Syst. Biol. 60(1):60–73, 2011
© The Author(s) 2010. Published by Oxford University Press on behalf of Society of Systematic Biologists.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
DOI:10.1093/sysbio/syq074
Advance Access publication on November 15, 2010

Quantifying the Impact of Dependent Evolution among Sites in Phylogenetic Inference

CHRIS A. NASRALLAH[1,*], DAVID H. MATHEWS[2], AND JOHN P. HUELSENBECK[1]

[1]*Department of Integrative Biology, University of California, Berkeley, 3060 Valley Life Sciences Building #3140, Berkeley, CA 94720-3140, USA; and*
[2]*Department of Biochemistry and Biophysics, University of Rochester Medical Center, 601 Elmwood Avenue, Box 712, Rochester, NY 14642, USA;*
**Correspondence to be sent to: Department of Integrative Biology, University of California, Berkeley, 3060 VLSB #3140, Berkeley, CA 94720-3140, USA; E-mail: nasrallah@berkeley.edu.*

Received 5 August 2009; reviews returned 4 January 2010; accepted 15 September 2010
Associate Editor: Marc Suchard

Abstract.—Nearly all commonly used methods of phylogenetic inference assume that characters in an alignment evolve independently of one another. This assumption is attractive for simplicity and computational tractability but is not biologically reasonable for RNAs and proteins that have secondary and tertiary structures. Here, we simulate RNA and protein-coding DNA sequence data under a general model of dependence in order to assess the robustness of traditional methods of phylogenetic inference to violation of the assumption of independence among sites. We find that the accuracy of independence-assuming methods is reduced by the dependence among sites; for proteins this reduction is relatively mild, but for RNA this reduction may be substantial. We introduce the concept of effective sequence length and its utility for considering information content in phylogenetics. [Continuous-time Markov model; distance methods; Independence; maximum likelihood; parsimony.]

One of the fundamental assumptions made by most methods of phylogenetic inference is that characters evolve independently. This is of course not the case in reality, and there has in recent years been an effort to develop models that more accurately reflect the various types of dependence among sites that have been observed in a biological context.

One kind of dependence is the correlation of rates of substitution at adjacent sites. Yang (1995) and Felsenstein and Churchill (1996) developed methods that allowed the rate of substitution at a given site to depend on the rates of substitution at neighboring sites. But it is important to note that in these models it is only the overall rate of substitution that is correlated among sites; substitutions under the model remain independent at different sites. We will focus on methods in which both the rate and types of changes observed at one nucleotide position are dependent upon the nucleotide observed at another position in the sequence.

An example of this kind of dependence, in which adjacent sites can influence not only the rate but also the types of substitutions that occur, is found in the triplet codon structure in protein-coding DNA. Certain substitutions may be less frequent at one site because a change at that site would alter the amino acid encoded by the three sites taken together. Muse and Gaut (1994) and Goldman and Yang (1994) developed codon-based methods to address these concerns, and Nielsen and Yang (1998) expressed the codon model in the form most commonly used today.

Dependence can also arise due to the secondary structure of RNA molecules. Particular attention has been paid to develop methods that address the pairing of nucleotides in RNA stem formations (Schöniger and von Haeseler 1994; Tillier 1994; Tillier and Collins 1995). Dependencies due to secondary structure are often more complicated than those at adjacent sites as the dependent positions may be quite far from each other in terms of sequence position. It should be noted that these models, like the codon models, are one-substitution-at-a-time models.

Codon models for protein-coding DNA and doublet models for RNA share in common a general approach for accounting for dependence: They expand the basic evolutionary unit in the model from the nucleotide to the triplet or to the doublet, respectively. Robinson et al. (2003) took this approach to its logical endpoint using the entire protein-coding DNA sequence as the unit of evolution. They considered dependencies resulting from amino acid interactions as well as those resulting from solvent accessibility, and in doing so they allowed the number of other sites on which a given site was dependent to vary across the sequence. Rodrigue et al. (2005) took a similar approach but using only the amino acid interactions, and Kleinman et al. (2006) showed that the model fit is much better when the solvent accessibility is included (see Anisimova and Kosiol 2009 for a review of several of these models of substitution).

Error in phylogenetic estimation due to dependent evolution has been detected in recent data set as well. Castoe et al. (2009) identified 13 mitochondrial protein-coding regions in squamates that they believe to be the result of strong nonneutral convergence. They argue that models of evolution that can account for convergence due to negative selection, such as those which consider the structure of a protein, might be useful for detecting similar cases that may otherwise strongly bias phylogenetic estimates.

Here we quantify how robust methods of phylogenetic inference are to violation of the assumption of independence. We use an evolutionary model similar to other sequence-based evolutionary models (Robinson et al. 2003; Rodrigue et al. 2005; Yu and Thorne 2006) to simulate sequence evolution under plausible dependent

constraints based on RNA and protein structures, and we evaluate the performance of traditional phylogenetic methods on these simulated data sets. We find that even small amounts of dependence in the data can lead to significant error in estimation of the true topology, and that this is especially true for RNA.

METHODS

General Strategy

We are interested in testing whether or not methods of phylogenetic inference that assume independent evolution at each site are robust to violation of that assumption. We are specifically interested in the ability to recover the correct tree topology rather than in accurately estimating branch lengths or other model parameters. The general strategy is as follows: 1) simulate an alignment under a known tree topology and set of branch lengths, with a known model of dependence; 2) estimate the tree from the simulated alignment using standard methods of phylogenetic inference, all of which assume independence of substitutions at different sites; and 3) assess the accuracy of the methods. The methods we will test are maximum likelihood (ML) using the general time-reversible model of substitution with gamma-distributed rate variation (GTR+Γ; Tavaré 1986; Yang 1993, 1994), neighbor-joining (Saitou and Nei 1987) using GTR+Γ distances, and parsimony as implemented in PAUP* 4.0b10x (Swofford 1998). We will be less interested in comparing these methods with each other than in examining the effect of dependence in the data on all of these methods.

The simplest case in which to study the effect of dependent evolution on phylogenetic inference is with the four-taxon tree and has been well-studied previously (Felsenstein 1978; Huelsenbeck and Hillis 1993). We largely focus on the four-taxon case in order to obtain a thorough understanding for how the tree length, tree topology, and varying levels of dependence affect inference. The tree we consider is shown in Figure 1. Two opposing terminal branches share a common length a, whereas the other two terminal branches and the internal branch share a common length b. The proportion a/b will be of great interest to us; when this quantity is larger the inference problem is increasingly difficult. We will also be interested in the total tree length (V), which allows the tree to be expanded or contracted while preserving the branch length proportions.

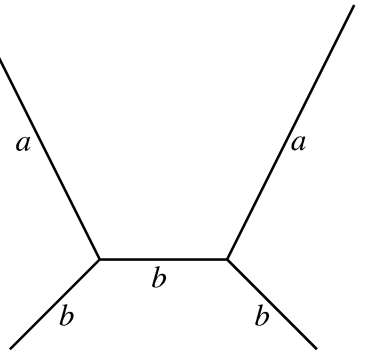

FIGURE 1. The four-taxon tree. The ratio of the branch lengths a/b and the total tree length V are the parameters of interest. As a/b becomes large, the inference problem becomes increasingly difficult.

Evolutionary Model

Calculation of the likelihood (or the parsimony score) of an alignment is greatly simplified by the independence assumption. If all sites are independent, then the probability of an alignment is simply the product of the probability of each column in the alignment (or the sum of the parsimony scores). To calculate this probability, the substitution process at a particular site is modeled as a continuous-time Markov chain. The process is governed by a rate matrix $\mathbf{Q} = \{q_{ij}\}$ where q_{ij} is the rate of change from state i to state j. This rate of change depends only on the current state i, and does not depend on what states may have been observed in the past (Markov property). Furthermore, the rate q_{ij} is agnostic to what is happening at every other site in the sequence. When dependence among sites is introduced it will not affect the Markov property, but the rate of change at a given site will depend on the state of the process at other sites.

The rate matrix $\mathbf{Q}$ can take many forms. For RNA-coding sequences, the matrix $\mathbf{Q}$ might be described by anything from the Jukes–Cantor model (Jukes and Cantor 1969) to the GTR model (Tavaré 1986), and does not in principle need to be time-reversible. For protein-coding sequences, $\mathbf{Q}$ could be described by various codon-based models (Muse and Gaut 1994; Goldman and Yang 1994) with different rates for synonomous/nonsynonomous sites as well as for transitions/transversions. These codon models typically restrict the possible changes from codon i to only those codons j that involve a single nucleotide substitution, and disallow stop codons.

