Model use in phylogenetics: nine key questions

Scot A. Kelchner and Michael A. Thomas

Department of Biological Sciences, Idaho State University, Pocatello, ID 83209-8007, USA

Models of character evolution underpin all phylogeny estimations, thus model adequacy remains a crucial issue for phylogenetics and its many applications. Although progress has been made in selecting appropriate models for phylogeny estimation, there is still concern about their purpose and proper use. How do we interpret models in a phylogenetic context? What are their effects on phylogeny estimation? How can we improve confidence in the models that we choose? That the phylogenetics community is asking such questions denotes an important stage in the use of explicit models. Here, we examine these and other common questions and draw conclusions about how the community is using and choosing models, and where this process will take us next.

Models in phylogenetics

No phylogeny estimation (see Glossary) is assumption free. To create an evolutionary tree of relationships, one must make assumptions about the evolutionary process that produced the observed data. Taken as a whole, these assumptions form a 'conceptual model' of character evolution with which estimates of evolutionary relationship are made. The conceptual model could include a mathematically explicit 'formal model' of character change, which is parameterized when applied to nucleotide or amino acid sequence data. It is this formal model that is referred to as the 'model' in phylogenetic literature (as it is here).

Phylogenetics cannot escape the use of conceptual models. Even those methods that do not formalize a model, and thus claim to be model-free (e.g. parsimony), make significant and sometimes incorrect assumptions about character evolution when estimating the amount of change between organisms [1–4]. The growth of formal model use in phylogenetics, however, has created a noticeable level of concern in the community. In particular, we think that confusion about certain model qualities is contributing to controversies about model 'accuracy', complexity and development.

Several reviews give practical introductions to formal models [5,6] and model selection methods [7,8] in molecular phylogenetics. Here, we focus instead on nine commonly voiced questions about the use of formal models in phylogenetics. We seek to clarify issues surrounding model interpretation, parameter choice and the need for adequate models when analyzing nucleotide sequence data.

Question 1: What are models in phylogenetics?

Formal models serve the same function in phylogenetics as they do in other fields of biology: they are tools that facilitate estimation. Only in the simplest of systems can a model be exact, and reality in most biological systems is complex. Therefore, we seek models that are good approximations of reality given that reality itself is unknowable. Proper modeling tells us what inferences the data support, rather than what full reality might be [9,10].

In molecular phylogenetics, models are most commonly used as estimators of evolutionary change at a given nucleotide site through time. The primary aim of molecular phylogenetic inference is to approximate the progression of lineage divergences that produced a group of observed sequences. In statistical terms, this makes the sample a collection of homologous nucleotide sites at a given moment (the aligned sequence matrix), and the population being estimated is the previous condition of these sites through time.

The sample (sequence matrix) can be used to formulate a model of site evolution that represents, for example, an expectation about how an A nucleotide might have replaced a G nucleotide at the same sequence position in the past. In probabilistic frameworks, such as maximum likelihood or Bayesian inference, most expectations of character evolution are formalized as parameters in a largely explicit model of character change [2,5,11]. Each estimable parameter is intended to represent a specific feature of sequence evolution, such as an unequal rate of evolution among sites (the α parameter) or an inordinate proportion of transition substitutions (the κ parameter). When created, such parameters are usually carefully considered and have a logical biological justification for their use in inferring change among sequence data. However, a parameter responds to patterns in the data that might or might not be the mutational process that we think we are modeling, particularly when the distribution assumed by the parameter is not a good match with that of the process it is intended to represent.

Conceptual models that are in general use (including the many forms of parsimony) share several assumptions in common that are not given formal parameters. Some of these assumptions are (i) i.i.d.: mutations are independent and identically distributed; (ii) tree-like evolution: lineages

Corresponding author: Kelchner, S.A. (kelchner@isu.edu).
Available online 17 October 2006.

www.sciencedirect.com 0169-5347/$ – see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2006.10.004

arise in a divergent manner without reticulation; (iii) stationarity: mutational processes are consistent through time; (iv) reversibility: mutations can revert to a previous state; and (v) Markov process: mutation events are not influenced by a previous mutation at that site. Such general assumptions are often violated in reality; for example, prokaryote groups frequently share genes among lineages via lateral gene transfer and, thus, do not evolve in a tree-like fashion.

For a phylogeny estimate to be accurate, the assumptions of the formal model and of the conceptual model must adequately represent the true evolutionary process that produced the observed data.

Glossary

Akaike information criterion (AIC): an estimator of the information-theoretic Kullback–Leibler distance between the true model and the estimated model. An AIC value is derived from a maximum likelihood estimate that is penalized for the number of estimable parameters in the model. Model selection by AIC can include nested and non-nested model comparisons. The AIC framework also enables the assessment of model selection uncertainty.

Bayes factors: a Bayesian analog to the likelihood ratio test. The model likelihoods being compared, however, are derived from integration over all possible parameter values, rather than from maximum likelihood estimates for the model. Bayes factors, AIC and BIC can be used to select a model during a Bayesian inference analysis instead of requiring a priori model selection.

