Modeling Structural Constraints on Protein Evolution via Side-Chain Conformational States

Umberto Perron,[1] Alexey M. Kozlov,[2] Alexandros Stamatakis,[2,3] Nick Goldman,*[,1] and Iain H. Moal*[,†,1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire, United Kingdom

[2]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

[3]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

[†]Present address: Computational and Modelling Sciences, GlaxoSmithKline Research and Development, Stevenage, United Kingdom

*__Corresponding authors:__ E-mails: goldman@ebi.ac.uk; moal@ebi.ac.uk.

Associate editor: Tal Pupko

Abstract

Few models of sequence evolution incorporate parameters describing protein structure, despite its high conservation, essential functional role and increasing availability. We present a structurally aware empirical substitution model for amino acid sequence evolution in which proteins are expressed using an expanded alphabet that relays both amino acid identity and structural information. Each character specifies an amino acid as well as information about the rotamer configuration of its side-chain: the discrete geometric pattern of permitted side-chain atomic positions, as defined by the dihedral angles between covalently linked atoms. By assigning rotamer states in 251,194 protein structures and identifying 4,508,390 substitutions between closely related sequences, we generate a 55-state "Dayhoff-like" model that shows that the evolutionary properties of amino acids depend strongly upon side-chain geometry. The model performs as well as or better than traditional 20-state models for divergence time estimation, tree inference, and ancestral state reconstruction. We conclude that not only is rotamer configuration a valuable source of information for phylogenetic studies, but that modeling the concomitant evolution of sequence and structure may have important implications for understanding protein folding and function.

Key words: **molecular evolution, phylogenetic estimation, phylogenetics, protein evolution, protein structure, rotamer, substitution model.**

Introduction

The development of evolutionary models is a prerequisite (albeit sometimes an implicit one) for many common bioinformatics tasks such as recognition of homologous sequences, phylogenetic tree estimation, evolutionary hypothesis testing, and protein structure prediction (Huelsenbeck and Rannala 1997; Felsenstein 2004; Koonin 2005; Ginalski 2006). Because of this, the development and improvement of model-based approaches to studying protein evolution is an area of research where advances have wide-spread benefits. Furthermore, high-resolution structural information from a variety of techniques is now available for large numbers of proteins and molecular assemblies (Milne et al. 2013; Carroni and Saibil 2016; Venien-Bryan et al. 2017), improving our understanding of how protein folding and residue function change over time.

Empirical Models of Amino Acid Replacement

When studying the evolution of amino acid sequences, substitutions are usually described using a continuous-time Markov model with the 20 amino acids as the states of the chain (Liò et al. 1998; Felsenstein 2004; Thorne and Goldman 2007; Perron et al. forthcoming). Models belonging to the empirical class are built by analyzing large quantities of sequence data (typically hundreds of protein alignments) and estimating relative substitution rates between all state (amino acid) pairs under a time-reversible model. Empirical models are typically assumed to be applicable to broad classes of proteins with little further parameter optimization aside from techniques that match amino acid frequencies to what is observed in a specific data set under study and allow for rate heterogeneity amongst sequence sites (Yang 1993).

The first empirical amino acid substitution model was introduced by Dayhoff and coworker in 1966 (Eck and Dayhoff 1966) and updated regularly until 1978 (Dayhoff et al. 1978). They compiled protein sequence alignments and tabulated amino acid substitutions along branches on the phylogenetic trees; from these data, an instantaneous rate matrix Q defining a continuous-time Markov model can be constructed (Kosiol and Goldman 2005). Since then, a number of alternative Q-matrices for amino acid substitution have been developed using similar approaches but taking advantage of more powerful model estimation techniques, and larger or more specific data sets. Examples include MTMAM (Yang et al. 1998) for mammalian mitochondrial proteins and GPCRtm for the transmembrane region of GPCRs (Rios et al. 2015); WAG (Whelan and Goldman 2001) and LG

(Le and Gascuel 2008) that were derived from larger, diverse databases of sequence alignments; and LG4X (Le et al. 2012) that aims to capture varying evolutionary dynamics at different sequence sites.

Role of Structural Information in Understanding Protein Evolution

Although considering particular proteins' distinct amino acid compositions and among-site rate variation improves a model's statistical fit to the data indicating a better description of evolutionary patterns, it seems clear, from considering protein structure and function, that at least some variability in the evolutionary process will be associated with the structural environment of a site (Thorne and Goldman 2007; Perron et al. forthcoming). For example, solvent-exposed residues evolve more rapidly than those buried in the protein interior ($\sim \times 2$) and exhibit different amino acid frequencies and substitution patterns, due to less steric constraint and the need to interact favorably with water. Similarly, secondary structure also influences substitutions, for instance disfavoring amino acids in α-helices that are incompatible with the canonical α-helical conformation due to disrupting backbone geometry or steric clashes arising from branching at the C^{β} atom. Models that account for some of these differences, for instance by using a separate 20-state model for different structural environments, have resulted in improved fit over those based on sequence alone (Goldman et al. 1998; Liò et al. 1998; Overington et al. 2008).

The tertiary and quaternary structures of proteins provide further constraints to evolution, in the form of natural selection operating on the specific interresidue interactions that stabilize the fold of the protein, as well as the need to avoid misfolding and aggregation (Overington et al. 1990; Shakhnovich et al. 1996). Although attempts to model some of these factors have been made (Bastolla et al. 2003, 2006; Robinson et al. 2003; Rodrigue et al. 2005, 2006; Arenas et al. 2017), this has proven difficult due to the challenging computational requirements of models that allow evolution at one position to be dependent upon the sequence at other positions. In addition to the difficulties that arise when site independence can no longer be assumed, these models are considerably more complex and require further assumptions such as 1) a constant tertiary structure, 2) approximate functions to map sequence to stability or misfolding propensity, and 3) additional approximate functions to map stability or misfolding propensity to rate effects. These approaches produce biologically plausible results and demonstrate the benefits of introducing explicit structural constraints to the evolutionary process. However, they are computationally demanding, difficult to use in an inference setting, and are not, unlike our model, readily integrated into commonly used software.

A substantially different approach to modeling how protein tertiary structure changes over evolutionary time was proposed by Challis and Schmidler (2012). In their approach, a protein's structural information is summarized via its α-carbon three-dimensional coordinates; the model then employs a time-reversible, continuous-time, and continuous-state Markov model to describe how the α-carbon coordinates constituting one protein can be transformed during evolution into the α-carbon coordinates that constitute a related protein. Implementations of the Challis–Schmidler model for phylogenetic inference purposes (Herman et al. 2014) are computationally tractable, at least for data sets of limited size, and have shown improvements over traditional models relying on sequence data alone. Specifically, the inclusion of structural information significantly reduced alignment and topology uncertainty and produced results that were more robust to the choice of data set.

One limitation of the Challis–Schmidler model is the assumption that, at equilibrium, the spatial locations of consecutive amino acids in a protein sequence would be independent of one another. In reality, such changes are correlated and constrained to specific torsional angles. Golden et al. (2017) have proposed a model of structural evolution that describes the evolution of protein tertiary structure using a specialized stochastic process that operates in dihedral angle space. The Golden model, although still quite different from traditional amino acid substitution models, is comparatively more realistic than previous stochastic models such as the Challis–Schmidler model and provides insights into the relationship between sequence and structure evolution.

An Evolutionary Model Based on Side-Chain Rotamer States

In this article, we present an evolutionary model that introduces structural information by accounting for the conformational state of each residue based on the atomic positions of its side-chain. Specifically, we split the traditional 20 amino acid states into discrete substates based on the χ_1 rotamer (short for "rotational isomer") configuration of their side-chains as defined in the Dunbrack rotamer library (Shapovalov and Dunbrack 2011). In this classification, each residue can adopt one of (typically) three discrete configurations (fig. 1 and supplementary fig. 1, Supplementary Material online) defined by the dihedral angle between the first two covalently linked carbons in the side-chain (C^{α} and C^{β}). These three stable rotamer configurations correspond to specific χ_1 dihedral angle values (~60°, ~−180°, and ~−60°) consistently across all residues; this means that residues sharing the same rotamer configuration (e.g., PHE1 and TRP1 in fig. 1) have side-chains that are similarly oriented with respect to the backbone. The adoption of one rotamer configuration over another is determined by their relative stability, a combination of the intrinsic stability of that state, local factors such as the backbone geometry and the position of atoms further along the side-chain, and the forces applied by the surrounding residues and the requirement to pack alongside them. Thus, they convey information about both the local structure as well as the interactions of the residue within the fold as a whole. As these states are discrete and finite, and each residue in a protein structure adopts exactly one χ_1 configuration, they can be readily incorporated into an expanded alphabet of amino acid states, maintaining the usual assumption of sitewise independence. This produces a

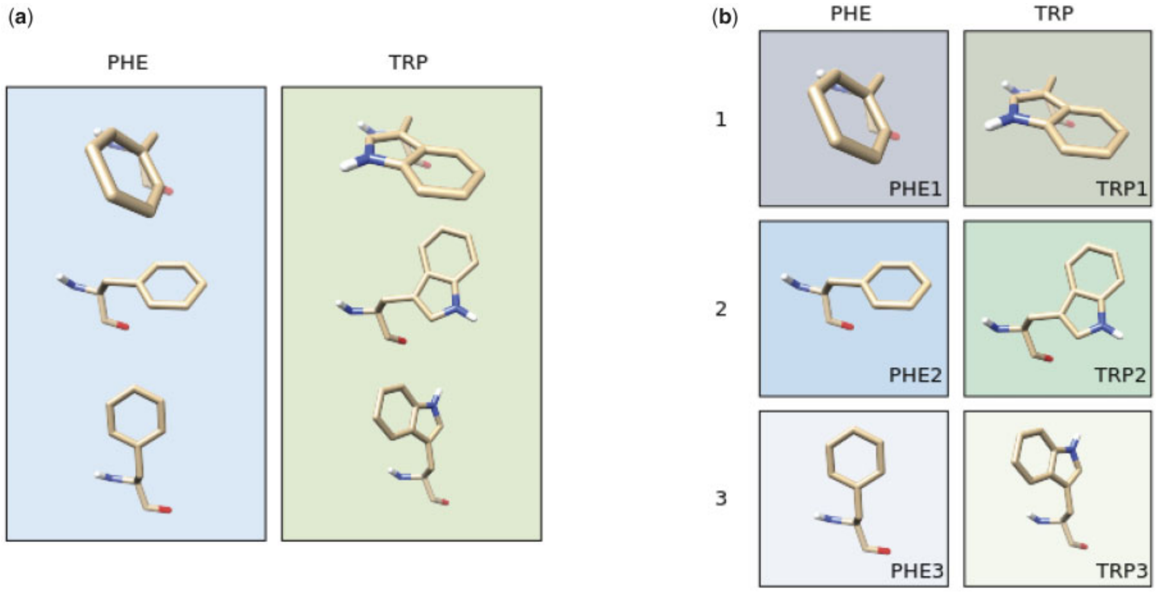

Fig. 1. Illustration of the rotamer configurations of phenylalanine (PHE) and tryptophan (TRP). (*a*) In traditional amino acid replacement models, their distinct χ_1 rotamer configurations are merged into a single amino acid state. (*b*) In our model, these states are split into three χ_1 configuration-specific states (1, 2, and 3) defined, as in the Dunbrack rotamer library, by the dihedral angle between the first two covalently linked carbons in the side-chain (C^α and C^β; see also supplementary fig. 1, Supplementary Material online).

scalable model that can be used in the same way as a traditional 20-state substitution model.

By compiling a large set of homologous sequences for which structural data are available, we develop a structurally aware "Dayhoff-like" substitution model based on an instantaneous rate matrix that uses an expanded state set composed of 55 states, each of which corresponds to the combination of a residue and its χ_1 configuration (table 1). Almost all amino acids show a significant, and often strong, conformational dependence in their substitution patterns, indicating that an amino acid can behave as a distinct entity depending on the orientation of its side-chain. Thus, our 55-state model (denoted "RAM55," for Rotamer-Aware Model) introduces valuable, biochemically plausible, structural information while retaining a classic architecture that can be readily implemented in widely used phylogenetic inference software such as RAxML-NG (Stamatakis 2014; Kozlov et al. 2019). This model improves our understanding of the relationships between protein sequence, structure, and evolution.

