Syst. Biol. 60(2):161–174, 2011

DOI:10.1093/sysbio/syq088
Advance Access publication on January 12, 2011

What's in a Likelihood? Simple Models of Protein Evolution and the Contribution of Structurally Viable Reconstructions to the Likelihood

Clemens Lakner[1,2,*], Mark T. Holder[3], Nick Goldman[4], and Gavin J. P. Naylor[2]

[1]Department of Biological Science, Section of Ecology and Evolution and [2]Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120, USA; [3]Department of Ecology and Evolution, University of Kansas, 6031 Haworth, 1200 Sunnyside Avenue, Lawrence, KS 66045; and [4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK;
**Correspondence to be sent to: Department of Biological Science, Section of Ecology and Evolution, Florida State University, Tallahassee, FL 32306-4120, USA; E-mail: lakner@scs.fsu.edu.*

Received 29 November 2009; reviews returned 22 March 2010; accepted 19 November 2010
Associate Editor: Olivier Gascuel

Abstract.—Most phylogenetic models of protein evolution assume that sites are independent and identically distributed. Interactions between sites are ignored, and the likelihood can be conveniently calculated as the product of the individual site likelihoods. The calculation considers all possible transition paths (also called substitution histories or mappings) that are consistent with the observed states at the terminals, and the probability density of any particular reconstruction depends on the substitution model. The likelihood is the integral of the probability density of each substitution history taken over all possible histories that are consistent with the observed data. We investigated the extent to which transition paths that are incompatible with a protein's three-dimensional structure contribute to the likelihood. Several empirical amino acid models were tested for sequence pairs of different degrees of divergence. When simulating substitutional histories starting from a real sequence, the structural integrity of the simulated sequences quickly disintegrated. This result indicates that simple models are clearly unable to capture the constraints on sequence evolution. However, when we sampled transition paths between real sequences from the posterior probability distribution according to these same models, we found that the sampled histories were largely consistent with the tertiary structure. This suggests that simple empirical substitution models may be adequate for interpolating changes between observed sequences during phylogenetic inference despite the fact that the models cannot predict the effects of structural constraints from first principles. This study is significant because it provides a quantitative assessment of the biological realism of substitution models from the perspective of protein structure, and it provides insight on the prospects for improving models of protein sequence evolution. [Ancestral state reconstruction; empirical amino acid models; maximum likelihood; phylogenetics; protein structure.]

Most commonly employed models of sequence evolution are based on the assumption that all sites are independent and identically distributed (i.i.d.). The i.i.d. component of stochastic substitution models is statistically convenient because it greatly reduces the number of parameters to be estimated from the data. It is also computationally useful because it allows us to calculate the likelihood for different sites separately (enabling the calculation of the likelihood to be completed in an amount of time that is proportional to the number of sites). Nevertheless, the i.i.d. assumption seems biologically implausible. If a protein must maintain its functional integrity over an evolutionary trajectory, then the set of acceptable substitutions must be restricted. It would seem that certain substitutions would be tolerated at some sites but not at others. Furthermore, the kinds of substitutions that would be tolerated at a site would depend on the residues' biochemical environment (commonly referred to as *context dependence*) and on the configuration of residues at other sites in the protein. Consequently, Wang and Pollock (2005) argue that "Most evolutionary analyses rely on the assumption that the probabilities of substitution at each site are independent of substitutions at other sites, although protein structure and function result from interactions among amino acids, and this assumption cannot be true in principle." Thus, it would seem that modeling the sequence of substitutions occurring over an evolutionary trajectory would involve computing a series of complex interacting dependencies that could not possibly be captured by simple i.i.d. models. Yet, this is exactly what is done, and in a great many cases, it yields phylogenetic trees that seem sensible. How could i.i.d. models be so effective in the face of what seems to be such a complex problem? One possible way out of the conundrum is to suppose that there are a few critically important sites that exhibit strongly nonlinear dependencies required to maintain the structural and functional integrity of proteins, but these are vastly outnumbered by sites that show no such dependency.

Several researchers have devised schemes to capture heterogeneous processes of substitution. For example, the gamma model introduced by Yang (1993, 1994) allows sites to evolve at different rates and almost always leads to significant improvements in likelihoods. Koshi et al. (1997) and Koshi and Goldstein (1998) proposed mechanistic models that consider the fitness effects of amino acid substitutions in different parts of the protein. Their models assume several *classes* of sites that evolve according to different substitution matrices. Similarly, the PASSML model ("Phylogeny and Secondary Structure using Maximum Likelihood;" Lio et al. 1998; Lio and Goldman 1999) is a hidden Markov model where the local structural environments are characterized by different substitution matrices. The probability that a site evolves according to certain process influences the corresponding probability at adjacent sites. However, spatial interactions between distant residues are ignored. Other studies have been conducted to try to identify pairwise dependencies between more distant residues that would serve to maintain structural and functional integrity of proteins. For example, Pollock and Taylor (1997) and Pollock et al. (1999) developed a phylogeny-based approach to identify correlated changes.

More recent efforts have approached the problem in an explicitly integrated way where the focus is no longer on modeling the evolution of individual sites, but on the viability of the entire string that makes up the folded protein. Notably, Robinson et al. (2003) and Rodrigue et al. (2005) pioneered a Bayesian approach that takes into account the compatibility of all implied ancestral sequences with the crystal structure of a protein. The underlying idea of their models is to constrain the implied ancestors to sequences that fit the tertiary structure. To do this, the substitution rate is decomposed into a mutation rate and a fitness term (for a review, see Thorne 2007). In other words, the rate of each mutation is weighted by a function of the difference in fitness before and after the mutation. The term "fitness" is used here as *compatible with the structure* as determined by an energy function. The evaluation of the fitness introduces dependence among sites because the energy function acts on the complete string. As a consequence, the states of the Markov chain are now complete sequences as opposed to individual amino acids. Rodrigue et al. (2006) found that the improved fit of the structure-aware models they used was mild when compared with rich independent sites models. Moreover, in a model testing comparison, they were outperformed by empirical amino acid models that allowed for among-site rate variation.

Empirical amino acid models constitute rate matrices and state frequencies that, in most cases, have been estimated via maximum likelihood (ML) for large representative data sets (for a Bayesian approach, see Huelsenbeck et al. 2008). The rate matrices reflect the 20 × 20 instantaneous rates of change between the amino acids. The rates are derived empirically and represent averages, scored over many different proteins, and within any given protein, over many different sites. Each of these sites likely has its own set of constraints and context dependencies. As such these average values, while providing accurate descriptions for hypothetical average cases, can be woefully inappropriate when applied to specific sites that have constraints that do not conform to the average. Examples of such matrices for water-soluble proteins are the PAM (Dayhoff et al. 1978, derived using maximum parsimony), JTT (Jones et al. 1992b), BLOSUM (Henikoff and Henikoff 1992), WAG (Whelan and Goldman 2001), and LG (Le and Gascuel 2008) models. Similarly, mtREV (Adachi and Hasegawa 1996) and mtMAM (Cao et al. 1998; Yang et al. 1998) were specifically devised for mitochondrial proteins.

In order to perform the likelihood calculations, most i.i.d. phylogenetic models use a matrix of instantaneous rates of change between residues to produce a transition probability matrix. Implicitly, the transition probabilities consider all possible pathways; for instance, at a single site, the probability associated with an $A \rightarrow G$ transition would consider all substitution histories—including those that involve more than one change (e.g., $A \rightarrow C \rightarrow G$). The probability of a path through sequence space with the two terminal sequences as beginning and end points depends on the substitution model. The likelihood is the integral over this distribution of path probabilities. Most substitution histories that are included in the integration are not biologically sensible because most of the paths through protein-sequence space will visit structurally inviable sequences (and thus require a number of selectively deleterious intermediates). For closely related sequences—that are truly only a few substitutions apart—most substitution histories that require a multitude of hidden changes will be assigned a very small probability and their contribution to the likelihood will be negligible.

Here, we set out to explore the question: "Are likelihood calculations under simple empirical i.i.d. models dominated by substitution histories that require ancestral states that result in unstable or potentially misfolded proteins?" We address this question by sampling substitution histories directly from their posterior distribution and subsequently evaluating the ancestral sequences for their sequence-to-structure fit. *Structural viability* is assessed with a protein threading approach. We make no attempt to test for function, but we assume that structural stability and specificity are minimal requirements for functioning proteins.

