

# Phylogenetic models and model selection for noncoding DNA

Scot A. Kelchner

Received: 5 October 2007 / Accepted: 6 December 2007  
© Springer-Verlag 2008

**Abstract** Awareness of the complex structure and evolutionary dynamics of noncoding DNA has improved both noncoding sequence alignment and the use of microstructural changes as characters in phylogenetic analysis. The next step is to consider improvements in the use and selection of phylogenetic models for noncoding sequence data. Models of character evolution are central to phylogeny estimation, but the use of an inadequate model can mislead topology selection and branch length estimations. This is particularly likely when sequence divergence is either limited (nearly invariable, as in population-level or species-level studies) or extreme (nearly saturated, as in deep-level studies that focus on conserved secondary structures). Noncoding data sets are often at these extremes, and they can be particularly awkward for model definition and model selection. This paper introduces the goals of model use in phylogenetics and identifies ten issues that arise from the application of models to noncoding sequence data. It is concluded that most of these issues derive from small data set sizes, very low or very high sequence variability, limitations of current phylogenetic models, and possibly character definition and nonindependence. Recommendations are made that should help to improve alignment, character quality, model selection, and phylogeny estimation based on noncoding sequence data.

**Keywords** Noncoding DNA · Model selection · Model adequacy · Sampling error · Intergenic spacers · Introns

## Introduction

Noncoding DNA is currently playing a major role in population genetics and molecular systematics research. Comparative sequence analysis of intergenic spacers and introns, particularly from organellar genomes, is widely employed for designating haplotypes, tracking changes in population structure, identifying genetic introgression, and creating networks and phylogeny estimations of organismal relationships. A decade of research has revealed much about the nature and evolution of noncoding loci, and assumptions about character evolution in noncoding sequences have changed as information becomes available on the structure and probable function of the best known regions.

It is now time to consider how this new information might help us improve comparative sequence analysis of noncoding data. When the degree of relationship is sought between two or more DNA sequences, a suite of assumptions are brought into the process. These assumptions may either be implicit in the method, or several may be explicitly stated in a formal mathematical model of character change. Understanding the role of these assumptions in estimating a phylogeny is the key to appropriate and efficient model use in phylogenetics.

Here, I provide a synthesis on the use of models in phylogenetics and their application to noncoding DNA sequence data. Much of the following discussion will be in the context of organellar noncoding DNA, which is a relatively simple system compared to nuclear and mitochondrial

---

S. A. Kelchner (✉)  
Department of Biological Sciences, Idaho State University,  
Pocatello, ID 83209-8007, USA  
e-mail: kelchner@isu.edu

regions and has a long history of use in plant population genetics and molecular systematics.

### Model use in phylogenetics

The rise of formal model use in phylogenetics has generated contention over the proper use of models and their overall adequacy for phylogenetic inference. Although a variety of in-depth reviews on model use and application are available (e.g., Swofford et al. 1996; Lewis 1998; Whelan et al. 2001; Posada and Buckley 2004; Sullivan and Joyce 2005), Kelchner and Thomas (2007) expressed the view that much of the controversy surrounding models in phylogenetics may be due to misunderstanding of certain model qualities as well as a confusion about the term ‘model’ itself. Here, a specific interpretation of the model in phylogenetics is presented, that of the model as a tool for estimation. Although there are likely to be other views on the nature of phylogenetic models, this interpretation can be used to address the various challenges noncoding DNA data presents to model use. The following summary is intended to be brief; the topic can be more fully explored in Posada and Buckley (2004), Sullivan and Joyce (2005), and Kelchner and Thomas (2007).

#### Conceptual versus formal models

Reconstructing the past is necessarily a process of estimation. In phylogenetics, this estimation is made using philosophical tenants (e.g., likelihood theory applies to aligned nucleotide sequence), a priori expectations (e.g., each nucleotide is an independent character), and rigorous statistical frameworks. In any of these approaches, it is important to remember that a resulting phylogeny is only an estimate of a set of evolutionary relationships, and rarely a fully confirmed series of actual evolutionary events. Therefore, the term phylogeny estimation will be used to describe the topology and its associated parameters, which are the usual results of the phylogenetic inference process.

Estimating an evolutionary tree of relationships requires that assumptions be made about the process of character evolution that produced the observed data. Hence, no phylogeny estimation is assumption free. Taken as a whole, the assumptions we make during phylogenetic inference can be considered a ‘conceptual model’ of character evolution (Kelchner and Thomas 2007). The conceptual model may or may not include a ‘formal model’ component, which is a term usually applied to a mathematically explicit, parameterized model of character evolution such as the Jukes–Cantor (JC) model or general-time-reversible (GTR) model for DNA sequence data.

Distinguishing between these two concepts of the model in phylogenetics is important in that it allows us to avoid the notion that a particular phylogenetic method can be truly “model free”. For example, a so-called model free method such as equally weighted parsimony (sensu Fitch 1971) lacks a parameterized formal model but still makes significant and sometimes incorrect assumptions about the process of character evolution underlying the data. Notably, so does an inadequate formal model that contains improper parameters for the data at hand. Employing the term ‘model’ more broadly than just in its formal sense helps us to understand that it is the assumptions of character evolution (and how these are handled in the analysis) that matter most in phylogenetic inference, and whether those assumptions are appropriate for estimating phylogeny in any particular case.

Choice of a methodological framework is also part of the conceptual model because the framework can include implicit assumptions about the process of evolutionary change. For example, using Occam’s Razor, we make an assumption that the simplest explanation of the observed data is most likely to be correct. In parsimony, this means that the tree with the shortest number of steps is our best explanation. By contrast, use of the maximum likelihood framework for phylogenetic inference assumes that an answer with the maximized likelihood value is most likely to be true (in the context of the data, choice of model, computational algorithm, and background knowledge). These optimality criteria are inextricably linked to other assumptions about the manner of character state transformation, such as whether all possible state changes are equally likely or whether each character has an equal rate of state change. Other non-parameterized assumptions are shared by most frameworks in general use and include expectations of a tree-like evolutionary divergence of lineages, mutations that are independent and identically distributed, and a Markov process of character state transformation that stipulates reversibility of transformations as well as stationarity of the mutational process through time.

As discussed below, many of these assumptions prove problematic for phylogenetic inference from noncoding DNA data, and DNA sequence data in general. Unless specifically noted, the term ‘model’ hereafter will refer to formal (statistical) models used as estimators, which are usually parameterized. The discussion will center on nucleotide sequence data, currently the most heavily used form of data in phylogenetic analyses but by no means the only data available for phylogenetic inference. The reason for the focus is that nucleotide substitution models are among the most well-developed and sophisticated for statistical frameworks at present. The interest of the systematics community in developing formal models for

other data sets (e.g., morphological, microstructural, or genome arrangement characters) will soon provide a broader context in which to discuss model adequacy and model selection for phylogenetic inference.

#### The purpose of parameterized models in phylogenetics

Model use in phylogenetics has the same aim as model use in other fields of biology: formal models are tools that facilitate the estimation process. Formal model use in phylogenetics often takes place in a statistical framework, so there is both a sample for which a model is selected, and a population for which estimations are being made. In sequence-based phylogenetics, this means the sample is a series of homologous nucleotide sites among taxa at a given moment (the sequence alignment), and the population being estimated is the previous condition of these sites through time (Kelchner and Thomas 2007). The process of estimating this population allows us to infer divergence of lineages and produce a tree of hypothesized evolutionary relationships among organisms. However, we should remain aware that the model is used to estimate change at sequence positions through time, and not the condition of remaining unsampled genome sequence or the relationships among the organisms themselves.

The sample (aligned sequence matrix) is used to formulate the model of site evolution that will help us estimate the condition of that site through time. An expectation, for example, that a C nucleotide could change to a T nucleotide at a certain position in a sequence constitutes an assumption about the character evolution process. In probabilistic frameworks such as maximum likelihood and Bayesian inference, several such expectations of character evolution are described by formal parameters in a model. Each parameter is intended to account for a specific feature of sequence evolution, such as the ratio of transition to transversion substitutions (the  $\kappa$  parameter) or an inequality of evolutionary rates among sites (the  $\alpha$  parameter).

Development of a parameter for phylogenetic models involves careful consideration of the distribution one might expect for a given evolutionary process, such as a gamma distribution for positional rate heterogeneity. If the parameter's expected distribution is a poor choice for the evolutionary process we intend to model, then we could be misled about the meaning of its estimated value at the end of the phylogenetic inference process. It is also conceivable that a single parameter may be accommodating multiple evolutionary processes when each process shares a similar distribution function. This might explain why  $\alpha$  is such an important parameter for most DNA sequence analyses, since a gamma distribution could represent a variety of evolutionary processes besides

positional rate heterogeneity (for instance, perhaps the covarion process; see Penny et al. 2001).

An adequate model does not unduly 'constrain' an analysis if appropriate parameters are included and the parameters themselves are estimated during phylogenetic inference. When the parameters are not given specific a priori values, the optimization process in probability frameworks will converge upon a value estimate for each parameter in the model. If a non-essential parameter is included in the analysis, it may very well return a value close to zero. The Bayesian framework is thought to be particularly good at dealing with non-essential model parameters, so much so that some researchers are proposing the widespread a priori selection of the most complex available model for Bayesian inference using DNA sequence data (Huelsenbeck and Rannala 2004). However, in other analytical frameworks (and perhaps even Bayesian), non-essential parameters have the potential to mislead phylogeny estimation, which obliges us to consider the fit of the model to each individual data set.

#### The importance of model selection

Selecting a model for a data set is a critical step in model use for statistical inference. The model should include the important features of a sample that represent the major processes which produced the observed data; it should not include any relatively extraneous features. This is because there is a cost to adding parameters to a model. Given a finite data set, adding more parameters to a model means that more estimates must be made from the same amount of data. This will improve the description of the sample, but it could lead to a loss of precision in the population estimate. Because each parameter estimate has variance associated with it, adding more parameters to a model can increase the total variance associated with the estimates (Huelsenbeck and Rannala 1997), one of which is the tree topology itself. So, even though a parameter-rich model might do an excellent job of describing the sample, it could perform poorly when estimating the population [see Posada and Buckley (2004) Kelchner and Thomas (2007) for more complete discussions].