Just as codon-based models expand the unit of evolution from the nucleotide to the codon, the model we consider further expands the unit of evolution from the nucleotide to the entire sequence. Consequently, we will be interested in sequence transition probabilities. More formally, we will consider a continuous-time Markov chain in which the state space is the set of all possible sequences of length N nucleotides. Let x and y be two such sequences. Then for all x and y, the matrix of rates of change from x to y can be defined as

$$\mathbf{R} = \{r_{xy}\} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ at 2 or more positions} \\ uq_{ij}E(x,y) & \text{if } x \text{ and } y \text{ differ at only 1 position} \\ -\sum_{x \neq y} r_{xy} & \text{if } x = y \end{cases}$$

where q_{ij} is, as described above, the rate of substitution under an independent-sites model for the position in the sequence that is changing, $E(x, y)$ compares the relative structural fitness of sequences x and y and is described in detail below, and u is a rate-scaling factor to ensure that branch lengths are interpretable in terms of average number of substitutions per site.

In the simulations of this paper, we specify the underlying $\{q_{ij}\}$ as follows. For RNA, we assume the K80 model of substitution (Kimura 1980) with the transition/transversion rate ratio $\kappa = 3$. For proteins, we assume a codon substitution model (Nielsen and Yang 1998) with transition/transversion rate ratio $\kappa = 1$, nonsynonomous/synonomous substitution rate ratio $\omega = 1$ and equal codon frequencies. These relatively simple substitution models were chosen to better examine the effects introduced by the dependencies due to structure, described below.

Energy of a Sequence

We will utilize a concept borrowed from structure prediction: a sequence folded into a particular structure will have a free energy associated with it. In structure prediction, the sequence is fixed and the structure of lowest energy is sought; here, we invert this problem by conditioning on the structure being fixed. We assume that all sequences share a common fixed structure that must be maintained to preserve functionality. It is this structure that determines the interactions among sites and their relative positions, and therefore determine their evolutionary interdependencies.

We will define the energy of a single sequence x as $E(x)$, but we find it useful to conceive of $E(x)$ not as an actual energy, rather as a measure of how well sequence x corresponds to the given structure, or as a kind of "structural fitness." If x could reasonably fold into the given structure, we expect $E(x)$ to be low, ideally negative. We can then calculate $E(y)$ for any sequence y as well. $E(x, y)$ then takes on the meaning of a comparison of the relative structural fitness of the two sequences. The precise form of $E(x, y)$ can in principle vary, and we will define $E(x, y)$ differently for RNA and for proteins.

For RNA, what we call energies are folding free energy changes (ΔG) predicted using the current nearest neighbor model of Turner and co-workers (Mathews et al. 2004). These free energy changes are predicted for a given base pairing structure using the efn2 model (Mathews et al. 1999). This approach utilizes information from both the base pairing and the coaxial stacking of nucleotides, allowing the potential to incorporate more information than a simple doublet model that considers doublets to be independent of each other. For RNA, we then define

$$E(x, y) = \mathrm{e}^{(E_z(x) - E_z(y))z}$$

where z is a free parameter determining the degree to which the difference in structural fitness affects the rate of substitution. Note that when $z = 0$, $E(x, y) = 1$ for all x and y, reducing the model to the independent-sites model specified by the single-site rate matrix $\mathbf{Q}$.

For proteins, we adopt the approach of Robinson et al. (2003) in simplifying the constraints governing the structure into two properties: energies due to pairwise interactions of amino acids and to solubility constraints (denoted $E_p(x)$ and $E_s(x)$, respectively). To do this, we utilize statistical potentials, which are pseudo-energy values associated with plausibilities of some aspect of the structure estimated from protein sequences of known structure. For pairwise interactions of amino acids, we can from the protein structure determine the relative positions of all amino acids in three-dimensional space, and declare two amino acids to be "in contact" if any of their non-hydrogen atoms are less than 4.5 Å apart (Bastolla et al. 2001). Pairs of amino acids whose three-dimensional proximity is due to sequential proximity (within three positions or less) are not considered to be in contact. Following Rodrigue et al. (2005), if our sequence is of length N we can describe a contact map as an $N \times N$ matrix $\mathbf{C}$ where

$$\mathbf{C} = \{c_{lm}\} = \begin{cases} 1 & \text{if positions } l \text{ and } m \text{ are in contact} \\ 0 & \text{if positions } l \text{ and } m \text{ are not in contact} \end{cases}$$

where l and m are indices of sequence position (Rodrigue et al. 2005).

Two aspects of this formulation should be noted. First, unlike RNA where sites can potentially pair, here a single site can be considered in contact with multiple other sites. Second, these interactions are all weighted equally regardless of actual physical distance, as long as they are sufficiently close. It would be straightforward to alter the latter such that the relative distance is preserved and certain interactions are more influential than others. As described by Rodrigue et al. (2005), we can now define the energy of the sequence x with respect to pairwise potentials as the sum of the pair potentials for all pairs of amino acids in contact:

$$E_p(x) = \sum_{1 \le l \le m \le N} c_{lm} b_{x_l, x_m}$$

where x_l and x_m are the amino acids of sequence x at positions l and m, respectively, and $\mathbf{B} = \{b_{x_l, x_m}\}$ is the pair potential matrix of Bastolla et al. (2001). To model solubility constraints on protein evolution, we follow Robinson et al. (2003), who used an analysis of a large number of proteins to estimate how frequently a particular amino acid is observed at different degrees of solvent accessibility [see also Jones et al. 1992 and Jones 1999]. From the protein structure, we determine the solvent accessibility of a particular amino acid position. The energy with respect to solubility of sequence x, $E_s(x)$, is then the sum across all sites of the plausibility of seeing the observed amino acid at that accessibility level,

$$E_s(x) = \sum_{1 \le k \le N} S(x_k, a_k)$$

where a_k is the degree of solvent accessibility of site k and $S(x_k, a_k)$ is the statistical potential for observing amino acid x_k at such a degree of solvent accessibility (Robinson et al. 2003). We can now define $E(x, y)$ for proteins in a similar form as for RNA:

$$E(x, y) = \mathrm{e}^{(E_p(x)-E_p(y))p+(E_s(x)-E_s(y))s}$$

where p and s are, like z in the case of RNA, parameters that control how much the difference in sequence energies affect the rate of substitution for pairwise potentials and solubility, respectively. Note again that when $p = s = 0$, $E(x, y) = 1$ for all x, y, reducing the model to one of independence among sites.

Simulation Procedure

There are a number of ways to simulate data at a single position under an independent-sites model. Some of these are not applicable for simulating data that are context dependent. We will discuss a few of these methods and their applicability. The first method (Fig. 2a) begins by drawing the nucleotide at the root node of the tree from the stationary distribution $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$. If we specify the rate matrix $\mathbf{Q}$ and a branch length t, we can calculate the transition probability matrix $\mathbf{P}(t) = \{p_{ij}(t)\} = \mathrm{e}^{\mathbf{Q}t}$. This provides, for all i and j, the probability that after a branch length of t, the descendent node is in state j given that our ancestral node was at state i. Each branch will have its own transition probability matrix because branch lengths may differ. We can then work our way up the tree starting from the root, choosing states at each node until we reach the tips. The usage of matrix exponentiation to calculate transition probabilities is attractive because it considers all the possible paths, or character histories, from i to j in time t. However, the matrix exponentiation becomes intractable when the rate matrix is large. This is the case with the dependent-sites model we have described, where the rate matrix $\mathbf{R}$ is $4^N \times 4^N$, and for any reasonable sequence length N the matrix is quite large indeed.