Methods

Data Preparation

We selected short to medium length (≤ 153 residues) monomeric proteins (Table 1) for which representative crystal structures are available in the Protein Data Bank (PDB; Berman et al. 2000). Specifically, we used vertebrate parvalbumin A, myoglobin, lysozyme c, and bacterial 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK). Table 1 details the PDB identifier, source organism, number of residues, and resolution for each structure. For each protein, we first aligned a pair of divergent sequences (Table 2): Leopard Shark (*Triakis semifasciata*) and Black Rat (*Rattus rattus*) for parvalbumin, American Alligator (*Alligator mississippiensis*) and Burchell's Zebra (*Equus burchelli*) for myoglobin, Human and Green Turtle (*Chelonia mydas*) for lysozyme, and *Vibrio cholerae* and *Salmonella typhi* for HPPK. The sequence pair for HPPK represents a subset of the data set used by Rodrigue et al. (2006). Due to the difficulties associated with incorporating gaps into the phylogenetic model and the energy calculations we restricted all analyses to gapless alignments.

In order to assess the impact of improved taxon sampling on the structural compatibility of the mappings, we augmented our pairwise alignments with additional sequences. Due to the computational burden, we only examined three- and four-taxon trees. The four-taxon case allowed for a basic analysis of the relationship between sequence-to-structure fit and tree topology estimation. The additional sequences (with

TABLE 1. Crystal structures used for the simulations

Protein	PDB ID	Source	Length	Resolution (Å)
Parvalbumin A	5PAL, 1RTP	Leopard Shark, Black Rat	109	1.54, 2.0
Myoglobin	1LHS	Loggerhead Sea Turtle	153	2.0
Lysozyme c	1JSF	Human	130	1.15
HPPK	1HKA	*Escherichia coli*	158	1.50

their UniProtKB or GenBank accession numbers) were *Amphiuma means* (PRVA_AMPME) and *Latimeria chalumnae* (PRVA_LATCH) for parvalbumin, *Struthio camelus* (MYG_STRCA) and *Rattus norvegicus* (MYG_RAT) for myoglobin, *Bufo andrewsi* (LYS_BUFAN) and *Camelus dromedarius* (LYSC_CAMDR) for lysozyme, and *Aeromonas salmonicida* (YP_001140682.1) and *Pectobacterium atrosepticum* (CAG76218.1) for HPPK. ML branch length estimates for all trees are provided in the online Supplementary material (available at www.systematicbiology.org). Alignments and trees are also available from TreeBase (Sanderson et al. 1994, study accession: http://purl.org/phylo/treebase/phylows/study/TB2:S11030).

Reference sequences.—One of the simplifying assumptions of the models proposed by Robinson et al. (2003) and Rodrigue et al. (2005) is that the protein structure remains constant across the tree. To assess the variation in sequence-to-structure fit that we should expect to see when we fit sequences onto a fixed structure, we evaluated sequences from several species in addition to the sequences from the organisms listed above. This allowed us to interpret the structural viability of the simulated ancestors in the context of a reference distribution. It is clear that this reference is incomplete, but it provides a sense of the range of scores that is *at least* displayed by real sequences that fold into the given structure. For HPPK, we chose the sequences from the data set in Rodrigue et al. (2006). For lysozyme (and possibly HPPK), the reference set contains paralogs. We found no indication that these sequences would preferentially fold into different structures. A complete list of reference-sequence identifiers can be found in the online Supplementary Material.

Likelihood Calculations and Data Augmentation

The models tested here were Poisson (Bishop and Friday 1987), WAG, and LG. The Poisson model is the amino acid equivalent of the Jukes and Cantor (1969) nucleotide substitution model where all substitution rates and equilibrium frequencies are equal. On a given tree ψ the actual likelihood $p(D|M,\psi)$ is related to the so called *augmented likelihood* $p(D,\phi|M,\psi)$ via (unobserved) *substitution histories* ϕ

$$p(D|M,\psi) = \int_{\Phi} p(D,\phi|M,\psi)\mathrm{d}\phi, \qquad (1)$$

where Φ denotes the set of all possible transition paths and M stands for the fixed parameters of the model (Mateiu and Rannala 2006; Rodrigue et al. 2007, 2008). We use the terms *substitution histories* (Rodrigue et al. 2005), *mappings* (Nielsen 2001), and *transition paths* (Jensen and Pedersen 2000; Pedersen and Jensen 2001; Robinson et al. 2003) synonymously. One mapping includes a sequence of events and associated times for *all* sites of the alignment. For a single branch, the likelihood is the probability of observing the root sequence s_0 (given the state frequencies of the model) multiplied by the probability of all possible transition paths to the descendant sequence s_1

$$p(s_0,s_1|M,\nu) = p(s_0|M)\int_{\Phi} p(s_1,\phi|s_0,M,\nu)\mathrm{d}\phi, \qquad (2)$$

where ν denotes the length of the branch.

Constrained simulations.—We followed the approach described in Nielsen (2001) to sample 500 substitution histories from the distribution $p(s_1,\phi|s_0,M,\nu)$. The process at each site of a substitution history is simply a realization of the continuous-time Markov chain that starts in the observed state of the ancestral sequence and ends in the state of the descendant sequence. Substitution histories for a branch were obtained as follows: first, events were sampled separately for all sites. Then, all site events were combined, resulting in histories that connect the terminal sequences through a series of one-step neighbors (Fig. 1). These ancestral sequences were subsequently evaluated for their compatibility with the crystal structure. In the four-taxon case, we sampled 500 histories for each of the three unrooted topologies.

For all models, the ML branch length was found to be a good estimator of the mean number of substitutions

TABLE 2. Sequence pairs used for the simulations

Protein	s_0	Swissprot accession number	s_1	Swissprot accession number	(%) sequence identity
Parvalbumin A	Leopard Shark	P30563	Black Rat	P02625	58.7
Myoglobin	American Alligator	P02200	Burchell's Zebra	P68083	63.4
Lysozyme c	Human	P84492	Green Turtle	P61626	62.3
HPPK	*Vibrio cholerae*	Q9KUC9	*Salmonella typhi*	Q8Z9C4	55.1

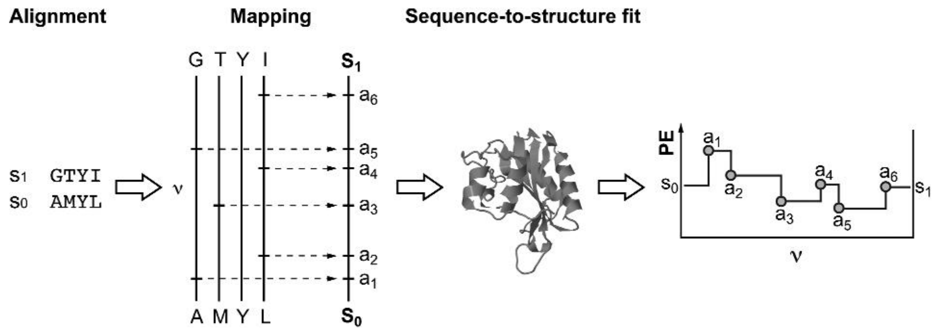

FIGURE 1. Example of a transition path between two hypothetical sequences s_0 (AMYL) and s_1 (GTYI). A mapping is generated for each site according to Nielsen (2001). Sorting all site events by their times results in a history that connects the terminals through a series of one step neighbors (a_1 to a_6). These ancestral sequences (along with s_0 and s_1) are subsequently evaluated for their compatibility with the crystal structure using a pseudo-energy (PE) score.

over mappings drawn from the posterior. Sampling from the joint posterior of branch lengths and substitution histories would have required us to specify a prior over branch lengths. To avoid this, we further simplified our approach by fixing the branch lengths to the ML estimates (it should perhaps be noted that the ML estimate of the branch length is not the same as the posterior mean of the number of substitutions for a given branch length). For each sequence pair and model, we used PAML 4 (Yang 2007) to infer the ML branch lengths (Table 3).

Unconstrained simulations.—In the constrained simulations, both endpoints of the branch were real sequences. To test the efficacy of the model to maintain structural compatibility without constraining the simulation to end in a predetermined end state (s_1 in Table 2), we sampled from the distribution $p(\phi|s_0,M,\nu)$. For each model, we performed 500 simulations with the same expected number of substitutions as above (Table 3). To test the behavior in the limit, when the sequence is effectively randomized (and the composition corresponds to the equilibrium amino acid frequencies implied by the model), 100 simulations were performed for each model on an unbounded branch of length 30.0. This describes the situation where each site underwent a large number of substitutions on the branch between two terminal sequences, effectively removing all phylogenetic signal.

