EMBL International PhD Programme

University of Cambridge

The Influence of Structural Constraints on Protein Evolution

Umberto Perron

This dissertation is submitted for the degree of

Doctor of Philosophy, Biological Sciences

EMBL-EBI & Clare Hall College,

May 2020

Declaration of Authorship

This thesis is the result of my own work and includes no collaborative contributions except where specified in the chapter preface and referenced in the text. It is not substantially the same as any work that has already been submitted before in whole or in part for any other degree or qualification in this, or any other university. This manuscript contains 26,439 words and does not exceed the prescribed word limit for the Faculty of Biology Degree Committee.

Umberto Perron

May 2020

Abstract

The Influence of Structural Constraints on Protein Evolution

by Umberto Perron

Few mathematical models of sequence evolution incorporate parameters describing protein structure, despite its high conservation, essential functional role and the increasing availability of structural data. The primary goal of my PhD project was to create a structurally aware amino acid substitution model in which proteins are represented using an expanded alphabet that relays both amino acid identity and structural information.

Each character in this alphabet 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. I generated a 55-state "Dayhoff-like" substitution model (RAM55) by assigning rotamer states in 79,558 structures (∼50% of all PDBe entries) and identifying substitutions between closely related sequences. RAM55's rotamer state exchange patterns clearly show that the evolutionary properties of amino acids depend strongly upon side chain geometry. Exploiting knowledge of these patterns assists in phylogenetic analyses: I show that RAM55 performs as well as or better than traditional 20-state models on simulated and empirical data for divergence time estimation, tree inference, side chain configuration prediction and ancestral sequence reconstruction.

Further, encoding observed characters in an alignment as ambiguous representations of characters in a larger state-space allows the application of RAM55 to 20-state amino acid data for which structures are not known. Adding structural information to as few as 12.5% of the sequences in an amino acid alignment results in excellent ancestral reconstruction performance compared to a benchmark that considers the full rotamer state information. This strategy significantly expands the applicability of RAM55 to real-world scenarios where structure might only be available for some of the sequences of interest.

Thus, not only is rotamer configuration a valuable source of information for phylogenetic studies, but modelling the concomitant evolution of sequence and structure may have important implications for understanding protein folding and function.

Contents

1. Introduction

The following material has been adapted and expanded from two first-author publications [133, 134]; the original manuscripts were written by me, edited collectively, and agreed upon by all co-authors.

Evolutionary models are a prerequisite for a number of common bioinformatic tasks such as identification of homologous sequences, phylogenetic tree inference, sequence and structure alignment, variant effect prediction, and protein structure modelling. Because of this, the development and improvement of model-based approaches to studying protein evolution has been a key area of research in bioinformatics from very early on and breakthroughs have had wide-spread, positive consequences. Although current models are far more realistic than the earliest ones, our understanding of protein evolution is still inadequate in many areas and the development of better evolutionary models is still an important area of research.

In this chapter I introduce the class of models most often used to describe the evolution of proteins with a brief overview of their development over the past 50 years. I then outline the main subject of my PhD project, the development of the RAM55 rotamer-aware model, and its applications for phylogenetic inference, ancestral sequence reconstruction, and homology modelling.

1.1 Evolutionary modelling approaches

Evolutionary models and sequence alignments have a central role in bioinformatics and are tightly interconnected, as pointed out early in their development [151]. In this context, it is important to distinguish between phylogenetic alignments and structural (or structure-aware) alignments of protein sequences. In the former, characters placed in the same column are believed to be evolutionarily homologous and share a common ancestor [116], whereas, in the latter, structural homology indicates that the 3-dimensional (3D) conformation of two residues and their positioning within the protein structure are highly similar [123]. This has obvious consequences for the interpretation of these two classes of alignments and the necessity to rely on phylogenetic alignments for informing evolutionary model inference.

Phylogenetic alignment strategies aim to represent common ancestry by arranging the amino acid sequences and inserting gaps between the residues so that identical or similar amino acids, which are considered to be descendent from a common ancestor, are placed in the same column and successive amino acid are placed in successive columns. In this process, amino acid similarity is quantified by some scoring scheme for amino acid matches and mismatches, and by a penalty that is applied when a gap is introduced in either sequence [42, 123]. Amino acid substitution matrices (also known as Q matrices), the core component of protein evolutionary models, are often used for phylogenetic alignment scoring as they describe general evolutionary preferences for certain substitutions over others. In a somewhat circular relationship, empirical substitution matrices are themselves estimated from large datasets of protein alignments.

When studying the evolution of amino acid sequences, substitutions are usually described using a continuous-time Markov model [47, 189]. Each state of the Markov chain represents one of the 20 standard amino acids at a given site in the alignment and at a particular position in time. The main feature of a Markov chain is that the probability with which the chain jumps into other states only depends on the current state and not on how the current state was reached (the Markovian property). In a Markov chain the

current state is observable, and therefore the state (i.e. the amino acid) transition probabilities are the only parameters [42]. In a hidden Markov model (HMM), the current state cannot be directly observed, but the output, which is dependent on the current state, is visible. Each hidden state generates a distinct probability distribution over all possible observable output states. Therefore, the sequence of observable outputs generated by an HMM contains information about the hidden sequence of states [42]. HMM hidden states can be used to represent protein structural features (e.g. secondary structure) that are not explicitly encoded in the amino acid sequence. Each secondary structure state, however, has a distinct amino acid transition probability distribution which can be detected from the visible amino acid output. This strategy allows us to generate an integrated model of sequence and structure evolution (see section 1.4).

General models of protein evolution are based upon the assumptions that all sites in a protein sequence evolve (1) independently and (2) identically, and (3) that protein sequences evolve independently in different lineages of their common phylogeny. These independence assumptions are convenient because they allow the likelihood of an aligned set of protein sequences to be written as a product of likelihoods for individual alignment columns (i.e. site-likelihoods) [189]. More realistically, there should be variability of amino acid replacement patterns among protein families, among sites within proteins, and among different lineages. In this chapter I focus primarily on models that attempt to describe sitewise heterogeneity, and site inswise dependence; as these provide the necessary background for my PhD project. A discussion of heterotachy [114], the variations of lineage-specific evolutionary rates over time, is beyond the scope of this introduction.

In traditional evolutionary models, site-likelihoods are calculated using Felsenstein's efficient pruning algorithm [48, 47]. In brief, the algorithm requires the calculation of transition probabilities: the probabilities of a possible state at the end of a branch on a phylogenetic tree, conditional upon the state at the beginning of the branch, the amount of evolution represented by the branch (i.e. the branch length), and the values of other model parameters, including the exchange rates from a substitution matrix. Computation of transition probabilities from underlying exchange rates of a Markov process in a maximum likelihood (ML) framework is a standard procedure (e.g. [111], see also section 4.2)).

For models that violate the site-independence assumption the pruning algorithm is often computationally intractable, resulting in limited model applicability.

Likelihoods from different evolutionary models can be compared in terms of fit for specific dataset of homologous sequences, with the assumption that the highest likelihood indicates the model that better approximates the observed evolutionary process [47, 189]. Information-theoretic scores such as the Akaike Information Criterion [3] are frequently used to account for different parametrization across models [189], and further corrections are required when comparing models with different state-spaces (see section 4.3)

1.2 Empirical models of amino acid replacement

Models belonging to the empirical class are built by analysing large quantities of sequence data (typically 100 or more protein alignments) and estimating relative substitution rates between amino acids under a time-reversible model. This means that the rate of exchange $q_{X,Y}$ from amino acids X to Y is assumed to satisfy the detailed-balance condition:

$$\pi_X q_{X,Y} = \pi_Y q_{Y,X} \text{ , for any } X \neq Y \tag{1.1}$$

where π_X and π_Y are the equilibrium frequencies of the amino acids ([189], p. 36). Another way of expressing the time-reversibility property is that the evolutionary process will behave identically when moving from the present state forward into a future state as when going from the present state backwards towards a past state. While there is no biological reason to expect the substitution process to be exactly reversible, reversibility is convenient for reasons of mathematical and computational tractability [189].

The first and most influential empirical amino acid substitution model was introduced by Dayhoff and colleagues in 1966 [45], and subsequently updated as additional protein sequence data became available [32, 33, 34]. Dayhoff and Eck estimated the exchange rates $q_{X,Y}$ via a technique that relied upon comparing closely-related (i.e. 85% or more sequence identity) protein sequences. An alignment of closely-related sequences will be identical at most positions. At the few positions where the sequences differ, the cause of

the difference will usually be a single amino acid replacement event as the chance that two or more replacement events occurred at a position will be negligible. Dayhoff and Eck assembled closely-related protein sequences available at the time, tabulated amino acid changes along branches of the phylogenetic tree and, from this collection of difference counts, were able to implicitly infer the relative rates of the possible different amino acid replacements [92]. Different methods for recovering the exchange rate matrix Q from results such as those presented by Dayhoff and Eck were subsequently proposed, and are reviewed by Kosiol and Goldman [92].

More recent techniques for estimating the exchange rate matrix (e.g. [181, 75, 104, 29]) have improved in their ability to deal with much larger training data sets, and make correct allowance for varying levels of sequence divergence and the consequent probabilities of multiple replacements at a position. The latter is particularly relevant since being able to handle more divergent sequences removes the need for high sequence identity thresholds and significantly expands the amount of available data for model estimation. Continuous-time models permit this and thus benefit from more training data while remaining just as applicable for closely or distantly related sequences.

A large number of alternative exchange rate matrices for amino acid substitution have been developed over the years, using approaches similar to Dayhoff's but taking advantage of larger or more specific datasets with the aim of developing more broadly-applicable general models as well as models tailored to specific contexts and constraints. Examples include the widely-used JTT model [80], several mutation matrices estimated by Gonnet and colleagues [62], MTMAM [191] for mammalian mitochondrial proteins, the WAG matrix [181] computed from nuclear protein alignments, LG [104] that uses a large and diverse database of sequence alignments, GPCRtm for the transmembrane region of GPCRs [140], and gene-specific substitution profiles to model antibody somatic hypermutation [157].

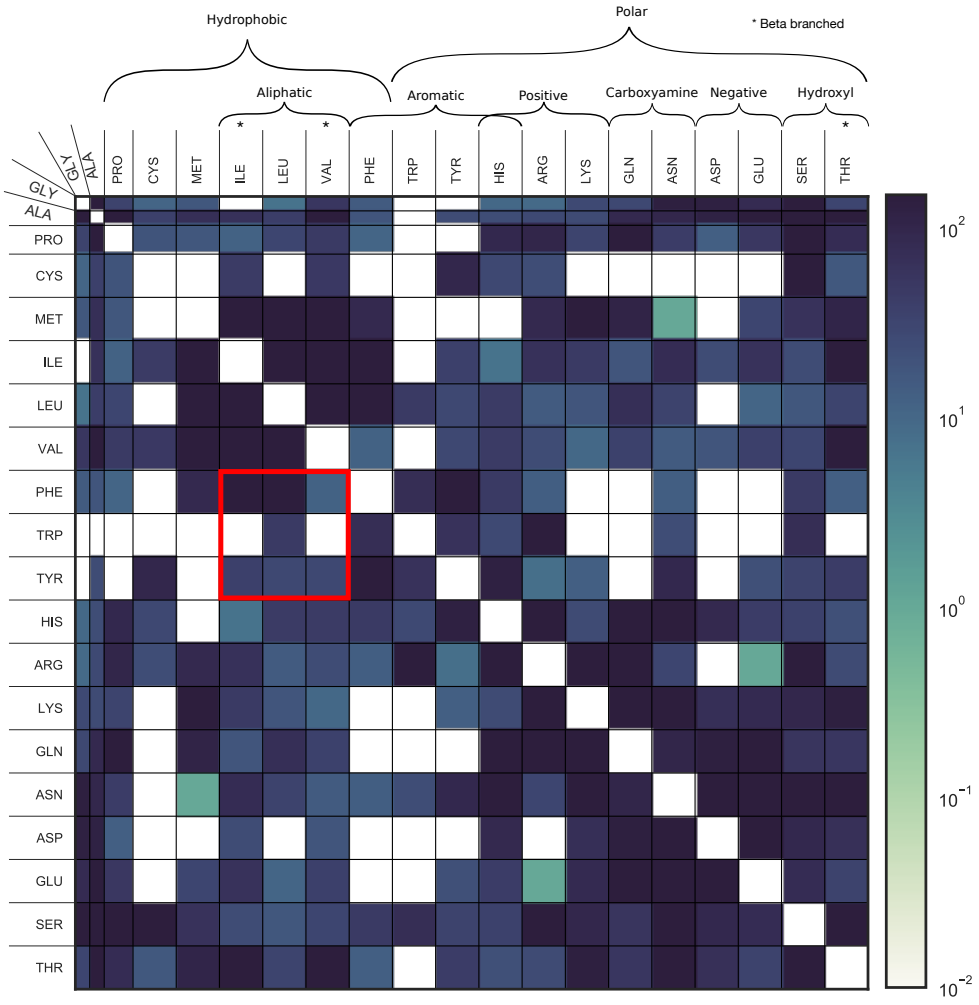

Figure 1.1: The Dayhoff-Eck model as updated by Dayhoff, Schwartz, and Orcutt [34] in 1978. Replacement rates are shown in heatmap form; amino acids are grouped by biochemical properties. Aromatic-aliphatic exchanges (in red) highlight how replacement patterns vary among physicochemically similar amino acids.