Instead of using a transition probability matrix to consider all the possible paths from state i to j over time t simultaneously, we could instead simulate a single character history (Fig. 2b). One of the properties of the continuous-time Markov chain is that if the process is in state i, the waiting time until we leave state i is an exponentially distributed random variable with rate $q_{ii} = -\sum_{j \neq i} q_{ij}$. This means we find our root node state i from the stationary distribution as before, but now draw an exponential random variable with rate $-q_{ii}$. If this time is less than t, we observe a change from i to some other state j. The particular state j is drawn with probability $p_{ij} = q_{ij}/-q_{ii}$. This procedure is repeated until the sum of the drawn waiting times exceeds the length of the branch t, at which point the state of the process is the state at the descendent node. This character history simulation is performed iteratively up the tree for all branches until we have our states at the tip nodes. This

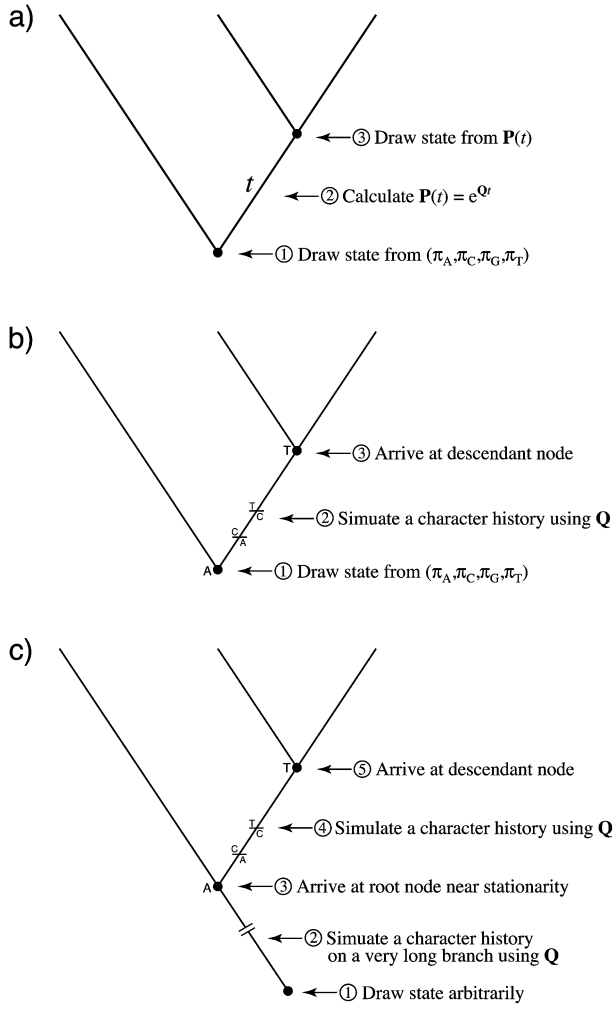

FIGURE 2. Three methods for simulating data under independence. a) Using matrix exponentiation is intractable for dependent data. b) Simulating a character history can be done with context dependency for an entire sequence, but drawing from the stationary distribution at the root node is still problematic. c) Evolve into stationarity by simulating a very long character history before reaching the root, then continuing up the tree as in (b).

method of drawing character histories has the benefit that it can be used under the dependent-sites model we have described. This is done by using the full sequence as the unit of evolution and replacing the site rate matrix $\mathbf{Q}$ with the sequence rate matrix $\mathbf{R}$, and then drawing a sequence history along the branch.

Both of these methods have assumed that we could draw the state at the root of the tree directly from the stationary distribution. This is not trivial under the dependent-sites model as the state space of all possible sequences is quite large (4^N possible sequences) when compared with independence (four possible nucleotides). However, the intuitive meaning of the process being at stationarity at the root is that the process has been underway for a long time before reaching the root of the tree, and we can simulate this directly (Fig. 2c). Under independence, if we pick any state

i as an ancestral state and then simulate its evolution along an exceedingly long branch before reaching the root, then the probability that we observe a particular state *j* at the root is the same as having drawn directly from the stationary distribution. This method can be used for the dependent-sites model described as well and is the method employed for all simulations in this study. We begin with an arbitrary sequence, not necessarily the one that would likely be sampled from the true dependent stationary distribution. We then evolve this sequence along a very long root branch under the model of dependence as described above, allowing the sequence to evolve into the one that would be sampled from the true dependent stationary distribution. The intuition should be clear: we need a sequence that corresponds to a fixed structure, so we choose a random sequence and allow it to evolve into the one that corresponds to the structure (directional selection). This yields a sequence at the root of the tree that corresponds to the structure, that can be used as a starting point for the simulation of the tree itself under continued structural constraint (stabilizing selection).

Structures Examined

In this study, we examine the effect of dependence introduced via structural constrain in both RNA and proteins. For RNA, we will focus on two structures: the *Bombyx mori* R2 element reverse transcriptase 3′ untranslated region, a 300-nucleotide structure previously examined by Mathews et al. (1997), and the eukaryotic 5S rRNA structure (119 nucleotides) examined by Yu and Thorne (2006). Each simulated parameter set using these structures include 400 and 1000 replicates, respectively. For proteins, we will also use two structures: mammalian myoglobin (*Physester catodon*; PDB code 1MBD; 459 nucleotides) and 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase (*Escherichia coli*; PDB code 1HKA; 474 nucleotides), both examined by Rodrigue et al. (2005). Simulations using these protein structures consist of 1000 and 500 replicates, respectively.

Energy at Stationarity

Because we have described the energy of a sequence as measuring how well a sequence fits a structure, we can visually inspect this process of approaching and sampling from the stationary distribution of sequences under the selective constraint by monitoring the energies of the sequences sampled. Figure 3a shows the energies of a sequence, initially sampled at random, evolving continuously under independence. As expected, the sequences sampled have similarly high energies because the vast majority of the 4^N possible sequences will not naturally fit the structure well. Contrast this with Figure 3b, which shows the energies of a sequence, similarly sampled at random originally, but evolving under the model of dependence. The sequences sampled converge to an area of the sequence space with much lower energies and remain there indefinitely. This indicates that the selective constraints of the structure limit the sequences that can be sampled to those that fit the structure reasonably well. That the chain fails to leave this area of the sequence space is an indication that we are in fact sampling sequences from the stationary distribution. In this manner, we can empirically determine the minimum branch

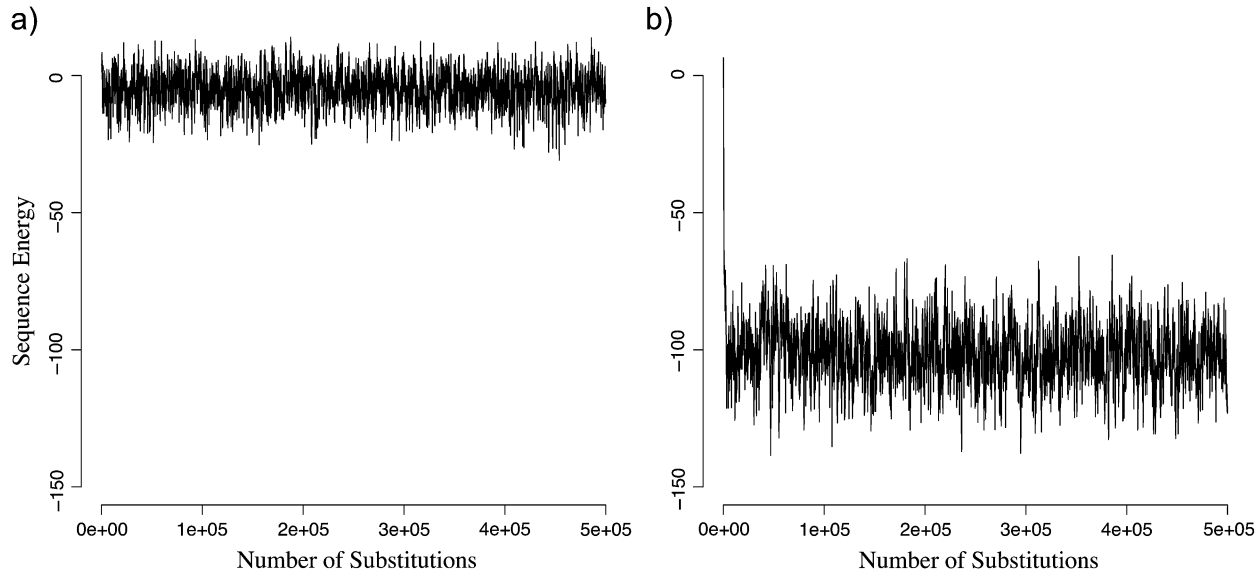

FIGURE 3. Energies sampled every 100 substitutions from a continuously evolving sequence. a) Independence among sites. Energies sampled are similar to that initially sampled at random. b) Dependence due to structural constraint. Low energies indicate that sequences sampled are those that fit the structure. The sequence evolves from a randomly sampled starting state of high energy to sample those states of low energy that correspond to the structure.

length necessary to sample from the stationary distribution with high probability prior to the simulation of sequences along the trees.