We further show that RAM55 results in a detectable improvement in model fit on simulated data, and on a number of diverse empirical data sets. It produces reliable tree topology and sequence divergence estimates. In addition, the RAM55 model also allows structurally aware reconstruction of both ancestral rotamer and amino acid states. This is of relevance to ancestral sequence reconstruction/resurrection used, for example, to investigate how the physical properties of proteins shaped their evolutionary process (e.g., Harms and Thornton 2013; Wheeler et al. 2016). We show that RAM55 can accurately reconstruct ancestral rotamer states from descendant protein sequences of known structure; it is also able to reconstruct ancestral amino acid states as well as or better than traditional 20-state models. Reconstructed rotamer states could help in resurrecting ancestral proteins by providing insight into their secondary structures as certain rotamer configurations are only allowed within a specific backbone geometry (Dunbrack and Cohen 1997; Lovell et al. 2000; Dunbrack 2002).

Table 1. Rotamer Configuration States.

Traditional State	Expanded States		
ALA	ALA		
ARG	ARG1	ARG2	ARG3
ASN	ASN1	ASN2	ASN3
ASP	ASP1	ASP2	ASP3
CYS	CYS1	CYS2	CYS3
GLN	GLN1	GLN2	GLN3
GLU	GLU1	GLU2	GLU3
GLY	GLY		
HIS	HIS1	HIS2	HIS3
ILE	ILE1	ILE2	ILE3
LEU	LEU1	LEU2	LEU3
LYS	LYS1	LYS2	LYS3
MET	MET1	MET2	MET3
PHE	PHE1	PHE2	PHE3
PRO	PRO1	PRO2	
SER	SER1	SER2	SER3
THR	THR1	THR2	THR3
TRP	TRP1	TRP2	TRP3
TYR	TYR1	TYR2	TYR3
VAL	VAL1	VAL2	VAL3

Note.—Left: 20 traditional states correspond to the 20 amino acids. Right: our 55-member expanded state set describes both the amino acid and χ_1 rotamer configuration for each constituent residue of a protein. Most amino acids have three possible χ_1 configurations corresponding to specific χ_1 dihedral angle values (~60°, ~−180°, and ~−60°) (see supplementary fig. 1, Supplementary Material online). Alanine (ALA) and glycine (GLY) have no side-chain and therefore no χ_1 configuration, whereas proline (PRO) only has two stable χ_1 configurations (~27°, ~−25°) because of steric requirements of its pyrrolidine ring.

Results and Discussion

Rotamer State Exchange Rates

We first investigate how rotamer states exchange over evolutionarily relevant time-spans by computing a replacement rate matrix derived from counting changes in homologous sites of proteins of known structure (see Materials and Methods: *Rotamer assignment and sequence alignments* and *Tabulating substitution counts*). Figure 2 shows these exchangeabilities in heat-map form for our 55-state model (RAM55), and for a 20-state "rotamer-unaware" empirical model (RUM20) we estimated from the same data set for comparison purposes. Our exchange rates show evidence of χ_1 configuration conservation, whereby the χ_1 configuration (R) is frequently conserved when the identity of the amino acid (A) changes (i.e., $(A,R) \leftrightarrow (A',R)$ with $A' \neq A$). This is visible in figure 2 where higher exchange rates are observed on the diagonal of many of the 3×3 submatrices (corresponding to changes in amino acid only) compared with the off-diagonal elements (changes in amino acid and rotamer configuration). This is particularly true of interchanges between biochemically similar amino acids: submatrices corresponding to aromatic–aromatic exchanges for example all have very distinct diagonal patterns, as do the exchanges between aspartic acid (ASP) and its derivative asparagine (ASN), and between serine (SER) and threonine (THR).

Overall, 111 of the 136 independent 3×3 submatrices show significant association among the interchanging states (see *Rotamer state exchangeability analysis*). To further quantify the strength of these submatrix patterns, we use Cramér's V ($\tilde{V}$), a measure of association between two categorical variables (here the χ_1 configurations of amino acids A and A'). Existence of strong association does not guarantee a diagonal pattern (rotameric state conservation); we therefore also consider the diagonal ratio for each submatrix, indicating the tendency of rates to lie on each 3×3 submatrix's diagonal. $\tilde{V}$ and diagonal ratio are shown in figure 3 for the 111 submatrices with significant associations.

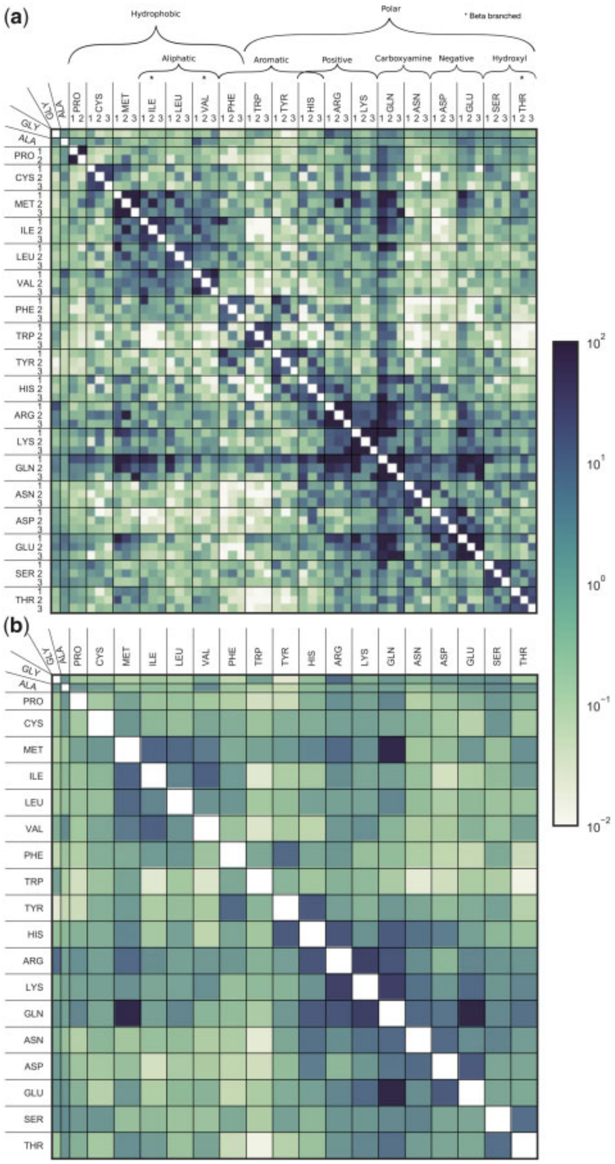

FIG. 2. Replacement models, with and without rotamer configuration information. Exchangeabilities ($s_{(A,R),(A',R')}$ and $s_{A,A'}$) are reported in heat-map form for (*a*) our 55-state model (RAM55) and (*b*) a 20-state model (RUM20) estimated from the same data set. Note that time-reversibility of the models means the exchangeabilities are symmetric (e.g., $s_{(A,R),(A',R')} = s_{(A',R'),(A,R)}$ for $(A,R) \neq (A',R')$).

All six aromatic–aromatic submatrices have high $\tilde{V}$ values and high diagonal ratios (fig. 3*a*, upper right), indicating a strong preference for conserving side-chain orientation. This exchange pattern might be capturing the effect of local constraints on how freely a bulky aromatic side-chain can be positioned without displacing or clashing with those of neighboring residues. A similarly strict configuration conservation can be observed for negative–negative and positive–positive exchanges; however, negative–positive exchanges have high $\tilde{V}$ but somewhat lower diagonal ratios (fig. 3*b*). These submatrices show significant association between specific configurations of the exchanging residues but no common pattern, possibly arising from the competing pressures to retain compatible geometries upon substitution but also to displace the charged moiety to a new location following a charge swap. It is also interesting to note that leucine has high diagonal score in exchanges with all the aromatics (aliphatic-aromatic comparisons, fig. 3*b*). In contrast, isoleucine and valine, both aliphatic and β-branched, have lower scores and show less tendency to conserve their side-chain orientation when exchanging to aromatic residues.

In addition to the conservation of the χ_1 configuration upon amino acid substitution, we also investigated the influence backbone geometry may have on the observed exchangeabilities. To do so we calculated, for each pair of rotamer states, the overlap between the bivariate joint distributions of their ϕ and ψ backbone dihedral angles. These overlap values correlate with the exchangeabilities, with a Spearman's ρ of 0.29 ($p = 1.7 \times 10^{-28}$), indicating that rotamer states exhibit a weak but highly significant preference to interchange with other rotamer states that occupy similar regions of the Ramachandran plot (see *Overlap of backbone distributions*). In some cases, strong nondiagonal patterns in the

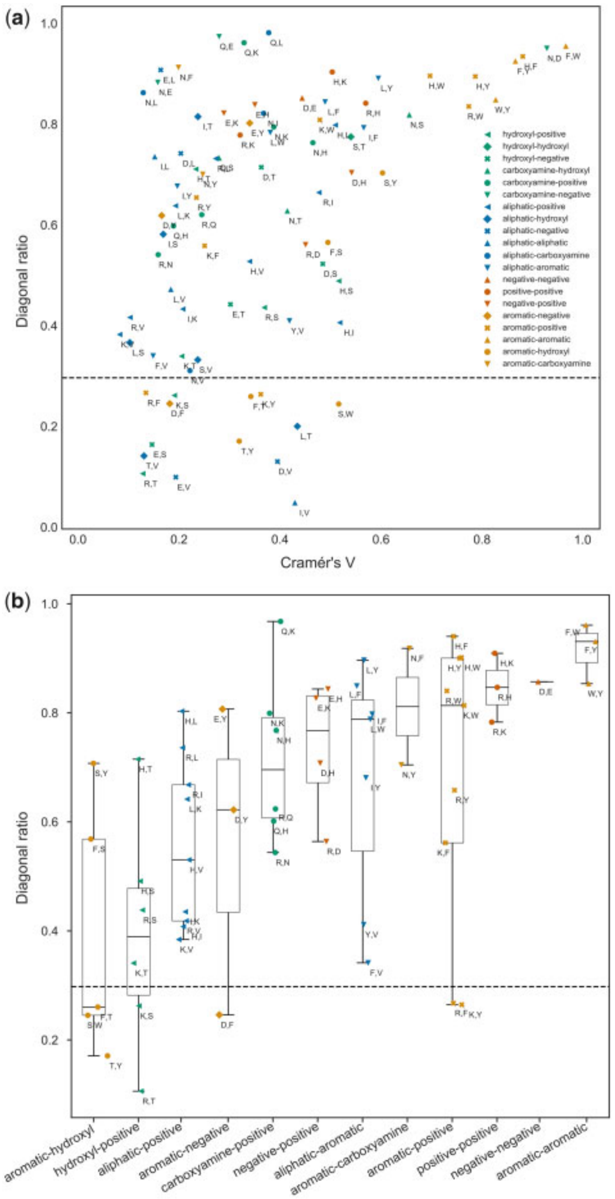

Fig. 3. Strength of association and diagonal ratio. Plots show pairs of residues whose 3 × 3 submatrices within the RAM55 *Q*-matrix achieve significant χ^2 statistic values. Pairs are labeled according to their component residues' biochemical properties. (*a*) Strength of association (Cramér's *V*, $\tilde{V}$) between the χ_1 configurations of residues composing each pair and diagonal ratio (measuring propensity to conserve χ_1 configuration). (*b*) Box plots show diagonal ratio values and medians for exchanges between pairs of residues grouped according to biochemical similarities.

exchageabilities of amino acid pairs correlate strongly with the overlap values, for instance for the exchange of threonine and leucine ($\rho = 0.85$), and indeed 76% (115 out of 153) of the 3 × 3 and 2 × 3 submatrices corresponding to changes in amino acid have a positive Spearman's ρ between overlap and exchangeability (see supplementary fig. 2, Supplementary Material online), indicating that for most amino acid pairs there is a tendency for evolutionary exchanges to be between side-chain geometries that accommodate similar backbone geometries.