Evaluation of Sequence-to-Structure Fit

An ideal protein design procedure should generate sequences that are compatible with the native conformation (*stability*) and incompatible with other structures (*specificity*, for a detailed treatment of the subject, see Koehl and Levitt 1999a,b). If the sampled ancestral sequences are to be compatible with the structure of the wild types, they must fulfill these criteria as well. Three approaches were taken to evaluate the compatibility of the sampled sequences with the crystal structures. First, we used empirically derived contact potentials to assess sequence-structure compatibility. These so-called *knowledge-based* potentials were used to calculate the pseudo-energy (PE), which is tightly correlated with the free energy of the sequence in the final folded form. Second, in order to test for structural specificity, we estimated the PE distribution for 10,000 shuffled sequences on each native structure (Table 1). Third, as an additional measure for specificity, we compared the PE of each sampled sequence on the native conformation to its PE scores on a set of misfolded decoys (see below). Sequence shuffling was used to test if a particular arrangement of residues had a significantly better fit to the structure than random sequences of the same composition. The decoys, on the other hand, were used to evaluate if a sequence was significantly more compatible with a particular structure than with other compact (but misfolded) conformations (local energy minima). For the shuffled sequences, we calculated the Z-score (in the sense of Bowie et al. 1991) only with respect to the native structure as

$$Z_s = \frac{\epsilon - \bar{x}_s}{\sigma_s}, \qquad (3)$$

where ϵ is the PE of the native sequence, $\bar{x}_s$ is the average PE of the shuffled sequences, and σ_s is the standard deviation of the PE scores of the shuffled sequences (all on the native structure).

We further define Z_d as the energy gap between a sequence on the native structure and its average PE on the set of misfolded decoys, expressed in standard deviations:

$$Z_d = \frac{\epsilon - \bar{x}_d}{\sigma_d}. \qquad (4)$$

Here, $\bar{x}_d$ denotes the average PE of a sequence on the set of decoys and σ_d is the standard deviation. Note that Z_s

TABLE 3. Branch lengths ν (expected number of substitutions per site) and likelihoods for the sequence pairs under different models

Protein	ν			−ln$L$		
	Poisson	WAG	LG	Poisson	WAG	LG
Parvalbumin A	0.54166	0.54033	0.58710	532.923	478.894	482.838
Myoglobin	0.46225	0.46228	0.49951	723.726	668.858	665.950
Lysozyme c	0.48017	0.51281	0.54236	619.853	594.483	601.541
HPPK	0.60856	0.66055	0.72945	791.087	741.261	735.662

is only defined with respect to the native conformation (but multiple sequences), whereas Z_d is defined for one sequence on a set of different conformations (see Chiu and Goldstein 1998, for a summary of the different uses of the term Z-score in the protein structure literature).

We used three coarse-grained, residue-level energy functions to calculate PE scores. First, we employed the pairwise contact potential derived by Bastolla et al. (2001), which was used as described in Rodrigue et al. (2005). The second energy function, previously used by Robinson et al. (2003), is based on several statistical potentials estimated for different degrees of separation within the sequence. In addition to the pair-potential term (E_{pair}), it also includes a solvent accessibility term (E_{solv}; Jones et al. 1992a; Jones 1998, 1999). We calculated the overall energy as a combination of the two terms as follows (based on Jones et al. 1992a):

$$\text{PE} = E_{\text{pair}} + wE_{\text{solv}}, \tag{5}$$

where $w = \sigma_{E_{\text{pair}}}/\sigma_{E_{\text{solv}}}$ denotes the ratio of the standard deviations of the terms over all decoys (PE_d), or over a large number of shuffled versions of a sequence (PE_s). Finally, we used the THOM2 threading model (Meller and Elber 2001), implemented in the software package LOOPP (version 2.000, http://www.cs.cornell.edu/home/meller/loopp/loopp-v2.000_doc.html). Side chain centers of mass were calculated for all residues of the misfolded decoys and stored in a coordinate-library input file for LOOPP. We modified the source code to incorporate our Z-score calculations. However, the actual energy evaluations of LOOPP were kept unchanged.

Decoy sets.—For each protein, 1000 independent conformations were sampled using the Rosetta (version 2.3, available at http://www.rosettacommons.org) *ab initio* folding algorithm (Simons et al. 1997; Rohl et al. 2004). In each case, predictions were based on the native amino acid sequences for the structures listed in Table 1. Fragments for the *ab initio* protocol were obtained from the *Robetta Full-chain Protein Structure Prediction Server* (http://robetta.bakerlab.org, Kim et al. 2004). The initial step of the Rosetta *de novo* folding algorithm involves sampling of conformational energy minima with a coarse-grained energy function (for an accessible description, see Das and Baker 2008). Decoys were refined in full-atom mode using the relax protocol (for details, see Misura and Baker 2005). During the full-atom refinement, Rosetta builds complete side chains and optimizes backbone torsion angles and side chain packing. The options used for Rosetta's relax mode were *−farlx*, *−fa_input*, *−ex*1, *−ex*2, and *−stringent_relax*. From the full-atom models, we computed the side chain centers of mass for the contact-distance thresholds used in the Jones and THOM2 potentials. Consequently, after fixing bond lengths and angles (*−idealize*), we applied the same optimization process to the wild-type structures.

Results

For all investigated sequence pairs, the ancestral sequences were largely compatible with the crystal structure when an empirical amino acid model was used. However, sequences resulting from the unbounded simulations lost their structural integrity relatively quickly. The difference between sequence pairs and unbounded simulations indicates that our i.i.d. models are much too simple to allow for robust extrapolation. However, for the tested degree of divergence, the wild-type sequences at the terminals constrain phylogenetic inferences in such a way that the implied ancestors still largely fit the three-dimensional structure. The results were qualitatively similar for all energy functions though only the results of the Jones energy function are reported here. All branches are scaled to a relative branch length of 1.0. Absolute branch lengths for sequence pairs are reported in Table 3. Edge lengths for all other trees can be found in the online Supplementary material. In the plots, black dots adjacent to the y-axis represent scores of reference wild-type sequences, and the horizontal lines indicate the minimum and maximum of these scores.

Sequence Pairs

Constrained simulations.—We first describe the results for the simulations on the branches that connect the sequence pairs listed in Table 2. Under the Poisson model, there was an obvious decrease in sequence-to-structure fit at the center of the branch (Fig. 2, left column). However, because the simulation was conditional on ending in the amino acid state of the second sequence, the energies and Z-scores (online Supplementary material) recovered toward the end of the branch. It seems reasonable to assume that in nature the ancestral states should be no less compatible with the structure than the terminals. Thus, we interpret an increase in the predicted PE of sequences as one approaches the middle of the branch as evidence for model misspecification. These sequences are more heavily influenced by the model than sequences near the tips because only sequences a few substitutions away from the empirically observed sequences will be seen near the end points of a branch (when the end points are constrained to match a known sequence). For WAG and LG (Fig. 2, middle and right columns, respectively), the plots resemble what one would expect from a structurally aware model (e.g., Robinson et al. 2003; Rodrigue et al. 2005)—ancestral sequences that are all largely compatible with the crystal structure. Figure 2 also shows the median and spread of the sample. For all models and data sets, the spread of the distribution was wider in the middle of the branch. It was also wider for the Poisson model than for the empirical models.