Results and Discussion

Rate Variation among Sites

We expect that the constraints imposed by structures will affect among-site rate variation; at stationarity, a site that is tightly constrained will experience a low rate of substitution relative to unconstrained sites. To confirm this, we simulated the evolution of a sequence at stationarity for varying levels of dependence and observed the number of changes occurring at different sites in the sequence. The results are shown in Figure 4. Under independence, RNA stem and loop positions observed similar rates of substitution (Fig. 4a), whereas under dependence ($z = 0.01$) the rate of substitution at stem position decreased and at loop positions increased (Fig. 4b). Further increasing the level of dependence did not seem to affect the change in substitution rate (data not shown). For proteins, the substitution process at a particular site can depend upon a number of other sites determined by the site's location in the folded protein. Whereas the rate of substitution observed was similar regardless of the number of contacted other sites under

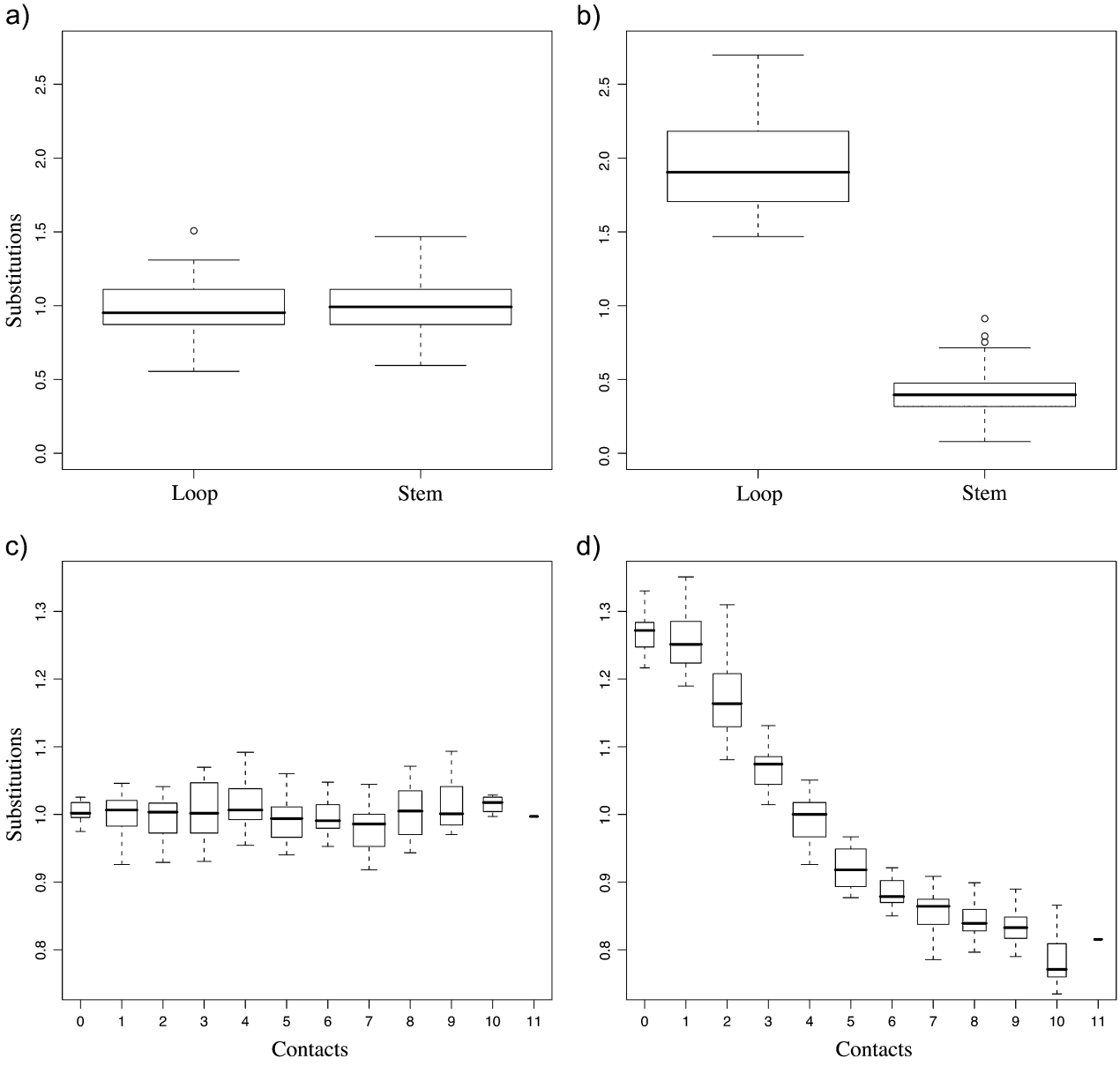

FIGURE 4. Number of substitutions at sites of varying constraint. Under independence RNA stem and loop sites experience similar rates of substitutions (a), but under dependence stem sites observe fewer and loop sites observe more substitutions (b). For proteins, under independence, all amino acid sites experience similar rates of substitution (c), whereas under dependence the rate of substitution is inversely proportional to the number of other sites with which the given site is in contact (d).

independence (Fig. 4c), under dependence we observed a clear negative correlation between the number of sites upon which a particular site is dependent and the rate of change that the site experienced (Fig. 4d). Similar results were obtained for all RNA and protein structures examined.

It is worth noting that the RNA model induces a higher rate of substitution among loop sites than among stem sites. This is quite different from what is observed in alignments of certain RNAs, in which loop regions are often highly conserved. This difference is in part because, whereas this model accounts for the dependencies introduced by the maintenance of the structure of the molecule, the model does not explicitly consider its function. If loop regions are conserved due to functional constraint of binding to another molecule, the dependence of these sites on their binding site is not captured by our model, which looks only at the structure of the single RNA. Clearly, both types of constraint are biologically relevant, and it would be straightforward to imagine expanding the model beyond a single sequence to consider two RNAs (or proteins) that interact, introducing dependencies both within and between the structures. It should also be noted that although our model does not capture stabilizing selection on RNA loop regions, neither do the independence-assuming models typically used for phylogenetic inference.

Effect of Dependence on RNA

We simulated sequences on the four-taxon tree using the predicted RNA structure of the *Bombyx mori* R2 element reverse transcriptase 3′ UTR (300 nucleotides) previously examined by Mathews et al. (1997). We did this for a constant tree size ($V = 1.75$) and for a range of branch length proportions at varying levels of dependence, and then estimated the topology from the data assuming independence. Figure 5a shows the accuracy of ML at estimating the true topology for these simulated data sets (400 replicates). The shorter the internal branch, the more difficult is the estimation problem. As expected, ML performs well for all tree shapes on data simulated with no dependence among sites. But as the level of dependence among sites increases the accuracy of ML decreases, particularly when the internal branch is short. Perhaps most striking is the decrease in accuracy resulting from even small levels of dependence in the data ($z = 0.1$), with accuracy falling to nearly 50% when the internal branch is short.

These simulations also provide a sense for just how short the internal branch must be before ML will begin to see a decrease in accuracy resulting from dependent evolution among sites. Whereas it might be encouraging if ML had difficulty only when the internal branch was quite short, this is not the case. Appreciable decreases in accuracy are observed over a wide range of internal branch lengths, indicating that the effects of dependent evolution on phylogenetic inference are not restricted to extreme topological cases.

It is important to note that although our structural model does induce rate variation among sites, this is at least partially accounted for in the GTR+Γ model used for analysis. This means that observed decreases in accuracy are more likely to result from differences resulting from the context-dependent nature of the substitution process induced by the model of structural constraint.

The decreased accuracy resulting from dependence in the data is not a particular property of ML however. Figures 5b,c show the analysis of the same data using neighbor-joining (with ML distances) and parsimony, respectively. Whereas the baseline expectations of how well the methods will perform when the data are independent differ, the trend is the same for all methods that assume independence: The effect of dependence is to reduce the accuracy of the methods, particularly when the problem is difficult, as is the case when branches differ markedly in length. We might note that neighbor-joining appears to do as well as ML in many cases, and in some cases seems to perform better. It would be tempting to attempt to draw broader conclusions from these simulations about the relative performance of these methods, but it must be remembered that we show here only a small portion of the possible parameter space of topologies, branch lengths, model parameters and have only shown a four-taxon case using a single structure. We refrain from drawing any such conclusions, and instead focus on the observation that all of these methods seem to suffer by failing to account for the dependence.

To examine whether these results were specific to the structure examined or more general, we simulated data using the eukaryotic 5S rRNA (119 nucleotides) as the reference structure. We did so on a slightly shorter tree length ($V = 1.0$) over the same range of branch length proportions and levels of dependence (1000 replicates). The analysis of these simulated sequence sets (Fig. 5d–f) are qualitatively consistent with the previous results: Methods that assume independence experience a reduction in accuracy over a wide range of branch length proportions as the level of dependence increases. This suggests that these decreases in accuracy are not specific to a single structure but are a more general property of the effect of dependent evolution in RNA.