Aside from those discussed above, there are several highly significant associations that have no obvious biochemical explanation. Indeed, some cases exhibit a tendency to avoid conserving χ_1 configurations during amino acid exchanges: for example, see the isoleucine–valine interchanges in figures 2*a* and 3*a*. Nevertheless, the strength of these associations indicates that our expanded state set incorporates valuable, biochemically plausible, structural information into the model. RAM55 (fig. 2*a*) can thus be considered a "high-resolution" version of RUM20 (fig. 2*b*) generated from the same data set. As we show in subsequent sections, this provides additional inferential power from the ability to distinguish states and state-interchanges according to χ_1 configuration.

Due to RAM55's expanded state space, the probability of observing any amino acid, given the initial state and a divergence time *t*, is different in the 55-state model than it is in the 20-states model. For instance, a histidine residue is more likely to be substituted with an asparigine when $\chi_1 \approx 60°$ than when in one of the other χ_1 configurations. In the RUM20 model the three χ_1 configurations are merged, and thus the amino acid probability distribution at time *t* corresponds to the weighted average of the three rotamer states. Thus, for each rotamer state, there is a divergence between the probability distributions of the amino acids states at time *t* using the RAM55 model when compared with that when using RUM20. Indeed, as RUM20 can be arrived at by merging states in RAM55, this divergence constitutes a loss of information regarding the amino acid probability distribution when RAM55 is approximated using RUM20. This can be quantified using the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951; see *KL divergence*). At $t = 0$, no loss occurs due to the amino acid sequence being fully known in both models. As $t \rightarrow \infty$, both models tend toward the equilibrium amino acid frequencies and the loss tends toward zero. The differences between the two models manifest in between these extremes. Figure 4 shows that average information loss for one state peaks at 0.0002 *bit* per site after 0.4 amino acid substitutions per site have occurred on average, although this can be much higher for certain rotamer states, and moreover indicates that the difference is most pronounced at the timescales at which evolutionary models are commonly applied: up to $t = 2.5$ which corresponds to ~20% amino acid sequence identity.

Model Benchmarking: Simulation

Here, we use simulated alignments to assess whether or not typical protein sequence data sets contain enough information to permit identification of the true generating model, in the case that the data were generated by the RAM55 rotamer state-aware model. We also investigate whether RAM55 affects our ability to infer phylogenies compared with models using the traditional 20-state space.

To examine our ability to detect χ_1 configuration-influenced evolution, we assessed our 55-state model's goodness-of-fit when analyzing alignments simulated using the model itself. These simulations use a variety of phylogenetic trees and branch scaling factors, to allow evaluation of model detection over a wide range of realistic conditions (see

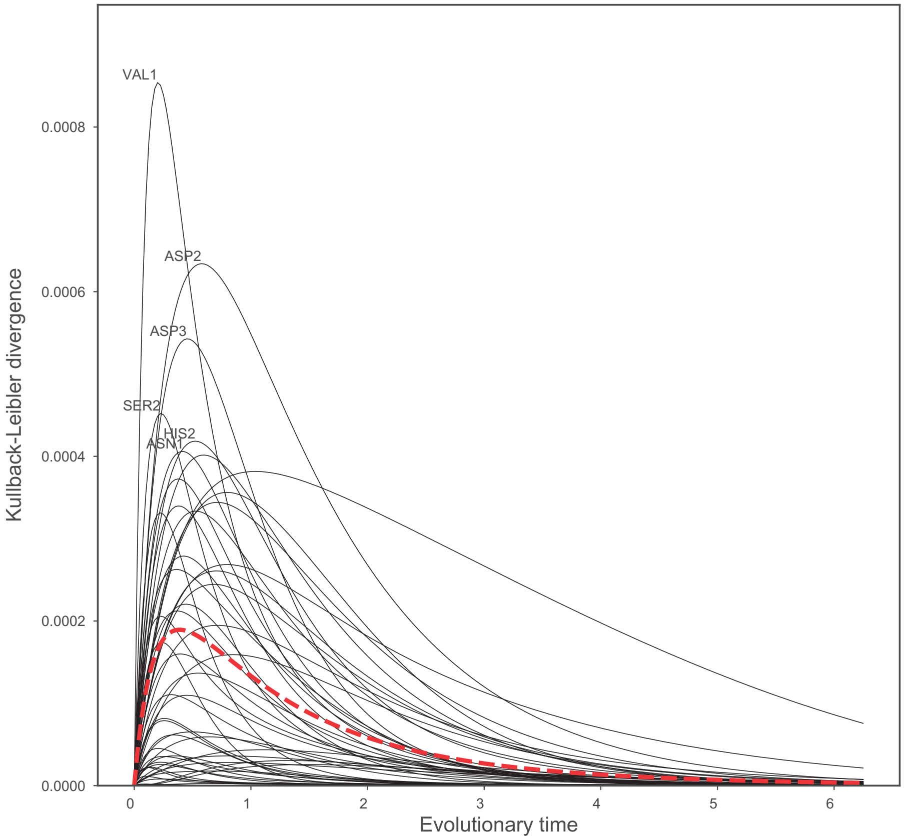

Fig. 4. KL divergence measures the amount of information lost when the structure-free RUM20 model is used to approximate the 55-state model, RAM55. KL divergence is computed for every pair of amino acid state and corresponding rotamer state as a function of evolutionary time t (expressed as expected number of amino acid substitutions). The overall information loss, computed by averaging over all state pairs' KL divergences and weighted by the rotamer state equilibrium frequencies, is shown in red.

Tree generation and alignment simulation). From these simulated alignments we then infer the corresponding phylogenies by maximum likelihood (ML) using RAM55 or other models that are widely used for phylogenetic analysis of amino acid sequences (see *Likelihood calculation and maximization over phylogenies*). Figure 5*a* compares Akaike Information Criterion (AIC) scores (Akaike 1974) or state-corrected AIC scores (see *Log-likelihood comparison across models*) of the inferred tree across multiple models. RAM55 consistently shows detectably better fit for the simulated data regardless of sequence divergence. Moreover, for most branch lengths and number of taxa, RAM55 has a lower AIC score than all other models for 100% of the simulations (see supplementary table 1, Supplementary Material online). At the extreme of trees with large tree lengths and low taxa number, the RUM20 model occasionally has a lower AIC score.

It is also interesting to note how LGexp, our version of the 20-state LG model "uniformly expanded" to 55 states but incorporating no structural information (see *Log-likelihood comparison across models*), fits the data worse than its "frequency-aware" counterpart (LGbyfreq-exp) whose AIC values are comparable with those of RUM20 and LG. This illustrates how adding noninformative complexity to a 20-state model penalizes its performance, while being correctly informed about each rotamer state's frequency but not specifically about its exchange rates still produces a worse fit than the full RAM55 model. These results confirm that, when the more complex RAM55 model matches the underlying process generating the input alignments, it is possible to detect an improvement in fit over simpler models.

As a further performance test, we also evaluated whether ML trees inferred under the RAM55 model are closer to the reference phylogeny used during the simulation process than those inferred with other models. For these comparisons we considered both 1) the Euclidean distance (Felsenstein 2004), a metric that accounts for both topological differences between trees and differences in branch lengths and 2) the lengths of individual branches. Under the former measure, RAM55 infers trees that are at least as close or closer to the reference phylogenies than those inferred by amino acid replacement models such as our RUM20 model or LG. Figure 5*b* compares the distributions of Euclidean distances between inferred and reference trees, estimated using the RAM55, RUM20, and LG models in simulations of 1,000 sites on a 64-taxon phylogeny. Shifts toward lower values for RAM55 indicate greater accuracy of trees inferred using this model. Similar results are obtained for other alignment lengths and numbers of taxa in model phylogenies, as well as when simulating over a larger, empirical tree (see supplementary figs. 3–5, Supplementary Material online).

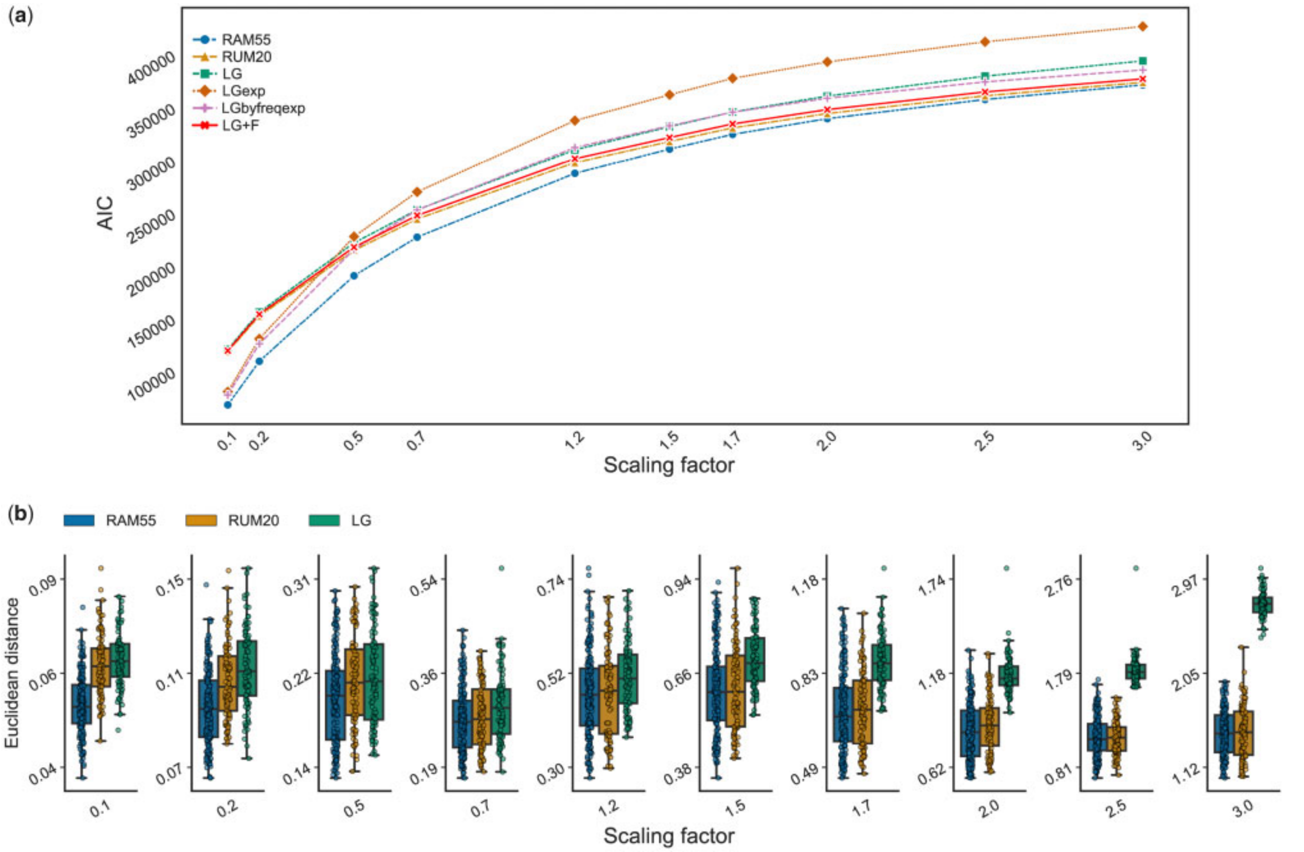

FIG. 5. (*a*) AIC values for competing models. Each data point corresponds to the mean AIC value of trees inferred from 100 simulated alignments. (*b*) Comparison of the RAM55 model (blue bars) against LG (green bars) and against our RUM20 model (orange bars) in terms of Euclidean distance of inferred trees from the reference phylogeny used to simulate the alignment. Box plots illustrate distance distribution and median (horizontal lines); scatter plot points represent individual distance values. Tree inference is performed on alignment data sets (1,000 sites, 64 taxa, and 100 replicates per scaling) simulated using RAM55 and a randomly generated reference phylogeny, scaled according to the factors on the *x* axis. Note the different *y* axis scales.