Unconstrained simulations.—We now describe the results for the simulations that were started from the same wild-type sequences as above but which were not

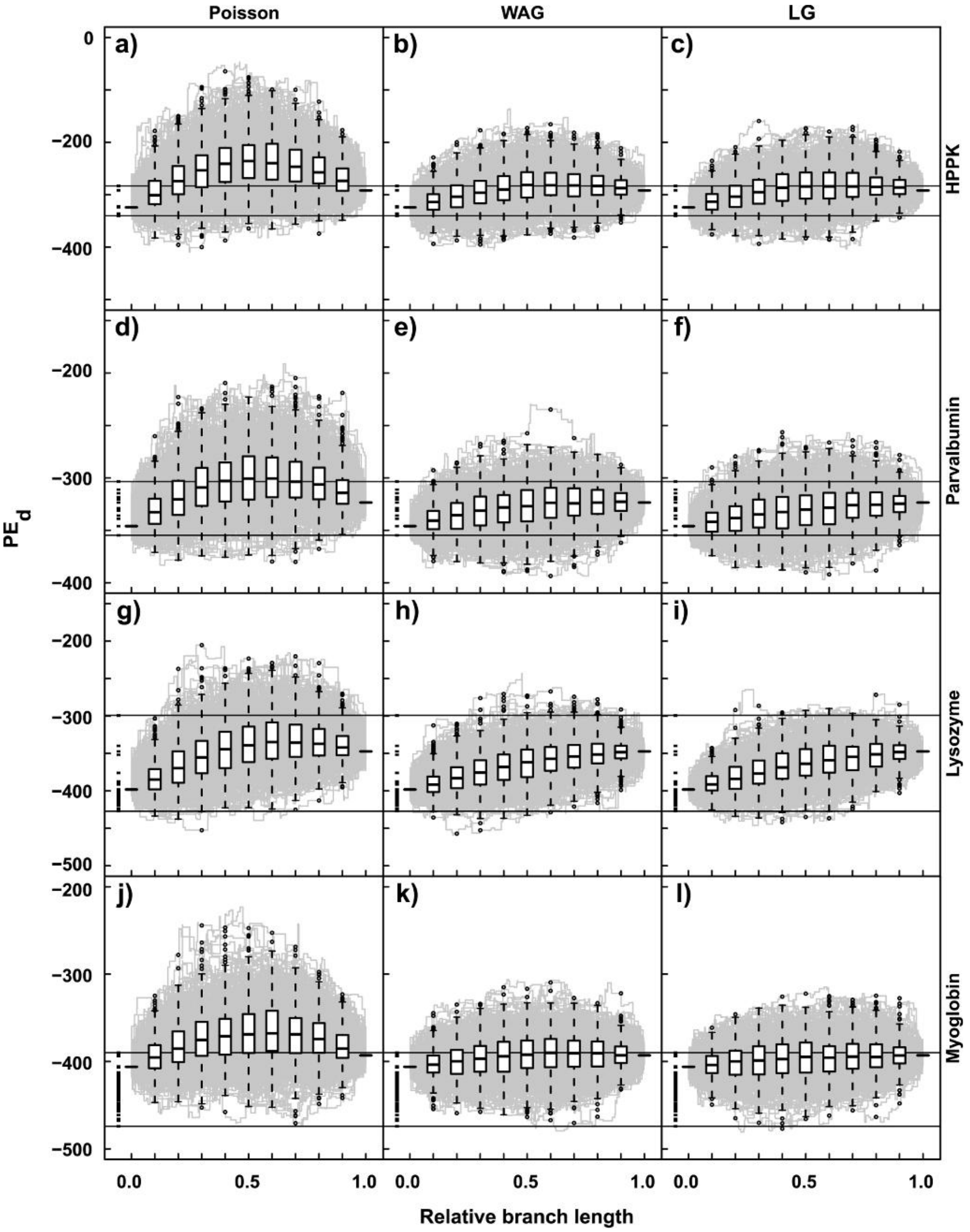

FIGURE 2. Pseudo-energies (PE_d) of 500 transition paths (gray). Each path consists of a series of one-step neighbors that connect the terminal sequences (Table 2). Black dots adjacent to the y-axis represent PE_d scores of reference wild-type sequences, and the black horizontal lines traversing each of the plots indicate the minimum and maximum of these scores. Branch lengths are scaled to 1.0, absolute values are listed in Table 3. Each row shows the results for a single protein: (a-c) HPPK, (d-f) parvalbumin, (g-i) lysozyme, and (j-l) myoglobin. Under the Poisson model (a, d, g, j), ancestors that were closer to the terminals were compatible with the structure, but the fit deteriorated toward the center of the branch. This effect was also observed under the WAG (b, e, h, k) and LG models (c, f, i, l), but to a lesser extent. For all models, the PE_d scores appeared to be "rescued" by the terminals. The spread was greater for the Poisson model than for the two empirical models. The box plots summarize the values at regular intervals along the relative branch lengths. Whiskers extend to the most extreme data point still within 1.5 times the interquartile range from the box. See text for details.

constrained to end in the second wild-type sequence after the same expected number of substitutions per site. This approach is commonly used to simulate phylogenetic data sets. Figure 3 shows that the average structural compatibility of the mutated sequences deteriorated steadily. The slope of the curve depended on the model: sequences obtained by introducing substitutions according to the Poisson model (Fig. 3, left column) lost their structural compatibility more quickly than when the WAG or LG models were used (middle and right columns, respectively). This pattern was also reflected in both the shuffling and the decoy Z-scores (online Supplementary material). This is in clear contrast to the results above where the simulations were

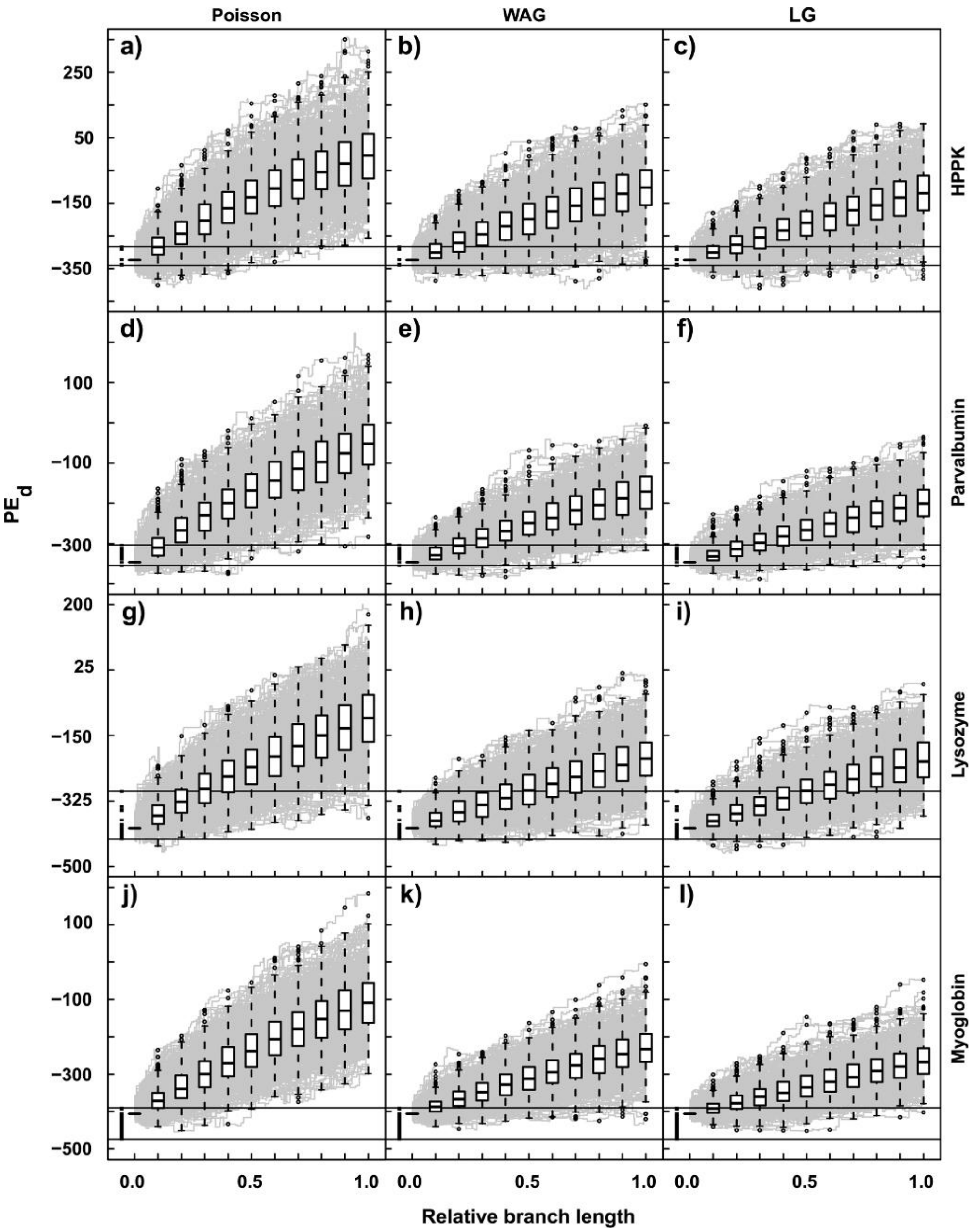

FIGURE 3. Pseudo-energies for 500 mappings where the end state was not constrained (gray). Each path consists of a series of one-step neighbors starting from the rst sequence of the pairs listed in Table 2. Under the Poisson model (a, d, g, j), sequence-to-structure fit deteriorated quickly. The same effect was observed for WAG (b, e, h, k) and LG (c, f, i, l). The slope was less steep under WAG and LG, which could indicate that these models are a better description of the real substitution process than Poisson. The spread was greater for the Poisson model than for WAG and LG. The situation displayed here reflects the structural implications of simulating data sets under such a model. As in Figure 2, black dots adjacent to the y-axis represent PE$_d$ scores of reference wild-type sequences, the horizontal lines indicate the minimum and maximum of these scores. Branch lengths are scaled to 1.0, absolute values are listed in Table 3, and whiskers extend to the most extreme data point still within 1.5 times the interquartile range from the box. See text for details.