The performance of phylogenetic methods assuming independence is also affected by the overall length of the underlying tree as well as its topology. Figure 6 shows the effect on accuracy of ML estimation using simulated R2 element RNA sequences over a range of branch length proportions on trees of total length 0.25, 1.0, and 1.75. The effect of a fixed level of dependence ($z = 0.5$) is to reduce accuracy relative to independence ($z = 0.0$) as shown before, but the effect is greater when the overall tree length is greater. On a larger tree (Fig. 6c) reductions in accuracy are observed at small branch length proportions, whereas on a small tree (Fig. 6a) the branch length proportion must be larger before reductions in accuracy are observed. This demonstrates how tree length and topology may interact to cause difficulties in estimation on dependence-containing data; dependence seems

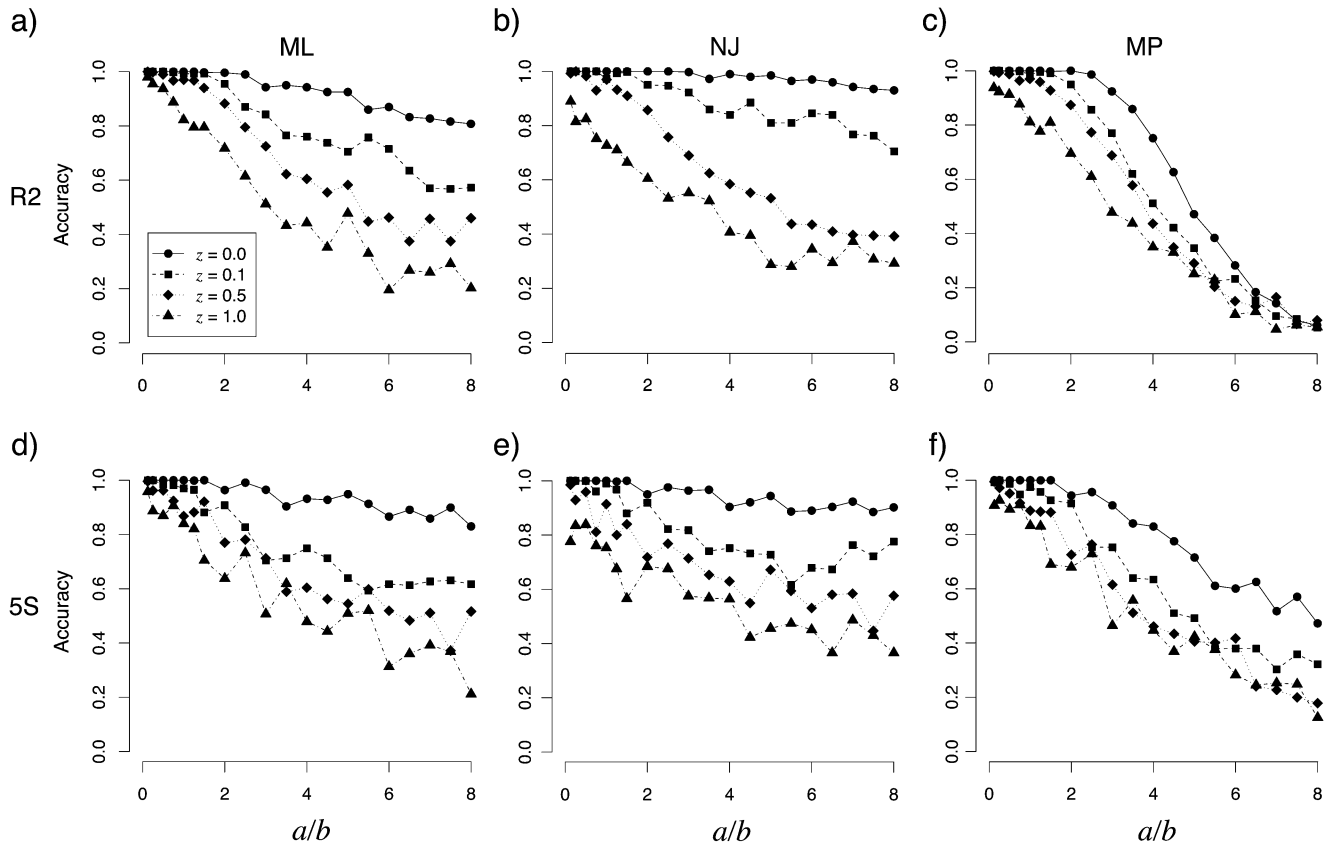

FIGURE 5. The accuracy of independence-assuming phylogenetic methods to infer the correct topology using RNA sequences constrained by structure simulated on a tree of total length $V = 1.75$. As the level of dependence in the data (z) increases, the methods are increasingly unable to infer the correct topology. This is especially true as the branch length ratio (a/b) becomes large and the problem becomes difficult. Structures: *Bombyx mori* R2 element reverse transcriptase 3′ UTR (R2) [300 nucleotides, 400 replicates] and 5S rRNA (5S) [119 nucleotides, 1000 replicates]. Methods: maximum likelihood GTR+Γ (ML), neighbor-joining using ML distances (NJ), parsimony (MP).

to have the greatest effect when the tree is very large and the internal branch is short. Results for neighbor-joining and parsimony were qualitatively similar, and for brevity we will largely focus the remainder of the four-taxon case discussion on results for ML, which are representative of trends observed using all methods examined.

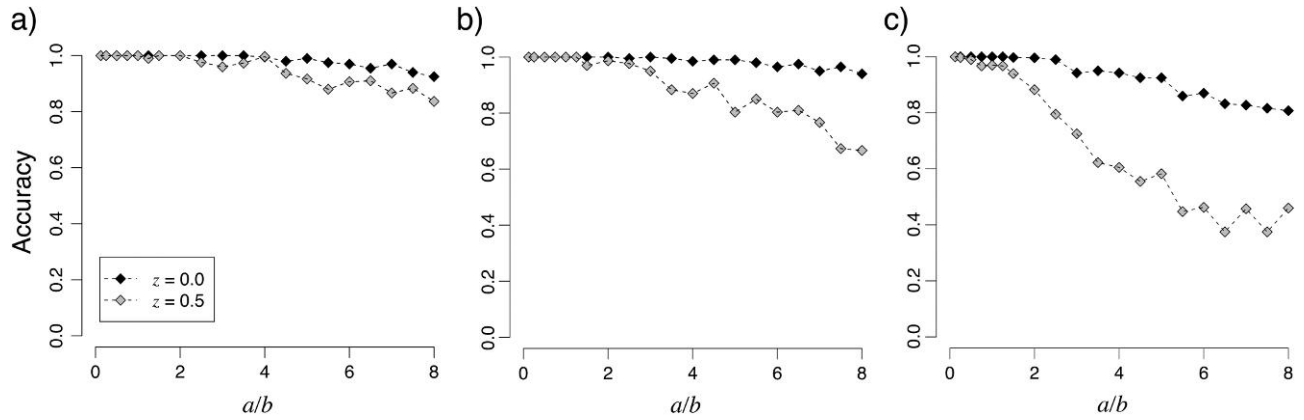

FIGURE 6. The total tree length (V) effects the accuracy of ML on simulated RNA sequences constrained by structure (R2). Dependence in the data (bottom curves) reduces the accuracy relative to independence (top curves), and this effect is more pronounced when the underlying tree is larger. a) $V = 0.25$. b) $V = 1.0$. c) $V = 1.75$. Qualitatively similar results were obtained for other independence-assuming methods and levels of dependence.

Effect of Dependence on Proteins

For proteins, the dependence involves two components: pairwise interactions and solubility constraints. To explore how each of these affect inference, we used the reference structure of mammalian myoglobin (*Physeter catodon*; PDB code: 1MBD; 459 nucleotides), previously studied by Rodrigue et al. (2005), to simulate data on a tree topology with $a/b = 5$ and a total length V of 1.3 (Fig. 7a) or 2.08 (Fig. 7b) across a wide range of dependence parameter values (1000 replicates). The larger tree shows the same trend as RNA: increased levels of dependence result in decreased accuracy. However, the effect seems to be less severe, particularly when the level of dependence is small. Furthermore, the dependence due to pairwise interactions has a much greater effect than dependence due to solubility constraints. Importantly, there is very little effect whatsoever observed when the tree length is small until levels of dependence become quite large indeed, even when the topology itself poses a moderately challenging problem.