1.3 Family-specificity and site heterogeneity

From observing any of the exchange matrices mentioned so far (e.g. the 1978 Dayhoff model, shown in Fig. 1.1), it is quite apparent that amino acid replacements tend to involve chemically similar residues. There have been numerous attempts to classify amino acids on the basis of their physicochemical properties, and physicochemical distances among amino acid types have been proposed (e.g. [63, 164]). However, modelling the evolutionary process by the physicochemical distances between residues alone is unlikely to fully reflect the complexity of observed replacement patterns [191, 195]. Already in the Dayhoff-Eck model, it is clear that amino acids sharing similar properties can behave as

distinct entities. Phenylalanine, tryptophan, and tyrosine all have aromatic side chains and yet display distinct patterns when exchanging with aliphatic residues: phenylalanine strongly favours exchanging with isoleucine and leucine over valine, tryptophan almost exclusively exchanges with leucine, and tyrosine exchanges with all three aliphatics almost equally (Fig. 1.1). The Dayhoff-Eck approach is appealing because its estimates of relative amino acid replacement rates are empirically derived and able to provide a more nuanced representation of the constraints acting on protein evolution. However, while the Dayhoff-Eck approach captures broadly-applicable amino acid exchange propensities, it describes evolution of the 'average site' in the 'average protein' at an 'average moment' in time.

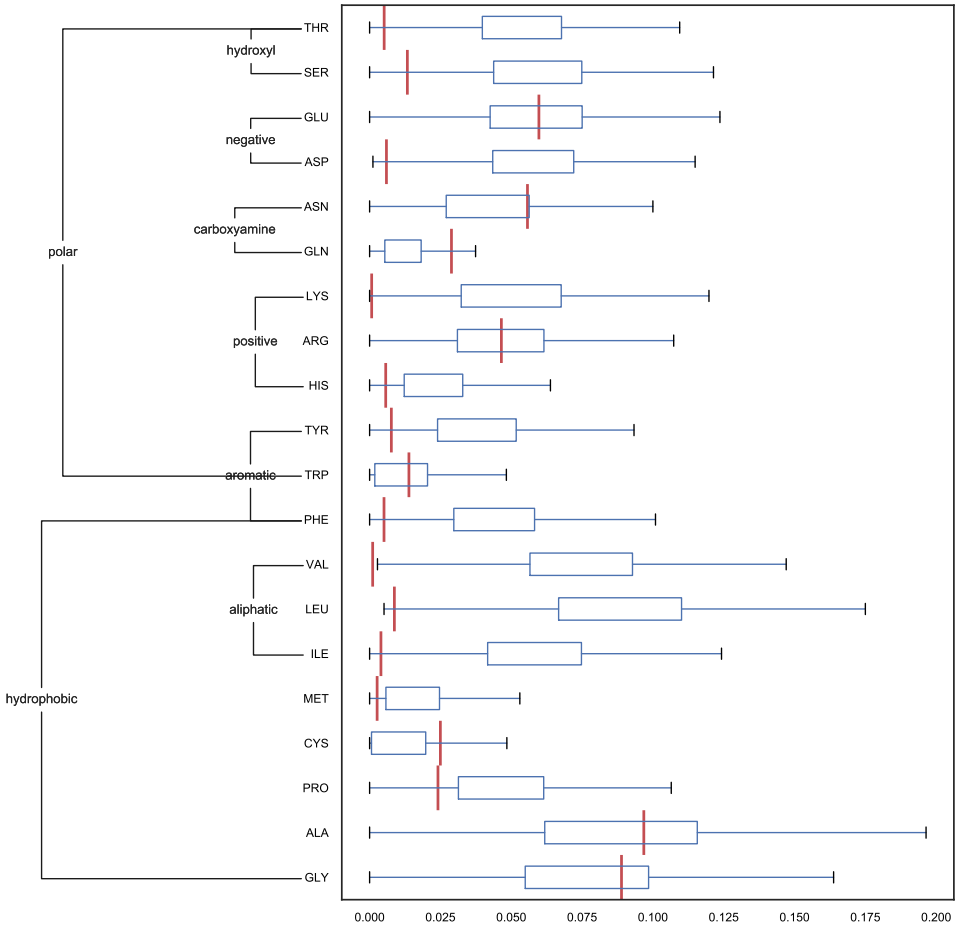

Figure 1.2: Amino acid composition varies across protein families. Box plots show the distribution of empirical amino acid equilibrium frequencies computed from 3061 Pfam [46] family alignments. Mean amino acid frequencies across all families are shown in red. The y-axis groups the amino acids according to their physicochemical properties. Some amino acid show much higher variability (e.g. glycine, alanine, cysteine, leucine, valine) while others (e.g. methionine, tryptophan, glutamine) are more consistent across families.

As mentioned above, researchers have inferred models for specific types or families of proteins, aiming for greater accuracy through greater specificity. For example Adachi and colleagues [1] derived amino acid replacement rate estimates from a data set consisting of all the 13 mitochondrial proteins from each of 20 different vertebrate species. They detected some substantial differences between their estimates and those derived from nuclear-encoded proteins. Similar findings were made over the years for protein evolution models derived from chloroplast genes [2], retroviral polymerase proteins [37], and influenza proteins [30].

A large portion of these differences has been attributed to variation in composition of amino acids across protein families (see Fig. 1.2), while exchange patterns for each amino acid are believed to remain largely unchanged across related organisms (e.g. mammals, vertebrates) or structurally and functionally similar proteins (e.g. transmembrane, globular proteins), but might vary between these [19, 112]. Dayhoff-type models can readily be modified to allow variation in composition of amino acids among proteins. Denoting by π_X and $q_{X,Y}$ the amino acid frequencies and replacement rates estimated with a Dayhoff-type approach, and by π'_X and $q'_{X,Y}$ the corresponding values we might choose to use in the analysis of a particular protein, then the π'_X can be treated as parameters specific to each protein of interest to allow variation of amino acid composition among proteins. This is achieved by forcing the $q'_{X,Y}$ for $X \neq Y$ to obey

$$q'_{X,Y} = \frac{\pi'_Y}{\pi_Y} q_{X,Y} \tag{1.2}$$

[19]. The resulting Markovian replacement process defined by the $q'_{X,Y}$ combines empirical information derived from databases (the $q_{X,Y}/\pi_Y$, also known as exchangeabilities [181]) with the amino acid frequencies for the particular protein under study (π'_X), now estimated from that data set.

The benefit of this hybrid parameterization can be a sizeable improvement in model fit [19]. It is also attractive because it maintains the biologically realistic property of allowing amino acid replacements to occur among chemically similar amino acids while estimating the 20 amino acid frequencies π'_A from individual protein alignments. The

result is both to increase our understanding of the importance of particular amino acid residues in different proteins and to improve robustness of phylogenetic inferences.

Not only is there variability of amino acid replacement among proteins, but there is also variability of amino acid replacement rates among sites within a single protein. Yang [188] introduced a practical method for incorporating heterogeneity of evolutionary rates among sites into models of nucleotide or amino acid substitution. Yang's innovation was to discretise a continuous gamma distribution of rates among sites into a relatively small number (C) of categories. Each site is then modelled as having its rate randomly drawn from this discretised gamma distribution. The discretisation itself is computationally convenient rather than biologically meaningful: evaluation of a likelihood with Yang's treatment of rate variation among sites requires an amount of computation that is approximately a factor C more than would be needed if all sites were assumed to share identical rates. Yang demonstrated that discretisation of a gamma distribution into a relatively small number of rate categories (i.e. four to six) generally fits data well, with more categories giving no substantial change in parameter estimates or improvement in terms of model fit [188].

It has been convincingly established that, for the majority of protein coding DNA sequences, allowing heterogeneity of nucleotide substitution rates over sites represents a great improvement over models that assume homogeneity of evolutionary rate, a fact partly attributable to the triplet coding nature of DNA and to the structure of the genetic code. It is also now known that amino acid replacement models often fit data much better when allowance is made for rate variation among amino acid positions (e.g. [**c**, 61, 103, 41]). This has confirmed that rate variation over protein sites is a widespread evolutionary phenomenon that correlates with several biophysical traits [43], and should be considered in all phylogenetic analyses of protein-coding sequences.

1.4 High-level structural context

Although allowing empirically for particular proteins' distinct amino acid compositions and among-site rate variation has provided large improvements in fits of models of protein

evolution, the models described so far shed little light on the cause of rate variation among sites. They also account for, but do not explain, different amino acid compositions between protein families, and different exchangeabilities between different amino acid pairs.

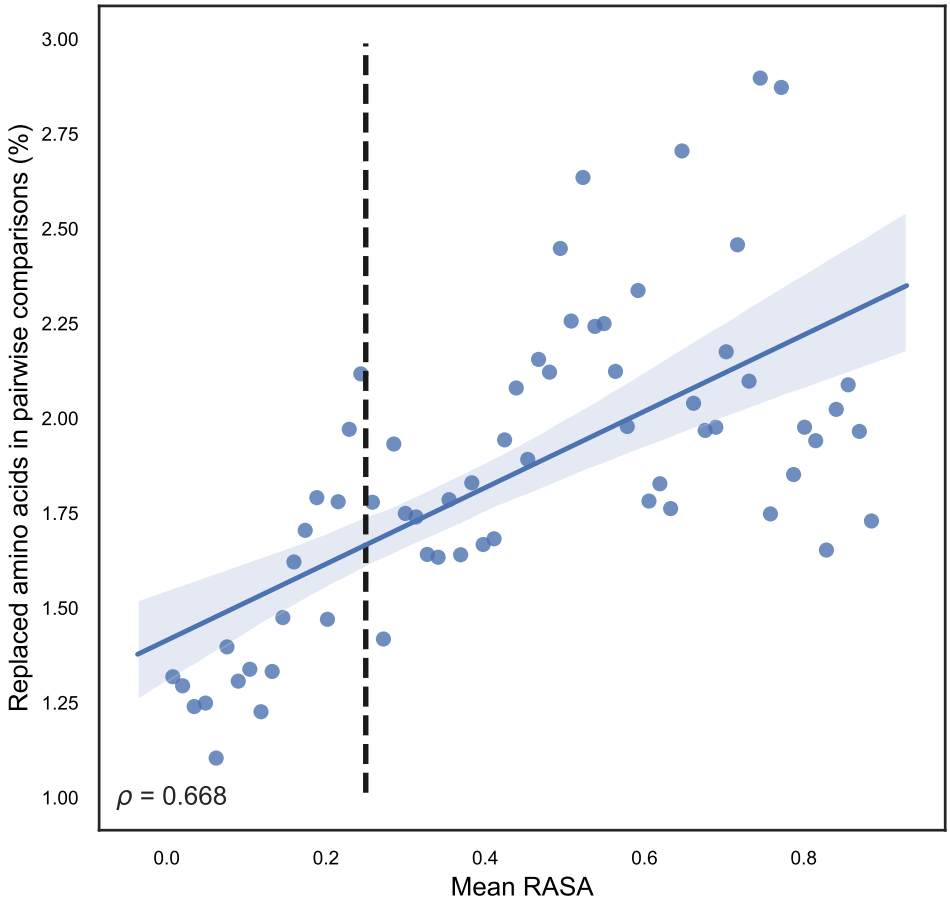

Figure 1.3: Solvent exposed sites are less conserved than buried sites. Relative accessible surface area (RASA) [168] and amino acid conservation are computed across 5659 Pfam [46] family alignments and 1.22×10^6 sites. RASA is divided in 64 bins and the mean RASA for each bin is shown on the x-axis. The y-axis shows the percentage of non-conserved sites in pairwise sequence comparisons for each RASA bin. The vertical dashed line represents the usual $< 25\%$ RASA threshold separating buried and exposed residues [108]. There is a strong tendency (Spearman's ρ 0.668) for more accessible amino acids to be less conserved.