Bayesian information criterion (BIC): similar to the AIC in that it compares transformed likelihoods, but differs by penalizing sample size as well as number of estimable parameters. The BIC can be an approximation of the natural log of the Bayes factor, making it computationally more tractable than Bayes factors for phylogenetic purposes. However, BIC is more likely to select less complex models than those selected by Bayes factors [7].

Branch length estimations: the amount of estimated change in each lineage, usually quantified as the number of substitutions per site. The degree to which the set of branch lengths approximates the actual number of substitutions is governed by the adequacy of the model.

Heterotachy: the property of change in evolutionary rate at a given sequence position through time. If uncorrected, heterotachy can mislead phylogeny estimation because of systematic error. Heterotachy is a growing concern in phylogenetics [65,66].

Likelihood ratio test (LRT): a method that compares the maximum likelihood estimates of two nested models given one data set. Significance is assessed by an arbitrary attained significance value (usually $p = 0.05$). Technically, the test requires that one of the models being compared is the true model given the data.

Model averaging: an information-theoretic technique that makes formal inferences based on a set of adequate models. It can be used when more than one model is a reasonable approximator given the data, as judged by the degree of difference in model AIC scores ('Akaike weights'). Model averaging incorporates uncertainty in model selection as well as uncertainty in parameter estimates.

ModelTest: a widely used computer program that rapidly fits one of 56 candidate models to a nucleotide data set using hierarchical likelihood ratio tests. ModelTest default settings will build a neighbor-joining Jukes-Cantor tree on which maximized model likelihoods are estimated using the program PAUP*. ModelTest also provides an AIC score for each model.

Nested models: models are considered nested when they are all special cases of the most general model in the candidate set. In ModelTest, the general model is GTR+I+Γ, which is the most parameter-rich model in the set of 56 compared. LRTs can only be applied to nested models, a requirement that prevents, for example, the comparison of the GTR model with a covarion model.

Phylogeny estimation: the various approaches for inferring evolutionary relationships among living organisms (i.e. a phylogeny). All methods for molecular data depend upon an implicit or explicit mathematical model describing the evolution of aligned nucleotide or amino acid sequence characters. A phylogenetic tree is an estimation of the true phylogeny of a group of organisms.

Systematic error: in phylogenetics, error owing to the use of an inappropriate model of character evolution. The error is accentuated by the addition of more data, making it 'systemic'. Both underfitting (not enough parameters) and overfitting (too many parameters) of a model can create systematic error. An inaccurate tree topology might be the most noticeable result, although any parameter value could be erroneously estimated.

Question 2: Must a model be 'exact' or merely 'good enough'?

Many researchers wonder how strictly a model must fit detectable patterns in a sample. It is widely (but incorrectly) assumed that the more closely a model fits a sequence alignment, the more accurate the phylogenetic inference. One can improve the fit of a model to the data by adding more parameters, which is why likelihood values for a data set improve with the increasing complexity of the model.

Better fit, however, does not necessarily improve accuracy of phylogeny estimation [12–14]. Why is this? One answer is that inference of the nature of sequence change is made from the sample at hand, a property that enforces certain limitations on model composition. There is a cost to improving model accuracy by parameter addition: the more parameters that must be estimated from a finite data set, the higher the overall variance associated with those parameters [15–17] (Box 1). Improvement of model fit to the data will result in a more accurate

Box 1. Tradeoff between bias and variance in a model

A tradeoff between bias and variance is expected in phylogenetic models (Figure I; adapted with permission from Ref. [9]). Theory predicts that, as additional parameters are included in a model, the chance of bias in the estimate will diminish, which is good. Simultaneously, there will be an increase in variance associated with the estimate, which is bad. Parameter-rich models must estimate more parameters from the same amount of data, a situation that can reduce the precision of parameter estimates. A potential optimum is the point where a model can minimize both bias and variance (the intersection of the two curves in Figure I).

The tradeoff concept is the primary justification for model selection methods, all of which seek to trim away unimportant parameters in an attempt to evade unnecessary variance. Reality might prove more complicated, however, than Figure I suggests. We think it unlikely, for example, that each parameter in a phylogenetic model has equivalent potential to reduce bias or increase variance. Also, the effect should vary between data sets, depending on the degree of variability, the amount of data available and the complexity of evolutionary processes that produced the sequences.

It is unknown where current phylogenetic models fall on this graph. The assumption seems to be that we are drifting far to the right with our general models (such as GTR+I+Γ), but this has been difficult to demonstrate. A general resistance among phylogeneticists to develop more complex models might be largely based on this assumption.

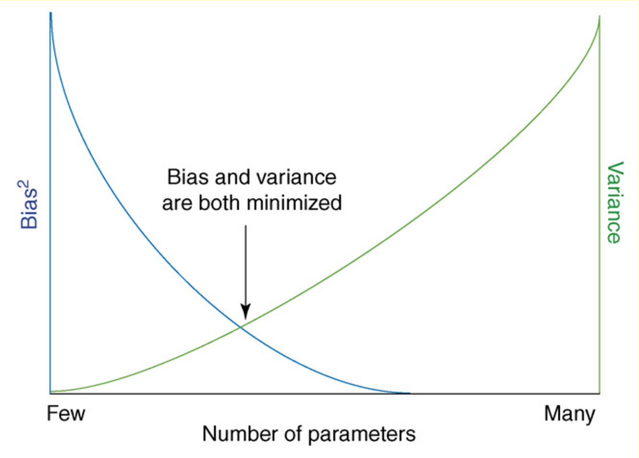

Figure I.

explanation of the sample (in this case, the sequence alignment) but a less precise estimate of the population (the condition of those sites through time). This means that any fine-scale parameterization of the variation present in a sequence alignment will make estimation of evolutionary history more difficult. Hence, a model must be 'good', but not 'exact'.