We then repeated these simulations using the reference structure of 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase (*Escherichia coli*; PDB code: 1HKA; 474 nucleotides), also examined previously by Rodrigue et al. (2005). Whereas the structures of these two proteins are quite different, the results using the two structures

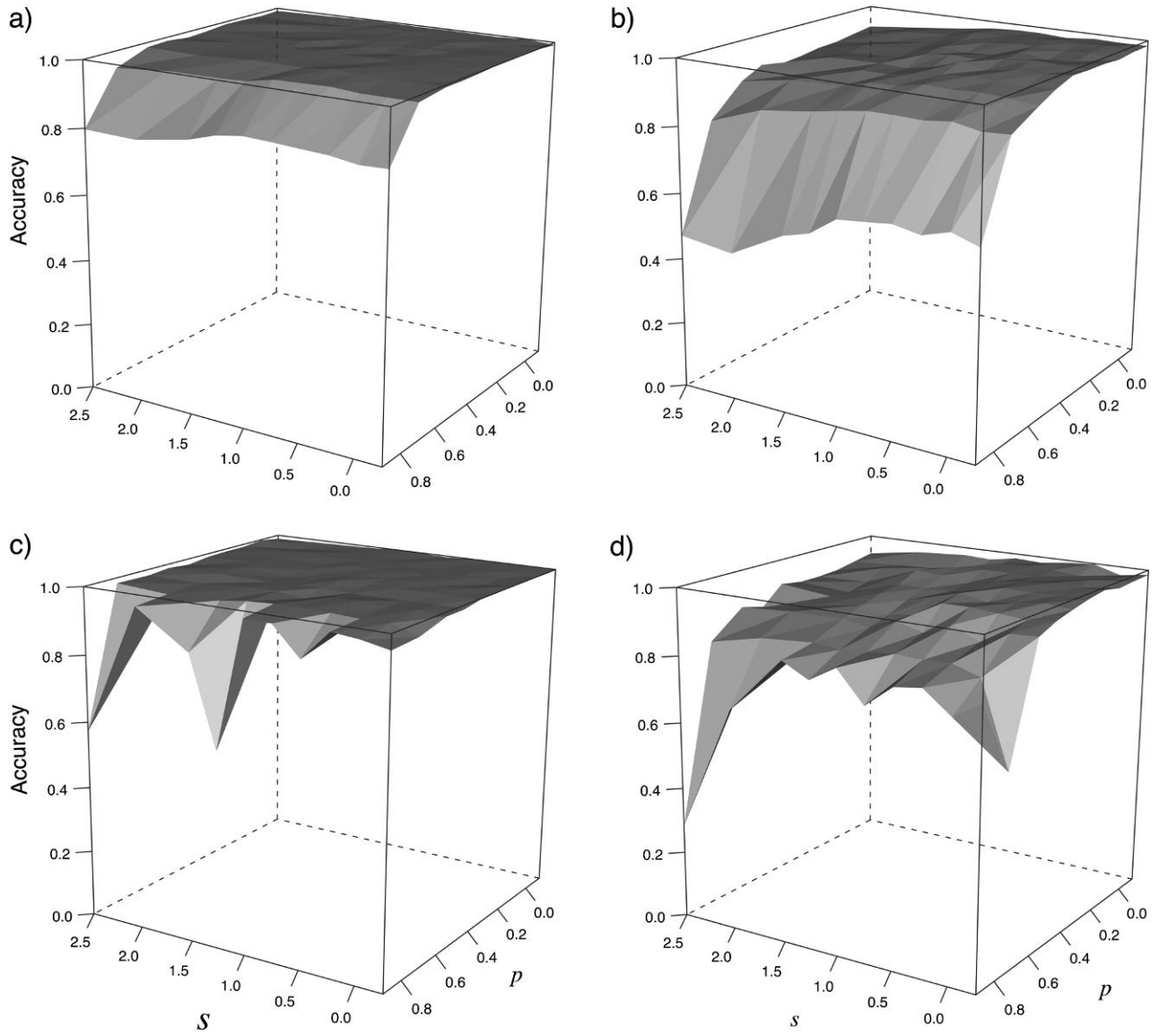

FIGURE 7. Accuracy of phylogenetic inference using ML using sequences generated under varying levels of dependence due to protein structure constraints: solubility (s) and pairwise interactions (p). Accuracy is reduced when dependence is strong and tree length is large. All panels represent the same tree topology ($a/b = 5$). Structures: mammalian myoglobin (MYO) [459 nucleotides, 1000 replicates] and 6-hydroxymethyl-7-8-dihydroxypterin pyrophosphokinase (PKA) [474 nucleotides, 500 replicates]. a) MYO, $V = 1.3$. b) MYO, $V = 2.08$. c) PKA, $V = 1.3$. d) PKA, $V = 2.08$.

are quite similar (Fig. 7c,d). There is some additional variance due to fewer replicates (500), but the trend is the same. This suggests that dependence among sites in proteins may have similar effects on phylogenetic inference regardless of the precise nature of the structure.

It is encouraging to see that for proteins, unlike RNA, small levels of dependence in the data do not seem to have a strong effect on the accuracy of phylogenetic methods. Estimates of the level of dependence in actual data will be considered below, but another consideration is whether or not the protein model, which reduces protein structure to two parameters of solubility and pairwise interactions, can adequately account for the complexity of actual protein structures. Vendruscolo and Domany (1998, 2000) and Park et al. (2000) have argued that there are limits to the utility of pairwise interaction potentials and hydrophobicity constraints in protein structure prediction. It is likely that the structural fitness of a sequence would be more accurately represented by the actual Gibbs free energy of the sequence, but at present this approach is computationally demanding. Although the simplified approach adopted here is well-justified, the conclusions drawn for proteins may not be the final word.

Effective Sequence Length

How phylogenetic methods behave when the data are neutral and independent may be used as a reference for describing how phylogenetic methods perform when ideal conditions are not met. We may consider the effective sequence length (L_e) as the length of independent neutral sequence that behaves in the same manner (in terms of phylogenetic accuracy) as our dependence-containing sequences. This is similar in spirit to the concept of an effective population size in population genetics. Because we expect dependence to introduce correlated substitutions, we expect the effective sequence length to be smaller than the actual sequence length (Huelsenbeck and Nielsen 1999). How much smaller is of interest and will depend on several factors including the actual sequence length, the nature of the structural constraints, the relative importance of the dependence, and the topology and length of the underlying tree.

Figure 8 quantifies the effective sequence length for one case examined. Each panel represents a different underlying tree topology (a/b) of the same overall tree length ($V = 1.75$). For each topology, we first simulated under independence sequences of different lengths and assessed the phylogenetic accuracy obtained by using these sequences. Shown in Figure 8 as the curves, these indicate the expected accuracy when using sequences of n independent neutral sites. For each topology, we then simulated RNA sequences of length 300 nucleotides (using the R2 reverse transcriptase structure) under dependence ($z = 0.1$) and assessed the accuracy using these dependent sequences, indicated by the horizontal lines. Where these observed (dependent) accuracies intersect our expected (independent) curve, we can project to the x-axis to estimate the effective sequence length for these dependence-containing data. The presence of dependence in the data results in a large decrease in effective sequence length, particularly for topologies in which the internal branch is relatively short.

One could argue that we might have easily predicted the effective length for RNA by simply considering paired sites to be as informative as a single unpaired site. In the structure used for the simulated RNA sequences, there were 168 stem and 132 loop positions, which by this method would predict an effective sequence length of 216 nucleotides. Alternatively, if all stem positions were considered to be invariable, the effective sequence length would be predicted to be 132 nucleotides. However, what we observe is that the dependence in our data leads to much lower accuracies, and subsequently much lower effective sequence lengths than both of these expectations, observing effective sequence lengths of less than 100 nucleotides. This implies that models simply accounting for covariation in the data are not accounting for all aspects of structural constraint and that these structural constraints lead to greater information loss.

Although we suggest that the concept of effective sequence length is useful for thinking about the effect of dependence on data, and is particularly useful for assessing these effects in our simulations, determining the effective sequence length requires knowledge about the true tree and importance of the structural constraints. The practicing systematist would therefore need to make some very strong assumptions in order to use the concept of effective sequence length to explicitly guide analysis.

Larger Data Sets

In order to understand how dependence among sites might affect phylogenetic inference we have focused on the four-taxon case using a single sequence/structure. This allowed us to thoroughly explore the relevant parameter space and gain some intuition for when we might expect error. However, using only four taxa or such a limited amount of sequence data is hardly something done in practice. It would therefore be useful to understand how the effects we have observed extend when the methods are presented with more taxa or more sequence data.