It is clear, from considering protein structure and function, that at least some of the variability will be associated with the structural environment of a site. A site on the surface of a globular protein will be exposed to solvent, typically water. Sites with high solvent accessibility therefore tend to be occupied by hydrophilic amino acids. In contrast, sites buried in the interior of a protein are less accessible to solvent and are likely to be occupied by hydrophobic amino acids. Further, exposed sites tend to evolve about

twice as fast on average as buried sites (see Fig. 1.3) [60], most likely because residues at buried sites interact with many neighbouring residues and thus replacements are more likely to disrupt the position of other neighbouring residues. The latter is especially true for replacements involving residues with distinct steric properties such as aromatics and aliphatics or possessing differently charged side chains.

One of the most striking findings in the study of protein structure has been that proteins' secondary and tertiary structure usually change very slowly during evolution [24, 147, 87]. In cases where homologous proteins are known to perform the same biological functions, it is often the case that their structures are very similar even when the sequences that code for them are quite diverged (e.g. [148]). In other words, protein sequence evolution tends in some sense to occur at a higher rate than the evolution of protein structure. This tendency can be exploited by models of amino acid replacement as a group of homologous protein sequences is likely to share a common underlying structure. Homologous amino acids thus are likely to be in a similar structural environment: for example they might all be part of an α-helix secondary structure. If the tertiary structure of one protein sequence in an alignment is available, then each alignment column can be assigned a structural environment and thus a separate 20-state model can be estimated for each environment (e.g. [127, 175, 90, 166, 60, 102, 140]). This strategy has often resulted in improved fit compared with standard general models such as JTT or WAG.

Further, the structural environment at one site in a protein is not independent of the environment at other sites. Functional sites with lower rates of evolution may cluster together in sequence space [76]. Secondary structure elements also induce clustering: for example, if a site is in an α-helix environment, then neighbouring sites along the sequence are also likely within the α-helix. It is possible to use this knowledge to further improve models of protein evolution. One approach has been the modelling of the organisation of structural environments along a protein sequence as a first-order Markov chain.

A hidden Markov model for the organisation of structural environments can be combined with the models of amino acid replacement for each structural environment to generate an integrated model of protein sequence evolution. This has been repeated for

several protein families [60, 112, 110], using structural categories based on the secondary structure and solvent accessibility status of protein residues. With this approach, each category of structural environment is modelled to have its own equilibrium amino acid frequencies and its own rates of amino acid replacement. Felsenstein's pruning algorithm [49] is used to compute likelihoods conditional on the (unobserved) states of the HMM, and the HMM's transition probabilities are incorporated to deal with uncertainty about to which structure state each site belongs to. The likelihood calculations now have a computational burden that is approximately a factor S more than without the HMM, where S is the number of distinct structure states considered: this remains fast enough to permit ML phylogenetic inference.

Leveraging the vastly increased number of resolved protein structures now available ($\sim$160000 on PDBe) [28], Lai et al. [99] developed a 7-state empirical evolutionary model based entirely on discrete secondary structure states. The authors demonstrate their model in the contexts of ML-based reconstruction of ancestral secondary structures, showing how their approach can highlight relationships that are rooted in the structural context and not evident in amino acid-based analyses.

Studies mentioned in this section illustrate the importance of structural environments to protein evolution. Different secondary structure elements exhibit distinct patterns of evolution, presumably related to different residues' propensities to adopt the local structures and biochemical properties required of protein functions. I will discuss the impact of these considerations on the development of structure-aware replacement models in later Chapters.

1.5 Site-specific structural and functional context

Some variation of amino acid composition and replacement rates among sites can be explained by consideration of high-level structural features, such as secondary structure and solvent accessibility. However, the specific biochemical function of a protein, and even of its individual sites, mean that the evolution of every site of every protein takes place under different constraints, potentially leading to different evolutionary patterns.

Bruno [18] made the first progress in this respect (also see [67]), devising a model that allowed each site in a protein to have its own equilibrium frequencies. This highly flexible approach suffers from potential overparameterization problems because incorporating 20 amino acid frequencies per site adds 19 degrees of freedom per site to the model and works best on datasets containing a large number of highly diverged sequences. Despite its limitations, this approach is biologically attractive as it allows for the exchanges involving physicochemically similar amino acids to be represented by similar site-specific frequencies as well as their generic replacement rates. These two processes are not mutually exclusive, but little is known about their relative importance.

A number of attempts to describe variation in amino acid frequencies over proteins' sites have followed Bruno's work. Wang et al. [176] utilised principal components analysis to infer the existence of at least four major site-specific amino acid frequency classes in a data set of about 20 large protein alignments. Constructing a mixture model accounting for these four amino acid classes, plus a fifth class of global frequencies, they demonstrated significant improvements in model fit for real data sets.

Adopting a Bayesian perspective, Lartillot and Philippe's 'CAT' model [100] largely overcomes the overparameterization concerns by assuming each site belongs to one of a restricted number of amino acid replacement process categories, but that the category is not directly observed. Different replacement categories share a common set of amino acid exchangeability parameters taken from existing standard models, but have different equilibrium amino acid frequencies. The number of categories and their frequencies are jointly determined by a Dirichlet process prior distribution – a continuous multivariate probability distribution commonly used as a prior in Bayesian statistics – and the set of 20 different amino acid frequencies for a given category has a Dirichlet prior as well. Lartillot and colleagues convincingly demonstrate that their Bayesian approach can be a substantial improvement over models that do not permit variation of preferred residues among sites. Although computationally intensive, implementations of the CAT approach and its variants are practical for reasonably-sized datasets.

Further work from Le and colleagues [106, 158] uses expectation–maximisation algorithms in a maximum-likelihood framework [36, 74, 86] to estimate a set of category-specific amino acid profiles and exchangeabilities from large alignment databases. As opposed to earlier models described in the previous section, categories do not explicitly map to structural classes (e.g. exposed, buried, α-helix, β-sheet, etc.). In these mixture-models, the first-level mixture of amino acid profiles is combined with a second-level mixture corresponding a gamma distribution of site rates using the approach outlined previously.

More recently, the focus has shifted towards simpler mixture-models that prioritise reasonable computing times and lower memory consumption. LG4X [103], for example, uses a multimatrix approach, similar to those in section 1.4, where sites are categorised depending on their evolutionary rate, and different replacement matrices are used for each of the four site categories. Rates and weights of the four matrices are estimated via expectation–maximisation from a very large number of alignments with no gamma distribution assumption.

Mixture-models have been shown to be flexible and accurate tools for phylogenetic inference, albeit more complex and computationally demanding ones. All modern mixture-models, including simplified models such as LG4X, show improved fit over general single-matrix models such as WAG and JTT, likely due to the mixture models' ability to implicitly capture partition-specific amino acid exchange propensities that are ultimately caused by local functional and structural constraints. In fact, some mixture categories appear to clearly correspond to buried or exposed contexts, while 'very slow' categories express the fact that even at slowly-evolving sites, substitutions between highly similar amino acids (both biochemically and in terms of the genetic code) are likely to occur [103]. For the purpose of understanding the interplay between sequence evolution and protein structure, however, mixture-models ultimately provide little additional insight over the multimatrix models described in section 1.4 as in both cases site-specific structural environments are summarised into a few coarse-grained replacement categories. Further, neither class of models explicitly considers residue positioning within a 3D protein structure, potentially missing out on the fact that distant sites in sequence space might be folded into the same

structural context. In the following section I describe a few approaches that attempt to account for folding constraints.

1.6 Stability- and structure-constrained models

The tertiary and quaternary structures of proteins provide further constraints to evolution, in the form of natural selection operating on the specific interactions between amino acids at different sites that stabilise the fold of the protein, as well as the need to avoid misfolding and aggregation [128, 154]. It follows that residue evolution at any site will affect, as a minimum, all other residues in the immediate 3D neighbourhood and their own evolution, and might, via knock-on effects, have large-scale effects on the overall protein conformation thus influencing the evolutionary process across the entire protein. This suggests that, ideally, models of protein evolution should explicitly include a representation of selective constraints on the structure and the stability of the native state. Conversely, covariation of residues at different sites may indicate, for example, that these sites are nearby in the tertiary structure of the protein, a strategy that has been used to make inferences about protein tertiary structure and function [131, 136, 169, 117, 152]. Approaches attempting to leverage dependence among sites to better model protein evolution can be broadly classified into two groups: structure-constrained models, and stability-constrained models [154, 12].

Structure-constrained models were first introduced by Parisi and Echave [130, 44] with their noteworthy technique for estimating the amount of structural change due to each mutation through an elastic network model, which represents the protein as a network of nodes placed at the α carbons and connected by oscillators [10, 11]. They begin their simulation with a known target structure, which is assumed not to change over time. They then introduce single random amino acid replacements generating a local perturbation which propagates through the elastic network and produces a collective structural deformation in the target structure. The extent of the perturbation depends both on how physicochemically similar the mutated amino acid and its replacement are, and on the overall properties of the network (i.e. the target structure). Each mutation is

then accepted or rejected according to the extent of the structural deformation it generates and a selection pressure parameter.

The underlying idea behind the Parisi and Echave technique is that a representation of the effect of mutations on structure can be employed to help parameterize rates of (non-synonymous) substitutions. In this model the change in stability is neglected, and the selective importance of the native structure is justified by the fact that it is the main determinant of protein function. In a more realistic setting, mutations modify both the structure and stability of the native state, and both of them are targeted by natural selection. Results obtained simulating sequence evolution using these structure-constrained models show strong agreement with observed replacement patterns across homologous proteins [44].

Following up on Parisi and Echave's work, simulated protein evolution with dependent change among positions has been widely employed to learn about the interplay between protein structure, expression, and evolution. Here I concentrate on the smaller number of structure-constrained evolutionary models that have been employed for likelihood-based inference.

A model of change permitting dependence among sequence positions poses significant challenges for conventional evolutionary inference procedures that use Felsenstein's [47] pruning algorithm. With a relatively general dependence structure, the number of rows and columns in the rate and transition probability matrices equals the number of possible sequences. As this grows exponentially with sequence length, Felsenstein's pruning algorithm becomes computationally intractable with dependent change among sequence sites unless sequence length is extremely short.

Fornasari, Parisi, and Echave [54] introduced one way to include amino acid replacement patterns observed in position-dependant simulations while maintaining the computational feasibility of likelihood-based inference with evolutionary independence among sites. They did so by performing large numbers of simulations with the aforementioned Parisi-Echave approach. They then tabulated the number of times in these simulations that each type of amino acid was replaced by each other amino acid at each protein

position. These replacement counts were used to derive position-specific replacement rate matrices that could be employed within a model which itself assumed independent change among protein sites. While this modelling approach is attractive and maintains the computational convenience of Felsenstein's pruning algorithm, the position-specific rate matrices are influenced by the amount of evolution that is simulated to obtain the replacement counts and the choice of this amount is somewhat arbitrary.

An alternative to Felsenstein's inference procedure was explored by Jensen and Pedersen [79, 132]. Their algorithm augments the observed sequence data at the tips of evolutionary trees with "sequence paths", which specify all of the changes that occurred on a each particular branch. For each change, the sequence path contains information about the time at which the change occurred, the sequence position at which the change occurred, and the nature of the change. Markov chain Monte Carlo techniques (MCMC) can be used to randomly sample possible sequence paths according to the probability density over paths, weighted by the correct probability of the time, position, and nature of the change to which the path corresponds. The big advantage of the sequence path approach to inference is that it is often computationally tractable in cases where Felsenstein's pruning algorithm is difficult to apply because transition probabilities cannot be calculated or well-approximated.

Combining the innovations of Parisi and Echave [130] with the augmentation strategy of Jensen and Pedersen [79, 132], Robinson and colleagues [142] designed a model of protein-coding DNA sequence evolution that assumes a globular protein's tertiary structure is known and shared by all the protein sequences being analyzed. Substitution rates in the Robinson model are intended to reflect both the mutation processes and the influence of protein structure on (non-synonymous) change. With the Robinson model, this influence was governed by the effect of mutations on the pairwise amino acid interactions and the change in hydrophobicity of the site that was modified. This influence could be assessed because the protein tertiary structure was assumed known and unchanging during evolution. The Robinson model yielded biologically plausible inferences regarding the evolutionary impact of hydrophobicity and solvent accessibility [142, 143, 23].