Estimates of individual branch lengths can be unbiased, or can be consistently over- or under-estimated depending on their location within a phylogeny. Nevertheless, RAM55 tends to more accurately estimate the correct evolutionary distance between sequences regardless of tree size (number of taxa), length of the examined branch or branch positioning in the tree. Figure 6—highlighting one internal branch for each of four topologies—illustrates this with branch length estimates from RAM55 having smaller variances and medians nearer to the reference values than estimates from LG; these results are representative of those obtained for other branches (results not shown). The additional χ_1 configuration information contained in RAM55 is thus allowing us to infer more-reliable phylogenies from alignments simulated under the 55-state model itself than does any of the 20-state models investigated.

Model Benchmarking: Empirical Alignments

We assessed RAM55's performance on three empirical amino acid alignments—with 13, 82, and 46 taxa, respectively—for which we can obtain corresponding structural information (see *Empirical alignments*), and compared goodness-of-fit and inferred phylogenies across models. RAM55 was used in ML analyses, and results compared with those derived using structure-free models such as the 20-state models LG, WAG, and our own RUM20, and the 55-state LGbyfreq-exp model which recognizes the frequencies of the 55 states but not their structural information content (see *Log-likelihood comparison across models*, *Likelihood calculation and maximization over phylogenies*).

Figure 7 shows the goodness-of-fit (measured by AIC values, see *Log-likelihood comparison across models*) for each empirical amino acid alignment under a variety of models. In all cases, RAM55 is a better fit for the data than all the other models used, indicated by the lower AIC values. Since our model is implemented in RAxML-NG (Kozlov et al. 2019), it was also straightforward to incorporate a discrete gamma model of rate heterogeneity (Yang 1993), ML estimation of equilibrium frequencies from the observed data, or both in combination (see *Likelihood calculation and maximization over phylogenies*). The corresponding models, denoted RAM55+G, RAM55+F and RAM55+G+F, resulted in further improvements in the model's fit, with RAM55+G+F performing best for all data sets. This empirical benchmark shows that RAM55 fits well when tested on three diverse data sets, and thus appears to be a valuable model of protein sequence evolution.

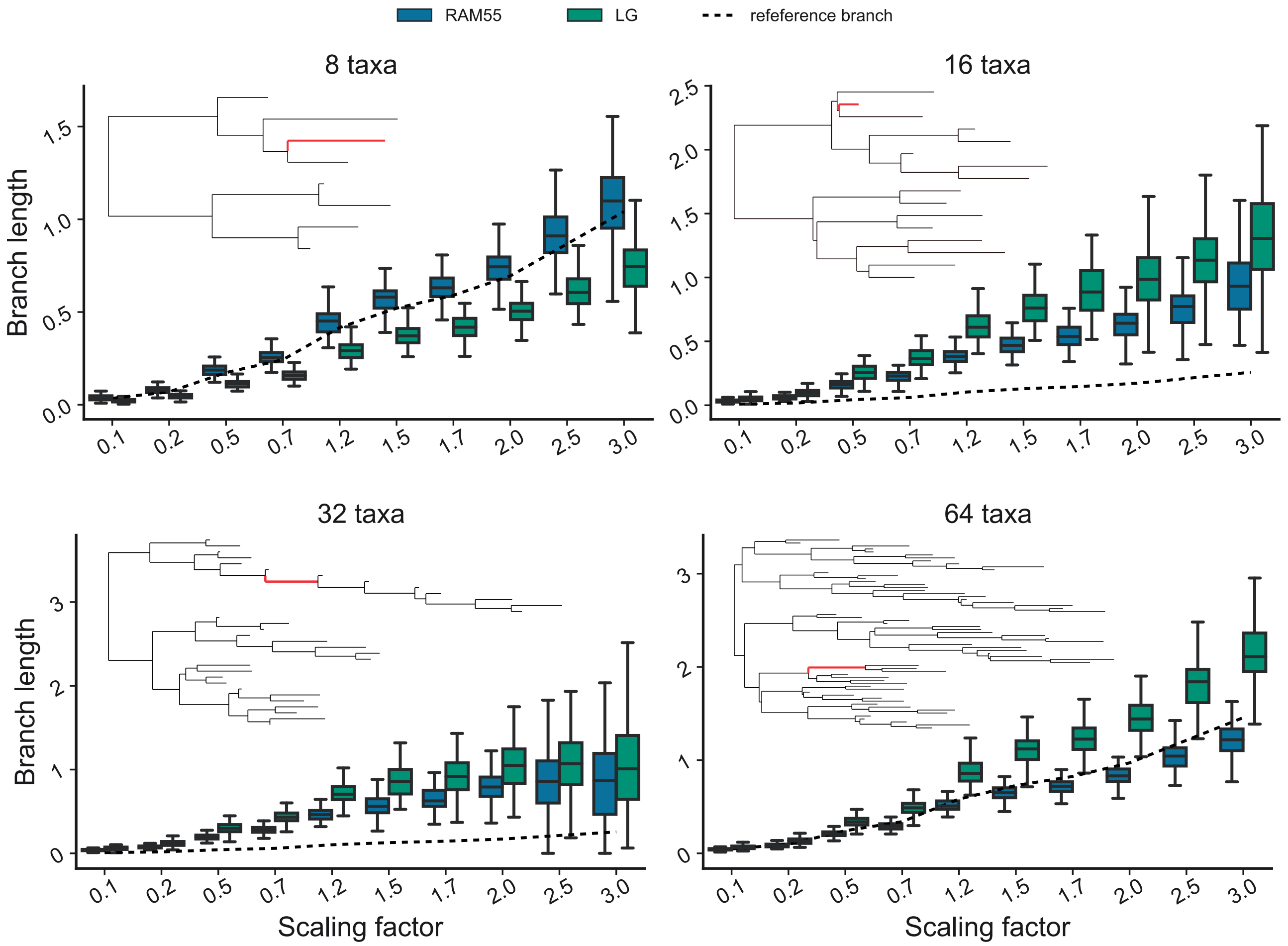

Fig. 6. Examples of individual branch length inferences illustrate the tendency for RAM55 to give estimates closer to the reference value. Trees inferred using RAM55 (blue) or LG (green), analyzing alignments (200 sites, 100 replicates per scaling) simulated using RAM55 and reference phylogenies of 8, 16, 32, and 64 taxa, scaled according to the factors displayed along the *x* axis. Highlighted internal branches (indicated in red) have true lengths indicated by the solid lines; distributions of inferred lengths are shown as box plots (evenly distributed horizontally and displaced for clarity).

Ancestral State Reconstruction

Having established RAM55's ability to infer reliable phylogenies when structural information is available, we evaluate whether it can be used for two further tasks. The first is the reconstruction of ancestral amino acid states, which can also be achieved using a standard substitution model in which side-chain configurations are not modeled. Second, we try to reconstruct the rotamer sequence in addition to the amino acid sequence, a capability which is unique to our model. We evaluated the performance on these tasks using both joint (Pupko et al. 2000) and marginal (Yang et al. 1995) reconstruction algorithms. The (55-state) RAM55 model can be applied to reconstruct plain (20-state) ancestral amino acid sequences, when present-day crystallography data are available, by first reconstructing ancestral rotasequences, and then simply masking the rotamer configuration information. The resulting ancestral amino acid sequences can then be compared with the known (masked) ancestor sequences in our simulations. We perform simulations as before under RAM55 using an 8-taxa reference topology, and then reconstruct ancestral amino acid states using RAM55 or LG and the joint reconstruction method (Pupko et al. 2000). Terminal amino acid sequences (fig. 8, nodes A and B) were also reconstructed in order to validate our "leave-leaves-out" (LLO) approach that serves as a proxy for ancestral sequence reconstruction when lacking a reference (see *Ancestral state reconstruction*). As shown in figure 8, it is possible to estimate terminal sequences with reasonable accuracy with this strategy (see also supplementary figs. 6 and 7, Supplementary Material online), suggesting this is a viable method to evaluate reconstruction accuracy on empirical alignments, where ancestral reference sequences (and structures) are unlikely to be available. Figure 8 shows that our model performs equally or slightly better than LG in terms of amino acid state reconstruction accuracy, particularly at longer evolutionary distances (see also supplementary fig. 7 and table 2, Supplementary Material online). Very similar results are achieved using the marginal reconstruction approach (supplementary figs. 6 and 8 and table 2, Supplementary Material online). These show that, in addition to exploiting information about χ_1 configuration evolution to assist with model selection and phylogeny inference, RAM55 can be used to reconstruct ancestral

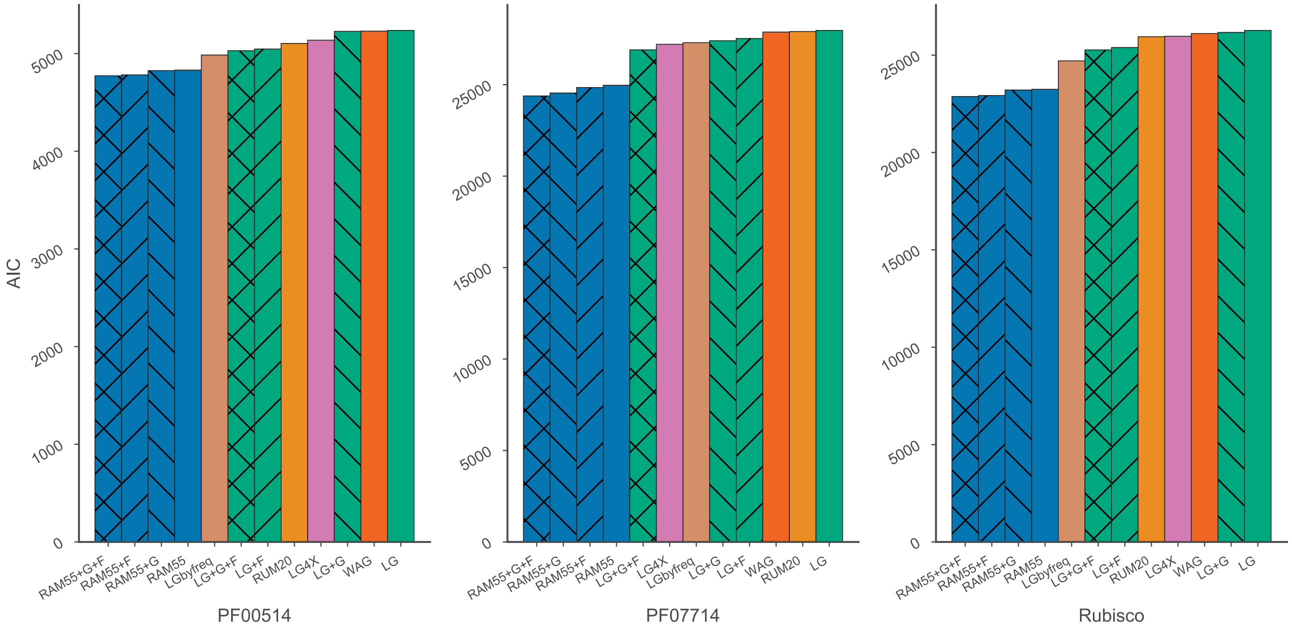

Fig. 7. A comparison of RAM55 variants against other models in terms of AIC on three empirical rotasequence alignments: PF00514 (β-catenin-like repeat), PF07714 (tyrosine kinase) and rubisco. "+G" models use a discrete gamma model of rate heterogeneity with four categories; "+F" models use ML-estimated state frequencies estimated from the observed data.

sequences as well as a 20-state model, whereas as the same time providing information about the side-chain conformation which is not possible by any other method.