"anchored" at both end points. Even when an empirical substitution model was used, the sequences quickly lost their compatibility with the structure. This means that good phylogenetic i.i.d. models may be appropriate for interpolating structurally viable ancestors between wild-type sequences, but that the situation is quite different for extrapolating what a sequence would look like after expected 0.5 substitutions per site.

Unbounded simulations were also performed on a longer branch that corresponded to 30 substitutions per site. When the substitution process is saturated, all amino acids occur according to their stationary state frequencies in the model at each sequence position. As the sequences reach a composition that is consistent with the stationary state frequencies of the model, the shuffling Z-scores are expected to approach zero: a random

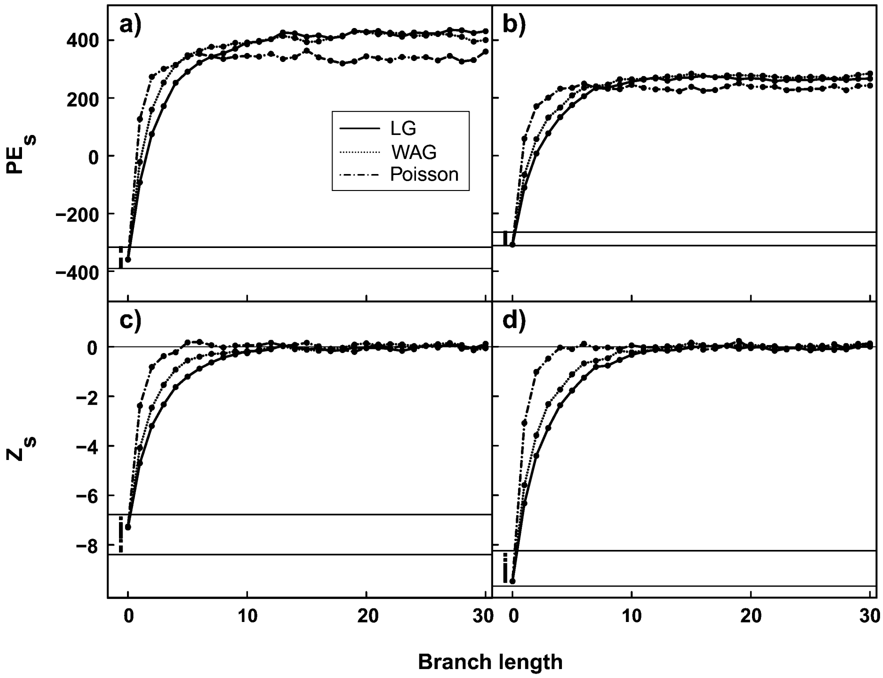

FIGURE 4. Average PE_s (a, b) and Z_s scores (c, d) for 100 samples from the posterior distribution of mappings where the end state was not constrained are shown for HPPK (a, c) and parvalbumin (b, d). The absolute branch length is 30 expected substitutions per site. At the end of the branch, the substitution process is saturated and all amino acids occur according to their stationary state frequencies in the model at each sequence position. The PE_s scores equilibrated at a value that depended on the substitution model and the length of the protein. As expected, when the composition corresponded to the stationary state frequencies of the model, the Z_s scores approached zero, irrespective of substitution model and sequence length.

sequence should not fit the native structure any better or worse on than any shuffled versions of that sequence. The mean values illustrated in Figure 4 show that the Z_s scores indeed approached zero for all models. At the point where the sequences were essentially random draws from the equilibrium frequencies of the model, all sequences were highly nonspecific. Similarly, the PE scores are expected to reach an equilibrium once the substitution process has reached saturation. It is clear from Figure 4 that this value depended on the substitution model (and the length of the protein).

Additional Taxa, Alternative Topologies

Our results for trees with three and four taxa clearly demonstrate the implications of improved taxon sampling. Additional sequences that are known to fold into the same structure improved the structural compatibility of the transition paths. In our examples, this held true despite the expected increase in patristic distances between the original sequence pairs. The added sequences provided information about other residues that may be observed at variable positions and thus informed the model about the possible nature of the unobserved substitutions. For all our three-taxon data sets, the transition paths were highly compatible with the structure when an empirical model was used (see online Supplementary material).

Figure 5 shows the improvement in PE_s scores for HPPK + Poisson for three taxa (b–e) compared with the initial two taxa (a). Although the sum of the branches between *Vibrio* and *Salmonella* increased (b), the PE_s scores of the ancestral sequences improved after adding *Aeromonas*. Panels (c) and (d) show the PE_s scores along the *Vibrio* and *Salmonella* branches, respectively. Panel (e) shows the PE_s scores for the branch leading to the newly added taxon *Aeromonas*.

On the four-taxon tree (f–k), where the fourth sequence (*Pectobacterium*) breaks the still relatively long branch to *Salmonella*, the sequence-to-structure fit of the substitution histories was improved further. Panels (f, g, h) show the PE_s scores along the path from *Vibrio* (f) to *Salmonella* (h). PE_s scores for the branches leading to *Aeromonas* and *Pectobacterium* are plotted in panels (i) and (k), respectively.

Because likelihood-based phylogenetic inference techniques base their calculations on partial likelihoods associated with ancestral nodes, the result that simple i.i.d. models can produce reasonable ancestral sequences is encouraging. The models are clearly not perfect, but they may be "good enough" to produce accurate tree inference. However, effects of unreasonable evolutionary trajectories on topology estimation are difficult to generalize. Although the mappings on all the ML trees (which were also the biologically most reasonable trees) showed improved sequence:structure fit, this effect was mild when compared with the alternative tree