Models, then, need to be adequate but not exact. What matters most during model selection is that the major processes of character evolution are accommodated by the model, enough so that a good estimation of the phylogeny can be made. Parameters that do not contribute substantially to the estimate can be trimmed from the model. A goal of model selection is therefore the determination of which parameters are required for the estimate given the data, and which can successfully be left out. To ensure proper estimation, however, at least one model in the candidate pool must be adequate for the data so that the

model selection process has the opportunity to choose an effective estimator.

Many applications for phylogeny estimations require adequate models because reasonable accuracy of the parameter estimates (which include branch lengths and tree topologies) is essential. Applications that rely on accurate estimates of branch lengths on a tree, such as molecular dating, molecular evolution studies, and ancestral character state reconstructions, are prime examples. Assessing the significance of topology differences or similarities, as is frequently done in alternative hypothesis testing, co-evolution studies, or detection of lateral gene transfer events, is also ultimately dependent on the adequacy of assumptions made about the character evolution process (Buckley et al. 2001a; Buckley 2002; Kelchner and Thomas 2007). Hence, appropriate model selection is a key to successful phylogenetic inference and its many uses in biology.

#### When do phylogenetic models ‘fail’?

Models for phylogenetic inference can fail under a variety of circumstances. These include (1) bias and systematic error due to the absence of one or more key parameters (‘underfitting’), (2) imprecise or inaccurate results due to a model that is too parameter-rich (‘overfitting’), (3) improper general assumptions of the methodological framework, and (4) random error due to small sample sizes, which make both the fitting and the estimation of model parameters subject to error.

In the first case, ‘underfitting’ a model by failing to include key parameters can lead to a strong bias in an analysis. When this bias gets stronger with the addition of more data, it is referred to as systematic error (i.e., ‘systemic’ to the process). Classic examples of bias and systematic error due to inadequate models include certain cases of long branch attraction (both simulated and empirical: e.g., Cunningham et al. 1998; Sullivan and Swofford 2001; Stefanovic et al. 2004) and base composition bias (Collins et al. 1994; Lockhart et al. 1994; Naylor and Brown 1998). Of more recent concern is bias due to heterotachy, in which a model of character evolution fails to accommodate temporal shifts in the evolutionary process (Philippe and Lopez 2001; Lockhart et al. 2006). Systematic error is a principle concern because it can strongly mislead a phylogeny estimation, particularly when such estimations have strong bootstrap support (Naylor and Brown 1998; Phillips et al. 2004; Brinkmann et al. 2005).

Alternatively, the inclusion of too many parameters with limited significance for the population estimate might create problematic levels of variance so that parameter estimates are imprecise, inaccurate, or untrustworthy. For instance, an a priori decision to use a parameter-rich model (such as GTR + I +  $\Gamma$ ) for a data set with little sequence

variation or small sample size could, in principle, introduce problematic levels of variance into the phylogeny estimation (Nei and Kumar 2000). In a maximum likelihood framework, the many parameters of the model will each have a variance associated with their estimates, and total variance could reach problematic levels so that a maximum likelihood estimate of the topology would represent one of many reasonable estimates for the data. If a single topology is returned from the search, the imprecision of the estimate may go unnoticed.

Other forms of inconsistent results in phylogeny estimation may be linked to inappropriate assumptions of the methodological framework itself when given a particular data set. For example, minimum evolution has been shown to be susceptible to strong base composition bias in a genomic-level analysis of yeast species (Phillips et al. 2004). In general, so is parsimony, neighbor-joining, and maximum likelihood when using inadequate models (Lockhart et al. 1994). It is known also that long branches in sister lineages might ‘repel’ each other under maximum likelihood when data are limited (Waddell 1995; Swofford et al. 2001). Each example is a case when one or more methods underperform for a problem that might be solved correctly with an alternative method, illustrating that the conceptual model of character evolution is also dependent on assumptions of the methodological framework to be accurate.

In this paper, I suggest that the most relevant categories of potential model problems for noncoding DNA are often small sample size and limited sequence variation. Small data sets are known to be problematic for parameter estimation, including the tree topology (e.g., Wakeley 1996; Swofford et al. 2001; Holland et al. 2003). Since most of the current uses of noncoding data involve the detection of nucleotide differences in one or a few intergenic spacers or introns, and these in closely related populations or species, short sequence alignments with limited sequence variation is a common problem. Such data sets are severely limited samples from which inference of the population are to be made, and determining what is a suitable model in these cases can be problematic. A complex model can sometimes be selected for a data set that shows little sequence variability, which could lead to overfitting of the model and unnecessary levels of variance associated with the parameter estimates (Nei and Kumar 2000). Alternatively, a small data set could prove difficult to select an appropriately complex model for, in which case a bias from the lack of a key parameter could potentially mislead results of the analysis. In either situation, the use of models for such data can create problems that arise fundamentally from random error due to an inadequate number of observations.

Practitioners of noncoding DNA phylogenetic analysis (or for that matter, any phylogenetic data set) should probably anticipate a level of error from both sampling and

bias issues. Steps should be taken to limit their potential effect when inferring phylogeny from aligned noncoding sequences. The following section explores potential difficulties for effective model use with noncoding DNA data. These difficulties are presented as ten related issues, many of which are derived from the joint problem of small sample size and lack of sequence variability that is found in many noncoding DNA studies. I discuss strategies that are presently available to address these issues and make suggestions on how best to proceed with phylogeny estimation when using noncoding sequence data.

### Challenges of model use for noncoding data

Noncoding sequences present several challenges for the implementation of models in the phylogeny estimation process. Most of the issues discussed below are not unique to noncoding data, although they may be more pronounced for such sequences. Issues 1 through 4 largely arise from the complicated nature of defining molecular characters (nucleotides, indels, etc.) and the lack of complete independence among at least some of the observed character transformations. Other problems (Issues 5 and 7) are created by the small number of characters and low-variability found in many noncoding data sets. Issues 6 and 8 address the difficulty in finding a model that appropriately represents character evolution in noncoding DNA. Finally, Issues 9 and 10 focus on the mechanics of creating an adequate pool of candidate models and reducing the bias that can occur during the model selection process.

#### Issue 1: locus homology

A good alignment is required when determining the most appropriate model to use for phylogenetic inference and then using this model for phylogeny estimation. A priori statements of positional homology must be adequate for the nucleotide data if the resulting tree-like structure is to represent evolutionary relationships. The first step in homology assessment is the verification of orthology among sequences.

The orthology of organellar noncoding sequences in a study was at one time widely and confidently assumed. Although multiple copies of mitochondrial and plastid genomes are present for each organelle type, and multiple organelles are usually present in each cell, organellar noncoding data were expected to behave as single-copy loci in the organism because there is typically only one gene copy per genome. However, this assumption has typically not been thoroughly tested in molecular systematics studies. It has become clear that paralogy can affect a phylogenetic analysis of organellar noncoding DNA (e.g., Pirie et al. 2007). A review by Wolfe and Randle (2004)

summarizes several studies that have reported complications to homology assessment in organellar genomes, including heteroplasmy, mobile elements, transfer of genes and noncoding DNA to other genomic compartments, and duplication events. These processes create either paralogous copies of the target locus or recombined forms of that locus which, if left unrecognized, could undermine the phylogenetic estimation process.

A noteworthy example of noncoding paralogy is the insertion of partial *trnF* genes in members of the Asteraceae, Juncaceae and Brassicaceae (Vijverberg and Bachmann 1999; Drábkova et al. 2004; Koch et al. 2006). Tandemly arranged sets of *trnF* domains can be found within the *trnL*–*trnF* intergenic spacer of these chloroplast genomes, sometimes numbering as high as 14 copies (Koch et al. 2006). Pirie et al. (2007) recently demonstrated a phylogenetic conflict among chloroplast loci that arose from a duplicate *trnL*–*F* copy in the nuclear genome of members of the Annonaceae, which had maintained conserved exons for an estimated 88 million years. The widespread use of the *trnL*–*trnF* region in plant molecular systematics raises concerns about whether pseudogene copies have gone previously unrecognized in this length-variable spacer region.

Chloroplast and mitochondrial noncoding regions are occasionally duplicated or transferred into other genomic compartments, either as large exported pieces of the genome, components of a polycistron, or as an intron that has often moved with its host gene. Plant mitochondria seem to be more susceptible to gain and loss of genetic material (reviewed by Knoop 2004) although there are multiple cases of chloroplast regions being transferred into the nuclear or mitochondrial genome (e.g., de Pamphilis and Palmer 1990; Cummings et al. 2003; Shahmuradov et al. 2003). There is evidence that the group I intron in chloroplast gene *trnL* has, on more than one occasion, moved between chloroplast and mitochondrial genomes in certain plant lineages. This direction of transfer is far more commonly observed than transfer of material into the chloroplast genome from the mitochondrion.

By contrast, it is relatively rare to find group II introns being copied or moved outside of their organelles in eukaryotes (Kudla et al. 2002; Won and Renner 2003; Meimberg et al. 2006). It may be that group II introns, which make up a high percentage of noncoding DNA in plant chloroplast genomes, offer the best chance for recovering an orthologous, single-copy region among organellar noncoding loci (Kelchner 2002).

#### Issue 2: character homology (DNA sequence alignment)

At a finer scale, homology of nucleotide sites must also be assessed, and it can be more difficult to establish site

homology in noncoding regions than in coding DNA. Length variation is common in intergenic spacers and introns, and gap placement is often a serious issue. Several papers in this volume (e.g., Morrison; Ochoterna) address the process of homology assessment and alignment, and the reader is referred to these authors for additional viewpoints. Here, only the topic of hand alignment is covered as a developing alignment technique for noncoding sequence data.