Employing similar parameterizations, Rodrigue and colleagues [144] demonstrated that evolutionary models with dependence among sequence positions due to tertiary structure are statistically superior to those without dependence. However, these authors also concluded that treatments of protein tertiary structure available at the time were not sufficient to produce satisfactory evolutionary models. They also found that adding rate heterogeneity among sites via Yang's discrete gamma technique generates model fit improvements beyond those realised solely by incorporating tertiary structure, meaning that "arbitrary" rate variation via aγ model proved more effective than their tertiary structure model.

Subsequent efforts to incorporate site interdependencies brought about by structural constraints have employed more sophisticated treatments of protein structure [85, 145, 16, 84]. In these cases, an energy-like scoring system for sequence–structure compatibility (often termed a statistical potential) is used to evaluate the probability of fixation of a given mutation, assuming an underlying protein structure that remains constant. Terms related to pairwise residue contact interactions, torsional angles, solvent accessibility and residue flexibility are included in the potentials, so as to study the effects of the main factors known to influence protein structure. However, these structurally-constrained models are often outperformed by some of the available site-independent models in terms of fit, possibly indicating that alternatives to coarse-grained statistical potentials should be explored to capture the full complexity of structural constraints on protein evolution [84].

More recent stability-constrained protein evolution models focus instead on the stability of the protein's native state (as opposed to unfolded and misfolded states) and attempt to estimate the folding free energy of the native state of a mutated protein under the assumption that each mutation maintains the same residue contacts as the original native state [7, 12]. Most of these models require the computation of protein effective free energy, approximated as a sum of pairwise contact interactions [13]. This, in turn, introduces probabilistic dependencies between protein sites, and, as discussed for structure-constrained models in this section, it enormously increases the complexity of likelihood computations.

Arenas and colleagues developed a mean-field (MF) substitution model that generates independent site-specific amino acid distributions with constraints derived from knowledge about inferred stability of protein structures as they evolve [7]. These in turn are based primarily on the number of other residues with which the residue at a particular site is in close contact. This model depends on amino acid composition (i.e. amino acid equilibrium frequencies) and one selection parameter that is estimated by maximising the likelihood of the protein sequence. When analysing under this model, the most variable sites are those with an intermediate number of native contacts. The MF model yields larger likelihoods than models that only consider the native state, and produces stable sequences for most proteins, with more realistic average hydrophobicity likely due to its explicit attempt to avoid highly-stable but misfolded conformations. Later work by Arenas et al. [8] represents a step toward a viable implementation of stability-constrained MF models for phylogenetic purposes, showing that ancestral protein sequences inferred under the MF model have significantly more realistic folding free energies than ancestral protein reconstructed with traditional empirical models.

Structure- and stability-constrained approaches described in this section yield models of protein evolution that are biologically meaningful and statistically valuable, demonstrating the benefits of introducing explicit structural constraints to the evolutionary process. However, in order to side-step the difficulties that arise when site independence can no longer be assumed, these models are very complex and require several 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. As is the case for mixture-models, there appears to be a trade-off between a model's ability to accurately describe complex evolutionary constraints and its viability for phylogenetic inference, ultimately leading to a simplification of the original approach (e.g. LG4X, MF) which, ideally, will still benefit from a more sophisticated modelling strategy.

1.7 Stochastic models of structural evolution

A substantially different approach to modelling how protein tertiary structure changes over evolutionary time was proposed by Challis and Schmidler [21]. In their approach, a protein's structural information is summarised 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. The Challis-Schmidler model of structure is quite simple, and assumes independent change in spatial locations among the amino acids. For each of these amino acids, it further assumes that three separate Ornstein-Uhlenbeck (OU) processes are operating, with these three processes independently affecting the three coordinates that specify the location of an α-carbon in three-dimensional Euclidean space. This model of spatial evolution of amino acids is computationally tractable because (independent) OU processes are time-reversible and have both normally distributed transition probabilities and normally distributed stationary distributions. Implementations of the Challis-Schmidler model for phylogenetic inference purposes [71] are practical, 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, the locations of consecutive residues are correlated and constrained to specific torsional angles. Golden et al. [59] have proposed a model of structural evolution that describes the evolution of protein tertiary structure using a specialised stochastic process that operates in dihedral angle space. The Golden model, while 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. Subsequent work by García-Portugués et al. [57] provides an improved probabilistic framework

that accounts for dependencies between sequence, structure and alignment uncertainty, and models local dependencies between aligned sites in both sequence and dihedral angle space. This model allows for both 'smooth' conformational changes and 'catastrophic' conformational jumps.

These recent stochastic approaches promise to deliver more biologically plausible models of protein evolution by simultaneously modelling sequence and structure change over time while accounting for alignment uncertainty and local dependencies among residues. For the purpose of phylogenetic inference, stochastic models are, however, currently limited by their inability to analyze more than a pair of proteins at the time for reasons of computational tractability. While Felsenstein's algorithm can be applied to discrete ancestral states in a computationally efficient manner, marginalizing the continuous ancestral dihedral angle states is likely to require more expensive MCMC algorithms.

1.8 A rotamer-aware model

From assessing the development of protein evolution modelling strategies over the past decades, I believe it clearly emerges that, rather than a comprehensive model that can only be applied to a minimal amount of data, it is preferable to concentrate on biologically meaningful improvements within the general framework of existing phylogenetic approaches; thus ensuring that any novel model can both be computed from, and applied to, modern large-scale datasets and tasks.

The primary goal of my PhD project was to develop a structure-aware model of protein evolution that was as simple as possible and as close as possible to the 'Dayhoff-Eck' empirical approach. This would ensure its compatibility with well-established phylogenetic inference strategies (ML-based inference, Felsenstein's algorithm) and allow for its implementation in popular phylognetics software, maximising its utility for the molecular evolution community. Computational complexity and reliance on MCMC algorithms have, in fact, limited several of the most promising approaches illustrated previously, specifically those attempting to constrain sequence evolution using structure or stability, and stochastic models attempting to model sequence and structure change simultaneously

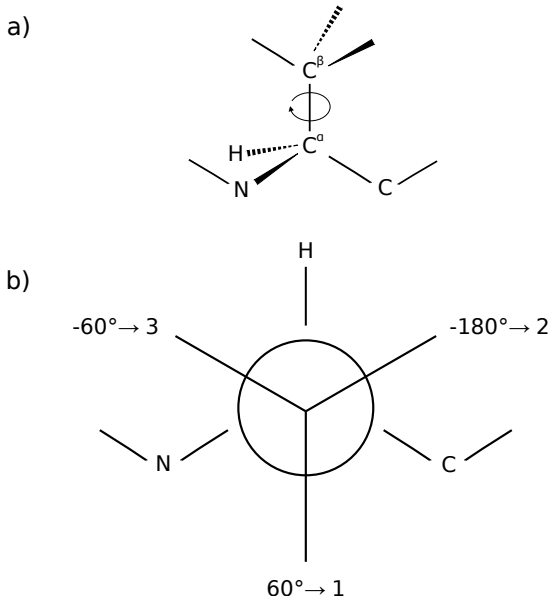

Figure 1.4: Definition of rotamer configurations. **(a)** A χ_1 rotamer configuration is defined by the dihedral angle generated by the rotation of the C^{α} - C^{β} bond (curved arrow). **(b)** The three typical stable configurations correspond to specific χ_1 dihedral angle values: $\sim$60° for configuration 1 (also known as gauche +), $\sim$180° for configuration 2 (trans), and $\sim -60°$ for configuration 3 (gauche −).

in continuous spaces. Despite their good statistical qualities, there is no widely-used implementation for these more complex models. In contrast, there are multiple popular packages implementing the simpler models including PAML [190], PhyML [65], IQ-TREE [126], MEGA [98], and RAxML-NG [94]. Can an informative structure-aware model retain site-independence and a discrete state space?

In order to accomplish this, my supervisors and I settled on using a single residue feature that could act as an information-rich structure proxy by having its state determined by local folding constraints on side chain orientation. We preferred this strategy over modelling backbone configuration change over time, despite its large contribution to protein folding and information content, because each residue could be easily assigned one of a small set of discrete, well-established side chain configuration states.

This was achieved by developing an evolutionary model that accounts for the conformational state of each residue based on the atomic positions of its side chain [134]. Specifically, we split the traditional 20 amino acid states into discrete sub-states based on the χ_1 rotamer (short for "rotational isomer") configuration of their side chains (Fig. 1.4)

as defined in the Dunbrack rotamer library [155]. In this classification, each residue can adopt one of (typically) three discrete configurations (Fig. 1.5) 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 ($\sim 60°$, $\sim -180°$and $\sim -60°$) consistently across all residues; this means that residues sharing the same rotamer configuration (e.g. PHE1 and TRP1 in Fig. 1.5) 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 energetic 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, permitting models that maintain the usual assumption of sitewise independence. This produces a scalable model that can be used in the same ways as a traditional 20-state substitution model.

While this model does not explicitly capture site-dependencies, it might do so, at least in part, implicitly. A change in side chain rotamer configuration when the amino acid does not change likely indicates a folding "adjustment", made necessary by another amino acid substitution at a different site. Conversely, a change of both amino acid and rotamer configuration might be either caused by a direct amino acid replacement at that site or by remote changes influencing this site's evolution. In the latter case the model will incorporate both into its general representation of the evolutionary process, without necessary being able to distinguish them.

By compiling a large set of homologous sequences for which X-ray crystallography structures are available, I develop a structurally-aware "Dayhoff-like" substitution model (see Chapter 2) 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 1.1). From analysing the RAM55 rate matrix I find that

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

Table 1.1: Rotamer configuration states. Left: 20 traditional states correspond to the 20 amino acids. Right: the 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 ($\sim 60°$, $\sim -180°$and $\sim -60°$) (see Fig. 1.4). Alanine (ALA) and glycine (GLY) have no side chain and therefore no χ_1 configuration, while proline (PRO) only has two stable χ_1 configurations ($\sim 27°$, $\sim -25°$) because of steric requirements of its pyrrolidine ring.

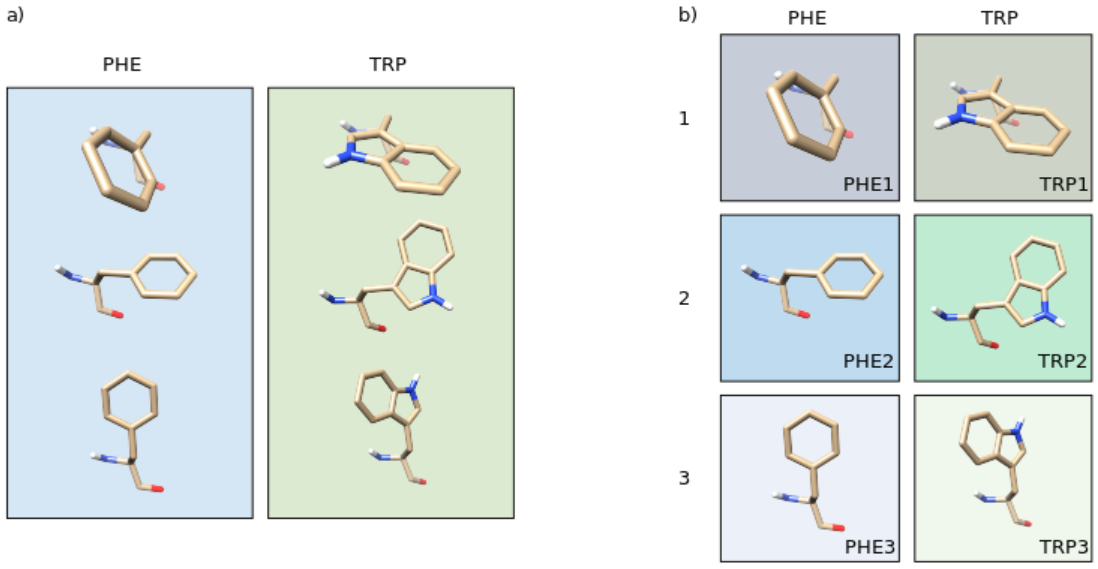

Figure 1.5: Illustration of the rotamer configurations of phenylalanine (PHE) and tryptophan (TRP). **(a)** In traditional amino acid replacement models their distinct χ_1 rotamer configurations are merged into a single amino acid state. **(b)** In our model these states are split into three χ_1 configuration-specific 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 Fig. 1.4).