Because models should be good value estimators of all parameters assessed for a data set (of which the tree topology is only one), reduction of parameters in a model must be done with care. A focus on models that only recover accurate topologies could yield models that prove poor estimators of other parameter values, such as transition: transversion ratios or the degree of positional rate heterogeneity.

Question 3: What phylogenetic applications rely on best-fit models?

Scientific conclusions that are based on phylogeny estimations will always depend on the assumptions of the formal and conceptual models. Therefore, it is worth considering what phylogenetic techniques or applications require a model that adequately covers the mutational processes generating the data.

Varying the assumptions about the manner of evolution at a nucleotide site can result in different branch length estimations, which can then alter our inference about the amount of mutational change that has occurred between two sequences. Any application or technique that relies on a correct assessment of the amount of evolutionary change among lineages will therefore need adequate models. When the aim is only to recover an accurate depiction of relationships (a correct topology), more flexibility in model composition can be allowed in some cases. So, although one can view the importance of model fit as a continuum, it is a continuum that is heavily skewed toward proper fit (Box 2).

Question 4: What happens when a model is 'wrong'?

Models can be 'wrong' in two ways, by either underfitting or overfitting the sample. When a model is a poor approximation of reality ('underfit' owing to absence of key parameters), the consequence can be systematic error, a condition in which a bias strongly influences the analysis, resulting in an inaccurate, but sometimes well supported, phylogeny estimation.

Long branch attraction is occasionally due to model underfitting [18–21], particularly when inadequate taxon sampling is coupled with faster rates of substitution in one or more lineages, a situation that is more likely to mislead an analysis when the model does not include a correction for positional rate heterogeneity [22–24]. Controversy over the primary split in the flowering plant lineage, for instance, has been identified as a long branch attraction issue on account of problems in sampling and model misspecification [25,26]. Another well known form of systematic error is composition bias, in which separate lineages experience a shift in relative base frequencies so that AT-rich lineages, for example, are mistakenly positioned together on a topology [27–29]. A case was recently reported [30] of compositional bias misleading a genome-scale minimum evolution analysis of yeast species, an example made even more interesting by the high bootstrap support on the misleading topology. These studies and others [13,31,32] have now demonstrated that underfitting a model can adversely affect phylogenetic accuracy.

Box 2. When models matter

The importance of model fit can be viewed as a continuum, depending on the phylogenetic technique or application being used and the questions being investigated (Figure I). For studies that focus only on relationship among organisms, the primary requirement is an accurate topology, rather than an accurate estimate of how much change has occurred on the tree. In certain cases, topology might not change over a wide range of suitable models, suggesting that topology itself is (in these cases) robust to mildly inadequate models [31,53]. This is particularly likely in data sets that show little observable sequence variation [54,55]. For cases in which resolution of deep nodes in a tree is difficult owing to saturation, lack of character change, or sampling issues, model choice can have a decisive role in phylogeny reconstruction or outgroup rooting, although not necessarily for the better [26,56,57].

For most other uses of phylogenies, branch lengths can have a crucial role. When the aim is to estimate divergence times among lineages on a tree, accurate branch-length estimates are more important [58,59], although a rate-smoothing algorithm [60] can ameliorate this effect to some degree.

Testing an alternative phylogenetic hypothesis (e.g. Kishino-Hasegawa, Shimodaira-Hasegawa and Incongruence Length Difference tests) also requires adequate models. To investigate significant differences between competing topologies (as is commonly done in studies of coevolution [61] or horizontal gene transfer), conclusions are usually based on significant likelihood differences or comparison of support values among incongruent topologies. Both likelihood comparisons and bootstrap values are functions of estimated branch lengths and ultimately, therefore, of the choice of models [12,46,62].

Given that the interpretation of mutational history at a specific site in a nucleotide sequence can change depending on assumptions about evolutionary processes [63], model adequacy is also central to studies of molecular evolution that use phylogenetic relationship among sequences (e.g. testing for the presence of a molecular clock, measuring modes of molecular evolution, or detecting positive or purifying selection).

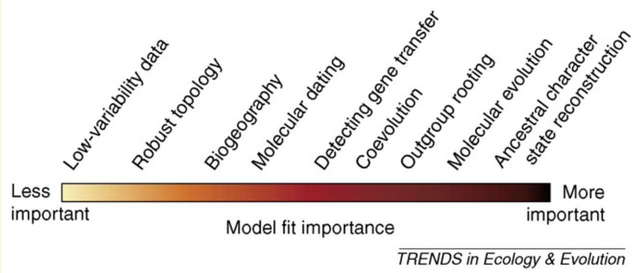

Figure I.