Improvements to homology assessment in length-variable regions incorporate biologically relevant information such as secondary structure and mutational mechanisms to place gaps and maximize reasonable a priori statements of homology (e.g., Kelchner and Clark 1997; Graham et al. 2000; Borsch et al. 2003). To date, this can only be achieved by careful assessment of sequence pattern and structure during hand alignment. Hand alignment of sequences using relevant biological information is a justifiable principle for making homology statements about nucleotide and indel characters. A computer algorithm that is not concerned with biological realism and the molecular evolution processes underlying the sequence data is unlikely to be a completely adequate estimator of positional homology (see also Kelchner 2000; Graham et al. 2000). Most practitioners implicitly recognize this and adjust computer-generated alignments to a greater or lesser degree.

It is fair to say that the ability to do hand alignments in a strictly imitable fashion ('repeatability' in its 'imitation' sense, Morrison 2006) by any practitioner is still lacking. Imitation of an alignment is something that computer algorithms do very well. The argument, however, that an alignment algorithm is therefore required for reproducibility in phylogenetics, and that hand alignment must be avoided, places far too high a premium on the ability of alignment algorithms to accurately capture positional homology, and far too little on the ability of the human brain to provide a good and reproducible alignment solution without the aid of software. That argument would make some sense if nucleotide alignments are (1) unable to be informed by knowledge of the biological processes contributing to the observed sequences, and (2) nucleotides are truly independent character units. Under such conditions, our a priori assessment of nucleotide homology might be no better than a guess, and we would be forced to rely solely on character state similarity and order to infer character homology. That has never been a productive approach with morphological characters in phylogenetics, so why should it be acceptable with molecular ones?

Rarely do we see discussion in the literature that morphological character statements must be strictly repeatable by anyone who is willing to try. Whether appropriate or not, it is expected that researchers have made biologically

informed decisions when demarcating homologous characters. A similar process is taking place in hand alignment if researchers use secondary structure, mutation mechanisms, and sequence patterns to make statements of nucleotide or indel homology. And as with most morphological studies, the principles used for homology decisions during sequence alignment can be, and should be, included in the paper: see, e.g., Graham et al. (2000), Borsch et al. (2003) and Löhne and Borsch (2005). Neither hand alignment nor morphological character designation guarantees that any person can exactly reproduce the homology statements used in a phylogenetic analysis, in the absence of background information.

Assessment of the proposed phylogeny estimation is therefore informed by a description of the principles used in the character homology statements. Just as scientists are free to disagree about homology assessment for morphological characters used in a phylogeny estimation, they are free to disagree on whether a sequence alignment is adequate for a phylogeny estimation. A researcher who claims in a study that a tree-like structure represents evolutionary history is usually well aware that the claim is fundamentally dependent on the quality of the alignment. If the sites used to produce the tree are not homologous, the tree-like structure is no more than a branching pattern recovered by the software's algorithms.

In both cases (morphology and nucleotides), the ability to reproduce the topology and other parameter estimates from the given data is important. This can be done without having to reconstruct from scratch the homology statements (positional alignments) used in the experiment: detailed explanation of the alignment principles used, and preferably the alignments themselves, can readily be provided to both facilitate a repeat of the same analysis and to study the homology statements that form the basis of the phylogeny estimation. Any realignment of the data constitutes a new analysis with, to a greater or lesser extent, 'new data' since the homology statements have been altered. If realignment produces a different tree, the result could be competing phylogeny estimations. Since both are simply hypotheses of sequence relationships, this is perfectly acceptable in the scientific framework and should lead to the testing of each alternative hypothesis to see which is the most likely representation of phylogeny.

### Issue 3: character definition

It seems straightforward to designate characters in a nucleic acid sequence: the sequence itself is simply a long series of nucleotides. Not surprisingly, nearly all algorithms for phylogenetic analysis treat an aligned nucleotide position (i.e., a column in an alignment matrix) as an independent character. For early computational tractability,

it was also convenient to (1) treat each character state transformation as being independent of all other transformations in the sequence, and (2) expect each nucleotide site to share one common model of character evolution, a property known as identical distribution. Together, these treatments enforce the assumption of IID (an acronym for independent and identically distributed) on nucleotide data, and the independence of the substitution process at each site is a requirement which is built into nearly all of our phylogenetic methods (Huelsenbeck and Neilsen 1999).

Although it is necessary to invoke the IID assumption with our current software, it is very unlikely that such conditions hold for the majority of sequence data. One nucleotide does not always equal one character. In noncoding sequence, for example, conserved secondary structures are common, as are repeat motifs and multinucleotide length mutations (reviewed by Kelchner 2000; Borsch et al., this issue). In the case of secondary structure, selection constraints can influence the type of substitution and length mutation that is observed in a stem, particularly in introns (Kelchner 2002; Quandt et al. 2004; Löhne and Borsch 2005). Some have argued that a character in a secondary structure is better represented by two pairing sites which evolve in concert to maintain a DNA or RNA helix (Dixon and Hillis 1993; Schöniger and von Haeseler 1994; Tillier and Collins 1995). Also, length variation in noncoding sequence is most often due to the addition or deletion of simple sequence repeats (SSRs) which involve a few to several adjacent nucleotides in a single mutational event (Graham et al. 2000; Hamilton et al. 2003; Borsch et al. 2007). Likewise, minute inversions are single event, which affect several nucleotides at once, and the occurrence of minute inversions is dependent on the presence of a hairpin structure (Kelchner and Wendel 1996; Mes et al. 2000; Graham and Olmstead 2000; Quandt et al. 2003; Kim and Lee 2005).

The outcome of such phenomena is that many mutations, at least in organellar noncoding DNA sequences, are now viewed as being neither completely independent nor identically distributed. This is a major shift from earlier expectations that noncoding regions evolved randomly and were under no selective constraints (i.e., every site was independent and equally likely to change). Both the advent and the persistence of a mutation now appear to be influenced by surrounding sequence pattern and structure (Kelchner 2000). This nonindependence among nucleotide characters might have repercussions for all aspects of phylogenetic analysis, including sequence alignment (Kelchner and Clark 1997; Kelchner 2000; Morrison 2006), parameter estimation (Steel et al. 2000), phylogeny estimation (Huelsenbeck and Neilsen 1999), calculation of bootstrap support (Sanderson 1995), and model selection (this paper). The degree that it affects

phylogeny estimation using noncoding data, however, has yet to be measured.

In model selection, character definition plays a little discussed but critical role. Our current models treat each nucleotide as an independent character. When selecting models for full phylogenetic estimation, the likelihood value (or information score) is calculated for each candidate model given the data at hand, and these values are compared to determine which of the candidate models is the best available estimator for a sample of nucleotide characters. One could expect a bias in model selection if nucleotides do not, in fact, represent independent characters (e.g., correlated sites in a conserved stem), because the selection process might choose an inadequate model when the candidate models do not properly take into account the heterogeneous assembly of characters.

Similarly, the ability to conduct a valid bootstrap analysis on sequence data requires that sampling has the IID property (Felsenstein 1985). Felsenstein (2004, pp. 343–344) notes that if enough data is available, the outcome is still likely to be IID in the bootstrap process, since rate differences, if randomly distributed, will all be represented in the sample. The conditions for his argument, however, may not hold for noncoding data, particularly in the data set size and sequence variability required. If that is the case, then Sanderson (1995) suggests that the problem of non-IID characters, at least in the bootstrap case, has two potential solutions: either the characters must be defined so that they represent IID conditions (which is currently not the case), or the sampling scheme must be modified so that the IID property is invoked. Altering the sampling scheme is not an easy task, however, given our current lack of knowledge about what constitutes an IID character in noncoding DNA. If such a scheme is ever developed for sequence data sets, which may prove too difficult due to the idiosyncrasies of evolution among regions, then bootstrap resampling could be conducted on these alternative characters in the standard manner that is used today. The problem, as both Sanderson (1995) and Felsenstein (2004) note, is not with the bootstrap method itself, but with the need for IID conditions to be created for the process.

Redefining characters in DNA sequence may be one approach to correcting for nonindependence of characters in phylogeny estimations, although it may prove as hard as the alternative approach of redefining the sampling scheme. A flexible technique for heterogeneous character discrimination in sequence data sets could better accommodate associated nucleotides, linked mutational events, and sites that truly are independent character units. Inference of mutational mechanisms and secondary structures could be automated by software, and decisions about character definition can be one result of this process. However, heterogeneous character definitions in a data set would also

require a change in the substitution models that we use, since these determine the probability of change at individual sequence positions and usually invoke a uniform model of stochastic change across all sites.

Perhaps the first question we should try to answer is whether violation of IID causes significant consequences for most phylogeny estimation using sequence data. Empirical attempts to investigate the affect are rarely conducted, although Huelsenbeck and Neilsen (1999) showed in simulations that phylogenetic methods can become less efficient when there is at least a temporal correlation among substitutions. Otherwise, we do not yet know how serious or widespread the problem is for alignment, model choice, or phylogeny estimation of DNA sequences.

#### Issue 4: microstructural changes as phylogenetic characters

Microstructural changes (indels and minute inversions) are becoming widely used as phylogenetic characters for noncoding data sets. The biological phenomena that create microstructural changes in organellar noncoding sequences are generally well known (Kelchner 2000, Borsch et al., this issue) and, barring cases of variation due to mobile elements or recombination (e.g., Laroche and Bousquet 1999), high quality characters are often produced for phylogenetic analysis when these mechanisms are used to inform homology statements for length mutations and inversions during sequence alignment (e.g., Graham et al. 2000; Ingvarsson et al. 2003; Wanke et al. 2007).