To address the question of how the methods perform on trees containing more than four taxa we simulated sequences on a 22-taxon tree using the R2 element RNA structure. In this tree (Fig. S1, available from http://www.sysbio.oxfordjournals.org/), all terminal branches are of the same length (0.05 expected substitutions per site) and are five times longer than internal branches (0.01 expected substitutions per site). In some sense, this makes for a relatively easy estimation problem: unlike the four-taxon case, all terminal branches are of equal length, and the overall tree length is quite

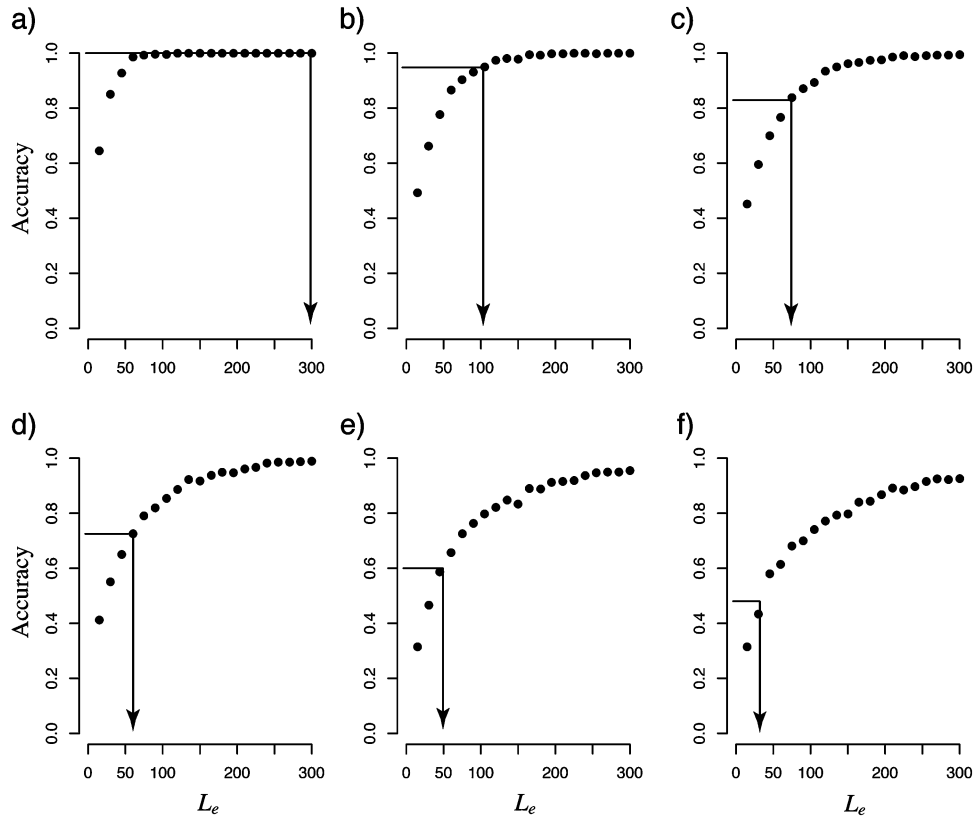

FIGURE 8. The effective sequence length (L_e) as a means of quantifying the phylogenetic information content of a sequence that contains dependence. All panels represent a fixed tree length ($V = 1.75$) and level of dependence ($z = 0.1$). a) $a/b = 0.5$. b) $a/b = 2$. c) $a/b = 3$. d) $a/b = 4$. e) $a/b = 6$. f) $a/b = 8$. The plotted curves indicate the accuracy of ML on these trees using independent data of varying lengths or the expected accuracy if the data were independent. The accuracy of ML on the simulated RNA sequences (R2, actual length = 300 nucleotides) on each topology is shown by the horizontal lines. Where these horizontal lines cross the curve, they drop to the x-axis to estimate the effective sequence length: the length of independent neutral sequence that displays the same amount of error in estimation that the actual dependence-containing sequence displays.

small. We simulated 500 RNA data sets on this tree for each of a range of levels of dependence and analyzed these data sets using the same methods used in the four-taxon case. We calculated the Robinson–Foulds metric (Robinson and Foulds 1981) to compare the estimated tree with the true tree, and the results are shown in

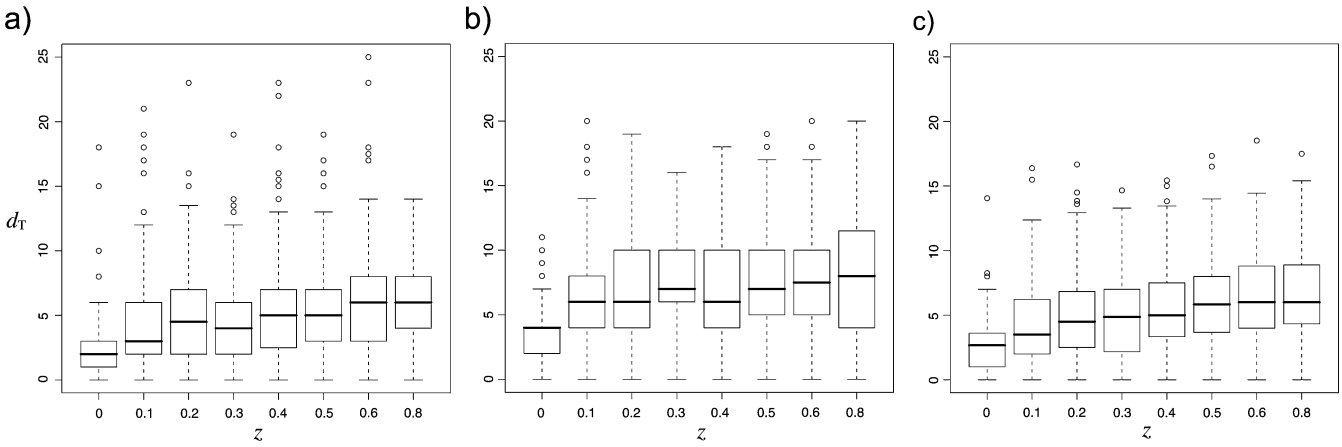

FIGURE 9. Accuracy of independence-assuming phylogenetic methods for 22-taxon simulations of RNA constrained by structure (R2; 500 replicates). The Robinson–Founds distance metric compares the estimated tree to the true tree for data sets under varying levels of dependence (z). For all methods, small amounts of dependence introduce error in tree estimation. a) ML, b) neighbor-joining, c) parsimony.

Figure 9. As expected, dependence in the data increases the amount of topological estimation error, in spite of the estimation problem not being an incredibly difficult one. Notably, even small amounts of dependence are sufficient to cause appreciable decreases in accuracy. We expect that on trees of greater length or containing variance in branch lengths might present more challenging problems and therefore be more sensitive to the effects of dependence. Although these simulations are hardly a thorough exploration of the space of possible trees larger than four taxa, they give a sense for how the problems observed might scale with the number of taxa.

Addressing the question of how very long dependence-containing sequences affect the analyses is not as straightforward. This is because one of the limitations of conditioning on an actual fixed structure is that the sequences are constrained to a fixed length. To test this question, we concatenated our R2 element data sets to create three very long ($\geq 30,000$ nucleotides) sets of sequences. We similarly concatenated 5S sequences to create two sets of sequences ($\geq 45,000$ nucleotides). We opted to use the same structure repeatedly to ensure that the same kind of dependence is introduced as there is no guarantee that different structures will not contain conflicting signal. The results are consistent with what was observed on shorter sequences (Fig. S2, available from http://www.sysbio.oxfordjournals.org/). When the dependence is large ($z \geq 0.5$), ML fails to estimate the correct topology when the problem is difficult. When the dependence is small ($z = 0.1$), ML is able to recover the true tree most of the time. Curiously, the neighbor-joining algorithm (using GTR+Γ distances) performs very well for all levels of dependence on all trees. Parsimony behaves qualitatively similar to the results on shorter sequences (Fig. 5c,f), Although these limited number of replicates are hardly conclusive, they give a sense for how these independence-assuming methods might handle a great deal of dependent sequence.

Estimates of Dependence

Our simulations have shown that failure to account for dependence among sites, such as dependence due to structural constraints, can greatly impair inference of the underlying tree topology. We have shown this for a wide range of levels of dependence, but it would be useful to have a sense for what might be reasonable levels of dependence to expect in actual data. Yu and Thorne (2006) estimated the level of dependence due to secondary structure for a set of eight 5S rRNA sequences to be 0.3661. In our simulations, we observe a significant impact on accuracy at lower levels of dependence than this ($z = 0.1$; see Fig. 5). This implies that failure to account for secondary structure of RNA may often lead to inaccurate inference of the true topology.