By contrast, a model might contain too many non-essential parameters that improve model fit but result in poor estimation of the parameters. Such overfitting of the model could, in principle, inflate the estimated degree of change for each lineage or invoke a set of shared characters for a clade that are not observable in a direct comparison of sequences. Suspected cases of overfitting empirical data, however, are rare [13,20,33].

Both overfitting and underfitting of models can produce topological error and poor branch length estimations. Because parameter values are usually estimated during analysis, the primary issue then is which parameters should be included for estimation when given a particular data set, a process referred to as model selection.

Question 5: How are models selected for nucleotide data?

Most model selection methods are statistical in nature, although a prominent exception is a priori model choice. A conceptual model is commonly selected for philosophical or operational reasons (e.g. the view that probability theory applies to sequence data, or that computational efficiency favors distance methods). By contrast, the a priori acceptance of a formal mathematical model is relatively rare, although some authors are now proposing the general use of the most parameter-rich model available when operating in the Bayesian framework [34].

Statistical methods of model selection are currently dominating model choice in phylogenetics, and most belong to three categories: frequentist (e.g. likelihood ratio tests); information-theoretic (e.g. Akaike Information Criterion); and performance-based (e.g. decision theory). There are recent in-depth reviews available that describe the details of each approach [7,8]; we present a descriptive overview of the methods in Box 3.

When a model is fit to an aligned sequence matrix, the choice of models relies on detectable patterns in the distribution of mutations. The process can therefore be misled when one attempts to fit a model to nucleotide data that show little or no sequence variation.

Question 6: What models are most frequently chosen for sequence data?

Given that nucleotide sequences are subject to an array of selective forces and mutational phenomena, one might reasonably expect that model choice would vary among sequence regions and data sets. Curiously, when using hierarchical likelihood ratio tests (hLRT), it does not. We surveyed 208 published data sets that used ModelTest [35] to select among 56 models. Our results show that only five relatively complex models make up nearly 80% of the observations, independent of genome, organism or region sequenced (Box 4; Online Supplementary Material).

The predilection of ModelTest to chose parameter-rich models for most data sets could reflect at least three probable explanations: (i) DNA sequences generally require parameter-rich models to accommodate features such as base composition, positional rate heterogeneity, unequal substitution class frequencies and perhaps unrecognized mutation processes; (ii) hLRTs are biased toward parameter-rich models, perhaps because of data violations of the i.i.d. hypothesis or the hierarchical fashion in which numerous models are being compared; and (iii) parameter-rich model selection by hLRTs is due to a combination of both complex molecular processes underlying the data and model selection bias. These possibilities suggest multiple lines of research for advances in model use for phylogenetics.

In the first case, the frequent occurrence across all categories of data of GTR+I+Γ (the general time reversible model with corrections for invariant characters and gamma-distributed rate heterogeneity) could suggest that even the most complex models in general use are not complex enough to capture all significant patterns in most data sets, a conclusion reached by other researchers [13,16,34]. If so, one path forward is to create and test novel parameters.

Alternatively, the observed parameter richness of chosen models might indicate a bias in the model selection process. For example, the relatively redundant inclusion of the invariant parameter (I) in the general model (GTR+I+Γ) is four times more common than the similar (and next most complex) GTR+Γ model in our survey. Although Γ and I are discrete parameters, they are strongly correlated [24,36] and Γ can accommodate for the absence of I in simulation studies (e.g. [31]). It seems

Box 3. Diversity of model selection methods

Although likelihood ratio tests are the most popular means of fitting a phylogenetic model to data, a variety of methods are currently in use.

Likelihood ratio test

Likelihood ratio tests (LRTs) compare nested candidate models in a pairwise fashion. Maximized likelihood estimates are obtained for each model, given the data and a 'reasonable' topology (often a neighbor-joining tree, as in ModelTest). Hypothesis testing takes the form of an alternative model being compared to a null model, and a p value is used to determine significance. Many have noted the probable inadequacy of the test for model comparison in phylogenetics (e.g. Refs [7,16]).

Bayes factors

Bayes factors are the Bayesian analog of the LRT; instead of using the maximized likelihood estimate for each model, the likelihoods being compared are derived from integrating over all possible parameter values within a Bayesian inference framework that includes prior probabilities (e.g. Refs [52,64]). Similar to BIC, Bayes factors can be used to select a model during Bayesian analysis, thereby combining the steps of model selection and phylogenetic inference.

Akaike information criterion

An information-theoretic approach, Akaike information criterion (AIC) is similar to LRT in that several candidate models are compared in the context of data and a reasonable topology. Maximized likelihoods for models, however, are penalized for the number of estimable parameters, effectively converting likelihoods into an estimate of informational distance. There is no hypothesis testing between models, and hence no p value of significance. Informed decisions about model adequacy are properly made by assessing the relative difference in the values of candidate models, and not (as is commonly done) by choosing the model with the highest ranking AIC score. More than one model can have similar scores, in which case each model should serve equally well as an approximator.

Bayesian information criterion

The Bayesian Information criterion (BIC) differs from AIC in that it accounts for sample size as well as the number of estimable parameters. By doing so, it approximates the log marginal likelihood of a model, assuming that priors are flat across models and parameters. The difference between two BIC scores is then a relatively quick estimate of the Bayes factor.