Many methods exist for coding inferred insertions and deletions as phylogenetic characters. At their foundation, most are a simple presence–absence technique for indels that are deemed potentially homologous (e.g., Golenberg et al. 1993; Kelchner and Clark 1997; Graham et al. 2000), a procedure that has been codified and converted to software for regions with overlapping gaps of different lengths (e.g., Simmons and Ochoterena 2000; Müller 2006). Tests of gap coding methods that use a wide range of branch lengths, data set sizes, tree topologies and symmetries are starting to appear in the literature (Ogden and Rosenberg 2007), although it is difficult at this time to determine which methods are optimal for any particular data set.

Despite the ready availability of software for the purpose, it is still critical to remember that indel coding methods are properly applied to a sequence alignment only after the gaps have been placed (i.e., after statements of character homology have been made for indels, and ambiguous regions have been removed). The methods can be applied just as easily to random computer-generated gaps in an alignment, independent of any homology assessment. Hence, the careful proposal of homology

statements for nucleotides in a length-variable region should precede the application of any indel coding method. Some authors have even included a table or appendix for the reader's consideration that lists each coded indel in a study (e.g., Golenberg et al. 1993; Kelchner and Clark 1997; Sang et al. 1997; Müller and Borsch 2005b; Borsch et al. 2007), which can be an important tool for assessing the value of specific indel characters in a phylogeny estimation.

Following the principle that indel characters, like nucleotide characters, should be supported by reasonable homology statements, length-variable regions that cannot be deciphered in terms of probable indel origins (so called 'ambiguous' alignment regions) should be excluded from a phylogenetic analysis (Kelchner 2000; Müller 2006). Although alignment algorithms might produce a tree-like signal from such data (e.g., Lutzoni et al. 2000), the resultant topology could simply be an artifact of the algorithm and not a true representation of phylogenetic signal.

Scored microstructural changes provide an unusual set of characters for our current phylogenetic models. Most indels have traditionally been scored as two-state characters (usually 0 = present, 1 = absent), although multistate character scoring has also been developed (e.g., Simmons and Ochoterena 2000) and revised (Müller 2006). The characters are then most commonly analyzed using equal character weighting in parsimony, or the binary model or standard discrete model in MrBayes (Ronquist and Huelsenbeck 2003). Likewise, a multistate Jukes–Cantor-type model proposed by Lewis (2001) could also be developed to analyze indels under similar assumptions within the maximum likelihood framework. In each of these approaches, however, the characters are treated as if each is independent and equally likely to show transformation. This is a straightforward way to treat indel characters but is very possibly inaccurate for estimating the evolutionary history of length mutations in noncoding sequences.

We know that microstructural changes can occur with unequal frequency. The clearest example is minute inversions, which are often the most homoplasious of scored microstructural mutations in a study (Kelchner and Clark 1997; Dumolin-Lapègue et al. 1998; Graham et al. 2000; Quandt et al. 2003). Similarly, simple sequence repeats are highly variable in their rate of occurrence, because the rate is thought to depend partly on the length of the repeat and partly on the number of tandemly arranged repeats in the template (van Ham et al. 1994; Kelchner 2000; Hamilton et al. 2003; Müller and Borsch 2005a; Borsch et al. 2007). Because repeats are often found in loops of helical DNA or RNA structures, functional constraints on loop size and content might also affect whether an insertion or deletion of a repeat is likely to persist.

In the absence of reasonable probability estimates for the occurrence of different microstructural changes, we continue to treat indel characters under parsimony or a simple multistate probability model. Interestingly, this seems to be causing little harm: CI and RI values for indel characters are often higher in parsimony analyses than those of nucleotide characters in the same data set (e.g., Graham et al. 2000; Ingvarsson et al. 2003; Müller and Borsch 2005a). However, development of probabilistic models for microstructural changes would still be useful for maximum likelihood and Bayesian analyses. Such models could enhance the estimation of phylogeny from a combined matrix of nucleotide and scored indel characters by facilitating the use of partitioned models.

#### Issue 5: data with low variability

Frequently, noncoding DNA is chosen for identifying genetic differences within and between populations and species. These studies often seek to establish haplotypes, develop networks of haplotype relationships, or estimate the phylogeny of the samples. Comparative sequence analysis in these cases often shows little variation. Although designation of haplotypes might be possible with such data, the creation of networks and phylogeny estimates will be more problematic when the observed distances between sequences are very small.

For data sets with low variability, alignment is usually not a problem since most sequences will not show a substantial amount of length variation. Selecting a model for the data, however, will be more challenging (Posada and Buckley 2004). Lack of observable variation among sequences can make model selection imprecise because the assessment of which parameters to include is hindered when there is an insufficient number of changes in the alignment matrix. Although the evolutionary processes that gave rise to the data may be complex, the lack of observed variation in the sample makes it difficult to accurately estimate those processes. The result is likely to be ambiguity in the model selection process and a tendency to choose 'simple' models with few free parameters. That may be fine for phylogeny estimation, however, if complex patterns of character evolution have not had time to eventuate in the population being estimated.

Certain features of the data may still be obvious for parameter inclusion. For example, base frequency inequalities can be reasonably estimated from the alignment, and some idea of the substitution-type frequencies might be derived (although poorly) from this estimate. Other features of character evolution, such as positional rate heterogeneity, will be nearly impossible to determine with precision. The variance associated with parameter estimates in a chosen model could be proportionately larger

for low-variability data than variance based on data sets with higher levels of sequence variation (e.g., Wakeley 1994, 1996), a problem that is statistical and due to an inadequate number of observations.

Predictions about the population can still be made with a small sample, but it is necessary to view such predictions with appropriate sobriety. A phylogeny estimate could be constructed, for instance, that has only one change supporting each resolved node. Bootstrap support will be marginal or non-existent because the data set is small and relatively invariable. Interpreting the topology as an estimate of phylogeny then becomes a matter of whether the researcher believes that infrequent mutations improve the chance that synapomorphies are correct (i.e., there has been little time for problematic levels of homoplasy to develop). It should be clear, however, that the level of evidence is marginal. Appealing to the selected model as support of a claim that a particular change is likely given the data would be, in this case, inappropriate since precision of the model selection process is largely undermined by the limited sequence variation in the data.

One approach for low-variability data sets is to produce a robust topology (e.g., Kelchner 2003). Robust topologies are tree structures that do not alter when assumptions change about the manner of character evolution (Penny et al. 1992). The approach is commonly used in systematics literature, but is rarely discussed as a method to seek out and avoid error in a phylogenetic analysis. When a full range of reasonable models each produce the same topology, one can at least rule out the possibility that an inadequate model from the candidate pool is biasing the phylogeny estimation. If minor changes in topology occur across models, the approach is then to collapse all branches that are susceptible to changes in assumptions about character evolution: what remains is a less resolved, but robust, topology (Kelchner 2003). Although potentially useful for making phylogeny estimations from small data sets with limited variability, the robust topology method does not protect from stochastic error, or any bias that is currently uncorrected by the models in the candidate pool. The strongest evidence that a robust topology represents phylogeny still comes from corroboration by independent data.

#### Issue 6: data with high variability

At the opposite end of the spectrum from population-level and species-level uses, noncoding DNA has recently found application at much deeper levels of organismal phylogeny than would have been predicted a decade ago (Borsch et al. 2003; Quandt et al. 2004; Löhne and Borsch 2005; Müller et al. 2006; Worberg et al. 2007). Successful use of noncoding DNA sequences for deep-level phylogenetics

requires careful consideration of alignment, data quality, and model adequacy because the sequences are likely to be highly variable in both length and localized hotspots of substitution.

Alignment is often performed in the context of secondary structure. Localized regions of unalignable length variation, usually in loops and interhelical sequence, are removed (e.g., Borsch et al. 2003; Quandt et al. 2004; Löhne and Borsch 2005; Neinhuis et al. 2005; Wanke et al. 2007). Secondary structures provide the biological information needed for positional homology assessment in introns and improve the chance that signal derived from the data set will reflect evolutionary history. However, non-coding sequence that evolves under structural constraints can limit the distribution and type of mutations that persist, increasing the chance of saturation in stem sites and greatly skewing the transition–transversion ratio towards transitions (Kelchner 2002; Quandt and Stech 2005). This might cause problems for phylogeny estimation if an inadequate model is used, or if model selection is poorly conducted (i.e., bias in the selection process, or an inappropriate pool of candidate models).

Model selection should be more precise with highly variable data than with low-variability data, since there are more observed changes in the sample as well as better resolution in the estimated topology used for likelihood ratio testing of candidate models. However, model selection is always limited by the pool of models that are included, so it is still important to recognize that one or more models selected for a study will only be the best-fit of those that were considered. If a key feature of character evolution has not been adequately addressed by any of the models, error in the estimation process could still result.

Support measures such as bootstrap values should be greater for highly variable data because resampling in bootstrap replicates is more likely to recover informative sites. If the values are not higher, then one should investigate whether poor homology assessment, saturation, data conflict, rapid radiation, or model inadequacy could be causing a lack of confidence in the apparent signal (e.g., Rodríguez-Ezpeleta et al. 2007). A robust topology analysis might also shed light on the potential causes for a tree with low bootstrap support: if it is found that altering the character evolution model also alters the phylogeny estimation, then there is an issue with either character conflict in the data or improper assumptions being made about character evolution. Conducting SH or AU tests (Shimodaira and Hasegawa 1999; Shimodaira 2002) on topologies that have alternative placements of the poorly supported branch could reveal that insufficient signal exists to resolve the branch's position (e.g., Buckley et al. 2001b).

## Issue 7: small data sets

A potential problem that is largely undiscussed with non-coding regions, particularly intergenic spacers from organellar genomes and spliceosomal introns from the nucleus, is that character matrices are often relatively small (usually less than 800 nucleotides per sequence). In some cases, researchers might be seeking a simple, quick, and inexpensive source of genetic characters at low taxonomic levels, and therefore require only a single noncoding locus to determine haplotypes for a network analysis. In other cases, sequences from a single noncoding region will be used as the primary molecular data set in a phylogenetic analysis that also includes scored morphological characters. More frequently, two short noncoding loci from different genomic compartments (e.g., a chloroplast spacer and a nuclear ITS region) will be used to compare phylogeny estimations from each genome, and apparent incongruence will be used to infer hybridization, gene transfer events, or failed coalescence.