We thus assessed our models' accuracy when joint reconstructing ancestral rotamer states simulated under the model itself and our 8-taxa phylogeny. Figure 9 shows that RAM55 is able to infer the correct ancestral side-chain configuration for residues belonging to internal sequences in almost all cases when the ancestral amino acid state is accurately reconstructed (as shown in fig. 8). Similar results are obtained for other alignment lengths and numbers of taxa in reference phylogenies (data not shown). These reconstructed ancestral rotamer states could be used to predict side-chain geometry for homology modeling of ancestral proteins, to assess which configuration better fits the evolutionary data.

Conclusions

We have created a Dayhoff-like continuous-time Markov model that accounts for structural constraints on protein evolution by employing an expanded state set where each state corresponds to an amino acid along with its side-chain's χ_1 configuration. The exchange rates of our 55-state model, RAM55, clearly capture effects of local steric constraints, for example those dictating how an aromatic side-chain can be positioned without displacing or clashing with neighboring residues. Other highly significant rotamer state exchange patterns, while still carrying valuable information for our model, have no obvious biochemical explanation and in some cases exhibit a tendency to avoid conserving χ_1 configurations during amino acid exchanges. These exchange patterns deserve further exploration, perhaps by relating 3D molecular descriptors to the exchange rates as has been attempted for amino acid exchange rates and 1D biochemical properties (Grantham 1974; Dayhoff et al. 1978; Zoller and Schneider 2013).

Using simulated data, we confirmed that our 55-state model captures enough information to detect the χ_1 configuration-aware expanded state space, and observed that it consistently offers detectably better fit to data compared with models that use the traditional 20-state space such as LG, WAG, and our RUM20. Further, RAM55 appears to infer equally or more reliable phylogenies than any of the 20-state models. This argues in favor of its consideration for phylogenetic analysis of protein sequences. Moreover, when applied to empirical data, the model provided a better fit than any of the traditional models evaluated.

Our model can also be applied to perform structurally aware reconstruction of ancestral sequences. Both amino acid and structural configuration states can be reliably inferred. Although there is little improvement in amino acid sequence reconstruction over traditional 20-state models, RAM55 could improve ancestral protein resurrection by 1) providing better phylogenies, which are valuable in themselves but also help toward 2) obtaining reconstructions of structural information, that is, χ_1 configurations, that are simply not possible by any other method.

More generally, inferred rotamer states could be used to predict side-chain geometry for homology modeling, to assess which configuration better fits the evolutionary data. In this article, we apply our model to empirical data where both amino acid sequences and the corresponding high-quality atomic coordinates are available. Although an increasingly large number of protein sequences can be associated with reliable X-ray crystallography data, and recent advances in cryo-electron microscopy promise to improve the resolutions

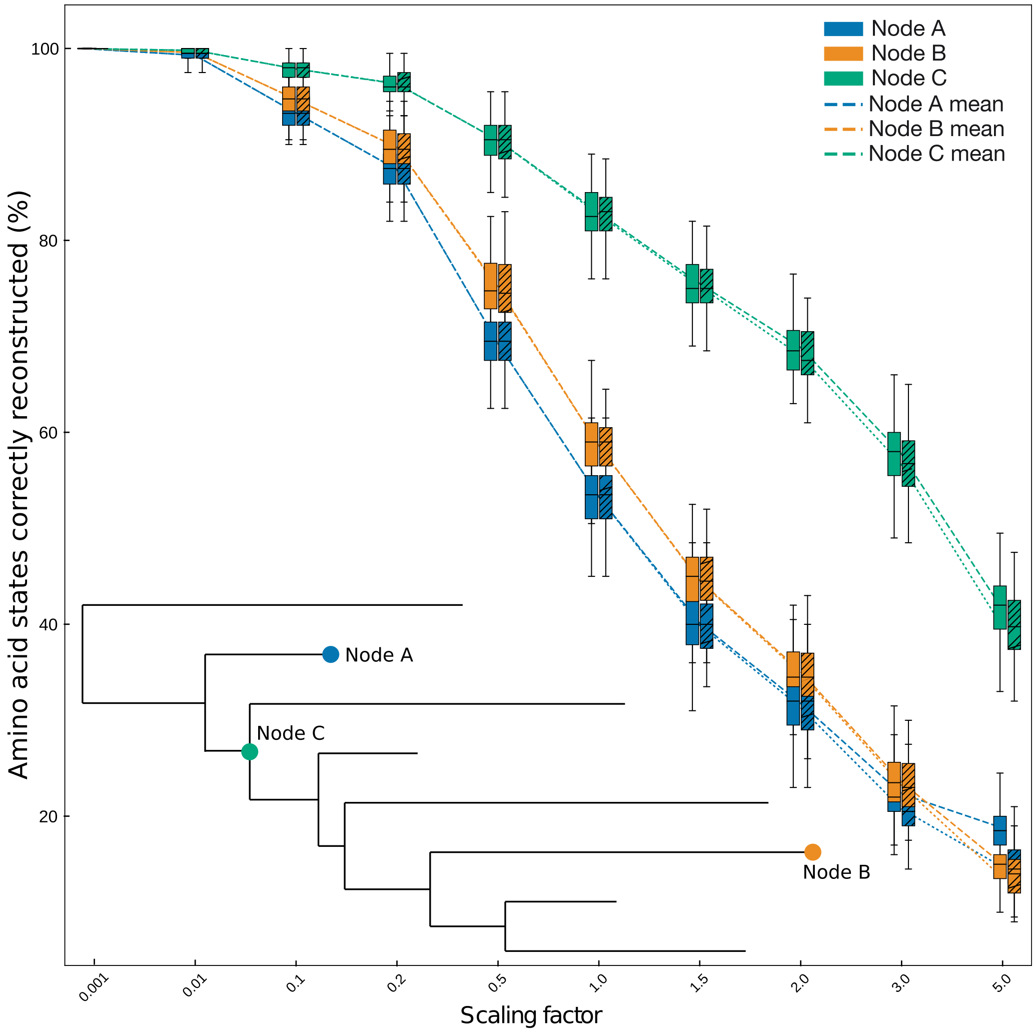

Fig. 8. Amino acid reconstruction accuracy. Amino acid states inferred using joint reconstruction from rotasequence alignments (200 sites) simulated under RAM55 using our 8-taxon reference phylogeny and scaling its branches according to the factors reported on the *x* axis (note the nonlinear scale used for clarity). The joint reconstruction algorithm is then employed along with RAM55 and the true phylogeny to reconstruct rotasequences at various internal (*C*) or terminal (*A*, *B*) nodes. The *y* axis indicates the proportion of amino acid states (i.e., masked rotasequence states) correctly reconstructed for each inferred sequence. Plain box plots indicate the distribution of percentages of sites correctly reconstructed (*y* axis) by this method; the same procedure is then repeated using LG on masked alignments (hatched box plots). Each box plot contains results from 100 simulation replicates for a given node.

that can be achieved for complex, dynamic molecular assemblies in their native state (Milne et al. 2013; Carroni and Saibil 2016; Venien-Bryan et al. 2017), many real-world applications of our approach might rely on data with a mixture of amino acid sequences and rotasequences, or amino acid sequences alone. Our model could be applied to this type of data by treating ambiguity regarding the rotameric state in the same manner in which sequencing errors and other forms of ambiguity are handled (Huelsenbeck 2002; Felsenstein 2004) thus allowing the information gained by accounting for the distinct evolutionary signatures of the rotamer states to be applied to all bioinformatics tasks relying on evolutionary modeling, as well as opening potential applications such as the prediction of side-chain geometries from amino acid sequences in the absence of other structural information.

In turn, rotamer information can find use in protein structure modeling. Two ways in which this could be achieved can be illustrated using the Rosetta modeling process (Leaver-Fay et al. 2011). During the conformational search, rotamer states are sampled, scored using an energy function, and accepted or rejected using Monte Carlo methods (Leaver-Fay et al. 2011). These states could be preferentially sampled according to their likelihood according to our RAM55 model. Alternatively, the scoring function could be adjusted by replacement of the rotamer probability term (which favors generally more-prevalent rotamer states) by a site-specific rotamer probability that also depends on the evolutionary context of that amino acid (Alford et al. 2017). Such an approach could also complement artificial intelligence-based developments in the exploitation of residue coevolution in

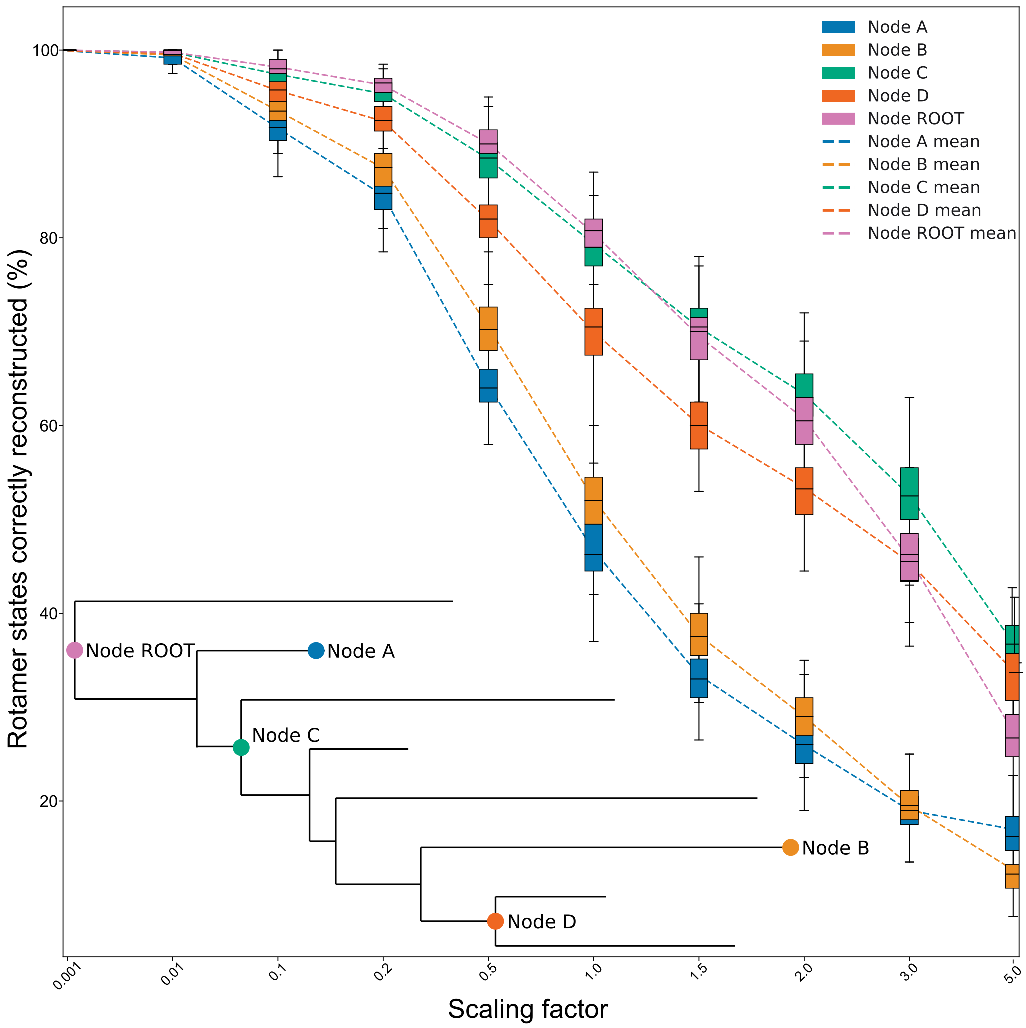

Fig. 9. Rotamer state reconstruction accuracy. Rotamer states inferred using joint reconstruction from rotasequences (8 taxa, 200 sites) simulated under RAM55 using our 8-taxon reference phylogeny and scaling its branches according to the factors reported on the *x* axis (note the nonlinear scale used for clarity). The joint reconstruction algorithm is then employed along with RAM55 and the true phylogeny to reconstruct rotasequences at various internal (ROOT, *C*, *D*) or terminal (*A*, *B*) nodes. The *y* axis indicates the proportion of rotamer states correctly reconstructed for each inferred sequence. Box plots indicate the distribution of percentages of sites correctly reconstructed (*y* axis) by this method. Each box plot contains results from 100 simulation replicates for a given node.

modeling the protein backbone (Wang et al. 2017; Liu et al. 2018; Service 2018; Xu 2018).