Decision theory

A relatively new approach involves decision theory [8,50,51]. The focus is on phylogenetic performance: in one method, model BICs are penalized relative to their degree of dissimilarity in branch length estimation. Hence, it is the ability of the model to estimate branch lengths that determines its quality, and not the overall fit of the model to the data.

Box 4. What is ModelTest telling us to use?

ModelTest is currently the most widely used program for fitting models to phylogenetic data using hierarchical likelihood ratio tests. We surveyed the first 137 phylogenetic publications from the first quarter of 2004 that cite ModelTest [35] in Thomson Scientific's Web of Science Citation Database (http://www.isiwebofknowledge.com/) and have an unambiguous description of how the model was selected (see Online Supplementary Material).

The 137 publications came from 43 scientific journals and yielded 208 data sets of aligned nucleotide sequences. The data sets are from nuclear, chloroplast and mitochondrial genomes, or are combined data sets from multiple genomes (Figure Ia). Sequence regions included rDNA, protein-encoding genes, introns and intergenic spacers, and no single region was represented in more than 11% of the data sets. Organismal diversity in the sample appears to be typical of the level of phylogenetic work being published for each taxonomic group, with the possible exception of prokaryotes, which are probably underrepresented. As might be expected, there is a high frequency of vertebrate studies, followed by arthropods and plants (Figure Ib).

Twenty of the 56 models examined by ModelTest are represented in our sample of phylogenetic literature (Figure Ic; model abbreviations follow Ref. [35]). Most of these models can be considered parameter rich, with 133 of 208 data sets (63.94%) requiring the three most complicated candidate models: GTR+I+Γ, GTR+Γ and TrN+I+Γ (ten, nine, and seven estimable parameters, respectively). Together with HKY+Γ and HKY+I+Γ (five and six estimable parameters, respectively), these five models make up nearly 80% of the samples in our survey.

The high frequency of GTR+I+Γ selection is consistent across all taxonomic groups with the exception of viruses, for which the model appears only once in ten instances. Not measured was the degree of similarity between sequences in each data set; if generally low, we could expect a skew of observations toward simpler models (i.e. to the left in Figure Ic).