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. I further show that, similarly to what happens for amino acids, rotamer state exchanges are influenced by the surrounding structural context, as evidenced when partitioning RAM55 according to solvent exposure, protein secondary structure and other structural features (see Chapter 3).

1.9 Applications of the rotamer-aware model

Models of protein evolution provide the foundations for a number of common bioinformatic tasks; among these some of the most prominent are sequence alignment, phylogenetic tree inference, ancestral sequence reconstruction, and protein structure homology modelling. The 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 [160, 94]. Similarly, it is possible to perform ancestral sequence reconstruction (ASR) under RAM55 using well-established reconstruction algorithms [186, 190]. These ASR techniques can be re-purposed to infer side chain configurations and assist homology modelling strategies.

After having computed RAM55 and examined its rotamer state exchange patterns (see previous section), I proceed to show that RAM55 results in a detectable improvement in model fit on simulated data, and on a number of diverse empirical datasets [134]. When inferring phylogenies, RAM55 produces reliable tree topology and sequence divergence estimates (see Chapter 4).

I then investigate the use of RAM55 for the purpose of ASR, a useful evolutionary biology technique with a wide variety of applications including assessment of how the physical properties of proteins shaped their evolutionary history [68, 180], vaccine development [6, 93] and resurrection of proteins from extinct organisms [58, 73]. The RAM55 model allows structurally-aware reconstruction of both ancestral rotamer and amino acid states. I show that RAM55 can accurately reconstruct ancestral rotamer states from descendant protein sequences of known structure [134]; it is also able to reconstruct ancestral amino acid states as well as or better than traditional 20-state models (see Chapter 5). 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 [40, 115, 39].

Prediction of side chain rotamer configurations is an important part of protein structure modelling and interaction modelling. For a given protein sequence of unknown structure, it is possible to construct a model of the target protein from its amino acid sequence and experimentally determined structures of related homologous proteins. This homology modelling strategy aims to predict both the main chain geometry and side chain configurations. In conserved regions, side chains can be modelled starting from configurations observed at corresponding sites in the nearest homologous structure. Further steps are then required, particularly to model non-conserved side chain configurations [178]. I show that side chain configurations can be predicted for an extant amino acid sequence using

RAM55 and a modified ancestral reconstruction algorithm (see Chapter 5) [179].

As decribed so far, an obvious limitation of RAM55 is its reliance on available protein structures for every protein sequence in a given dataset of interest, in order to compute the χ_1 dihedral angles and assign the rotamer states. While 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 that can be achieved for complex, dynamic molecular assemblies in their native state [120, 20, 174], many real-world applications of RAM55 might rely on data with a mixture of amino acid sequences and rotamer sequences, or even amino acid sequences alone.

It is possible to expand RAM55's applicability across these realistic "mixed input" scenarios by repurposing well-established methods that allow the handling of data for which only part of the state-space information is available [179]. This is achieved by encoding observed states as ambiguous representations of characters in a larger state-space (e.g. $V \rightarrow A, C, G$ for nucleotides). This application of standard statistical theory for missing data has been used previously in phylogenetics (e.g. [189], p. 110–112), however its use is not limited to uncertainty representation and it can allow rigorous statistical inference if implemented correctly. A notable example is covariotide modelling, conceptualised by Fitch and Markowitz [51], where each nucleotide may be in an evolutionary "on" or "off" state that cannot be directly observed [77]. Can these principles be applied more broadly to allow inference from "partial" observations? In Chapter 5 I illustrate that it is possible to infer information about evolutionary processes that occurred in an expanded state-space (e.g. rotamer states) using only partial data (e.g. amino acid states), taking advantage of established methods for handling ambiguity in sequence alignments. The ability to model sequences in a state-space with a larger set of characters allows RAM55 to reconstruct ancestral amino acid rotamer configurations using only amino acid sequences. Using input data consisting of a mixture of rotamer and amino acid sequences further improves these reconstructions.

1.10 Conclusion

In this chapter I have outlined some of the most relevant protein evolution modelling strategies that have been developed over the years starting from Dayhoff and Eck's original 1966 paper. From this review it is easy to notice how models benefit from explicitly accounting for at least some of the structural constraints on amino acid replacement. While more sophisticated models are able to closely replicate the observed interplay of sequence and structure evolution, their applicability to real world problems is severely limited by computational tractability issues. This observation has guided the main part of my PhD project: the development of a novel structure-aware model that accounts for residue side chain configuration and is compatible with widely-used phylogenetic inference algorithms and software. The RAM55 model can be applied to many bioinformatics tasks relying on evolutionary modelling, and can be expanded to accept "mixed" data without having to rely on available protein structures. Here I have introduced all RAM55 applications I explored so far, including phylogenetic tree inference, ancestral sequence reconstruction and homology modelling, which I will discuss in detail in the following Chapters. The detailed description of my work begins in Chapter 2 with a full explanation of the data collection and analysis generating the RAM55 model which forms the main focus of my research.

2. Data collection and analysis

This chapter has been adapted and expanded from a first-author publication [134]. I performed all the data collection and analysis; the original Perron et al. [134] manuscript was written by me, edited collectively, and agreed upon by all co-authors.

In this chapter I describe how I generated the structure-aware RAM55 model using a "Dayhoff-like" approach: the exchange rates were computed from counts of observed changes in alignments of closely-related protein sequences mapped to known high-quality X-ray structures [134]. Our analysis of the RAM55 rate matrix, which I further expand here, highlights that almost all amino acids evolve differently depending on the rotamer configuration of their side chains.

2.1 Assigning rotamer states

In order to tabulate substitution events, we required a data set of aligned amino acid sequences annotated with their χ_1 rotamer state. This was obtained from the Pfam database [46], one of the largest curated collections of protein family multiple sequence alignments, by first selecting those aligned sequences that are mapped to a high resolution crystal structure, in the Protein Data Bank in Europe (PDBe) [28] to ensure we only retrieve those structures that are likely to be reliable. For this step I relied on Pfam's own mapping of its aligned domains onto PDBe structures (see Code and Data Availability). Structure resolution, which is proportional to the disorder in the protein crystal, represents the average positional uncertainty for all atoms in a protein structure: here we selected proteins with resolution $<$2.5Å which is generally considered to be indicative of well a resolved structure [118].

For each amino acid in each sequence, I assigned a χ_1 rotamer configuration according to the value of the dihedral angle defined by the coordinates of the four atoms that compose χ_1 (N, C^α, C^β, and C^γ for most residues) in the corresponding PDBe structure, and the Dunbrack rotamer library [155] (see Code and Data Availability). The final result are aligned sequences of rotamer states, which are referred to as "rotasequences".

On a more specific level, the B-factor (also called temperature factor or temperature value) measures the positional uncertainty for each atom in a structure [171]. B-factors $< 30\text{Å}^2$ signify good confidence in the atom's position, while values $> 60\text{Å}^2$ indicate strong uncertainty [118]. To ensure that rotamer state assignments were based on unambiguous electron densities and not modelling artefacts, I removed residues with an average B-factor $> 30\text{Å}^2$ for the four atoms defining χ_1 (N, C^α, C^β and C^γ for most amino acids). Further factors such as thermal fluctuations, crystal packing forces and ligand binding might also confound RAM55 by creating differences between the structures of homologous proteins that are not due to evolution. The B-factor filtering also addresses these. Chapter 3 contains an in-depth analysis of rotamer state variability across homologous proteins.

I also removed non-standard residues (e.g. seleno-cysteine, seleno-methionine), disordered residues and those with peptide bonds exceeding 1.8Å, the last to ensure a continuous polypeptide. Cysteines involved in disulfide bonds are treated in the same manner as unbonded cysteines and I do not explicitly filter for multiple-occupancy atoms. Finally, I further cross-referenced these rotamer assignments against PDBe's own rotamer assignments for the corresponding residues, which are validated by the MolProbity pipeline [22], and discarded all mismatches (see Code and Data Availability). In this study we consider only χ_1 configurations, and not those of rotable bonds further along the side chains, 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 chains atom positions as all other side chains atoms depend upon it; it gives us a manageable number of states; and it always connects two sp^3 hybridised atoms, and thus is strictly rotameric and has exactly three conformational states [39] although one is inaccessible in proline.

These quality filtering steps resulted in high-quality rotasequence alignments from 5,659 Pfam families, including 79,558 unique PDBe structures (∼50% of all PDBe entries as of 22/1/2020) and 1,215,571 alignment sites covering 2,552 species across 12 kingdoms. Thus, a general rotamer state replacement model estimated from this collection of alignments is likely to provide a broadly applicable description of the evolutionary process, similarly to what is achieved (albeit on smaller alignment datasets) by general amino acid replacement models such as WAG and LG. Such a rotamer model could then be tailored to any specific dataset of interest through a manipulation of its equilibrium frequency and gamma distribution parameters as described in section 1.3.

2.2 Tabulating substitution counts

The amino acid sequences and corresponding χ_1 configuration annotation are combined into rotasequences which use an expanded alphabet (see Table 1.1): to emphasise the distinction between RAM55's rotamer states and traditional amino acid states, here and in the following chapters I employ a different notation than the one in eqn. 1.1. Each rotasequence consists of multiple symbol pairs (A, R), each of which specifies a rotamer state comprising the amino acid A (as employed by traditional 20-state models) and a χ_1 rotamer configuration R (1, 2 or 3). Proline only has two stable χ_1 configurations, and thus only two rotamer states (i.e. PRO1, PRO2) while for alanine and glycine, which lack a side chain, only the amino acid symbol (A) is used (i.e. ALA, GLY).

For each family I then performed a sequence alignment-guided pairwise comparison of rotasequences. I used Pfam's original domain alignment to construct a neighbour joining (NJ) phylogenetic tree [149] using MAFFT [81], and then iteratively tabulated differences between pairs of rotasequences by taking a circular tour through the NJ tree using an efficient algorithm analogous to the one described by Korostensky and Gonnet [89] (see Fig. 2.1) , which results in each change in the sequence being counted, at most, twice. While comparing pairs of rotasequences, I omitted those with a rotasequence identity $< 75\%$ (corresponding to ∼80% amino acid sequence identity, see section 3.3), to minimise the risk of multiple substitution events at the same site being incorrectly tabulated as a

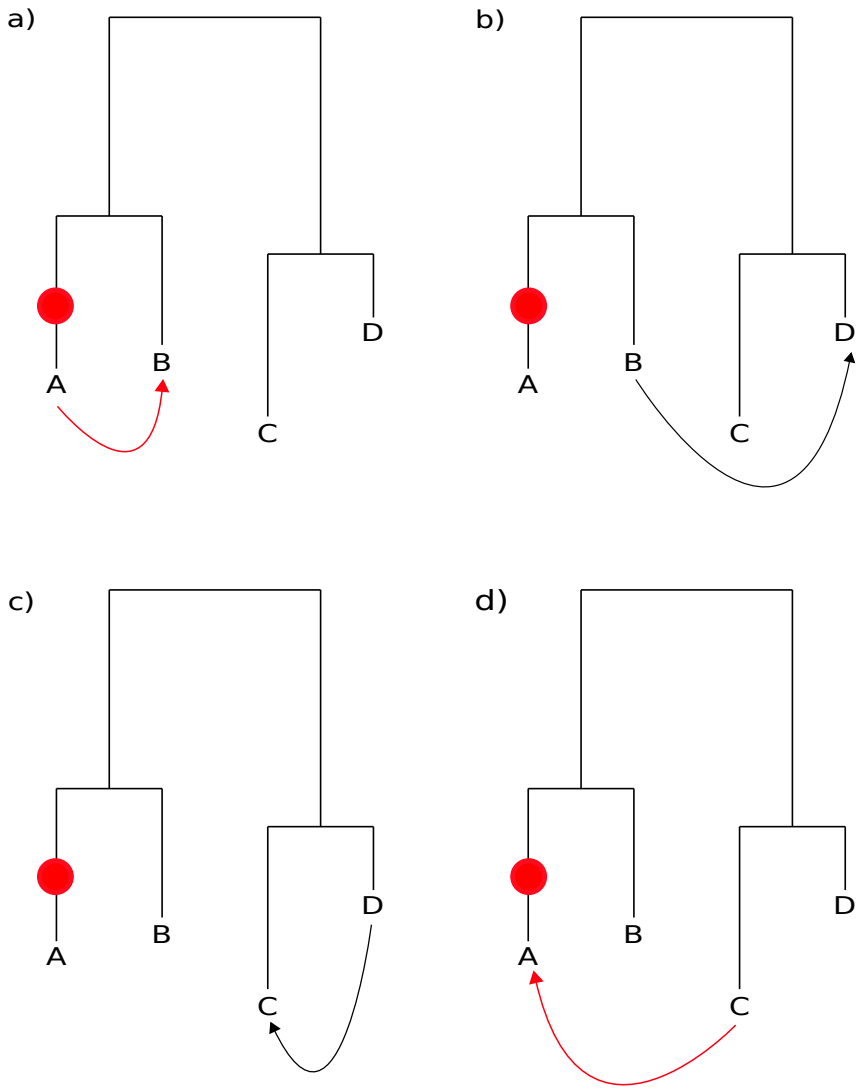

Figure 2.1: Efficient pairwise comparison of sequences according to tree distance and topology. The algorithm takes a circular tour of the terminal nodes of a tree by **(a)** first comparing the closest pair of sequences (A-B). Then **(b)** it compares B against its next-closest neighbour D, followed by **(c)** D against its closest neighbour C, and finally **(d)** C against A. Each sequence (e.g. A) is paired with a maximum of two others (e.g. A-B, A-C) and each change in the sequence (red dot) is thus counted, at most, twice (red arrows). This strategy is analogous to the one described by Korostensky and Gonnet [89].

single observed difference. I tabulated 35,079,033 counts, corresponding to 2,523,650 rotamer state substitutions and 32,555,383 instances of rotamer state conservation, of the former 1,991,505 (∼79%) involved a χ_1 configuration change while the amino acid was conserved, 442,020 (∼18%) an amino acid substitution while the χ_1 configuration was conserved, and 90,125 (∼3%) a change of both amino acid and side chain configuration.