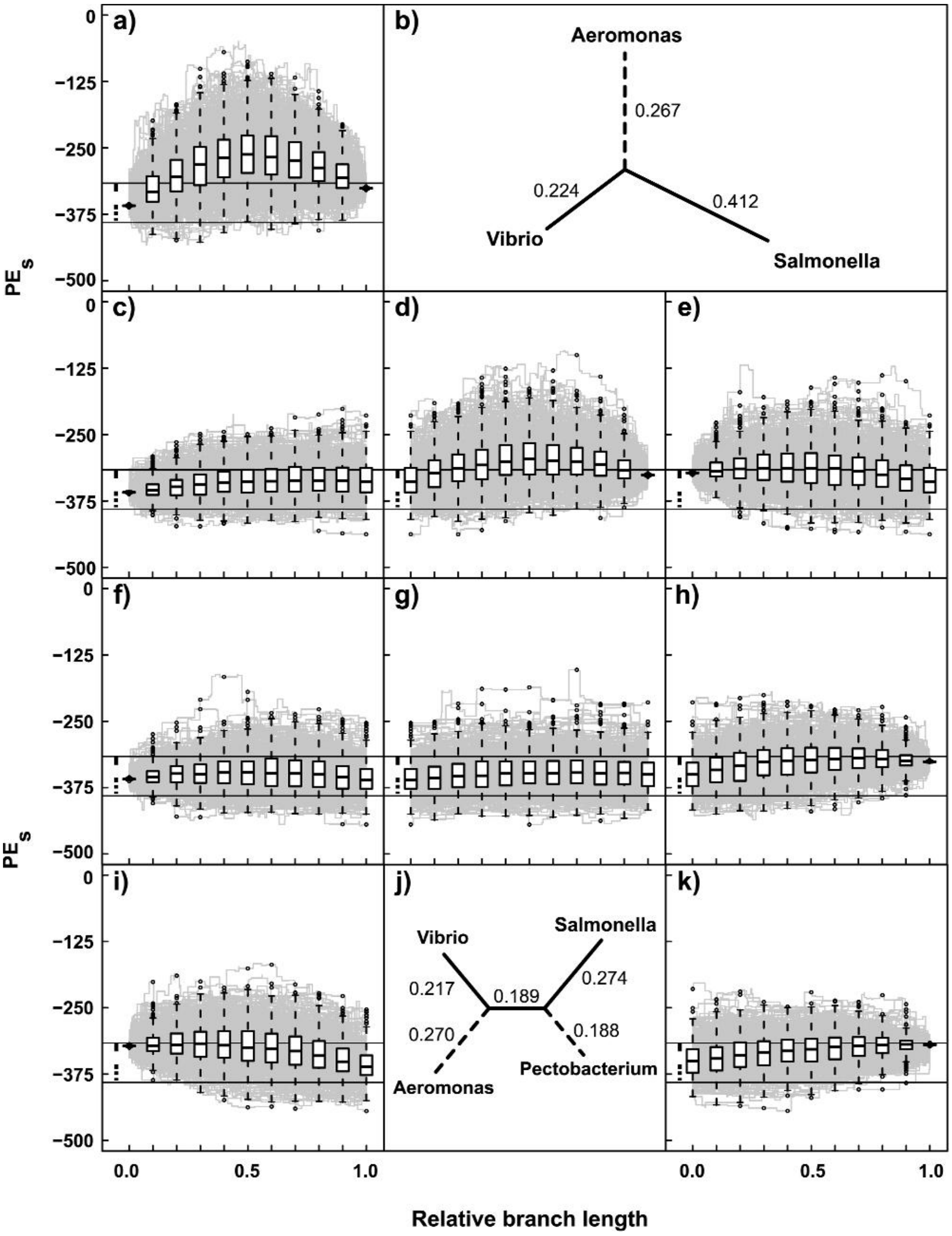

FIGURE 5. The effect of breaking long branches through improved taxon sampling for HPPK under the Poisson model. a) The case for two taxa: the PE$_S$ scores for the *Salmonella-Vibrio* branch (ML branch length 0.609 expected substitutions per site) are shown. (b-e) The case for three taxa: panel (b) illustrates the ML branch lengths when the *Aeromonas* sequence was added to the two-taxon data set. PE$_S$ scores for the branches from the internal node to *Vibrio* (c), *Salmonella* (d), and *Aeromonas* (e) are shown. (f-k) The case for four taxa: panel (j) depicts the ML tree with branch lengths when *Pectobacterium* was added to the three-taxon data set. The remaining panels show the PE$_S$ scores for the internal branch (g) and the branches leading to *Vibrio* (f), *Salmonella* (h), *Aeromonas* (i), and *Pectobacterium* (k). Even though the patristic distance between *Vibrio* and *Salmonella* increased as taxa were added, the sequence-to-structure fit of the transition paths improved. The additional sequences informed the model about the nature of the unobserved changes. In keeping with previous gures, black dots adjacent to the y-axis represent PE$_S$ scores of reference wild-type sequences, the horizontal lines indicate the minimum and maximum of these scores. Branch lengths are scaled to 1.0, and whiskers extend to the most extreme data point still within 1.5 times the interquartile range from the box. See text for details.

topologies. As a general result, we observed that for trees that were based on the same data and the same model, structural compatibility simply tended to decrease on longer branches. In our examples, the trees with lower ML scores had extremely short internal branches and long terminal edges. It is conceivable that an incorrect topology comprising short branches may induce a higher percentage of structurally viable reconstructions than a correct topology. Preferring tree estimates based on structural compatibility could unduly

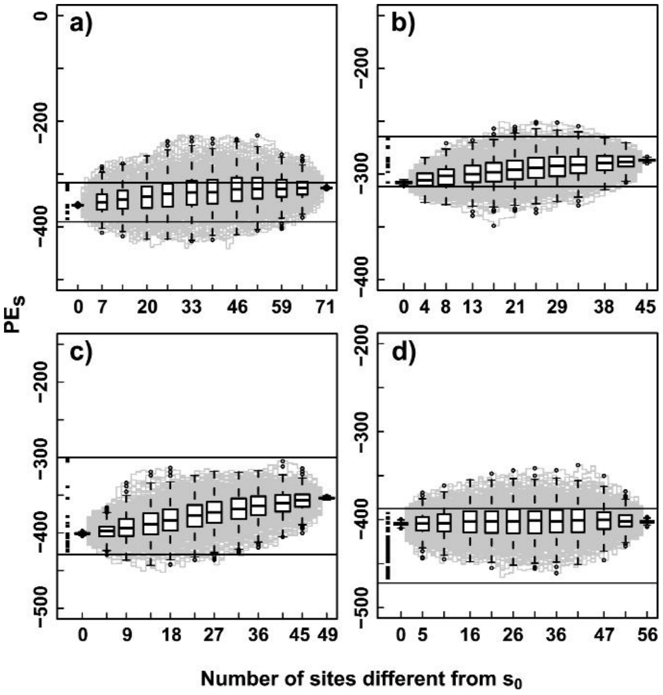

FIGURE 6. PE_s scores of 500 most parsimonious mappings. a) HPPK. b) Parvalbumin. c) Lysozyme. d) Myoglobin. The box plots summarize the values at regular intervals and indicate that the sequence-to-structure fit did not systematically get better (or worse) toward the center of the branch. Whiskers extend to the most extreme data point still within 1.5 times the interquartile range from the box.

favor tree estimates with short branch lengths. However, methods (such as parsimony) that place a premium on short tree lengths are prone to tree estimation errors.

Figure 6 shows a sample of 500 most parsimonious paths connecting s_0 and s_1 (Table 2). Sequence-to-structure fit did not systematically get worse (or better) toward the center of the branch, and the most parsimonious paths were mostly comprised of sequences that fitted the structure well. This suggests that the energetically unfavorable sequences from the model-based mappings were primarily due to amino acids that were not observed at any of the terminals (at the corresponding sites). In the remainder, we will refer to these amino acids as “extra” or “unobserved” amino acids.

To further investigate the importance of the states the model suggested for the hidden substitutions, we grouped individual sequences from the mappings by the number of extra amino acids they contained (Fig. 7). For two taxa, it is apparent that for both Poisson (a) and LG (d), the number of extra amino acids is negatively correlated with sequence:structure fit. From a structural perspective, LG clearly resulted in better choices than Poisson.

For LG, and to a lesser extent Poisson, the PE_s scores of sequences with the same number of extra amino acids improved when taxa were added (b, e: three taxa. c, f: four taxa). Adding evenly sampled sequences added information about the capacity of a site to tolerate changes. For two taxa, the capacity of a site to tolerate substitutions may be over- or underestimated: some sites appeared variable with two taxa, but the added taxa implied that the site was actually evolutionarily conserved (and therefore perhaps structurally important). Thus, for more taxa mappings at these sites underwent fewer substitutions to extra amino acids. On the other hand, some sites appeared conserved for two taxa but were variable for four taxa for which mappings often involved more substitutions to unobserved residues. These are likely to be sites that are more tolerant to changes. The PE_s scores presented in the upper row of Figure 7 correspond to the ones shown in Figure 5.

It can also be seen that, across the tree, more extra amino acids co-occurred in sequences on the two-taxon tree than on the four-taxon tree. This was partly because there were now three (panels b and e) or four (c and f) choices at each site, but it was likely also a consequence of tree structure. Even though, overall, there were more hidden changes on the bigger tree, many of them occurred at only a few very variable sites dispersed across the tree. Others preferentially occurred on different branches of the tree. For four taxa, we used the mappings on the ML trees whose topology was the same for all models. See Fig. 5 (j) for Poisson and online Supplementary material for LG branch lengths.

DISCUSSION

Our study suggests that wild-type protein sequences that are as little as 55% identical can constrain phylogenetic likelihood inference to implied ancestors that are largely compatible with the three-dimensional structure. This is especially true for some of the more recently proposed empirical amino acid models.