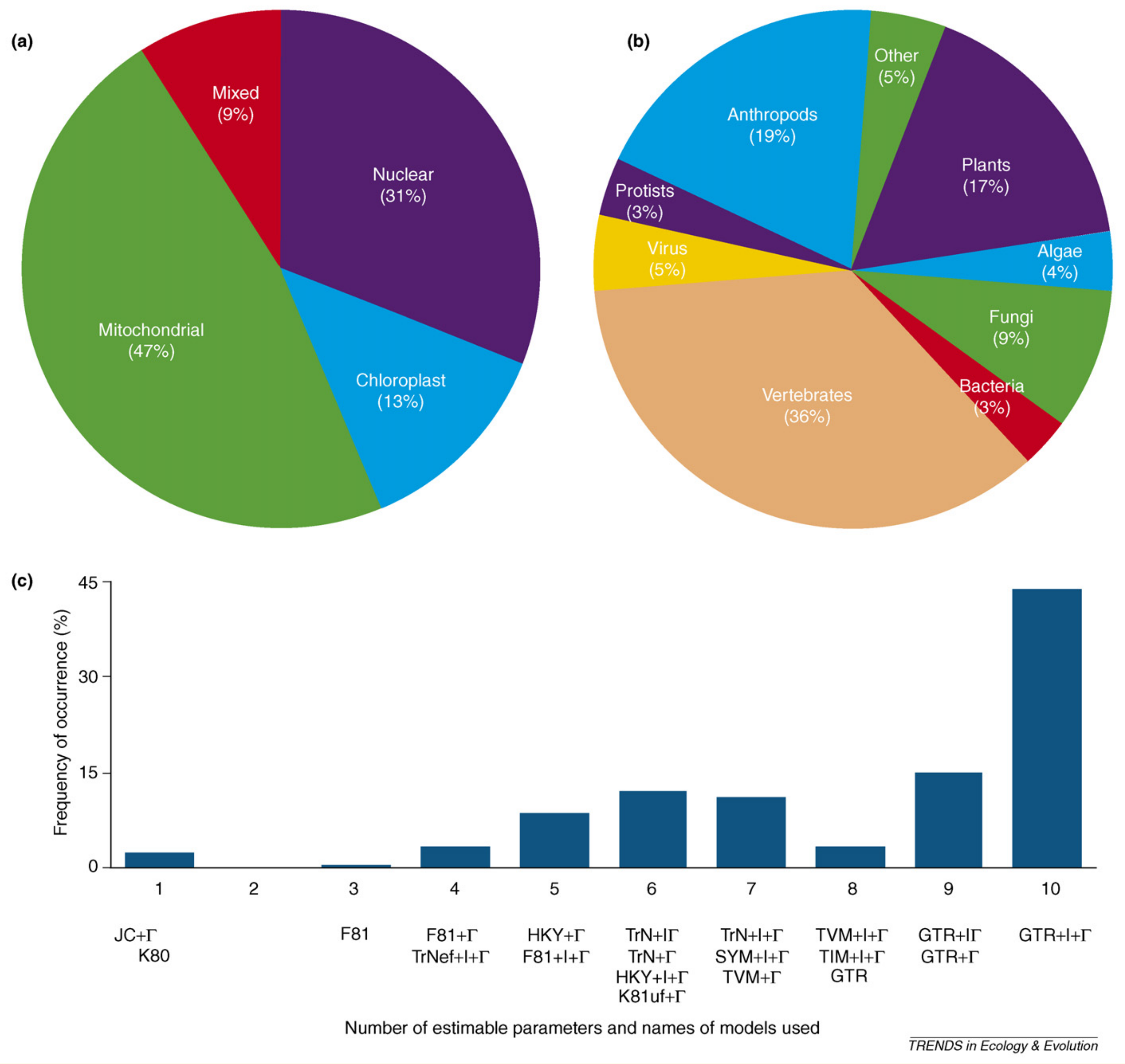

Figure I.

odd to us that the inclusion of I is essential in these data sets; perhaps its prevalence is an example of a model selection bias in which multiple hierarchical comparisons among too many models (56 in this case) leads to the incorporation of extra, largely spurious parameters. Another explanation might be that violation of the character independence assumption is making hLRT-based selection for model fit problematic. In our opinion, resolving the question of potential bias in hLRT-based model selection would be of general interest and is worth further research.

Question 7: How can model selection methods be improved?

Three main problems of valid inference using mathematical models are: (i) selecting the model; (ii) estimating the model parameters; and (iii) determining the precision of the estimate [9]. Phylogeneticists have made significant progress on the first two issues but have not yet widely used a method to assess precision of the model estimate. In our view, this is an important exception in an otherwise rigorous system and should be rectified (see also Ref. [7]).

'Precision of the estimate' should include precision in model selection as well as in parameter estimates, only one of which is the tree topology. Is the best-fit model the only adequate model for a given data set, or could other models also be reasonable choices? Can this uncertainty about the model be included in confidence measures of the phylogeny estimation? Underutilized tools exist for assessing model uncertainty, one of which is a goodness-of-fit test, which involves the simulation-based method of parametric bootstrapping [6,37]. In the Bayesian framework, a comparable approach simulates new data under the conditions indicated by the marginal posterior distributions obtained from the actual data [38,39].

An information-theoretic approach to model choice largely solves the problem of assessing the precision of the estimate. In the case of many models providing nearly equivalent information about the data, model averaging is an appropriate solution [7,9,40]. The technique is likely to increase the variance of the subsequent estimates because uncertainty about the choice of models is added to the uncertainty of the phylogeny estimation. This is appropriate when one is unsure of the correct model for the data at hand. Model averaging could have an important role in cases where different models are adequately informative yet produce alternate topologies.

To help limit bias in the model selection process that could be due to multiple model comparisons, we see a need for a fundamental shift in terms of the number of models being compared [9]. For example, the candidate pool for ModelTest could be limited to between five and ten well chosen models, instead of the 56 special cases of the general time reversible model with I and Γ, which are themselves subjectively chosen from among the 203 possible models of GTR alone. Limiting the number of candidate models would not alter the performance measure of either the LRT, AIC or BIC process, but the choice of models for comparison could vary depending on the data set, particularly when little sequence variation is present.

Question 8: Are all parameters equally important?

What parameters are strong performers in our current models? Our survey of model choice across organisms, genomes and sequence regions (Box 4) suggests that a few parameters are of consequence for most nucleic acid sequence comparisons. Primary among these is α (the shape parameter for a gamma distribution of positional rate heterogeneity), which appears in 98% of the models selected for in our 208 data sets. Correction for base frequency inequalities (*uf* in ModelTest) is almost as common, appearing in 95% of selected models. Nearly 80% of selected models required α, *uf* and a correction for different substitution frequencies (special cases of the GTR model).

Our observations are in accord with several empirical and simulation studies [13,24,31–33,41]. These parameters are not present in most forms of parsimony or in simpler distance and likelihood models, such as Jukes-Cantor or Kimura-2-parameter, which are still widely used as a priori models in molecular phylogenetics.

Question 9: Will phylogenomics eliminate the need for model selection?

If small data sets can create excessive levels of variance when estimating many parameters, would it be possible to increase the amount of data so that parameter-rich models can be used? The assembly of increasingly massive data sets might partly reflect this notion. The 'phylogenomics' approach, however, has not allowed researchers to sidestep the model selection process. Although gigantic data sets are likely to overcome statistical issues related to small sample size (sampling error), they continue to be susceptible to systematic error from inadequate models [42–44].

In addition, bootstrap values should be high in phylogenomic studies because sampling error has largely been eliminated. This can be problematic in the known cases where an alternative model produces a different topology that also exhibits high bootstrap values [26,27,45]. The important lesson from such findings is that genome-scale analyses are not impervious to systematic error owing to poor model selection, even when high bootstrap values are observed. Confidence in the topology should come from additional sources to the bootstrap, such as corroboration, sensitivity analysis and quality estimates of model adequacy [40,43,44].