We believe that future models of protein evolution will benefit from being informed about multiple structural constraints, and will do so by integrating a number of structural features. These will include some of the ones previously proposed by others, and we propose χ_1 configuration, with its implementation and computational advantages, as another candidate. The process of exploring the best combinations, and indeed devising practical algorithms and computational strategies to implement them, is for future studies.

Materials and Methods

Rotamer Assignment and Sequence Alignments

In order to tabulate substitution events, a data set of aligned amino acid sequences annotated with their χ_1 rotamer state was required. This was obtained from the Pfam database (Finn et al. 2014) by first selecting those aligned sequences that are mapped to a high-resolution crystal structure (<2.5 Å) in the Protein Data Bank in Europe (Velankar et al. 2010) to ensure we only retrieve those structures that are likely to be reliable. For each amino acid in each sequence, we assigned a χ_1 rotamer configuration based on atomic coordinates, as defined in the Dunbrack rotamer library (Shapovalov and Dunbrack 2011). We removed residues with an average *B*-factor (Trueblood et al. 1996) >30 for the four atoms defining χ_1 (N, C^α, C^β, and C^γ for most amino acids), to ensure that rotamer state assignments were based on unambiguous electron densities and not modeling artefacts.

Factors such as thermal fluctuations, crystal packing forces and ligand binding might confound our model by creating differences between the structures of homologous proteins that are not due to evolution. Our *B*-factor filtering also addresses these. In a study of 63 pairs of structures of the same protein, one in the apo state the other in the holo state, 95% of side-chains did not change rotamer (Zavodszky and Kuhn 2005). Of the residues which did change rotameric state, the majority adopted the same χ_1 configuration, indicating that thermal motions are concentrated away from the backbone. Similarly, Najmanovich et al. (2000) investigated 221 bound/unbound pairs and found that 94% of residues retained their rotameric configuration, with 40% of proteins having no residues with altered χ_1 state. Furthermore in a study of 123 pairs of structures, of the residues that did alter their χ_1 state, most were solvent exposed and thus not restricted by the need to pack into the protein interior (Zhao et al. 2001). These residues are characterized by diffuse electron densities and high thermal *B*-factors, which we remove with the above filter. Neither Najmanovich et al. (2000) nor Clark et al. (2019) found a correlation between side-chain conformational change and backbone conformational change. Collectively, these studies indicate that although side-chain movement is important when comparing structures of identical sequence, especially with respect to ligand binding, these movements tend to be within-rotamer changes, concentrated away from the backbone, and in surface-exposed residues which are removed by the *B*-factor filtering. Thus, the effects of thermal motion and variations in crystallization conditions are negligible with regard to the substitution rates presented herein.

We also removed nonstandard residues, disordered residues and those with peptide bonds exceeding 1.8 Å, the last to ensure a continuous polypeptide. In this study we consider only χ_1 configurations, and not those of rotable bonds further along the side-chain, for a number of reasons: χ_1 is present across all residues with the exception of glycine and alanine; it is closest to the backbone and thus usually better resolved in terms of atom positions; it conveys the most information about side-chain atom positions as all other side-chain atoms depend upon it; it gives us a manageable number of states; and it always connects two sp^3 hybridized atoms, and thus is strictly rotameric and has exactly three conformational states (Dunbrack 2002) although one is inaccessible in proline. These quality filtering steps resulted in alignments from 3,646 Pfam families, including 31,801 unique Uniprot entries, 251,194 PDBe structures, and 81,523, 991 residues.

Tabulating Substitution Counts

We combined the amino acid sequences and rotamer state sequences to produce sequences in an expanded alphabet (see table 1), which we refer to as "rotasequences." Each rotasequence consists of symbol pairs (*A*, *R*), each of which specifies a state comprising the amino acid *A* (as employed by traditional 20-state models) and a χ_1 rotamer configuration *R* (1, 2, or 3). For each family (see supplementary files, Supplementary Material online), we then performed a sequence alignment-guided pairwise comparison of rotasequences. We used Pfam's original domain alignment to construct a NJ phylogenetic tree (Saitou and Nei 1987) using MAFFT (Katoh and Standley 2013), and then iteratively tabulated differences between pairs of rotasequences by taking a circular tour through the NJ tree using an algorithm analogous to the one described by Korostensky and Gonnet (2000). Although comparing pairs of rotasequences, we omitted those with a rotasequence identity $<$ 75%, to minimize the risk of multiple substitution events at the same site being tabulated as a single observed difference. This approach results in an efficient set of pairwise comparisons using all leaves of the trees with each observed difference counted at most twice. We tabulated 30,439,912 counts, corresponding to 4,508,390 rotamer state substitutions and 25,931,520 instances of rotamer state conservation.

We then computed the observed number of occurrences of sites in all aligned sequence pairs with rotamer states (*A*, *R*) in one sequence and (A', R') in the other as $n_{(A,R),(A',R')}$. Although these counts could be used directly to calculate an instantaneous rate matrix (IRM), this would result in biases arising from the filtering procedures described above. For example, because alanine and glycine can never be filtered out by *B*-factor, these residues are overrepresented. Further, some amino acids, such as those commonly well packed into the core of the protein, are better resolved and have lower *B*-factors than those more commonly found at the protein surface, and are thus also overrepresented (supplementary fig. 9, Supplementary Material online). To account for this, we also compute substitution event counts $(n_{A,A'})$ for a Dayhoff-like 20-state empirical model (RUM20) using the same Pfam-based data set, but ignoring χ_1 configurations and without performing *B*-factor filtering. Our normalized rotamer state exchange count matrix $\hat{N}$, recovering the actual observed residue frequencies, then becomes

$$\widehat{n}_{(A,R),(A',R')} = \frac{n_{A,A'}}{\sum\limits_{\substack{r\in R_A,\\ r'\in R_{A'}}} n_{(A,r),(A',r')}} . n_{(A,R),(A',R')} \quad , \tag{1}$$

where $R_A = \{R : (A, R) \in S_{55}\}$ is the set of rotamer configurations *R* such that the corresponding pair (*A*, *R*) is a member of S_{55}, the set of all 55 possible combinations shown in table 1.

The 55-state IRM $\hat{Q}$ is computed from these normalized counts as described by Kosiol and Goldman (2005): the instantaneous rate of change of (*A*, *R*) into (A', R') (with $(A, R) \neq (A', R')$) is given by the number of such events as a proportion of all observations of (*A*, *R*):

$$\widehat{q}_{(A,R),(A',R')} = \frac{\widehat{n}_{(A,R),(A',R')}}{\sum\limits_{(a,r)\in S_{55}} \widehat{n}_{(A,R),(a,r)}} . \tag{2}$$

As usual, diagonal elements $\hat{q}_{(A,R),(A',R')}$ are set so that row sums of the IRM equal 0 (Kosiol and Goldman 2005). The 20-state RUM20 IRM was similarly obtained from the unfiltered 20-state counts $n_{A,A'}$, for comparative purposes.

Rate Scaling

Times and branch lengths are typically measured as the expected number of substitutions per site (Felsenstein 2004). Our rates were therefore first scaled according to ρ so that, at equilibrium, they will result on average in one rotamer state ($(A,R) \longrightarrow (A',R')$ with $\{(A,R),(A',R')\} \in S_{55}$) substitution per unit of time (Liò and Goldman 1998):

$$\hat{Q}^* = \frac{1}{\rho}\hat{Q} \tag{3}$$

with

$$\rho = \sum_{(A,R)\in S_{55}} \pi_{(A,R)}\widehat{q}_{(A,R),(A,R)}, \tag{4}$$

where $\pi_{(A,R)}$ is the equilibrium frequency of rotamer state (A, R) obtained from the normalized counts ($\hat{N}$). Because we have an expanded state set, $\hat{Q}^*$ results in one rotamer state substitution per unit time but less than one amino acid state substitution. We therefore perform a further scaling step in order to allow direct comparison of branch lengths estimated with our 55-state model and any 20-state model, that is, in terms of number of amino acid state substitutions. This additional scaling factor ρ^* is defined as

$$\rho^* = \sum_{(A,R)\in S_{55}} \pi_{(A,R)} \sum_{\substack{(A',R')\in S_{55} \\ A'\neq A}} q^*_{(A,R),(A\prime,R\prime)} \tag{5}$$

and corresponds to the proportion of rotamer state changes where the amino acid changes, irrespective of χ_1 configuration. Then the "superscaled" IRM is given by

$$\hat{Q}^{**} = \frac{1}{\rho^*}\hat{Q}^*. \tag{6}$$

This matrix, at equilibrium, has on average $1/\rho^* = 1.79$ state changes per unit time, consisting of 1 amino acid state substitution plus 0.79 (i.e., $(1-\rho^*)/\rho^*$) χ_1 configuration changes that are invisible to traditional 20×20 models. This means that branch lengths are directly comparable to those under 20-state structure-free models that can only detect amino acid changes. RAxML-NG's implementation of RAM55 provides output in these superscaled time units (comparable to traditional amino acid distances) and also in the units of equation (3).

From $\hat{Q}^{**}$, final scaled exchangeabilities (available in supplementary files, Supplementary Material online) were obtained as

$$s_{(A,R),(A',R')} = \frac{\hat{q}^{**}_{(A,R),(A\prime,R\prime)}}{\pi_{(A',R')}} \tag{7}$$

(Liò et al. 1998; Whelan and Goldman 2001). Exchangeabilities computed from a general data set can be combined with state frequencies estimated from any particular data set under study, and a data set-specific IRM can thus be obtained by inverting equation (7). This hybrid parametrization procedure, denoted by adding "+F" to a model name, can produce a significant improvement in model fit (Thorne and Goldman 2007; Perron et al. forthcoming).

Rotamer State Exchangeability Analysis

Each pair of different amino acids corresponds to a 3×3 submatrix (with the exception of pairs including alanine, glycine and proline) in $\hat{N}$. Since $\hat{N}$ is symmetric there are 136 unique 3×3 submatrices. For each of these, we computed the Pearson's χ^2 statistic and P value (with Bonferroni correction) for the hypothesis test of independence of the observed χ_1 rotamer configuration change frequencies, where the expected frequencies are computed based on the marginal sums under the assumption of independence. Pairs of residues with Bonferroni P value <0.05 show significant association among their rotamer states; only these are considered for further analysis (e.g., fig. 3).

We assessed the strength of association between the χ_1 rotamer configurations of each pair of residues using Cramér's V ($\tilde{V}$) with bias correction (Bergsma 2013). We also computed the proportion of counts that lie on each 3×3 submatrix's diagonal ("diagonal ratio"). The latter is a measure of the tendency of a pair of exchanging residues to conserve their χ_1 rotamer configuration. To better assess trends in diagonal ratio and $\tilde{V}$, residues with three available χ_1 rotamer configurations (excluding methionine) are then classified into six groups depending on the biochemical properties of their side-chains: aliphatic (isoleucine, leucine, valine), aromatic (phenylalanine, tryptophan, tyrosine), positive (arginine, lysine, histidine), carboxylamine (asparagine, glutamine), negative (aspartic acid, glutamic acid), and hydroxyl (serine, threonine). While this is an imperfect classification, as residues do not fit unambiguously into distinct, discrete groups and have multiple salient features, it nonetheless helps up to better visualize how side-chain properties influence exchange rates.