Small character sets create significant issues for model selection and model use in biology (Burnham and Anderson 2002), including phylogenetic inference. In terms of model selection, small data sets as well as larger, less variable data sets can both have small 'effective sample sizes', because both are limited in the number of observable changes that can be used for determining the adequacy of a model. If we envision model selection as a process that efficiently detects patterns in a data set and then chooses key parameters to cope with those patterns, then the assessment of whether a parameter is needed for improving the likelihood of a model is hampered when there are not enough observations available for the pattern to become evident. If little sequence variation is also a feature of the data, the problem of fitting a model to small data sets should be exacerbated because even less pattern will be detectable in short stretches of nearly invariable sequence. The issue is a classic statistical one of making estimates from a small sample size.

Simulations that use complex forms of character evolution and problematic trees often require a large number of nucleotide sites to recover a correct phylogeny 100% of the time, even when the same model is used for estimation that was used for generating the data. Sullivan and Swofford (2001) found, under the conditions of their simulation study, that long branch attraction and repulsion in their four-taxon and eight-taxon trees were most quickly overcome by a combination of an adequate model and at least 5,000–10,000 nucleotide characters with sufficient variation. Holland et al. (2003) investigated cases in which a large number of characters were required to overcome sampling error when a long branch is rooted to a short internode. Swofford et al. (2001) showed the effect of data

set size on the ability of maximum likelihood to correctly resolve a long branch ‘repulsion’ problem (the inverse-Felsenstein zone): under their simulation conditions, the proportion of correctly estimated trees had only reached ~85% when using their largest data sets of 100,000 nucleotides, although it was clearly trending toward complete accuracy if more data were available. Even though the primary goal of each of these studies was not to investigate the minimum amount of data needed in model-based phylogenetics, together they illustrate the potential for stochastic error to mislead a phylogenetic analysis of smaller data sets.

Avoiding sampling error in the statistical estimation of phylogenies, then, requires large effective sample sizes. The trouble is, how does one properly increase the effective size of a target noncoding locus? One could try choosing very large targets, but it is rare to find a non-repeating, contiguous stretch of organellar noncoding sequence that is longer than 2,000 nucleotides. A popular alternative is to combine data from multiple noncoding loci, a procedure that could readily assemble a data set of sites that evolve under strongly heterogeneous processes of character evolution. Although this can violate the IID assumption and add another level of complexity to the model selection process (Issue 8), it is not clear whether strong heterogeneity causes significant problems in most phylogeny estimations.

The likelihood that each locus evolved under a similar set of selective constraints is probably low (Bull et al. 1993), even when there is no significant incongruence detected among loci. One possible exception is group II intron sequences, which maintain a shared and conserved secondary structure that is essential for splicing reactions. Kelchner (2002) proposed that group II introns offer the best chance for combining multiple data sets evolving under the same general process, particularly when the combined data are then treated as character partitions that correspond with the shared structural features of these molecules. On the whole, however, we still lack sufficient information about the selective constraints operating in most intergenic spacers and some introns to make useful categories for process partitioning and subsequent partitioned model analysis. It is at present difficult to determine whether combining certain sequence regions will appropriately boost sample size without creating hazards for model selection and the estimation process.

#### Issue 8: heterogeneous data and partitioning

Although a single noncoding locus may occasionally show enough character variation to diagnose clades within the study group, it is unusual to find in the literature examples of a single noncoding locus providing a fully resolved and

supported phylogeny estimation (unless the divergences being estimated occurred a very long time ago: e.g., Borsch et al. 2003; Worberg et al. 2007). A common practice among molecular systematists has been to combine sequences from different noncoding loci to increase the number of variable characters in an analysis and achieve more resolution in the phylogeny estimation. Combining noncoding data was not an issue in the literature when the community still shared an assumption that all noncoding characters evolved in a neutral fashion, both randomly and independently. These regions, however, are now recognized as highly structured and functional sequence in many cases. Previously straightforward data combination (i.e., all noncoding DNA can be assembled into a single data block) has given way to partitioning strategies for grouping sequence segments that are expected to share similar selective constraints.

Partitioning is usually conducted in one of two ways: either the individual loci are recognized as separate process partitions (e.g., intergenic spacer vs. intron), or features of secondary structure within and among noncoding data sets are used as broad categories of related character evolution (e.g., pairing nucleotides in a helix vs. unpairing nucleotides in a loop). Little is known about most intergenic spacers and their specific functions in regulation, transcription, and cellular processes. The nuclear internal transcribed spacers of ribosomal genes are better understood because their secondary structure is highly conserved for reasons related to transcription and translation (Baldwin et al. 1995; Alvarez and Wendel 2003). However, we are still investigating the structure and function of many widely used intergenic spacers in organellar molecular systematics (e.g., Quandt and Stech 2004; Won and Renner 2005), so for now character partitions are usually made between different loci because detailed knowledge of intralocus heterogeneity is currently limited.

Group I and group II introns are probably the best understood noncoding regions in terms of structure, function, and evolution. For this reason, sequence data from these introns is usually much easier to partition in the context of conserved secondary structure (Kelchner 2002; Quandt and Stech 2005). Although character partitioning can sometimes correspond to intron domain helices, it is more reasonable to employ two simple partition categories: pairing and nonpairing nucleotides. Kelchner (2002) found that finer-scale partitioning of the nonpairing category in Myoporaceae *rp16* intron sequences (where nonpairing sites were further categorized as bulge, loop, and interhelix) showed less difference in mutation patterns among nonpairing categories than did patterns between the more general nonpairing and pairing partitions. This suggested that just two character partitions (pairing and nonpairing) were reasonable subsets for the *rp16* intron at that

taxonomic level. The question of whether a pairing/non-pairing partitioning scheme can be uniformly applied across group II intron data sets has not been sufficiently explored.

For model use in phylogenetics, heterogeneous data is challenging to both the accuracy of model selection and the adequacy of a chosen model for phylogeny estimation (Wilgenbusch and de Queiroz 2000; Nylander et al. 2004; Lemmon and Moriarty 2004; Brown and Lemmon 2007). There is a concern that the application of a single model of character evolution to heterogeneous data might result in the selection of an inadequate model for most characters in the data set (e.g., Bull et al. 1993), or produce poor estimates of parameters in the model (e.g., Sullivan et al. 1995; Yang 1996; Kelchner 2002). The advent of partitioned models, currently available in a Bayesian framework, tries to address these concerns by allowing each defined character partition to be fit independently for a model. The partitioned model can therefore be a combination of several models, each one specifically selected for a certain subset of data.

What remains to be determined is whether partitioned models are in general a significant improvement over using a single complex model across a combined, heterogeneous data set. One preliminary study (S.A. Kelchner, J. Wilgenbusch, and D.L. Swofford, abstract 126 of the Botany 2004 meetings at Snowbird, Utah) found little difference among the two approaches in their ability to recover accurate topologies for simulated multilocus data sets of various sizes, although this maximum likelihood experiment needs to be repeated with a more in-depth analysis of branch lengths and tree topologies. Others have found that a partitioned gamma parameter provides improvements for model fit of combined data in the likelihood framework (e.g., Pupko et al. 2002). Similarly, simulation studies (e.g., Brown and Lemmon 2007) and empirical studies (e.g., Nylander et al. 2004) using Bayesian inference have found model partitioning to improve both the rate of convergence and parameter precision for combined data sets.

The relative importance of partitioning the model may be linked to data set size (see also Pupko et al. 2002). Character partitioning reduces the effective size of a data set for which model selection and phylogenetic analysis will take place. By attempting to fit a model to subsets of the data, we are in effect dismantling a larger, combined data set into several smaller segments, each of which must be large enough and variable enough to support quality model fitting and analysis. Hence, partitioning can mean that even though we have assembled a large data set of combined loci, we might still experience some of the limitations for model use that we find in a small data set. If the statistical variance associated with model selection and subsequent parameter estimation for each of these small data sets becomes too much, we would expect a tradeoff

between improving overall model fit to the data and losing precision in the estimation process. At its worst, partitioned models could be a case of overfitting the model to the data, particularly in highly partitioned data sets, although this effect has not yet been observed at problematic levels in Bayesian simulations (Lemmon and Moriarty 2004).

If we accept the view that the process of reconstructing a phylogeny is actually a model-based attempt to estimate evolutionary history, then what should primarily concern us is the adequacy of a model, and not a fine-tuned exactness, for the data at hand. This might mean that a single, parameter-rich model will be sufficient to address the heterogeneity in some combined data sets, whereas partitioning will be essential in other cases to cover strong differences in character evolution among, or within, loci. Because data set size, degree of variability among sequences, and the strength of heterogeneity in the data will all play a role in appropriate model selection and subsequent estimation, it is likely that the adequacy of a model will have to be assessed in a case-by-case basis, at least within the maximum likelihood framework.

Finally, although some might argue that partitioning based on dissimilar evolutionary constraints is a ‘slippery slope’ that could ultimately lead to each nucleotide receiving its own character evolution model, I suggest that such an outcome makes little sense in a phylogenetic inference framework. In the context of the discussion on model fit, one can envision a case where separate models for each nucleotide would lead to a significant problem of overparameterization, and cases of overfitting for such models have been observed with empirical data (Cunningham et al. 1998; Buckley and Cunningham 2002). Fine-scaled partitioning greatly reduces the number of observations available in each data partition, making model selection far more difficult because of the small sample sizes. Small data partitions could also lead to unmanageable levels of variance associated with the parameter estimates. I suggest that the solution, as with determining parameter inclusion in model selection, might rest on producing an adequate partitioning scheme that represents the major categories of selective constraint without reducing each partition size to a level that makes sampling error a strong force in the model selection and phylogeny estimation processes. Toward this end, a mixture model (Pagel and Meade 2004) may prove useful within the Bayesian framework, which determines heterogeneity classes of characters during the analysis and fits a rate matrix to each estimated partition.