As mentioned in the previous section, I only considered structures with resolution <2.5Å when assembling the RAM55 dataset. This threshold represents a compromise between dataset size and dataset quality, since determining side chain atom positions (and thus χ_1 configurations) becomes increasingly more difficult as resolution increases. A potential consequence of "noisier" structures with resolutions approaching 2.5Å might be an artificial increase in the number of observed χ_1 configuration changes (i.e. $(A, R) \rightarrow (A, R')$) due to the same amino acid being erroneously assigned different χ_1 configurations in related structures. Here, I investigated this issue by evaluating the rotamer states replacement rate at different structure resolutions in the RAM55 dataset. When considering all possible rotamer state changes (i.e. amino acid, χ_1 configuration, and their combinations), only a modest increase in rotamer state replacement rate can be observed as the average resolution of the two structures being compared increases (see Fig. 2.2, in blue). Further, if I exclusively consider changes involving only the χ_1 configuration (i.e. $(A, R) \rightarrow (A, R')$), the increase in resolution appears to affect the rotamer state replacement rate even less (see Fig. 2.2, in orange). This suggests that the previously described quality filtering steps are largely able to reduce the number of poorly resolved side chains which could lead to ambiguous side chain orientation assignments, and that structures with resolution as high as 2.5Å could still provide useful evolutionary information.

I 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 $n_{(A,R),(A',R')}$. While these counts could be used directly to calculate an instantaneous rate matrix (IRM), this would result in biases arising from the structure quality filtering procedures described in the previous section. For example, because alanine and glycine can never be filtered out by B-factor, all other residues are under-represented compared to them. Further, some amino acids, such as those commonly well-packed into the core of the protein, are

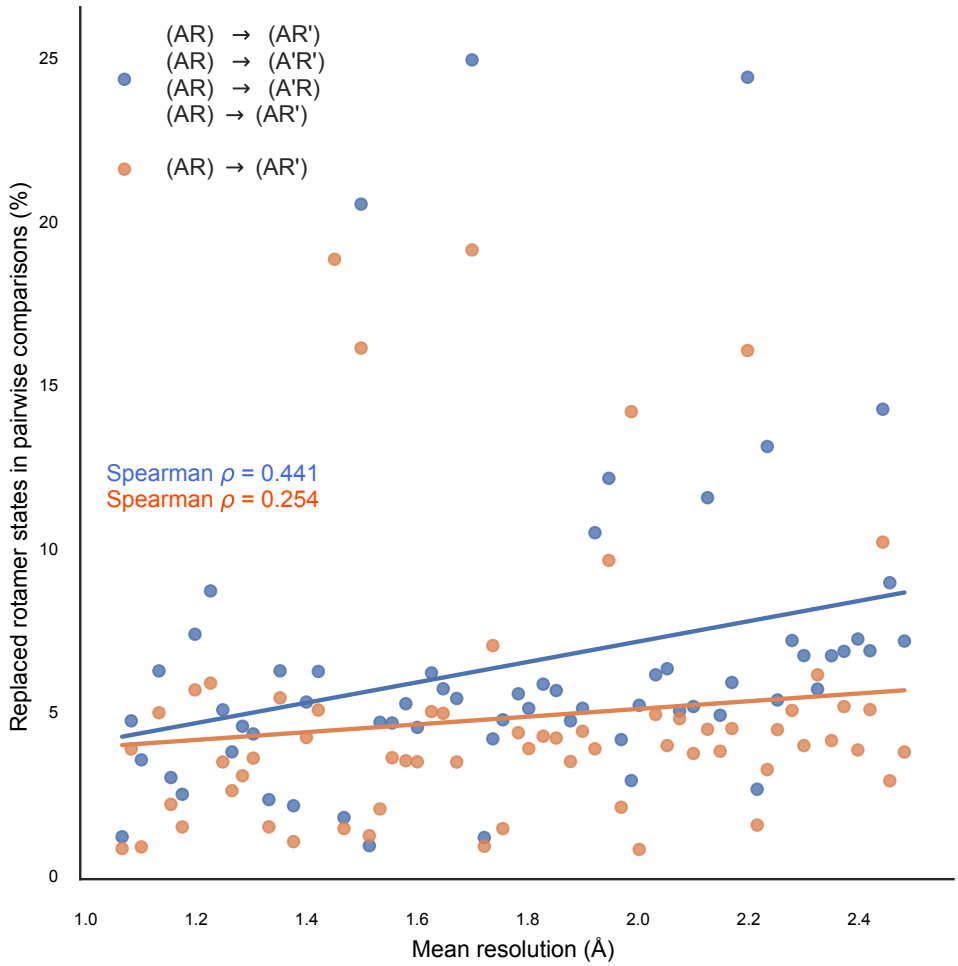

Figure 2.2: The relationship between structure resolution and χ_1 configuration conservation. 8000 structures are randomly sampled from the 79,558 included in the RAM55 dataset. For each pairwise comparison involving two of the sampled structures, the average structure resolution is computed as well as the percentage of replaced rotamer states and the percentage of rotamer states where only the χ_1 configuration is replaced (i.e. $(A, R) \rightarrow (A, R')$). The data is then divided into 64 bins according to the average structure resolution (shown on the x-axis) and the average percentage of rotamer substitutions is reported on the y-axis. While there is some increase in the overall percentage of rotamer state substitutions (in blue), the percentage of changes involving only χ_1 (in orange) appears to grow less rapidly than all replacements types taken together.

better resolved and have lower B-factors than those more commonly found at the protein surface, which are thus also under-represented in the RAM55 dataset (Fig. 2.3). Finally, some residues are intrinsically harder to crystallise: cysteine residues can cause disulfide-mediated aggregation, and flexible residues such as lysine, glutamate, and glutamine tend to have higher side chain atom B-factors as shown by their labelled B-factors distributions in Figure 2.3 which are noticeably shifted to the right.

To account for this I also compute substitution event counts ($n_{A,A'}$) for a Dayhoff-like

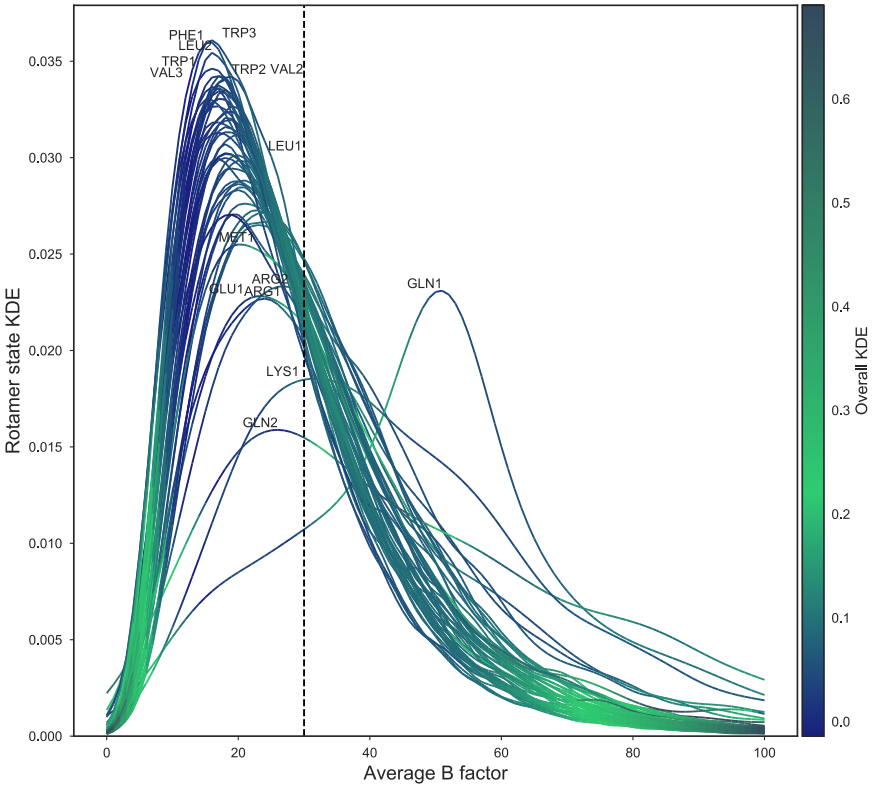

Figure 2.3: Structural data quality distributions across residues. Kernel density estimate (KDE) for the average B-factor for each rotamer state (excluding alanine and glycine) across all residues in the unfiltered dataset. Average B-factor is computed over the four atoms (N, C^{α}, C^{β} and C^{γ} for most residues) that constitute the dihedral angle defining the χ_1 rotamer configuration. A threshold of B-factor < 30 is then applied (dashed line) to ensure only highly reliable atomic coordinates are used to assign rotamer states. Only outlier density plots are labelled, for clarity; the color scheme represents overall plot density distribution at each point along the x-axis. Areas under the curves to the right of the dashed line indicates the proportions of each rotamer state in the alignments that were removed by the B-factor filtering.

20-state empirical model – which is named RUM20 for "Rotamer Unaware Model" – using the same Pfam-based dataset, but ignoring χ_1 configurations and without performing B-factor filtering. The normalised rotamer state exchange count matrix $\widehat{N}$, recovering the actual observed residue frequencies while retaining the relative ratio of substitution

computed from the filtered data, then becomes:

$$\widehat{n}_{(A,R),(A',R')} = \frac{n_{A,A'}}{\sum\limits_{\substack{r\in R_A,\\ r'\in R_{A'}}} n_{(A,r),(A',r')}} . n_{(A,R),(A',R')} \tag{2.1}$$

where for a given amino acid A, $R_A = \{R : (A,R) \in S_{55}\}$ 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 1.1.

2.3 Computing and scaling the exchangeabilities

The 55-state instantaneous rate matrix (IRM) $\widehat{Q}$ is then computed from these normalized counts as described by Kosiol and Goldman [92]: the 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):

$$\widehat{q}_{(A,R),(A',R')} = \frac{\widehat{n}_{(A,R),(A',R')}}{\sum\limits_{(a,r)} \widehat{n}_{(A,R),(a,r)}} \tag{2.2}$$

RUM20's IRM can be similarly obtained from the unfiltered 20-state counts.

These rates are first scaled according to ρ so that, at equilibrium, they will result on average in one rotamer state substitution per unit of time:

$$\rho = -\sum_{(A,R)} \pi_{(A,R)} \widehat{q}_{(A,R),(A,R)} \tag{2.3}$$

where $\pi_{(A,R)_i}$ is the equilibrium frequency of rotamer state (A,R) obtained from the raw counts. The scaled rate matrix then is:

$$\widehat{Q}^{\star} = \frac{1}{\rho}\widehat{Q} \tag{2.4}$$

where $\widehat{Q}$ is the unscaled IRM composed of the elements $\widehat{q}_{(A,R),(A',R')}$.