Overlap of Backbone Distributions

The global structure of a protein is determined largely by the configuration of the peptide backbone onto which the side-chains are bonded, which can be characterized by the dihedral angle of the two rotatable bonds, ϕ and ψ, of each amino acid. Steric effects determine which combinations of ϕ and ψ are allowed, and which are favored, as commonly visualized using a Ramachandran plot, which shows permitted regions and observed distributions over ϕ and ψ (Ramachandran et al. 1963). As these steric effects arise, in part, from the side-chain, and the rotameric states of the side-chain are influenced by the conformation of the backbone, each rotamer state has its own probability distribution on the Ramachandran plot (Dunbrack and Karplus 1993). To test the hypothesis that rotamer states preferentially exchange with other rotamer states with similar backbone dependencies, we calculated the overlap between the (ϕ,ψ) distributions of pairs of rotamer states (i, $j \in S_{55}$) from their probability density functions, $f_\cdot(\phi,\psi)$, as estimated from rotamer counts in the Dunbrack backbone-dependant rotamer library Shapovalov and Dunbrack (2011), using the equation:

$$O_{ij} = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \min\{f_i(\phi,\psi), f_j(\phi,\psi)\}\, d\phi\, d\psi\,. \tag{8}$$

KL Divergence

We measured the amount of information lost, regarding the amino acid sequence, when a 20-state model is used to approximate our 55-state model by computing the KL divergence in *bits* (Kullback and Leibler 1951) for each rotamer state (A, R) and its corresponding amino acid state A. This metric measures the divergence between the amino acid probability distribution at time t, when starting with rotamer state (A, R), between the RAM55 model in which both A and R and considered and the RUM20 model in which only A is used. The KL divergence is computed as a function of evolutionary time t using:

$$D_{\mathrm{KL}}(P_{\mathrm{RAM55}}(t,(A,R))||P_{\mathrm{RUM20}}(t,A)) = \sum_{a\in S_{20}}\left(\sum_{r\in R_a} P_{\mathrm{RAM55}}(t,(A,R),(a,r))\log_2 \frac{\sum_{r\in R_a} P_{\mathrm{RAM55}}(t,(A,R),(a,r))}{P_{\mathrm{RUM20}}(t,A,a)}\right) \tag{9}$$

with S_{55} and S_{20} being the 55- and 20-state spaces, R_a the χ_1 configurations of amino acid a, $P_{\mathrm{RUM20}}(t)$ and $P_{\mathrm{RAM55}}(t)$ the probability matrices of the respective models at time t (see eq. 10 below), and for example, $P_{\mathrm{RAM55}}(t,(A,R),(a,r))$ the $((A,R),(a,r))$ element of $P_{\mathrm{RAM55}}(t)$.

Likelihood Calculation and Maximization over Phylogenies

We implement our models using ML methods applied to multiple sequence alignments (Felsenstein 2004). This standard approach searches for the tree T that maximizes the likelihood function with substitutions being modeled by a Markov process. Markovian state substitutions over time t are described by a probability matrix defined by

$$P(t) = e^{tQ}\,, \tag{10}$$

where Q is the IRM of the Markov process. The likelihood of T (including tree topology and branch lengths) given data (alignment) X and IRM Q can then be computed as

$$L(T|Q,X) = \prod_{i} L(T|Q,X_i)\,, \tag{11}$$

where $L(T|Q,X_i)$ corresponds to the likelihood of T given the states observed at site i of X (site independence assumption). $L(T|Q,X_i)$ is computed by applying equation (10) to each tree branch and using the pruning algorithm (Felsenstein 1981). Maximizing L over T provides estimates $\hat{T}$ and thus the most likely phylogeny given the observed data and the current substitution model. The "+F" approach to matching the model's state frequencies to the observed data can be implemented by simultaneously maximizing L over these frequencies.

It is also generally acknowledged that sites do not evolve at the same rate, due to various evolutionary constraints. The most common way of accounting for this heterogeneity is to assume that rates across sites follow a discretized gamma distribution (Yang 1994). The shape parameter of the gamma distribution, α, is usually estimated by ML along with T as it is considered specific to each data set. Models using the gamma distribution to model rate heterogeneity are denoted "+G."

In this study, all ML tree inferences were performed using RAxML-NG (Kozlov et al. 2019) which, following the needs of this study, now has functionality allowing custom state spaces and rate matrices of any size, permitting us to use our 55-state model RAM55 to infer tree topology, branch lengths and likelihoods that can be used for model fitting and comparisons. It also permits the +F and +G variants of substitution models through its "+FO" and "+G" options. Our expanded state space has some inevitable repercussions for CPU time, not least because 20-state models benefit from a highly optimized likelihood computation in RAxML-NG, whereas the RAM55 model currently works with general kernels that are less efficient. Nevertheless, computation times remain acceptable, tending to be 5–10 times longer than using 20-state models (supplementary fig. 10, Supplementary Material online).

Tree Generation and Alignment Simulation

We simulated sequence alignments under RAM55 using four randomly generated trees (8, 16, 32, or 64 taxa; branch lengths $\in [0.01, 0.5]$; see supplementary fig. 11 and supplementary files, Supplementary Material online) as guide and a substitution simulation approach based on Method 1 of Fletcher and Yang (2009), modified for our expanded state set. Additionally, a pruned (mammals) and scaled version of the Ensembl-compara species tree (Herrero et al. 2016; see supplementary files, Supplementary Material online) was also used. To allow investigation of a realistic range of sequence divergences (around 10–85%) while maintaining consistent tree topologies, all branches of our trees were scaled according to a set of 10 scaling factors: {0.1, 0.2, 0.5, 0.7, 1.2, 1.5, 1.7, 2, 2.5, 3} for model benchmarking simulations, or {0.001, 0.01, 0.1, 0.2, 0.5, 1, 1.5, 2, 3, 5} for ancestral reconstruction simulations. For each scaled tree, we generated 100 rotasequence alignments of a realistic length (200 or 1,000 sites) using the strategy detailed above. These combinations of simulation parameters were designed to generate a broad range of evolutionary scenarios that might be encountered in empirical studies.

Under 55-state models, simulated alignments were analyzed in this form; their constituent rotasequences were converted into amino acid sequences for inference under 20-state models by masking the rotamer configuration component of their rotamer states (i.e., $(A,R) \rightarrow A$).

Log-Likelihood Comparison across Models

When selecting the best fitting model for a specific data set, an information-theoretic score such as the AIC is frequently used (Akaike 1974; Sullivan and Joyce 2005). This approach fits well with our comparison where many models of interest are nonnested and, in cases, have different state spaces. The AIC score is defined as

$$\mathrm{AIC} = 2k - 2\log\hat{L}\,, \tag{12}$$

where k is the number of estimated parameters in the model and $\hat{L}$ is the maximized value of the likelihood function of equation (11). However, the likelihood function depends upon a model's state space (Anderson and Burnham 2002): 20-state models of amino acid substitution cannot be directly compared with our 55-state model as they exist in different state spaces. Whelan and colleagues have developed a generalized “correction” allowing the comparison of likelihoods between state spaces (Whelan et al. 2015). This strategy is applicable to any two state spaces (D, C) providing that 1) each state in D maps to a single state in C and each state in C maps to a unique set of states in D and 2) both likelihoods are obtained from the same original alignments, X^C and X^D, with X^C being the “compounded” version of X^D following the state mapping. The corrected likelihood of the distinct model (D) can then be expressed in terms of the compound model (C) likelihood and an adapter function as

$$L(X^D|\theta^C) = L(X^C|\theta^C)\sum_{\text{taxa}\ p}\ \sum_{\text{sites}\ q}\frac{\pi^D_{d(p,q)}}{\pi^C_{c(p,q)}}, \tag{13}$$

where θ^C and θ^D represent the totality of parameters from C and D; $d(p,\,q)$ and $c(p,\,q)$ are the distinct and compound states observed for taxon p at site q; and $\pi^D_{d(p,q)}$ and $\pi^C_{c(p,q)}$ are these states' equilibrium frequencies in their respective substitution models. In our application of this approach, the distinct model D corresponds to RAM55, whose states can be uniquely compounded into amino acid states (e.g., TRP3 → TRP), and the compound model C corresponds to a 20-state amino acid model (e.g., WAG, LG or our RUM20) whose states can be mapped to a unique set of rotamer states (e.g., TRP → {TRP1, TRP2, TRP3}).

As an independent approach to test the contribution of knowledge of rotameric configuration-state substitutions, we generated 55-state models that were expanded versions of the 20-state LG model (Le and Gascuel 2008). Likelihoods are directly comparable to RAM55's since they share the same state space. This model expansion operation was performed with the introduction of no information about the additional states (LGexp model) or, alternatively, by accounting for just the observed frequencies of these additional states in our data set (LGbyfreq-exp). In each case, we started from LG's exchangeabilities and reconstructed a raw substitution count matrix by reversing 20-state versions of equations (7) and (2). For LGexp, this reconstructed counts matrix N was then expanded into a 55-state counts matrix ($\bar{N}$) according to

$$\bar{n}_{(A,R),(A',R')} = \frac{n_{A,A'}}{|R_A|.|R_{A'}|}, \tag{14}$$

where $|R_A|.|R_{A'}|$ is the product of the dimensions of a submatrix in $\bar{N}$ corresponding to a single cell of N. Equations (2) and (7) are then applied to $\bar{N}$ to derive the IRM for the LGexp model. This expanded model represents the “most-uninformed” expression of a 20-state model in a 55-state space, introducing rotamer states but no information about their relative frequencies or replacement rates.

Alternatively, for LGbyfreq-exp, N was expanded according to

$$\bar{n}_{(A,R),(A',R')} = \pi_{(A,R)}\pi_{(A',R')}n_{A,A'}\,, \tag{15}$$

where $\pi_{(A,R)}\pi_{(A',R')}$ is the product of RAM55's equilibrium frequencies for states $(A,\,R)$ and $(A',\,R')$. The LGbyfreq-exp expanded model's rates are therefore informed about each rotamer state's frequency, but not the relative rates of replacement between them observed in real protein sequences. We can thus compare all our models in term of their fit for a specific data set using equation (12) with the likelihood term corresponding either simply to $\hat{L}$ for 55-state models (RAM55, LGexp, LGbyfreq-exp) or to the state-corrected likelihood obtained from equation (13) for 20-state models (RUM20, LG, WAG). The latter is referred to as a “state-corrected AIC score.”

Empirical Alignments

We assessed RAM55's goodness-of-fit and performance on empirical data using rotasequence alignments (available in supplementary files, Supplementary Material online) that can be masked by removing the rotamer configuration information in order to convert then to amino acid sequences for comparison inferences with 20-state models. Alignments PF00514 and PF07714 correspond to two Pfam family alignments and their corresponding structural information from PDBe: β-catenin-like repeat and tyrosine kinase, respectively. We followed the same procedure previously used (see *Rotamer assignment and sequence alignments*) to assign rotamer states, using Pfam's domain alignment and mapping of sites to PDBe residues. These alignments are relatively short—13 taxa and 334 sites for PF0054, 82 taxa and 345 sites for PF07714—as they only include those portions of sequences Pfam recognizes as part of that family's domain. The third alignment was obtained by querying Uniprot (UniProt Consortium 2017) with the term “rubisco” and obtaining the corresponding PDBe entries with no reliance on Pfam domain alignments. Rotamer states were then assigned as described when estimating RAM55; however, in this case we did not limit ourselves to Pfam's definition of a family domain and this results in a longer alignment (46 taxa, 681 sites).

Ancestral State Reconstruction

There are two broad categories of approaches to the problem of ancestral sequence reconstruction. Marginal reconstruction assigns the most likely state to each ancestral sequence at a given site independently of the states reconstructed for other ancestral sequences at that site. Joint reconstruction instead finds an assignment of ancestral states throughout the tree that jointly maximizes the likelihood of the observed data at that site (Yang et al. 1995; Yang 2007). We used both the marginal reconstruction algorithm (Yang et al. 1995) and Pupko et al.'s implementation of the joint reconstruction algorithm (Pupko et al. 2000) to infer ancestral rotamer and amino acid sequences; we adapted both algorithms to fit our expanded state space.