#### Issue 9: limits of current models

Each of our current phylogenetic models is unlikely to be the ‘true model’ representing the evolution of characters in

a data set (Posada and Buckley 2004). A ‘true model’ is not necessary if an adequate model is available, but there is reason for concern that our best-fit models may not be adequate models in many of our studies. It is possible that our current models are limited in important ways for phylogeny estimation of sequence data in general, and noncoding data in particular. Here, a few phenomena are discussed that may not yet be adequately covered.

Heterotachy is a growing concern in molecular systematics (Philippe and Lopez 2001; Lopez et al. 2002; Lockhart et al. 2006). The term heterotachy refers to a shift through time of selective constraints on sites in a DNA sequence. Applying a single model to such data would therefore inadequately represent the changes in character evolution that may have occurred at those sites across the sample. Possible solutions for heterotachy might be (1) creating a new parameter with a distribution that matches the pattern created by heterotachy in DNA sequences, or (2) accommodating that pattern, in certain cases, with an available model (Galtier 2001) or parameter (e.g., possibly  $\alpha$ ; Penny et al. 2001).

Heterotachy is usually discussed in the context of protein evolution and it is not clear whether the effect would be as pronounced in noncoding DNA. For example, it seems unlikely that a group II intron would experience a significant shift through time in selective constraints on its principle structural domains, which are shared across all complete group II introns and are necessary for proper intron splicing. However, the loops in these domains can be quite extensive, and certain tertiary interactions involving these loops may not be as strongly conserved. Hence, loop regions could conceivably undergo shifts in selective constraints through time, as might secondary structures in intergenic spacers. Our knowledge of how heterotachy might affect noncoding regions is still quite poor.

Nonindependence of mutations is another concern. Models that are widely used in molecular systematics expect independence of nucleotide characters, which is unlikely to be the case in noncoding DNA (Issue 2). Some specific parameters and models have been proposed to compensate for certain forms of nonindependence, and they might rationally be applied here. These include the  $\lambda$  parameter for nucleotide pairing in secondary structures (Muse 1995), the ‘doublet model’ in MrBayes for ‘auto-correlated’ sites in stems (based on Schöniger and von Haeseler 1994), models for Markov-dependent rates at adjacent sites (Yang 1995), and perhaps even a compound Poisson process model (Huelsenbeck and Nielsen 1999), which was first created to investigate standard model performance when site mutations are temporally correlated.

Despite a general worry that models are becoming over-parameterized, it may be that our parameter-rich models are still in need of further development for most nucleic

acid sequences (Sanderson and Kim 2000). Kelchner and Thomas (2007) surveyed model selection in published literature and revealed that ModelTest (Posada and Crandall 1998) most frequently selects parameter-rich models, such as GTR + I +  $\Gamma$ , for every class of sequence data reported. If this is not a case of model selection bias (Issue 10), then it suggests that even the most complex models available in ModelTest cannot be trimmed of parameters for most data sets. Our models might therefore require further parameterization to cover all significant processes in nucleotide character evolution. A cautious exploration of new model parameters is probably warranted, particularly with larger and more variable noncoding data sets such as those that are being used for estimating relationships across angiosperm lineages.

Indel evolution is at present a poorly modeled phenomenon. Improvements are desirable that will efficiently incorporate microstructural characters into a probabilistic framework and facilitate their coestimation with nucleotide substitutions. Currently, the ‘standard discrete’ model can be used for microstructural character partitions in the program MrBayes, which treats the character data in a manner akin to a multistate Jukes–Cantor model. Otherwise, a parsimony approach in PAUP (Swofford 1998) is commonly used, which allows coestimation of indel and nucleotide characters only within the parsimony framework. Because the parsimony model may sometimes be inadequate for both indel and nucleotide characters, it is inconvenient that a probabilistic approach has not yet been sufficiently developed for the inclusion of microstructural changes in phylogeny estimation.

Finally, many of the issues discussed above are specific to data sets with a large enough sample size and sequence variation to make systematic error a potential problem. It is important to remember that model adequacy for nucleotide and indel data may not often be an issue in cases where data sets are small or show little sequence variation. Determining whether a new parameter is required or not when selecting a model is largely an irrelevant point if even the recognition of haplotypes is difficult in a particular sequence alignment. The limitations of candidate models for noncoding DNA, however, still need to be investigated in most studies because the chance of error is always present.

#### Issue 10: model selection and model selection bias

Model selection is an essential component of model-based phylogenetics and cannot be avoided if a statistical approach to phylogeny estimation is desired. Methods of model selection vary considerably and are limited in their effectiveness by the availability of adequate models in the selection process. With empirical data, e.g., it is unlikely

that any model we currently select is going to be the ‘true model’, so our candidate pool of models in the selection process is already misspecified (Sanderson and Kim 2000; Posada and Buckley 2004). This is a limitation to model use that cannot be overcome at the present time and is perhaps an inevitable consequence of estimating phylogeny with models of character evolution.

Kelchner and Thomas (2007) recognize four general approaches to model selection: (1) frequentist, such as likelihood ratio testing, (2) information-theoretic, such as Akaike information criterion (AIC), (3) performance-based, which includes decision-theory methods, and (4) a prior acceptance of a model. Methods 1–3 are statistical in nature; method 4 is often philosophical. Posada and Buckley (2004) and Sullivan and Joyce (2005) provide detailed introductions to the topic of model selection and its techniques, so it will not be reviewed here other than to discuss a few important issues that warrant greater attention.

When a model is selected for a given data set, it is only the best-fit of those models that were considered (in other words, those in the pool of candidate models). A total of 56 models are included in the candidate pool in ModelTest, the most widely used program for likelihood ratio testing of phylogenetic models (Posada and Crandall 1998). These nested models range from JC (zero free parameters; Jukes and Cantor 1969) to GTR + I +  $\Gamma$  (ten free parameters; based on Rodríguez et al. 1990), but they represent only a few of the more than 200 possible nested models of GTR alone. When ModelTest uses likelihood ratio testing, the nesting of candidate models is essential, which means that non-nested models such as covarion or parsimony cannot be included in the selection process.

ModelTest has a predilection to select parameter-rich models for most sequence data, independent of organism, locus, or genome compartment used in the study (Kelchner and Thomas 2007). As discussed above, this might indicate a general need for more parameters in our models. It may also indicate, however, that a bias exists in the way ModelTest chooses a best-fit model. ModelTest uses a hierarchical likelihood ratio test, which means that models are selected through a series of predetermined comparisons that progress from simple to complex models (models that have no free parameters to models that are parameter-rich). Such comparisons in regression statistics can create bias in the determination of parameter importance (Sanderson and Kim 2000; Burnham and Anderson 2002; Pol 2004), sometimes leading to the inclusion of extraneous parameters that have little to do with model adequacy. Whether ModelTest, as currently implemented, is suffering from such a bias is unknown, although some studies have shown that model selection by likelihood ratio tests depends on the order in which the comparisons are made (Cunningham

et al. 1998; Pol 2004). The relatively simple expedient of using fewer candidate models and dynamical model comparisons should reduce the chance of bias when selecting models by likelihood ratio tests (Kelchner and Thomas 2007).

The Akaike information criterion (AIC) is an information-theoretic approach that seems to be the second most popular means of choosing a phylogenetic model in current literature. Unlike the hierarchical likelihood ratio test implemented in ModelTest, information-theoretic approaches do not require nested models and do not use hypothesis testing to determine if certain parameters should be included. Instead, AIC requires that informed decisions be made about model adequacy based on the relative scores of the candidate models. Although it is frequently reported that the model with the best AIC score (‘highest ranking’ model) is chosen, this is in fact an incorrect use of the AIC approach. What matters more is how distant the next best model is in terms of information content, which is assessed by comparing the relative difference in AIC scores among the models (their  $\Delta$  values). Burnham and Anderson (2002) give a ‘rule of thumb’ for cases when sample size is large and many reasonable candidate models are available: all models with AIC differences  $\leq 2$  should be considered as having ‘substantial support’ as best-fit models for the data (see also Posada and Buckley 2004). In many cases, this will mean that more than one model will be a reasonable choice as the AIC “best-fit”. Whether parameter estimates (including tree topology) change across these best-fit models is worthy of investigation in a manner similar to that used in determining a robust topology. If the topology or another parameter of interest changes across these models, then multimodel inference (when parameter estimates are made across multiple adequate models) is a reasonable solution to accommodate model uncertainty in the phylogeny estimation (Posada and Buckley 2004; Kelchner and Thomas 2007).

The mention of multimodel inference brings us to the issue of model selection uncertainty in general, and the largely unrecognized potential for bias to influence the model selection process in phylogenetics. Briefly, in an otherwise rigorous system of model use, the assessment and incorporation of model selection uncertainty into phylogeny estimation is lagging. We could routinely be measuring how much uncertainty is involved in the selection of an adequate model for a given data set because this is an important component of the overall uncertainty associated with any phylogeny estimation (Sullivan and Joyce 2005). Goodness of fit measures can be used in a likelihood framework for assessing the quality of model choice for the data (e.g., Goldman 1993; Whelan et al. 2001; Sullivan et al. 2000). In the AIC framework, the mathematics is already in place for the incorporation of

model selection uncertainty into the overall estimation (Burnham and Anderson 2002; Posada and Buckley 2004). However, as with many methods in model-based phylogenetics, the incorporation of model selection uncertainty will probably be delayed until software that facilitates its application is both readily available and commonly used.