The next step: new models, new methods

The importance of models and model selection in modern phylogenetics continues to motivate research on their proper use and improvement. We see three categories of model study that are, or shortly will be, dominating research of model application in phylogenetics.

Novel parameters

A model chosen in any analysis is only the best fit of those models that were tested. If a model exists, or could exist, that performs better as an approximator for a data set, then it should be developed and included in the model-testing process. This could include models that are more parameter rich than the general GTR+I+Γ model. There are few suspected cases where possible overfitting has been detrimental to phylogeny estimation, and thus we advocate a cautious exploration of novel parameters and model applications. Mutational phenomena that are not adequately covered in standard models include heterotachy, composition bias and violations of the i.i.d. hypothesis. If a distribution can be identified for each of these patterns, parameters could be developed to correct for bias that they might impart in phylogenetic analyses.

Relaxing current parameters

An alternative to novel parameter development is to increase the complexity of the model by relaxing certain parameter constraints. Partitioned models, for example, enable improved fit of the parameters to subsets of the data that evolve under different selective constraints [46,47]. Partitioning can increase the fit and quality of parameter estimation in a Bayesian framework [33] and has potential for widespread application in phylogenomics [43]. Mixture models, which permit a substitution rate matrix to vary among sites, offer an additional improvement in model fit by compensating for heterogeneity of substitution processes [48]. Mixture models can be implemented so that rate and pattern heterogeneity classes are estimated during analysis, an idea that lends itself readily to Bayesian inference methods, which are less sensitive to parameter richness [49]. It is not yet known whether partitioned models and mixture models suffer generally from increased levels of variance in parameter estimation.

Model selection

Model selection methods are rapidly evolving beyond hierarchical likelihood ratio tests. Bayes factors have a prominent role in the new approaches, which include decision theory methods [50,51] and model selection during Bayesian analysis [33,52]. Dynamical (instead of hierarchical) likelihood ratio tests using a reduce number of candidate models should also be considered, which would require a simple change to the ModelTest framework, or independent calculation using PAUP* or other phylogenetics programs.

There is already a push to better incorporate information-theoretic approaches that facilitate model averaging in cases where more than one model (nested or otherwise) contains similar approximating power [7,40]. Model averaging should better accommodate the total variance of a phylogeny estimation, perhaps tempering the current enthusiasm for trees that have high bootstrap support values.

Summary

All phylogeny estimation requires that assumptions be made about evolutionary processes. These assumptions are the conceptual model (either formal or implicit) that is used to estimate parameter values, such as branch lengths and topology, in a phylogenetic analysis. Although models are imperfect representations of reality, they need only be good approximators. Because many scientific conclusions are based on phylogeny estimations, and given that we have no certain knowledge of the true evolutionary relationships between organisms, confidence in the performance of a model is necessarily a function of confidence in the suitability of the model for the data at hand. Researchers should therefore seek the best possible model in most phylogenetic applications, a process that is still in development, particularly in the assessment of model selection uncertainty. Nucleotide data are generally complex, and parameter-rich models are most frequently chosen for comparative sequence analysis, suggesting that further model development and methods of application are warranted.

Acknowledgments

We thank Rick Williams, Peter Sheridan, Amanda Fisher, Olaf Bininda-Emonds, Thomas Buckley and two anonymous reviewers for comments that improved the article. S.A.K. was supported in part by National Science Foundation grant DEB-0515828 and a Visiting Fellowship at The Australian National University. M.A.T. was supported by a grant from the PhRMA Foundation and by the Idaho INBRE grant from the National Institutes of Health.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tree.2006.10.004.

References

1 Penny, D. *et al.* (1992) Progress with methods for constructing evolutionary trees. *Trends Ecol. Evol.* 7, 73–79

2 Steel, M. and Penny, D. (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17, 839–850

3 Tuffley, C. and Steel, M. (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.* 59, 581–607

4 Felsenstein, J. (2004) *Inferring Phylogenies,* Sinauer

5 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Hillis, D.M. *et al.*, eds), pp. 407–514, Sinauer

6 Whelan, S. *et al.* (2001) Molecular phylogenetics: state of the art methods for looking into the past. *Trends Genet.* 17, 262–272

7 Posada, D. and Buckley, T.R. (2004) Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53, 793–808

8 Sullivan, J. and Joyce, P. (2005) Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 36, 445–466

9 Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,* (2nd edn), Springer-Verlag

10 Johnson, J.B. and Omland, K.S. (2004) Model selection in ecology and evolution. *Trends Ecol. Evol.* 19, 101–108

11 Lewis, P.O. (1998) Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. In *Molecular Systematics of Plants II: DNA Sequencing* (Soltis, D.E. *et al.*, eds), pp. 132–163, Kluwer

12 Buckley, T.R. *et al.* (2001) Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst. Biol.* 50, 67–86

13 Buckley, T.R. and Cunningham, C.W. (2002) The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Mol. Biol. Evol.* 19, 394–405

14 Phillips, M.J. and Penny, D. (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* 28, 171–185

15 Nei, M. (1991) Relative efficiencies of different tree-making methods for molecular data. In *Phylogenetic Analysis of DNA Sequences* (Miyamoto, M.M. and Cracraft, J., eds), pp. 90–128, Oxford University Press