To test whether our RAM55 model allowed us to correctly infer unobserved ancestral states starting from data simulated under the model itself, rotasequence alignments were simulated using RAM55, trees with fixed topology (supplementary fig. 11, Supplementary Material online) and branch lengths scaled, in turn, according to a set of factors: {0.001, 0.01, 0.1, 0.2, 0.5, 1, 1.5, 2, 3, 5}. For each scaled tree, 100 replicate alignments were generated: These included internal node rotasequences to be used as references against which inference accuracy was assessed. Masked versions of all sequences were also created, to allow 20-state model inference (i.e., inference of ancestral amino acid sequence alone) using LG. The phylogenies from these simulations were then employed alongside RAM55 or LG to reconstruct ancestral states. Finally, reconstructed rotasequences or amino acid sequences were compared position-by-position against the simulated reference sequences and the results reported in terms of percent sequence identity (percent correct inference).

We then investigated RAM55's performance when reconstructing ancestral rotasequences or amino acid sequences from empirical rotasequences. Ideally, this would be performed by comparison of inferred ancestral rotasequences with known ancestral structures. Although an increasing number of resurrected ancestral protein structures have been resolved (e.g., Konno et al. 2011; Ingles-Prieto et al. 2013; Hart et al. 2014; Risso et al. 2014; Clifton et al. 2018), their rarity, combined with the fact that most of these studies reconstruct ancestral amino acid sequences from alignment of present-day proteins that in many cases lack high-quality structural information, do not allow a systematic comparison of our reconstructed rotasequences with reference ancestral rotasequences obtained from deposited structures. To overcome this, we employed a LLOapproach (supplementary fig. 12, Supplementary Material online) in which we remove a pair of terminal sibling nodes from the alignment and proceed to reconstruct all internal nodes including one of the aforementioned pair of taxa according to the marginal or joint algorithms. (Pairs of sibling terminal nodes (*A*, *B*), as opposed to single terminal nodes (*A*), were removed as otherwise a remaining close neighbor of *A* could allow for easy reconstruction of *A*'s sequence.) LLO allows us to compare the inferred terminal sequence against the known original, as a proxy for the desired comparison. This approach was first validated on terminal sequences simulated under RAM55 and then used on empirical sequences from the PF00514 alignment; in this case the phylogeny inferred using RAxML-NG and RAM55 is used for the reconstruction process along with RAM55 or LG.

Code Availability

Code used to generate random trees and simulate substitutions along their branches (see *Tree generation and alignment simulation*) is available at: https://bitbucket.org/uperron/ram55; Last accessed 22nd May 2019. This repository also includes our implementations of the joint and marginal reconstruction algorithms (see *Ancestral state reconstruction*), as modified for our expanded state set. In addition, we provide an example of how to obtain a rotasequence alignment, suitable for tree inference with RAxML-NG and RAM55, using a user-submitted set of Uniprot IDs and the PDBe API.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Nicola De Maio and Claudia Weber at European Molecular Biology Laboratory European Bioinformatics Institute who provided insight and expertise that greatly assisted this work. This work was supported by European Molecular Biology Laboratory (U.P. and N.G.), the Klaus Tschira Foundation (A.M.K. and A.S.), and the Biotechnology and Biological Sciences Research Council (Future Leader Fellowship BB/N011600/1, to I.H.M.).

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 19(6):716–723.

Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, et al. 2017. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput*. 13(6):3031–3048.

Anderson DR, Burnham KP. 2002. Avoiding pitfalls when using information-theoretic methods. *J Wildl Manage*. 66(3):912–918.

Arenas M, Weber CC, Liberles DA, Bastolla U. 2017. ProtASR: an evolutionary framework for ancestral protein reconstruction with selection on folding stability. *Syst Biol*. 66(6):1054–1064.

Bastolla U, Porto M, Roman HE, Vendruscolo M. 2003. Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *J Mol Evol*. 56(3):243–254.

Bastolla U, Porto M, Roman HE, Vendruscolo M. 2006. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. *BMC Evol Biol*. 6(1):43.

Bergsma W. 2013. A bias-correction for Cramér's *V* and Tschuprow's *T*. *J Korean Stat Soc*. 42(3):323–328.

Carroni M, Saibil HR. 2016. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods* 95:78–85.

Challis CJ, Schmidler SC. 2012. A stochastic evolutionary model for protein structure alignment and phylogeny. *Mol Biol Evol*. 29(11):3575–3587.

Clark JJ, Benson ML, Smith RD, Carlson HA. 2019. Inherent versus induced protein flexibility: comparisons within and between apo and holo structures. *PLoS Comput Biol*. 15(1):e1006705.

Clifton BE, Kaczmarski JA, Carr PD, Gerth ML, Tokuriki N, Jackson CJ. 2018. Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nat Chem Biol*. 14(6):542–547.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Vol. 5, chapter 22. Silver Spring (MD): National Biomedical Research Foundation.

Dunbrack RL. 2002. Rotamer libraries in the 21st century. *Curr Opin Struct Biol*. 12(4):431–440.

Dunbrack RL, Cohen FE. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*. 6(8):1661–1681.

Dunbrack RL, Karplus M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol*. 230(2):543–574.

Eck RV, Dayhoff MO. 1966. Atlas of protein sequence and structure. Silver Spring (MD): National Biomedical Research Foundation.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.

Felsenstein J. 2004. Inferring phylogenies. 1st ed. Sunderland (MA): Sinauer Associates.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res*. 42(D1):222–230.

Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*. 26(8):1879–1888.

Ginalski K. 2006. Comparative modeling for protein structure prediction. *Curr Opin Struct Biol*. 16(2):172–177.

Golden M, García-Portugués E, Sørensen M, Mardia KV, Hamelryck T, Hein J. 2017. A generative angular model of protein structure evolution. *Mol Biol Evol*. 34(8):2085–2100.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149(1):445–458.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862–864.

Harms MJ, Thornton JW. 2013. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet*. 14(8):559–571.

Hart KM, Harms MJ, Schmidt BH, Elya C, Thornton JW, Marqusee S. 2014. Thermodynamic system drift in protein evolution. *PLoS Biol*. 12(11):e1001994.

Herman JL, Challis CJ, Novák Á, Hein J, Schmidler SC. 2014. Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Mol Biol Evol*. 31(9):2251–2266.

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SM, Amode R, Brent S, et al. 2016. Ensembl comparative genomics resources. *Database* 2016: bav096.

Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. *Mol Biol Evol*. 19(5):698–707.

Huelsenbeck JP, Rannala B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276(5310):227–232.

Ingles-Prieto A, Ibarra-Molero B, Delgado-Delgado A, Perez-Jimenez R, Fernandez JM, Gaucher EA, Sanchez-Ruiz JM, Gavira JA. 2013. Conservation of protein structure over four billion years. *Structure* 21(9):1690–1697.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Konno A, Kitagawa A, Watanabe M, Ogawa T, Shirai T. 2011. Tracing protein evolution through ancestral structures of fish galectin. *Structure* 19(5):711–721.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 39(1):309–338.

Korostensky C, Gonnet GH. 2000. Using traveling salesman problem algorithms for evolutionary tree construction. *Bioinformatics* 16(7):619–627.

Kosiol C, Goldman N. 2005. Different versions of the Dayhoff rate matrix. *Mol Biol Evol*. 22(2):193–199.

Kozlov A, Darriba D, Flouri T,Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics, in press. https://doi.org/10.1093/bioinformatics/btz305.

Kullback S, Leibler RA. 1951. On information and sufficiency. *Ann Math Stat*. 22(1):79–86.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol*. 29(10):2921–2936.

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

Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, et al. 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*. 487:545–574.

Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. *Genome Res*. 8(12):1233–1244.

Liò P, Goldman N, Thorne JL, Jones DT. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* 14(8):726–733.

Liu Y, Palmedo P, Ye Q, Berger B, Peng J. 2018. Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst*. 6(1):65–74.e3.

Lovell SC, Word JM, Richardson JS, Richardson DC. 2000. The penultimate rotamer library. *Proteins* 40(3):389–408.

Milne JL, Borgnia MJ, Bartesaghi A, Tran EE, Earl LA, Schauder DM, Lengyel J, Pierson J, Patwardhan A, Subramaniam S. 2013. Cryo-electron microscopy—a primer for the non-microscopist. *FEBS J*. 280(1):28–45.

Najmanovich R, Kuttner J, Sobolev V, Edelman M. 2000. Side-chain flexibility in proteins upon ligand binding. *Proteins* 39(3):261–268.

Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 2008. Environment-specific amino-acid substitution tables—tertiary templates and prediction of protein folds. *Protein Sci*. 1(2):216–226.

Overington J, Johnson MS, Sali A, Blundell TL. 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc R Soc Lond B* 241(1301):132–145.

Perron U, Moal I, Thorne J, Goldman N. Forthcoming. Probabilistic models for the study of protein evolution. In: Balding D, Moltke I, Marioni J, editors. Handbook of statistical genetics. 4th ed. New York:Wiley-Interscience.

Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*. 17(6):890–896.

Ramachandran GN, Ramakrishnan C, Sasisekharan V. 1963. Stereochemistry of polypeptide chain configurations. *J Mol Biol*. 7:95–99.

Rios S, Fernandez MF, Caltabiano G, Campillo M, Pardo L, Gonzalez A. 2015. GPCRtm: an amino acid substitution matrix for the transmembrane region of class A G Protein-Coupled Receptors. *BMC Bioinformatics* 16(1):206.

Risso VA, Gavira JA, Gaucher EA, Sanchez-Ruiz JM. 2014. Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins* 82(6):887–896.

Robinson DM, Jones DT, 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.

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

Service R. 2018. Google's DeepMind aces protein folding. Science. doi:10.1126/science.aaw2747.

Shakhnovich E, Abkevich V, Ptitsyn O. 1996. Conserved residues and the mechanism of protein folding. *Nature* 379(6560):96–98.

Shapovalov MV, Dunbrack RL. 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 19(6):844–858.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu Rev Ecol Evol Syst*. 36(1):445–466.

Thorne J, Goldman N. 2007. Probabilistic models for the study of protein evolution. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics, chapter 14. 3rd ed. New York:Wiley-Interscience.

Trueblood K, Bürgi H-B, Burzlaff H, Dunitz J, Gramaccioli C, Schulz H, Shmueli U, Abrahams S. 1996. Atomic dispacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. *Acta Crystallogr A Found Crystallogr*. 52(5):770–781.

UniProt Consortium 2017. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*. 45(D1):D158–D169.

Velankar S, Best C, Beuth B, Boutselakis CH, Cobley N, Sousa Da Silva AW, Dimitropoulos D, Golovin A, Hirshberg M, John M, et al. 2010. PDBe: Protein Data Bank in Europe. *Nucleic Acids Res*. 38(Database issue):D308–D317.

Venien-Bryan C, Li Z, Vuillard L, Boutin JA. 2017. Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery. *Acta Crystallogr F Struct Biol Commun*. 73(Pt 4):174–183.

Wang S, Sun S, Li Z, Zhang R, Xu J. 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol*. 13(1):1–34.

Wheeler LC, Lim SA, Marqusee S, Harms MJ. 2016. The thermostability and specificity of ancient proteins. *Curr Opin Struct Biol*. 38:37–43.

Whelan S, Allen JE, Blackburne BP, Talavera D. 2015. ModelOMatic: fast and automated model selection between RY, nucleotide, amino acid, and codon substitution models. *Syst Biol*. 64(1):42–55.

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.

Xu J. 2018. Distance-based protein folding powered by deep learning. bioRχiv. doi:10.1101/465955.

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, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141(4):1641–1650.

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.

Zavodszky MI, Kuhn L. A. 2005. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci*. 14(4):1104–1114.

Zhao S, Goodsell DS, Olson AJ. 2001. Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation. *Proteins* 43(3):271–279.

Zoller S, Schneider A. 2013. Improving phylogenetic inference with a semiempirical amino acid substitution model. *Mol Biol Evol*. 30(2):469–479.