## Recommendations

**Homology.** Model use for phylogeny estimation requires that the sample (the sequence alignment) is valid for estimating the population (the condition of the sequences through time). If homology statements are generally inaccurate, the resulting topology will not represent evolutionary history of the sample sequences. The previously held expectation that organellar noncoding regions are effectively single-copy in an organism is no longer a safe supposition. Attempts should be made to verify single-copy status for target loci prior to phylogeny estimation, and to identify and correct probable cases of recombination. Sequence reads (e.g., '.abi' or '.scf' trace files) should be checked for multiple signal. BLAST searches can assist with finding potential homologs and paralogs that are in the GenBank databases and help identify the origin of sequence segments that deviate considerably from those in the alignment.

**Hand alignment (for now).** Positional homology of nucleotides among sequences is crucial for proper statistical inference of evolutionary history. Careful assessment of positional homology for sites in length-variable regions is especially important when indels are to be used as coded characters for phylogeny estimation. Relying on current software to position gaps is allowing homology assessment to be determined by an algorithm that is incapable of recognizing structural constraints and mutation mechanisms in most noncoding sequences. Careful attention to biologically relevant information during hand alignment should help in identifying probable homology of microstructural characters as well as assist in determining which length-variable regions must be excluded due to lack of sufficient evidence of homology.

**Indel characters.** Although it is unlikely that current indel character models are adequate representations of microstructural change probabilities, the generally high information content of hand-aligned indel characters argues for their continued inclusion in phylogenetic analyses of noncoding sequence data. Indel character coding should only take place after biologically reasonable homology statements have been made for inferred microstructural changes, since most indel coding methods assume homology of aligned length variants. Careful attention to homology assessment, then, should result in

higher quality indel characters and increased power of coded indels in the phylogeny estimation.

**Partitioning.** Most noncoding loci are relative short in sequence length and can be considered heterogeneous assemblages of characters evolving under different selective constraints. Both the combination of multiple noncoding loci into a single data set and the partitioning of a single locus into subsets of characters can cause problems for model use. In the first case, the sample size is larger but the candidate pool of models may not contain an adequate model for the heterogeneous data. In the second case, overpartitioning of the data will greatly reduce the size of each sample to which a model is being fit and is likely to decrease the precision of model selection for each partition. Perhaps the goal with heterogeneous data, like that of model selection itself, should be to select adequate partitions that reflect major categories of selective constraints. Representing intron data as pairing versus nonpairing nucleotide sites, for example, might maximize partition sample sizes (which reduces sampling error) while compensating for two considerably different sets of selective constraints (which reduces the chance of bias in the model selection process). How this would be done in a rigorous framework is unclear. One available solution might be mixture models, which are intended for discovering character partitions during a Bayesian analysis, although the performance of these models is still largely untested.

**Caution when choosing a model.** The process of model selection is also subject to bias and error. Likelihood ratio testing would be better conducted in a dynamical fashion, with a smaller pool of reasonable candidate models. AIC selection should consider all models that have substantial support: that is an AIC difference of  $\Delta \leq 2$  (recommended by Posada and Buckley 2004), or more conservatively my recommendation of  $\Delta \leq 4$  for small data sets. The quality of the phylogeny estimation is dependent on the adequacy of models in the candidate pool, and if we have doubt about the adequacy of our current models for noncoding data then we should also have lower confidence in our phylogeny estimations.

**Corroboration.** The best evidence that a topology represents phylogeny is corroboration of the topology from multiple data sources, even when the data sets are limited in size (Miyamoto and Fitch 1995). If independent corroboration does not exist for a phylogeny estimation, then we are forced to rely on confidence that our model is adequate for the data and that error and bias did not adversely influence the process. Exploratory analyses involving multiple partitioning schemes, robust topology techniques, bootstrapping, and testing of alternative hypotheses are the best that can be done at present to increase the confidence of a phylogeny estimation when corroboration is unavailable. However, corroboration should be sought out when

the consequences of a phylogeny estimation are substantial (e.g., a taxonomic reclassification, or a conservation decision).

**Acknowledgments** Financial support during the development of this article was provided by National Science Foundation award DEB-0515828 to S.A. Kelchner. The author thanks Sean Graham for his careful comments and prompting during the formation of the manuscript, Amanda Fisher and Chang Liu for their fine-tuning of the text, an anonymous reviewer for insightful suggestions that clarified several of the issues discussed, and Thomas Borsch, Dietmar Quandt and the German Science Foundation for their kind invitation to present this topic at the 17th International Symposium on Biodiversity and Evolutionary Biology in Bonn, Germany (September 2006).

## References

- Alvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molec Phylogenet Evol* 29:417–434
- Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ (1995) The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Ann Missouri Bot Gard* 82:247–277
- Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C, Barthlott W (2003) Noncoding plastid *trnT-trnF* sequences reveal a well-resolved phylogeny of basal angiosperms. *J Evol Biol* 16:558–576
- Borsch T, Hilu KW, Wiersema JH, Löhne C, Barthlott W, Wilde V (2007) Phylogeny of *Nymphaea* (Nymphaeaceae): evidence from substitutions and microstructural changes in the chloroplast *trnT-trnF* region. *Int J Pl Sci* 168:639–671
- Brinkmann H, Van der Giezen M, Zhou Y, de Raucourt GP, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* 54:743–757
- Brown JM, Lemmon AR (2007) The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Syst Biol* 56:643–655
- Buckley TR (2002) Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst Biol* 51:509–523
- Buckley TR, Cunningham CW (2002) The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Molec Biol Evol* 19:394–405
- Buckley TR, Simon C, Chambers GK (2001a) 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
- Buckley TR, Simon C, Shimodaira H, Chambers GK (2001b) Evaluating hypotheses on the origin and evolution of the New Zealand alpine cicadas (Maoricicada) using multiple-comparison tests of tree topology. *Molec Biol Evol* 18:223–234
- Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ (1993) Partitioning and combining data in phylogenetic analysis. *Syst Biol* 42:384–397
- Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York
- Collins TM, Wimberger PH, Naylor GJP (1994) Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst Biol* 43:482–496
- Cummings MP, Nugent JM, Olmstead RG, Palmer JD (2003) Phylogenetic analysis reveals five independent transfers of the chloroplast gene *rbcL* to the mitochondrial genome in angiosperms. *Curr Genet* 43:131–138
- Cunningham CW, Zhu H, Hillis DM (1998) Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* 52:978–987
- de Pamphilis CW, Palmer JD (1990) Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. *Nature* 348:337–339
- Dixon MT, Hillis DM (1993) Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Molec Biol Evol* 10:256–267
- Drábková L, Kirschner J, Vlcek C, Páček V (2004) *TrnL-trnF* intergenic spacer and *trnL* intron define major clades within *Luzula* and *Juncus* (Juncaceae): importance of structural mutations. *J Molec Evol* 59:1–10
- Dumolin-Lapègue S, Pemonge M-H, Petit RJ (1998) Association between chloroplast and mitochondrial lineages in oaks. *Molec Biol Evol* 15:1321–1331
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791
- Felsenstein J (2004) Inferring phylogenies. Sinauer, Sunderland
- Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Zool* 20:406–416
- Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol* 18:866–873
- Goldman N (1993) Statistical tests of models of DNA substitution. *J Molec Evol* 36:182–198
- Golenberg EM, Clegg MT, Durbin ML, Doebley J, Ma DP (1993) Evolution of a non-coding region of the chloroplast genome. *Molec Phylogenet Evol* 2:52–64
- Graham SW, Olmstead RG (2000) Evolutionary significance of an unusual chloroplast DNA inversion found in two basal angiosperm lineages. *Curr Genet* 37:183–188
- Graham SW, Reeves PA, Burns ACE, Olmstead RG (2000) Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *Int J Pl Sci* 161:S83–S96
- Hamilton MB, Braverman JM, Soria-Hernanz DF (2003) Patterns and relative rates of nucleotide and insertion/deletion evolution at six chloroplast intergenic regions in the New World species of Lecythidaceae. *Molec Biol Evol* 20:1710–1721
- Holland BR, Penny D, Henny MD (2003) Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—a simulation study. *Syst Biol* 52:229–238
- Huelsenbeck JP, Neilsen R (1999) Effect of nonindependent substitution on phylogenetic accuracy. *Syst Biol* 48:317–328
- Huelsenbeck JP, Rannala B (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276:227–232
- Huelsenbeck JP, Rannala B (2004) Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst Biol* 53:904–913
- Ingvarsson PK, Ribstein S, Taylor DR (2003) Molecular evolution of insertions and deletion in the chloroplast genome of *Silene*. *Molec Biol Evol* 20:1737–1740
- Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123
- Kelchner SA (2000) The evolution of noncoding chloroplast DNA and its application in plant systematics. *Ann Missouri Bot Gard* 87:482–498
- Kelchner SA (2002) Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *Amer J Bot* 89:1651–1669