However, it is important to note that these kinds of models allow two methods of specifying the importance of structural constraint. One is an explicit level of dependence as specified by the tuning parameters discussed here (z for RNA, p and s for proteins). Another form of constraint is more implicit, namely how much flexibility is allowed in the structure. Here, we have presented a model in which the (implicit) requirements of the structure are strict, but the (explicit) level of dependence has been varied. Yu and Thorne (2006), however, allowed more internal flexibility in the structures they examined. This implies that the explicit level of dependence in a model such as what we have presented might be lower than what Yu and Thorne (2006) presented because the implicit constraint is greater. How much lower is a reasonable question and will be important in determining the level of decreased phylogenetic accuracy to be expected as a result.

For proteins, the story is also complicated. Although it is clear that there is dependence due to secondary structure in proteins (Thorne et al. 1996; Goldman et al. 1998), estimates of the level of dependence vary considerably. The model we have described and the model under which these estimates were obtained utilized similar levels of implicit flexibility, so we focus on the estimates themselves. Rodrigue et al. (2005) a model that involved the same pair potentials we employ, but it did not utilize solubility constraints. They estimated levels of dependence due to pairwise interactions to be in the range of $0.36 - 0.70$. Robinson et al. (2003) used a model that included both pair potentials and solubility and obtained estimates of pairwise dependence an order of magnitude less than Rodrigue et al. (2005) ($0.028 - 0.038$) while also estimating the dependence due to solubility ($0.88 - 0.95$). The large difference in pairwise interaction estimates could be due to differences in the modeling of the pairwise interactions or because the Rodrigue et al. (2005) model lacked solubility constraints. Choi et al. (2007) used the Robinson et al. (2003) model to estimate pairwise and solubility dependence for a wide range of proteins (Choi et al. 2007; Fig. 1), which not surprisingly agree with the Robinson et al. (2003) estimates. The difference between the Robinson et al. (2003) and Choi et al. (2007) estimates and the Rodrigue et al. (2005) estimates is an important one. As we have shown, pairwise interactions of the level Robinson et al. (2003) describe have little effect on our ability to estimate the true topology in spite of our assumptions of independence. If however pairwise dependence is of the level Rodrigue et al. (2005) describe, the impact on phylogenetic estimation is quite large.

Use of Energy as Fitness

The use of the energy of a sequence on a particular structure is but one possible surrogate for the fitness of a sequence and may have its limitations. It is possible, for example, that a given sequence might be able to fold well into many possible structures; that although a given sequence might have a low energy on the structure of interest, it might have an even lower energy on an alternate structure. This implies that this sequence would in reality spend more time folded in the alternative structure than the one of interest. In this case, we

might argue that the sequence energy itself is not a good proxy for the fitness of the sequence. A better surrogate for fitness in this case might be the probability that a sequence will fold into the desired structure. However, this would involve considering the energy of a sequence on all its possible structures, and as we are allowing the sequence itself to change this becomes computationally prohibitive, particularly as the sequence length increases.

Additional Model Limitations

The model we have presented is one in which dependence among sites results from the existence of a structure that must be maintained in order to perform some function. One limitation to this is that we do not allow the structure itself to evolve along the tree. This might be reasonable for short phylogenetic distances, but the fact remains that even closely related sequences vary widely in their structural homology across taxa. Accounting for variance in the structural constraints across the tree will be a challenge for future research.

Another way in which the kind of model we have described might be developed is to allow for more than one substitution at a time. Huelsenbeck and Nielsen (1999) developed a compound Poisson model that allows for this, and it might be a natural pairing with the type of model described here; evaluating the energy/fitness of a sequence two substitutions away is a straightforward extension. Allowing more than one substitution at a time might be particularly useful when the intrinsic constraints are very strong, enabling sequences to cross fitness valleys more easily.

CONCLUSIONS

We have shown that failure to account for dependence among sites due to secondary and tertiary structure can lead to inaccurate estimation of the underlying tree topology. This is particularly true when the dependence is strong as may be the case with RNA, when the internal branch is relatively short, and when the overall tree length is large. These findings have direct implications for anyone interested in phylogenetic estimation or analyses dependent thereupon. We have also shown that the effect is stronger than might have been expected under simpler models of dependence, such as considering paired RNA sites as one. This indicates that there is room for improvement in phylogenetic methods by accounting for the nature of the dependencies in the data. We have introduced the concept of an effective sequence length as an intuitive means of quantifying the effects of dependence and have presented a general method of simulating data on phylogenetic trees under complex models of evolution.

Although in this paper we have focused on RNA and protein structures to introduce the dependencies among sites, the findings here may extend to the general case in which there may be dependence among characters. Morphological characters, for example, may contain large amounts of dependence, although it may be much more difficult to model the particular nature thereof. But our findings that the presence of dependence in the data, if unaccounted for, may lead to error in phylogenetic estimation should hold regardless of how well we understand the nature of the dependence itself. This suggests that in cases where we may be unable to model the dependence, being able to simply detect the presence of dependence in the data might be valuable.

SUPPLEMENTARY MATERIAL

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

FUNDING

This work was supported by the National Institutes of Health (MCB-0075404 to J.P.H. and R01GM076485 to D.H.M.).

ACKNOWLEDGMENTS

We thank Jack Sullivan, Marc Suchard, and two anonymous reviewers for their constructive critiques of this manuscript, Bastien Bousseau and Weiwei Zhai for helpful discussions, and Sang Chul Choi for sharing details on protein solubility constraints.

REFERENCES

Anisimova M., Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic substitution models. Mol. Biol. Evol. 26:255–271.

Bastolla U., Farwer J., Knapp E.W., Vendruscolo M. 2001. How to guarantee optimal stability for most representative structures in the protein data bank. Proteins. 44:79–96.

Castoe T. A., de Konig A. P. J., Kim H.-M., Gu W., Noonan B. P., Naylor G., Jiang Z. J., Parkinson C. L. , Pollock D. D. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. U.S.A. 106:8986–8991.

Choi S. C., Hobolth A., Robinson D. M., Kishino H., Thorne J. L. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. Mol. Biol. Evol. 24:1769–1782.

Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401–411.

Felsenstein J., Churchill G. A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13: 93–104.

Goldman N., Thorne J., Jones D. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 149:445–458.

Goldman N., Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11: 725–736.

Huelsenbeck J. P., Hillis D.M. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

Huelsenbeck J. P., Nielsen R. 1999. Effect of non-independent substitution on phylogenetic accuracy. Syst. Biol. 48:317–328.

Jones D. 1999. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287:797–815.

Jones D.T., Taylor W.R., Thornton J.M. 1992. A new approach to protein fold recognition. Nature. 358:86–89.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N, editor. Mammalian protein metabolism. San Diego (CA): Academic Press. p. 21–123.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Kleinman C., Rodrigue N., Bonnard C., Philippe H., Lartillot N. 2006. A maximum likelihood framework for protein design. BMC Bioinformatics. 7:326.

Mathews, D.H., Banerjee A., Luan D., Eickbush T., Turner D. 1997. Secondary structure model of the rna recognized by the reverse transcriptase from the r2 retrotransposable element. RNA. 3:1–16.

Mathews, D.H., DisneyM.D., Childs J.L., Schroeder S.J., Zuker M., Turner D.H. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. Proc. Natl. Acad. Sci. U.S.A. 101:7287–7292.

Mathews D.H., Sabina J., Zuker M., Turner D.H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288:911–940.

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

Nielsen R., Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–93.

Park K., Vendruscolo M., Domany E. 2000. Toward and energy function for the contact map representation of proteins. Proteins. 40:237–248.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Robinson D.M., Jones D.M., Kishino H., Goldman N., Thorne J.L. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20:1692–1704.

Rodrigue N., Lartillot N., Bryant D., Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 347:207–217.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–425.

Schöniger M., von Haeseler A. 1994. A stochastic model and the evolution of autocorrelated dna sequences. Mol. Phylogenet. Evol. 3:240–247.

Swofford D.L. 1998. PAUP*: phylogenetic analysis using parsimony and other methods. Sunderland (MA): Sinauer Associates, Inc.

Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures in Mathematics in the Life Sciences 17:57–86.

Thorne J., Goldman N., Jones D. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. 13:666–673.

Tillier E.R.M. 1994. Maximum likelihood with multiparameter models of substitution. J. Mol. Evol. 39:409–417.

Tillier E.R.M. Collins A. 1995. Neighbor joining and maximum likelihood with rna sequences: addressing the interdependence of sites. Mol. Biol. Evol. 12:7–15.

Vendruscolo M., Domany E. 1998. Pairwise contact potentials are unsuitable for protein folding. J. Chem. Phys. 109:11101–11108.

Vendruscolo M., Domany E. 2000. Protein folding using contact maps. Vitam. Horm. 58:171–212.

Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics. 139:993–1005.

Yu J., Thorne J.L. 2006. Dependence among sites in RNA evolution. Mol. Biol. Evol. 23:1525–1537.