16 Sanderson, M.J. and Kim, J. (2000) Parametric phylogenetics? *Syst. Biol.* 49, 817–829

17 Yang, Z. *et al.* (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* 44, 384–399

18 Felsenstein, J. (1978) Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410

19 Kim, J. (1996) General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Syst. Biol.* 45, 363–374

20 Cunningham *et al.* (1998) Best-fit maximum likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* 52, 978–987

21 Anderson, F.E. and Swofford, D.L. (2004) Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Mol. Phylogenet. Evol.* 33, 440–451

22 Kuhner, M.K. and Felsenstein, J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468

23 Gaut, B.S. and Lewis, P.O. (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* 12, 152–162

24 Yang, Z. (1996) Among-site rate heterogeneity and its impact on phylogenetic analysis. *Trends Ecol. Evol.* 11, 367–372

25 Soltis, D.E. and Soltis, P.S. (2004) *Amborella* not a 'basal angiosperm'? Not so fast. *Am. J. Bot.* 91, 997–1001

26 Stefanovic, S. *et al.* (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol. Biol.*, DOI: 10.1186/1471-2148/4/35

27 Collins, T.M. *et al.* (1994) Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.* 43, 482–496

28 Lockhart, P.J. *et al.* (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11, 605–612

29 Galtier, N. and Guoy, M. (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15, 871–879

30 Phillips, M.J. *et al.* (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21, 1455–1458

31 Sullivan, J. and Swofford, D.L. (2001) Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution patterns are violated? *Syst. Biol.* 50, 723–729

32 Lemmon, A.R. and Moriarty, E.C. (2004) The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53, 265–277

33 Nylander, J.A.A. *et al.* (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53, 47–67

34 Huelsenbeck, J.P. and Rannala, B. (2004) Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 53, 904–913

35 Posada, D. and Crandall, K.A. (1998) ModelTest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818

36 Waddell, P.J. *et al.* (1997) Hadamard conjugations and modeling sequence evolution with unequal rates across sites. *Mol. Phylogenet. Evol.* 8, 33–50

37 Goldman, N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36, 182–198

38 Huelsenbeck, J.P. *et al.* (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310–2314

39 Bollback, J.P. (2002) Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* 19, 1171–1180

40 Alfaro, M.E. and Huelsenbeck, J.P. (2006) Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst. Biol.* 55, 89–96

41 Wakeley, J. (1994) Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* 3, 436–442

42 Naylor, G.J.P. and Brown, W.M. (1998) Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* 47, 61–76

43 Delsuc, F. *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375

44 Brinkmann, H. *et al.* (2005) An empirical assessment of long-branch attraction artifacts in deep eukaryotic phylogenomics. *Syst. Biol.* 54, 743–757

45 Philippe, H. *et al.* (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.* 22, 1246–1253

46 Pupko, T. *et al.* (2002) Combining multiple data sets in a likelihood analysis: which models are the best? *Mol. Biol. Evol.* 19, 2294–2307

47 Wilgenbusch, J. and de Queiroz, K. (2000) Phylogenetic relationships among the phrynosomatid sand lizards inferred from mitochondrial DNA sequences generated by heterogeneous evolutionary processes. *Syst. Biol.* 49, 592–612

48 Pagel, M. and Meade, A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence of character-state data. *Syst. Biol.* 53, 571–581

49 Holder, M. and Lewis, P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* 4, 275–284

50 Minin, V. *et al.* (2003) Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.* 52, 674–683

51 Abdo, Z. *et al.* (2005) Accounting for uncertainty in the tree topology has little effect on the decision theoretic approach to model selection in phylogeny estimation. *Mol. Biol. Evol.* 22, 691–703

52 Suchard, M.A. *et al.* (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* 18, 1001–1013

53 Rokas, A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804

54 Rzhetsky, A. and Nei, M. (1995) Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* 12, 131–151

55 Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics,* Oxford University Press

56 Kelsey, C.R. *et al.* (1999) Different models, different trees: the geographic origin of PTLV-I. *Mol. Phylogenet. Evol.* 13, 336–347

57 Graham, S.W. *et al.* (2002) Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Mol. Biol. Evol.* 19, 1769–1781

58 Sanderson, M.J. and Doyle, J.A. (2001) Sources of error and confidence intervals in estimating the age of angiosperms from *rbcL* and 18S rDNA data. *Am. J. Bot.* 88, 1499–1516

59 Kishino, H. *et al.* (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* 18, 352–361

60 Sanderson, M.J. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* 14, 1218–1231

61 Clark, M.A. *et al.* (2000) Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. *Evolution* 54, 517–525

62 Buckley, T.R. (2002) Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51, 509–523

63 Ronquist, F. (2004) Bayesian inference of character evolution. *Trends Ecol. Evol.* 19, 475–481

64 Brandley, M.C. *et al.* (2005) Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Syst. Biol.* 54, 373–390

65 Philippe, H. and Lopez, P. (2002) On the conservation of protein sequences in evolution. *Trends Biochem. Sci.* 26, 414–416

66 Lockhart, P. *et al.* (2006) Heterotachy and tree building: a case study with plastids and eubacteria. *Mol. Biol. Evol.* 23, 40–45