- Kelchner SA (2003) Phylogenetic structure, biogeography, and evolution of Myoporaceae. The Australian National University, Canberra
- Kelchner SA, Clark LG (1997) Molecular evolution and phylogenetic utility of the chloroplast *rpl16* intron in *Chusquea* and the Bambusoideae (Poaceae). *Molec Phylogenet Evol* 8:385–397
- Kelchner SA, Thomas MA (2007) Model use in phylogenetics: nine key questions. *Trends Ecol Evol* 22:87–94
- Kelchner SA, Wendel JF (1996) Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Curr Genet* 30:259–262
- Kim K-J, Lee H-L (2005) Widespread occurrence of small inversions in the chloroplast genomes of land plants. *Mol Cells* 19:104–113
- Knoop V (2004) The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr Genet* 46:123–139
- Koch MA, Dobes C, Kiefer C, Schmickl R, Klimes L, Lysak MA (2006) Supernetwork identifies multiple events of plastid *trnF(GAA)* pseudogene evolution in the Brassicaceae. *Molec Biol Evol* 24:63–73
- Kudla J, Albertazzi FJ, Blazevic D, Hermann M, Bock R (2002) Loss of the mitochondrial *cox2* intron 1 in a family of monocotyledonous plants and utilization of mitochondrial intron sequences for the construction of a nuclear intron. *Molec Genet Genomics* 267:223–230
- Laroche J, Bousquet J (1999) Evolution of the mitochondrial *rps3* intron in perennial and annual angiosperms and homology to *nad5* intron 1. *Molec Biol Evol* 16:441–452
- Lemmon AR, Moriarty EC (2004) The importance of proper model assumption in bayesian phylogenetics. *Syst Biol* 53:265–277
- Lewis PO (1998) Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. In: Soltis DE, Soltis PS, Doyle JJ (eds) *Molecular systematics of plants II: DNA sequencing*. Kluwer Academic Publishers, Boston, pp 132–163
- Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol* 50:913–925
- Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Molec Biol Evol* 11:605–612
- Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T (2006) Heterotachy and tree building: a case study with plastids and eubacteria. *Molec Biol Evol* 23:40–45
- Löhne C, Borsch T (2005) Molecular evolution and phylogenetic utility of the *petD* group II intron: a case study in basal angiosperms. *Molec Biol Evol* 22:317–3332
- Lopez P, Casane D, Philippe H (2002) Heterotachy, and important process in protein evolution. *Molec Biol Evol* 19:1–7
- Lutzoni F, Wagner P, Reeb V, Zoller S (2000) Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Syst Biol* 49:628–651
- Meimberg H, Thalhhammer S, Brachmann A, Heubl G (2006) Comparative analysis of a translocated copy of the *trnK* intron in carnivorous family Nepenthaceae. *Molec Phylogenet Evol* 39:478–490
- Mes THM, Kuperus P, Kirschner J, Stepanek J, Oosterveld P, Storchova H, den Nijs JCM (2000) Hairpins involving both inverted and direct repeats are associated with homoplasious indels in non-coding chloroplast DNA of *Taraxacum* (Lactuceae: Asteraceae). *Genome* 43:634–641
- Miyamoto MM, Fitch WM (1995) Testing species phylogenies and phylogenetic methods with congruence. *Syst Biol* 44:64–76
- Morrison DA (2006) Multiple sequence alignment for phylogenetic purposes. *Austral Syst Bot* 19:479–539
- Müller K (2006) Incorporating information from length-mutational events into phylogenetic analysis. *Molec Phylogenet Evol* 38:667–676
- Müller K, Borsch T (2005a) Phylogenetics of *Utricularia* (Lentibulariaceae) and molecular evolution of the *trnK* intron in a lineage with high substitutional rates. *Pl Syst Evol* 250:39–67
- Müller K, Borsch T (2005b) Phylogenetics of Amaranthaceae based on *matK/trnK* sequence data—evidence from parsimony, likelihood, and Bayesian analyses. *Ann Missouri Bot Gard* 92:66–102
- Müller K, Borsch T, Hilu KW (2006) Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK*, *trnT-F*, and *rbcL* in basal angiosperms. *Molec Phylogenet Evol* 41:99–117
- Muse SV (1995) Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics* 139:1429–1439
- Naylor GJP, Brown WM (1998) Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst Biol* 47:61–76
- Nei M, Kumar S (2000) *Molecular evolution and phylogenetics*. Oxford University Press, Oxford
- Neinhuis C, Wanke S, Hilu KW, Müller K, Borsch T (2005) Phylogeny of Aristolochiaceae based on parsimony, likelihood, and Bayesian analyses of *trnL-trnF* sequences. *Pl Syst Evol* 250:7–26
- Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL (2004) Bayesian phylogenetic analysis of combined data. *Syst Biol* 53:47–67
- Ogden TH, Rosenberg MS (2007) How should gaps be treated in parsimony? A comparison of approaches using simulation. *Molec Phylogenet Evol* 42:817–826
- Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol* 53:571–581
- Penny D, Hendy MD, Steel MA (1992) Progress with methods for constructing evolutionary trees. *Trends Ecol Evol* 7:73–79
- Penny D, McComish BJ, Charleston MA, Hendy MD (2001) Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J Molec Evol* 53:711–723
- Philippe H, Lopez P (2001) On the conservation of protein sequences in evolution. *Trends Biochem Sci* 26:414–416
- Phillips MJ, Delsuc F, Penny D (2004) Genome-scale phylogeny and the detection of systematic biases. *Molec Biol Evol* 21:1455–1458
- Pirie MD, Vargas MPB, Botermans M, Bakker FT, Chatrou LW (2007) Ancient paralogy in the cpDNA *trnL-F* region in Annonaceae: implications for plant molecular systematics. *Amer J Bot* 94:1003–1016
- Pol D (2004) Empirical problems of the hierarchical likelihood ratio test for model selection. *Syst Biol* 53:949–962
- Posada D, Buckley TR (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
- Posada D, Crandall KA (1998) ModelTest: testing the model of DNA substitution. *Bioinformatics* 14:817–818
- Pupko T, Huchon D, Cao Y, Okada N, Hasegawa M (2002) Combining multiple data sets in a likelihood analysis: which models are the best? *Molec Biol Evol* 19:2294–2307
- Quandt D, Stech M (2004) Molecular evolution of the *trnT(UGU)-trnF(GAA)* region in bryophytes. *Pl Biol* 6:545–554
- Quandt D, Stech M (2005) Molecular evolution of the *trnL-UAA* intron in bryophytes. *Molec Phylogenet Evol* 36:429–443
- Quandt D, Müller K, Huttunen S (2003) Characterization of the chloroplast DNA *psbT-H* region and the influence of dyad symmetrical elements on phylogenetic reconstructions. *Pl Biol* 5:400–410
- Quandt D, Müller K, Stech M, Frahm J-P, Frey W, Hilu KW, Borsch T (2004) Molecular evolution of the chloroplast *trnL-F* region in land plants. *Monogr Syst Bot Missouri Bot Gard* 98:13–37

- Rodríguez F, Oliver JL, Marín A, Medina JR (1990) The general stochastic model of nucleotide substitution. *J Theor Biol* 142:485–501
- Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H (2007) Detecting and overcoming systematic error in genome-scale phylogenies. *Syst Biol* 56:389–399
- Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574
- Sanderson MJ (1995) Objections to bootstrapping phylogenies: a critique. *Syst Biol* 44:299–320
- Sanderson MJ, Kim J (2000) Parametric phylogenetics? *Syst Biol* 49:817–829
- Sang T, Crawford DJ, Stuessy TF (1997) Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Amer J Bot* 84:1120–1136
- Schöniger M, von Haeseler A (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Molec Phylogenet Evol* 3:240–247
- Shahmuradov IA, Akbarova YY, Solovyev VV, Aliyev JA (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. *PL Molec Biol* 52:923–934
- Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492–508
- Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molec Biol Evol* 16:1114–1116
- Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. *Syst Biol* 49:369–381
- Steel M, Huson D, Lockhart PJ (2000) Invariable sites models and their use in phylogeny reconstruction. *Syst Biol* 49:225–232
- Stefanovic S, Rice D, Palmer JD (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol Biol* 4:35–54
- Sullivan J, Joyce P (2005) Model selection in phylogenetics. *Annual Rev Ecol Evol Syst* 36:445–466
- Sullivan J, Swofford DL (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
- Sullivan J, Holsinger KE, Simon C (1995) Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmodontine rodents. *Molec Biol Evol* 12:988–1001
- Sullivan J, Arellano E, Rogers DS (2000) Comparative phylogeography of Mesoamerican highland rodents: concerted versus independent response to climatic fluctuations. *Amer Naturalist* 155:755–768
- Swofford DL (1998) PAUP\*. Phylogenetic analysis using parsimony (\* and other methods). Sinauer, Sunderland
- Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable B (eds) *Molecular systematics*. Sinauer, Sunderland, pp 407–514
- Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol* 50:525–539
- Tillier ERM, Collins RA (1995) Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Molec Biol Evol* 12:7–15
- van Ham RCHJ, t'Hart H, Mes THM, Sandbrink JM (1994) Molecular evolution of non-coding regions of the chloroplast genome in the Crassulaceae and related species. *Curr Genet* 25:558–566
- Vijverberg K, Bachmann K (1999) Molecular evolution of a tandemly repeated *trnF* (GAA) gene in the chloroplast genome of *Microseris* (Asteraceae) and the use of structural mutations in phylogenetic analysis. *Molec Biol Evol* 16:1329–1340
- Waddell PJ (1995) Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood. Massey University, Palmerston North, New Zealand
- Wakeley J (1994) Substitution-rate variation among sites and the estimation of transition bias. *Molec Biol Evol* 11:436–442
- Wakeley J (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol Evol* 11:158–163
- Wanke S, Jaramillo MA, Borsch T, Samain M-S, Quandt D, Neinhuis C (2007) Evolution of Piperales—*matK* gene and *trnK* intron sequence data reveal lineage specific resolution contrast. *Molec Phylogenet Evol* 42:477–497
- Whelan S, Lio P, Goldman N (2001) Molecular phylogenetics: state of the art methods for looking into the past. *Trends Genet* 17:262–272
- Wilgenbusch J, 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
- Wolfe AD, Randle CP (2004) Recombination, heteroplasmy, haplotype polymorphism, and paralogy in plastid genes: implications for plant molecular systematics. *Syst Bot* 29:1011–1020
- Won H, Renner SS (2003) Horizontal gene transfer from flowering plants to *Gnetum*. *Proc Natl Acad Sci USA* 100:10824–10829
- Won H, Renner SS (2005) The chloroplast *trnT-trnF* region in the seed plant lineage Gnetales. *J Molec Evol* 61:425–436
- Worberg A, Quandt D, Barniske A-M, Löhne C, Hilu KW, Borsch T (2007) Phylogeny of basal eudicots: insights from non-coding and rapidly evolving DNA. *Organisms Diver Evol* 7:55–77
- Yang Z (1995) A space-time process model for the evolution of DNA sequences. *Genetics* 139:993–1005
- Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 11:367–372