The scaling step described above ensures that the rate of rotamer state substitution in RAM55, at equilibrium, is on average one per unit of time. However, this means that

the rate of amino acid substitution in RAM55 will be < 1, because some rotamer state substitutions only consist of a χ_1 configuration change (i.e. $(A, R) \longrightarrow (A, R')$). This is potentially confusing, and I preferred to use a timescale where t measures the number of amino acid changes on average at equilibrium, as it is common practice for protein substitution models. In order to do this, I need the overall replacement rate in RAM55's IRM increased according to an additional scaling factor ρ^* which is defined as:

$$\rho^* = -\sum_{(A,R)} \pi_{(A,R)} \sum_{\substack{(A',R'),\\ A' \neq A}} q^*_{(A,R),(A',R')} \tag{2.5}$$

where ρ^* represents the fraction of rotamer state substitutions where the amino acid (A) is replaced by a different one (A').

Then the "superscaled" rates are:

$$\widehat{Q}^{\star\star} = \frac{1}{\rho^\star}\widehat{Q}^\star = \frac{1}{\rho\rho^\star}\widehat{Q} \tag{2.6}$$

this matrix, at equilibrium, will result on average in one amino-acid state substitution per unit of time and allows a direct comparison of branch lengths estimates between RAM55 and any 20-state model.

From $\widehat{Q}^{\star\star}$, exchangeabilities can finally be obtained as:

$$s_{(A,R),(A',R')} = \frac{\widehat{q}^{\star\star}_{(A,R),(A',R')}}{\pi_{(A',R')}} \tag{2.7}$$

The 55-state model now allows for likelihood-based computations on phylogenetic trees, permitting inference of tree topology, branch lengths and likelihoods that can be used for model fitting and comparisons (see sections 4.2, 4.3).

2.4 Rotamer state exchangeability analysis

With the exchange rate matrix computed from observed changes in homologous sites of proteins of known structure, it is now possible to investigate how rotamer states exchange over evolutionarily-relevant time-spans. For the purpose of this analysis, the latter refer to

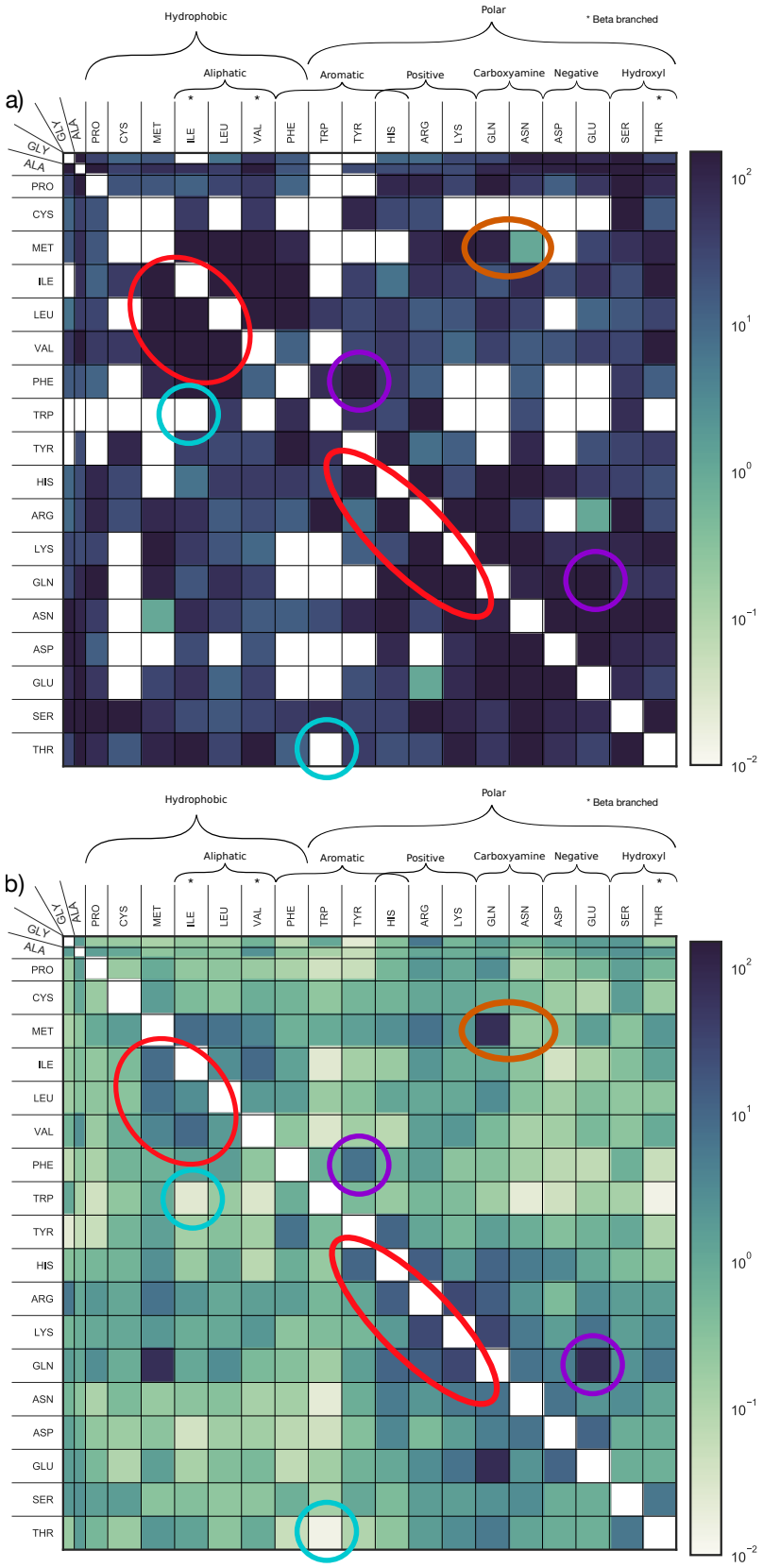

Figure 2.4: Similarities between two empirical amino acid replacement models. Exchangeabilities ($s_{A,A'}$) are reported in heat-map form for **(a)** the 1978 Dayhoff model [34] and **(b)** a 20-state model (RUM20) estimated from the same dataset as RAM55. Note that time-reversibility of the models means the exchangeabilities are symmetric (e.g. $s_{A,A'} = s_{A',A}$ for $A \neq A'$. Amino acid are grouped by biochemical properties and the most notable replacement patterns present in both matrices are highlighted (see text for details).

timescales at which evolutionary models are commonly applied for phylogenetic inference tasks: up to $t = 2.5$ which corresponds to $\sim 20\%$ amino acid sequence identity.

First, however, it is prudent to asses how the changes I observe in the RAM55 dataset compare to a well-established empirical model. Figure 2.4 shows amino acid exchangeabilities in heat-map form for the 1978 Dayhoff model (Fig. 2.4a), and for RUM20 which I estimated from the same dataset as RAM55 for comparison purposes (Fig. 2.4b) [134]. Observing the RUM20 exchange rates, it is easy to notice strong similarities with the Dayhoff model: these include 1) rapid exchanges among the aliphatic residues block and methionine and 2) a characteristic set of amino acids exchanging with their immediate biochemical neighbour (both highlighted in red); 3) methionine exchanging frequently with glutamine and more rarely with asparagine (in orange); 4) high exchangeabilities for phenylalanine - tyrosine and glutamic acid - glutamine pairs (in violet); and 5) low exchangeabilities for tryptophan - isoleucine and tryptophan - threonine (in light blue). Most of these exchange patterns are also shared by general amino acid replacement matrices developed after Dayhoff's (e.g. those in section 1.2).

Here I further examined the amino acid composition of the RAM55 dataset by comparing RUM20's equilibrium frequencies – already factored out of the exchangeability matrix (see eq. 2.7) – against those of LG [105]: a modern empirical model, and the Dayhoff model. Figure 2.5 highlights how these frequencies are quite similar for most amino acids, despite the fact that these models have been estimated from different protein alignment datasets, with RUM20's and LG's being much larger than Dayhoff's. One noticeable difference is that for glutamine both LG and Dayhoff show a much higher frequency than RUM20. This is likely due to a combination of 1) glutamine having a highly flexible side chain which makes it more difficult to crystallise and resolve, and 2) glutamine being rarer than other flexible residues such as lysine. Because of this, my normalisation strategy (see section 2.2) appears to be less effective in accounting for structure quality filtering for this residue. Finally, it is worth keeping in mind that the subset of proteins which can be mapped to known structures is not necessarily representative of all proteins, both because some proteins might be harder to crystallise and solve, and because of higher interest in certain cellular functions and pathways.

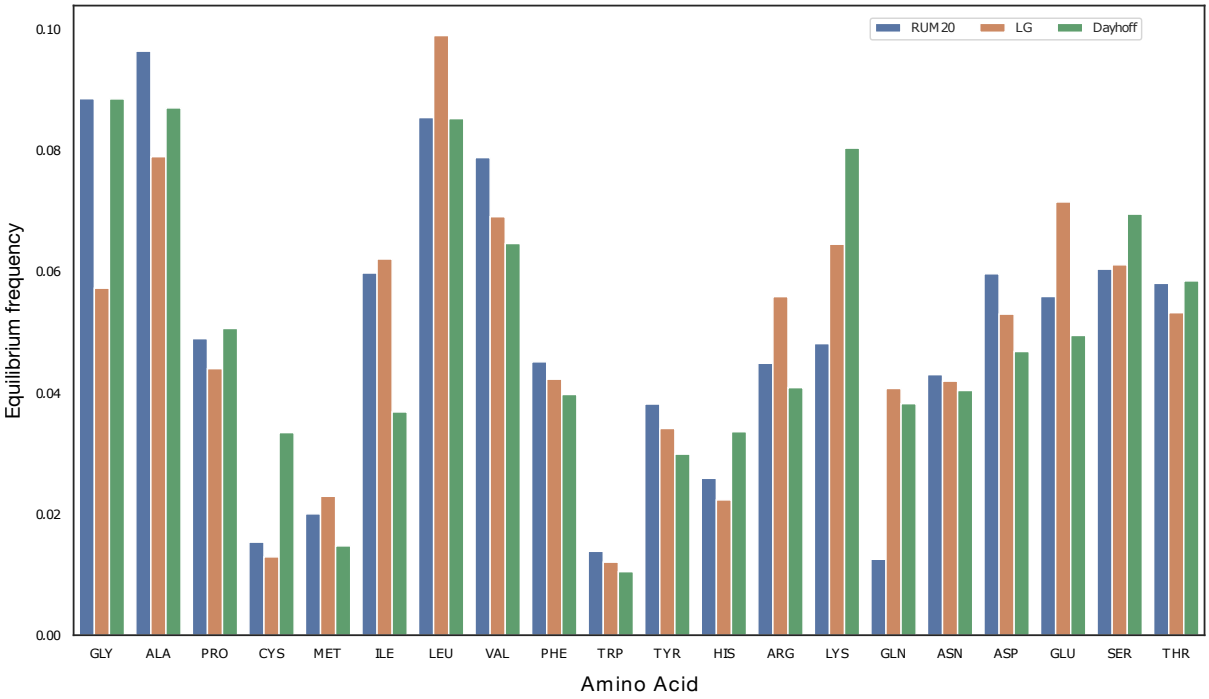

Figure 2.5: Amino acid equilibrium frequencies from three empirical replacement models. The y-axis shows the equilibrium frequency for each amino acid across three empirical models: RUM20, LG [105], and the Dayhoff model [34]. These models are estimated using different protein alignment datasets; much larger ones in the case of LG and RUM20. Despite this, the equilibrium frequencies are quite similar for most amino acids across these three models.

All of these results suggests that the RAM55 dataset can provides an adequate representation of the amino acid replacement process. What happens to the exchange rates when we split 20 amino acid states into 55 rotamer states?

The RAM55 exchangeability matrix represented in Figure 2.6a recapitulates the same "amino acid level" exchangeability trends described above for RUM20 (Fig. 2.6b). It also shows 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. $(A, R) \longleftrightarrow (A', R)$ with $A' \neq A$). This is visible in Figure 2.6a where each pair of different amino acids corresponds to a 3×3 sub-matrix (with the exception of pairs including alanine, glycine and proline). Higher exchange rates are observed on the diagonal of many of the 3×3 sub-matrices (corresponding to changes in amino acid only) compared to the off-diagonal elements (changes in amino acid and rotamer configuration). This is particularly true of interchanges between biochemically similar amino acids: sub-matrices corresponding to

aromatic-aromatic exchanges for example all have very distinct diagonal patterns, as do the exchanges between aspartic acid and its derivative asparagine (highlighted in orange), and between serine and threonine (highlighted in red) which both have an hydroxyl group in their side chains.