In our examples, given the same data and inference model, structural compatibility of the mappings was negatively correlated with branch length. If a conventional model were to be simply guided by structural compatibility, it would preferentially assign higher probabilities to trees without long edges. This could potentially exacerbate phylogenetic pathologies like long branch attraction. This problem is most pronounced in maximum parsimony. The observation that most parsimoniuous ancestral reconstructions of distantly related sequences are often found to be compatible with the structure is likely not because they are correct, but rather because they are combinations of states that are observed at the terminals, which are inherently predisposed to be structurally viable (Williams et al. 2006). We believe that likelihood models can be improved by incorporating site-specific constraints imposed by the local biochemical environment directly into the substitution process (e.g., Le and Gascuel 2010).

A number of empirical studies have found ancestral protein sequences inferred with maximum parsimony

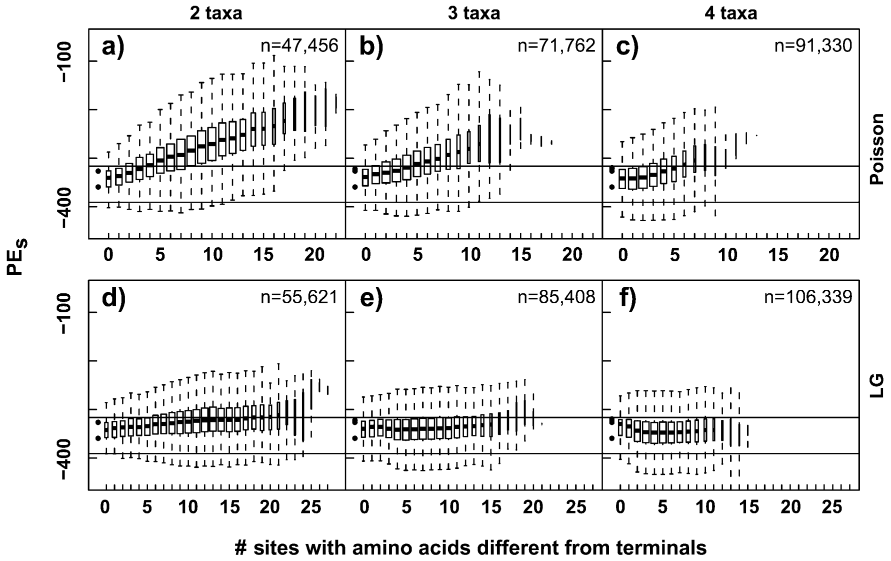

FIGURE 7. Individual sequences from the 500 mappings were grouped by the number of positions whose amino acids differed from those observed at any of the terminals ("extra" amino acids) and plotted against their PE$_s$ scores. (a, d) Two taxa. (b, e) Three taxa. (c, f) Four taxa. Both Poisson (a, b, c) and LG (d, e, f) showed a negative correlation between sequence-to-structure fit and the number of extra amino acids in a sequence. However, from a structural perspective, LG suggested more favorable substitutions than Poisson. As taxa were added, sequences with the same number of extra amino acids tended to exhibit better sequence-to-structure fits. In addition, there was a tendency for the number of extra amino acids to diminish as the number of taxa was increased. These issues are further developed in the text. Box widths are proportional to the square roots of the number of sequences in a group, and *n* denotes the total number of sequences collectively implied over all 500 mappings. For four taxa, we used the ML trees whose topology was the same for both models. Whiskers extend to the most extreme data point still within 1.5 times the interquartile range from the box (outliers not shown). Results are only shown for HPPK; similar patterns were seen for the other data sets examined.

or ML methods to be functional (e.g., Chang et al. (2002)). The accuracy of ancestral-state inference methods has been a subject of critical evaluation (e.g., Collins et al. 1994; Krishnan et al. 2004; Pollock and Chang 2007). Williams et al. (2006) simulated thermodynamically stable sequences on a phylogeny and subsequently compared the structural compatibility of reconstructed ancestral sequences with the "true" simulated ancestral sequences. They found that parsimony and ML reconstructions were biased toward higher thermostability. In our study, different draws from the posterior distribution of mappings tended to result in sequences with reasonable free energies (compared with extant sequences). This result is in line with what Williams et al. (2006) observed for Bayesian reconstructions of ancestral protein sequences. Our results also suggest that most parsimonious reconstructions are likely to be structurally viable. However, even if this holds up under a more discerning measure of sequence to-structure-fit, and if these sequences would indeed turn out to be functional, realistic interpolation of unobserved states can only be accomplished with model-based methods. Ideally, such a model would be able to predict states that should be *expected* at certain positions (which may later be supported when new sequences are added to the tree).

Perhaps counterintuitively, preliminary results for four-taxon trees suggest that even though accounting for rate hetogeneity improves the likelihood, it may be slightly detrimental for sequence-to-structure compatibility. This effect was very mild and needs further investigation but can be explained by the fact that rate variation led to increased branch lengths and thus more multiple substitutions (see Fig. 35 of the online Supplementary material, results were similar for all trees and proteins). Site-specific rates were estimated with PAML (using a discretized gamma distribution with 30 categories) with the empirical Bayesian approach described in Yang and Wang (1995). If this is corroborated by further work, then the better fit to the data associated with gamma distributed rates may result in a more realistic description of the number of changes but compromise the structural integrity of the molecule. This is a consequence of the discrepancy between model assumptions and the criteria used to evaluate the biological integrity.

To rigorously test if a sequence would fold into a given structure would require much more careful evaluation than the relatively simple residue-based energy functions used here. It is possible that the current approach is too coarse-grained and thus not discerning enough to pick up the negative fitness effects of some of the substitutions. If this is the case, then residue-based contact potentials may not be sufficiently powerful to significantly improve phylogenetic inference. All atom energy functions have been shown to be more accurate; however, they require computationally expensive modeling of protein structure. The burden introduced by computing transition rates between full sequences is already prohibitive for topology inference (the transition kernel for amino acid sequences becomes a $20^N \times 20^N$ matrix). As stated by Koehl and Levitt (1999b) to rigorously address this problem, simultaneous exploration of both the sequence space and the conformation space is required. Evaluating all sequences for their specificity is an additional expensive step, which has so far not been implemented in any of the structure-aware phylogenetic models.

Clearly, the degree of sequence divergence was limited in our study. It is reasonable to assume that structure-aware models could potentially improve inferences for very distantly related sequences. However, the applicability of such models for highly divergent data sets is met by a few serious obstacles. Above all, sequence-length polymorphism is difficult to accommodate because it invariably leads to unequal lengths of the chosen representative structure and some members of the data set. In phylogenetic analyses, *missing data* are commonly treated as *ambiguous data*. This means that gaps could be "filled" with any of the residues with equal probability. However, it would clearly be inappropriate to introduce a "dummy" residue for every gap when calculating sequence-to-structure energies. When sequences are evaluated for their structural fitness, they need to be threaded onto the crystal structure. Sometimes gaps are modeled as an additional character state (e.g., Rivas and Eddy 2008) but this is controversial because the molecular evolutionary causes of insertion and deletion are diverse and need to be modeled accordingly.

Furthermore, computationally expensive (and potentially inaccurate) modeling may be needed to "insert" or "delete" residues from the structure. Decisions about how to deal with additional sites will have to be made if structure-aware models are used for more divergent data sets. Choi et al. (2008) recently addressed the problem of calculating equilibrium frequencies for sequences of different lengths in the context of residue interdependence, but the structural aspect of the problem remains. It seems plausible that the analysis could be restricted to subsets of the structure that are not interacting with each other (this problem is related to the fact that interactions with other proteins are not considered by looking at the crystal structure alone). It should be noted that for deeper divergences another fundamental assumption, that the overall structure remains constant across the subtree, may be violated.

It has been demonstrated (e.g., Hillis 1998; Pollock et al. 2002; Zwickl and Hillis 2002; Hillis et al. 2003) that dense taxon sampling significantly improves phylogenetic estimates. Our study underscores the importance of dense taxon sampling by providing evidence that the data at the terminals have an important effect on the structural viability of the implied ancestors. In general, for shallow divergences, the inference will mainly be driven by the data. However, as sequences become more divergent, the model becomes more and more important because it provides a weighting for the different paths the evolutionary process could have taken through sequence space. Our results show that simple empirical models alone are not good enough to allow for reliable extrapolation. If sequences are truly separated by a large number of substitutions, the capability to interpolate realistic sequences along the internodes is likely to be similarly compromised.

