W. *et al.* (1998) Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol. Evol.* 13, 361–366
26 Pagel, M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. London B Biol. Sci.* 255, 37–45
27 Qiang, J. *et al.* (1998) Two feathered dinosaurs from northeastern China. *Nature* 393, 753–761
28 Lutzoni, F. and Pagel, M. (1997) Accelerated evolution as a consequence of transitions to mutualism. *Proc. Natl. Acad. Sci. U. S. A.* 94, 11422–11427
29 Larget, B. and Simon, D.L. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16, 750–759
30 Mau, B. and Newton, M.A. (1997) Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat.* 6, 122–131
31 Mau, B. *et al.* (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* 55, 1–12
32 Newton, M. *et al.* (1999) Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. In *Statistics in Molecular Biology* (Seillier-Moseiwitch, F. *et al.*, eds), Institute of Mathematical Statistics
33 Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* 43, 304–311
34 Yang, Z.H. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol.* 14, 717–724
35 Metropolis, N. *et al.* (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092
36 Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109
37 Huelsenbeck, J.P. *et al.* (2000) A compound Poisson process for relaxing the molecular clock. *Genetics* 154, 1879–1892
38 Thorne, J.L. *et al.* (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15, 1647–1657
39 Felsenstein, J. (1985) Confidence intervals on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791

Intraspecific gene genealogies: trees grafting into networks

David Posada and Keith A. Crandall

Intraspecific gene evolution cannot always be represented by a bifurcating tree. Rather, population genealogies are often multifurcated, descendant genes coexist with persistent ancestors and recombination events produce reticulate relationships. Whereas traditional phylogenetic methods assume bifurcating trees, several networking approaches have recently been developed to estimate intraspecific genealogies that take into account these population-level phenomena.

During the past decade, the explosion of molecular techniques has led to the accumulation of a considerable amount of comparative genetic information at the population level. At the same time, recent advances in population genetics theory, especially coalescent theory, have generated powerful tools for the analysis of intraspecific data. These two developments have converted intraspecific phylogenies into useful tools for testing a variety of evolutionary and population genetic hypotheses. Several phylogenetic methods, especially NETWORK (see Glossary) approaches, have been developed to take advantage of the unique characteristics of intraspecific data. In this article, we summarize some population genetics principles, explain why networks are appropriate representations of intraspecific genetic variation, describe and compare available methods and software for network estimation, and give examples of their application.

Gene genealogies

Given a sample of GENES, the relationships among them can be traced back in time to a common ancestral gene. The genealogical pathways interconnecting the current sample to the common ancestor constitute a GENE TREE or gene genealogy. A gene tree is the pedigree of a set of genes and exists independently of potential mutations. The only portion of a gene tree that can generally be estimated with genetic data is that portion marked by the (potential) mutational events that define the different ALLELES (Box 1). This lower resolution tree is the allele

http://tree.trends.com 0169–5347/01/$ – see front matter PII: S0169-5347(00)02026-7