Finally, we think that it is important to further pursue the incorporation of structural information into phylogenetic models—especially when simulating realistic protein sequences. In principle, models that accurately evaluate the fitness of complete sequences should be the most realistic attempt to explain molecular evolution. Given the complexities of such models and the good performance of simpler models demonstrated in this study, we suspect that a directed effort in modeling site-specific constraints imposed by local biochemistry is a promising avenue for phylogenetic tree inference.

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found at http://www.sysbio.oxfordjournals.org/.

Funding

This work was supported by a Marie Curie Fellowship to C. L. and National Science Foundation award DEB 1036500 to G. J. P. N.

Acknowledgments

We thank Tracy Heath for helpful comments on an early version of the manuscript and Douglas Robinson for discussions about the Jones energy function. The insightful comments and suggestions of Olivier Gascuel, Richard Goldstein, Andrew Roger, Fredrik Ronquist, and an anonymous reviewer greatly improved this article. Sequence-to-structure compatibility computations were carried out at the Florida State University shared High-Performance Computing facility, and graphs were generated in R (http://www.R-project.org/).

References

Adachi J., Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Biol. 42(4): 459–468.

Bastolla U., Farwer J., Knapp E.W., Vendruscolo M.H. 2001. How to guarantee optimal stability for most representative structures in the protein data bank. Proteins. 44(2):79–96.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. 2000. The protein data bank. Nucleic Acids Res. 28(1):235–242.

Bishop M.J., Friday A.E. 1987. Tetrapod relationships: the molecular evidence. In: Patterson C., editor. Molecules and morphology in evolution: conflict or compromise? Cambridge, UK: Cambridge University Press. p. 123–140.

Bowie J.U., Lüthy R., Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science. 253(5016):164–170.

Cao Y., Janke A., Waddell P.J., Westerman M., Takenaka O., Murata S., Okada N., Pääbo S., Hasegawa M. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. 47(3):307–322.

Chang B.S.W., Jönsson K., Kazmi M.A., Donoghue M.J., Sakmar T.P. 2002. Recreating a functional ancestral archosaur visual pigment. Mol. Biol. Evol. 19(9):1483–1489.

Chiu T.L., Goldstein R.A. 1998. Optimizing potentials for the inverse protein folding problem. Protein Eng. 11(9):749–752.

Choi S.C., Redelings B.D., Thorne J.L. 2008. Basing population genetic inferences and models of molecular evolution upon desired stationary distributions of DNA or protein sequences. Philos. Trans. R Soc. Lond., B. Biol. Sci. 363(1512):3931–3939.

Collins T., Wimberger P., Naylor G.J.P. 1994. Compositional bias, character state bias, and character state reconstruction using parsimony. Syst. Biol. 43(4):482–496.

Das R., Baker D. 2008. Macromolecular modeling with Rosetta. Annu. Rev. Biochem. 77:363–382.

Dayhoff M.O., Schwartz R.M., Orcutt B.C. 1978. A model for evolutionary change in proteins. In: Dayhoff, M.O., editor. Atlas of protein sequence and structure. Washington (DC): National Biomedical Research Foundation, p. 5:345–352.

Henikoff S., Henikoff J.G. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89(22): 10915–10919.

Hillis D.M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.

Hillis D.M., Pollock D., McGuire J., Zwickl D. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.

Huelsenbeck J.P., Joyce P., Lakner C., Ronquist F. 2008. Bayesian analysis of amino acid substitution models. Phil. Trans. R Soc. B. 363:3941–3953.

Jensen J.L., Pedersen A.M.K. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. 32:499–517.

Jones D.T. 1998. THREADER: protein sequence threading by double dynamic programming. In: Salzberg, S., Searls D., Kasif S., editors. Computational methods in molecular biology. Chapter 13. Amsterdam (The Netherlands): Elsevier Science, p. 285–312.

Jones D.T. 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287(4):797–815.

Jones D.T., Taylor W.R., Thornton J.M. 1992a. A new approach to protein fold recognition. Nature. 358(6381):86–89.

Jones D.T., Taylor W.R., Thornton J.M. 1992b. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8(3):275–282.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kim D.E., Chivian D., Baker D. 2004. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32(Web Server issue):W526–W531.

Koehl P., Levitt M. 1999a. De novo protein design. I. In search of stability and specificity. J. Mol. Biol. 293(5):1161–1181.

Koehl P., Levitt M. 1999b. De novo protein design. II. Plasticity in sequence space. J. Mol. Biol. 293(5):1183–1193.

Koshi J.M., Goldstein R.A. 1998. Models of natural mutations including site heterogeneity. Proteins. 32(3):289–295.

Koshi J.M., Mindell D.P., Goldstein R.A. 1997. Beyond mutation matrices: physical-chemistry based evolutionary models. Genome. Inform. Ser. Workshop Genome. Inform. 8:80–89.

Krishnan N.M., Seligmann H., Stewart C.-B., de Koning A.P.J., Pollock D.D. 2004. Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. Mol. Biol. Evol. 21(10):1871–1883.

Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25(7):1307–1320.

Le S.Q., Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst. Biol. 59(3):277–287.

Lio P., Goldman N. 1999. Using protein structural information in evolutionary inference: transmembrane proteins. Mol. Biol. Evol. 16(12):1696–1710.

Lio P., Goldman N., Jones D.T. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics. 14(8):726–733.

Mateiu L., Rannala B. 2006. Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. Syst. Biol. 55(2):259–269.

Meller J., Elber R. 2001. Linear programming optimization and a double statistical filter for protein threading protocols. Proteins. 45(3):241–261.

Misura K.M.S., Baker D. 2005. Progress and challenges in high-resolution refinement of protein structure models. Proteins. 59(1): 15–29.

Nielsen R. 2001. Mutations as missing data: inferences on the ages and distributions of nonsynonymous and synonymous mutations. Genetics. 159(1):401–411.

Pedersen A.M.K., Jensen J.L. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. 18(5):763–776.

Pollock D., Taylor W.R., Goldman N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. 287(1):187–198.

Pollock D., Zwickl D., McGuire J., Hillis D.M. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51(4):664–671.

Pollock D.D., Chang B.S.W. 2007. Dealing with uncertainty in ancestral reconstruction: sampling from the posterior distribution. In: Liberles D.A., editor. Ancestral sequence reconstruction. Chapter 8. Oxford: Oxford University Press. p. 85–94.

Pollock D.D., Taylor W.R. 1997. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. Protein Eng. 10(6):647–657.

Rivas E., Eddy S.R. 2008. Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput. Biol. 4(9):e1000172.

Robinson D., Jones D.T., Kishino H., Goldman N., Thorne J.L. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20(10):1692–1704.

Rodrigue N., Lartillot N., Bryant D., Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 347(2):207–217.

Rodrigue N., Philippe H., Lartillot N. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol. Biol. Evol. 23(9):1762–1775.

Rodrigue N., Philippe H., Lartillot N. 2007. Exploring fast computational strategies for probabilistic phylogenetic analysis. Syst. Biol. 56(5):711–726.

Rodrigue N., Philippe H., Lartillot N. 2008. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics. 24(1):56–62.

Rohl C.A., Strauss C.E.M., Misura K.M.S., Baker D. 2004. Protein structure prediction using Rosetta. Meth. Enzymol. 383:66–93.

Sanderson M.J., Donoghue M.J., Piel W., Eriksson T. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Am. J. Bot. 81(6): 183.

Simons K.T., Kooperberg C., Huang E., Baker D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. 268(1):209–225.

Thorne J.L. 2007. Protein evolution constraints and model-based techniques to study them. Curr. Opin. Struct. Biol. 17(3):337–341.

Wang Z.O., Pollock D.D. 2005. Context dependence and coevolution among amino acid residues in proteins. Meth. Enzymol. 395:779–790.

Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18(5):691–699.

Williams P.D., Pollock D.D., Blackburne B.P., Goldstein R.A. 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput. Biol. 2(6):e69.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10(6):1396–1401.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39(3):306–314.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24(8):1586–1591.

Yang Z., Wang T. 1995. Mixed model analysis of DNA sequence evolution. Biometrics. 51(2):552–561.

Yang Z., Nielsen R., Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15(12):1600–1611.

Zwickl D., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51(4):588–598.