Since $\widehat{N}$ is symmetric, there are 136 unique 3×3 sub-matrices. For each of these I computed the Pearson's χ^2 statistic and p-value (after 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 are showing significant association among their rotamer states; only these are considered for further analysis (e.g. Fig. 2.7). Overall, 111 of the 136 independent 3×3 sub-matrices show significant association among the interchanging states. To further quantify the strength of these sub-matrix exchange patterns I use Cramér's V (also written as $\widetilde{V}$) with bias correction [14], 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 (rotamer configuration conservation); I therefore also consider the diagonal ratio for each sub-matrix, indicating the proportion of rates that lie on each 3×3 sub-matrix's diagonal. $\widetilde{V}$ and diagonal ratio are shown in Fig. 2.7 for the 111 sub-matrices with significant associations.

All six aromatic-aromatic sub-matrices have high $\widetilde{V}$ values and high diagonal ratios (Fig. 2.7a, upper right), indicating a strong preference for conserving side chains orientation. This exchange pattern might be capturing the effect of local constraints on how freely a bulky aromatic side chains can be positioned without displacing or clashing with those of neighbouring residues. A similarly strict configuration conservation can be observed for negative-negative and positive-positive exchanges; however, negative-positive exchanges have high $\widetilde{V}$ but somewhat lower diagonal ratios (Fig. 2.7b). These sub-matrices 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

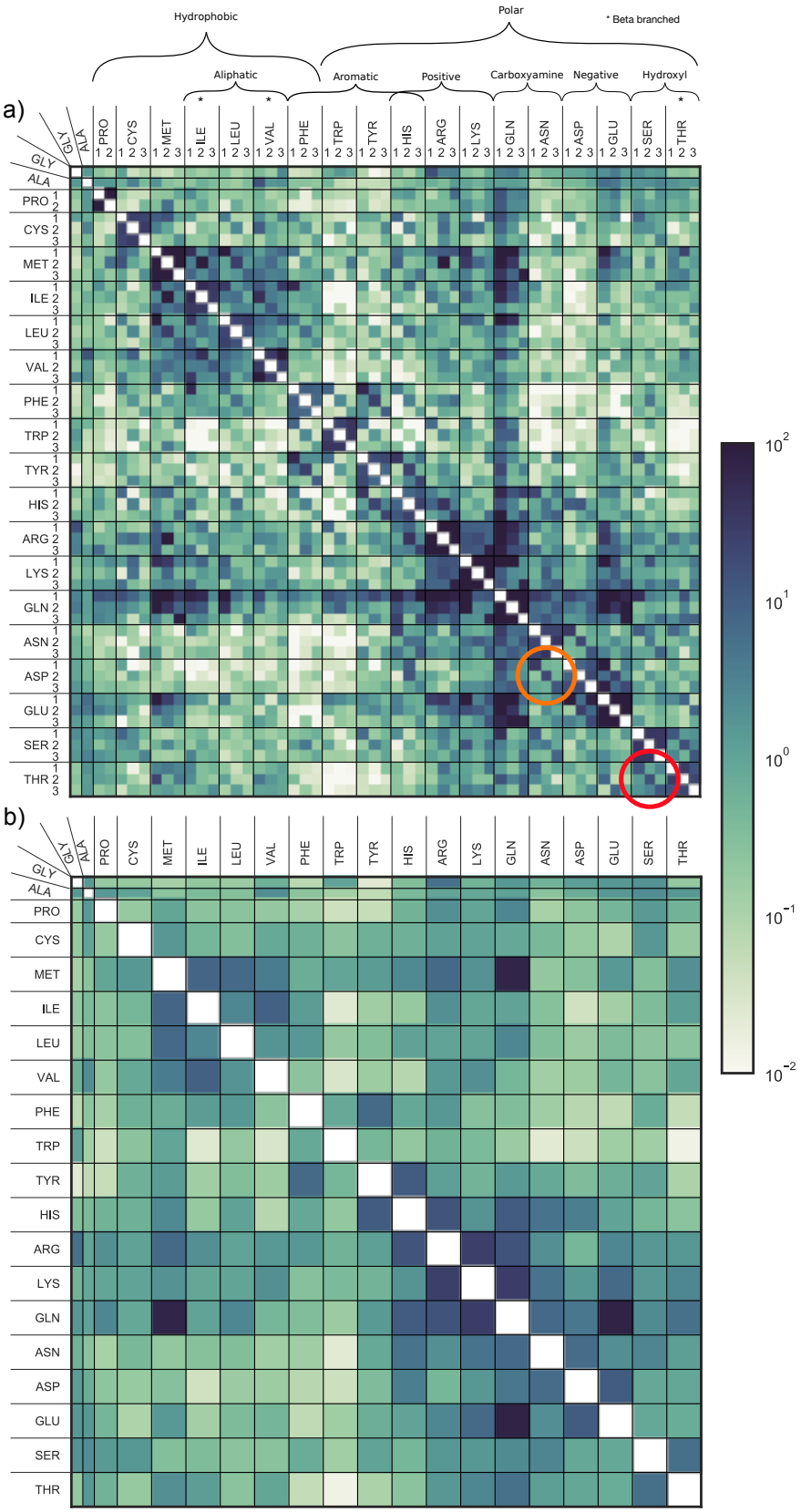

Figure 2.6: Replacement models, with and without rotamer configuration information. Exchangeabilities ($s_{(A,R),(A',R')}$ and $s_{A,A'}$) are reported in heat-map form for **(a)** the 55-state model (RAM55) and **(b)** the 20-state model (RUM20) estimated from the same dataset. Note that time-reversibility of the models means the exchangeabilities are symmetric (e.g. $s_{(A,R),(A',R')} = s_{(A',R'),(A,R)}$ for $(A,R) \neq (A',R')$).

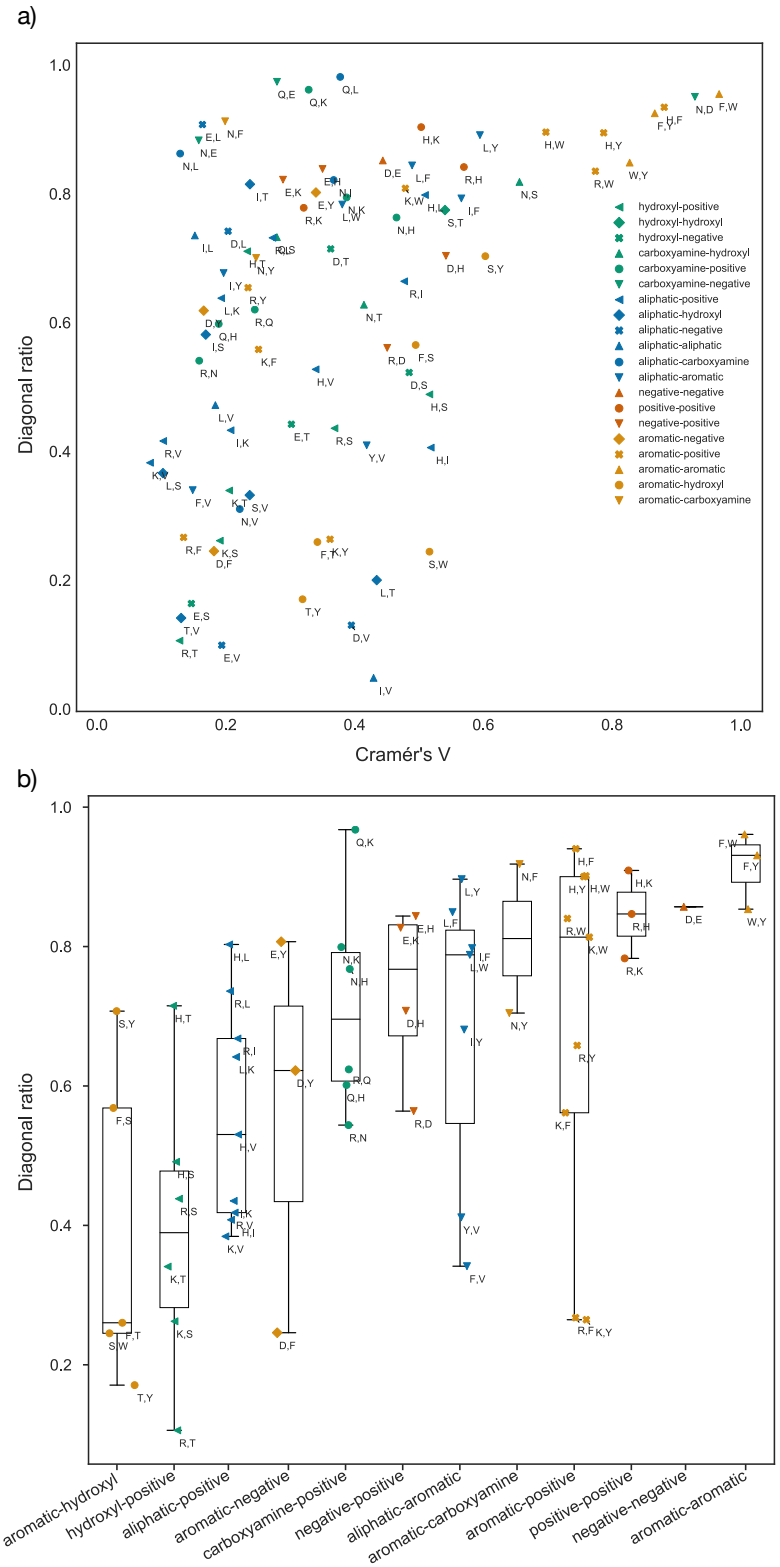

Figure 2.7: Strength of association and diagonal ratio. Plots show pairs of residues whose 3×3 sub-matrices within the RAM55 Q-matrix achieve significant χ^2 statistic values indicating association of substitutions involving these amino acids. Pairs are labelled according to their component residues' biochemical properties. **(a)** Strength of association (Cramér's V, $\widetilde{V}$) 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.

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. 2.7b). In contrast, isoleucine and valine, both aliphatic and β-branched, have lower scores and show less tendency to conserve their side chains orientation when exchanging to aromatic residues.

All these results suggest that, by splitting amino acids into rotamer states, RAM55's exchange rates capture additional information which cannot be extracted from traditional 20-state models.

2.5 Rotamer state evolution

As described above, rotamer state replacement patterns are influenced by the biochemical similarity of their amino acids. There is often a strong tendency to conserve χ_1 configuration when exchanging between similar amino acids, particularly so for those with bulky aromatic side chains or charged side chains. Here I further investigate differences in the evolutionary process for individual rotamer states sharing the same amino acid. Figure 2.8 illustrates in four examples how rotamer states can evolve (i.e. be replaced and thus no longer be conserved) at different rates as sequences diverge, depending on their amino acid and χ_1 configuration. For valine (Fig. 2.8a) VAL1 is fastest, followed by VAL3 and VAL2; for lysine (Fig. 2.8b) LYS1 is much faster than LYS2 and LYS3; for threonine Fig. (2.8c) THR1 and THR3 are replaced almost at the same rate while THR2 is replaced much more quickly; and for phenylalanine there is little difference among its three rotamer states. For valine and threonine it is necessary to keep in mind that, because of the former's symmetric side chain, the C^γ atoms assignment is somewhat arbitrary and thus χ_1 configurations 1, 2, and 3 of valine can also be equivalent to 2, 3, 1 of threonine. Ambiguity between VAL1 (Fig. 2.8a, tan) and THR2 (Fig. 2.8c, green) likely accounts, along with threonine's hydroxyl group, for such a distinct behaviour in two C^β-branched residues.

Different replacement rates compound with different equilibrium frequencies to generate unique rotamer state exchange patterns (eq. 2.7). For most amino acids, χ_1 configuration 3 ($-60°$) is the most abundant in the RAM55 alignment dataset (see Fig. 2.9a, b), this trend is observed in most rotamer libraries (e.g. [38, 115, 155]), and is particularly strong for residues with longer side chainss at positions that fall outside canonical α helices (see Chapter 3). There are, however, a few notable exceptions including the C^B-branched residues (Fig. 2.9c, d), with the same "C^γ ambiguity" caveat mentioned previously likely meaning that valine's frequencies pattern might be actually closer to threonine's.

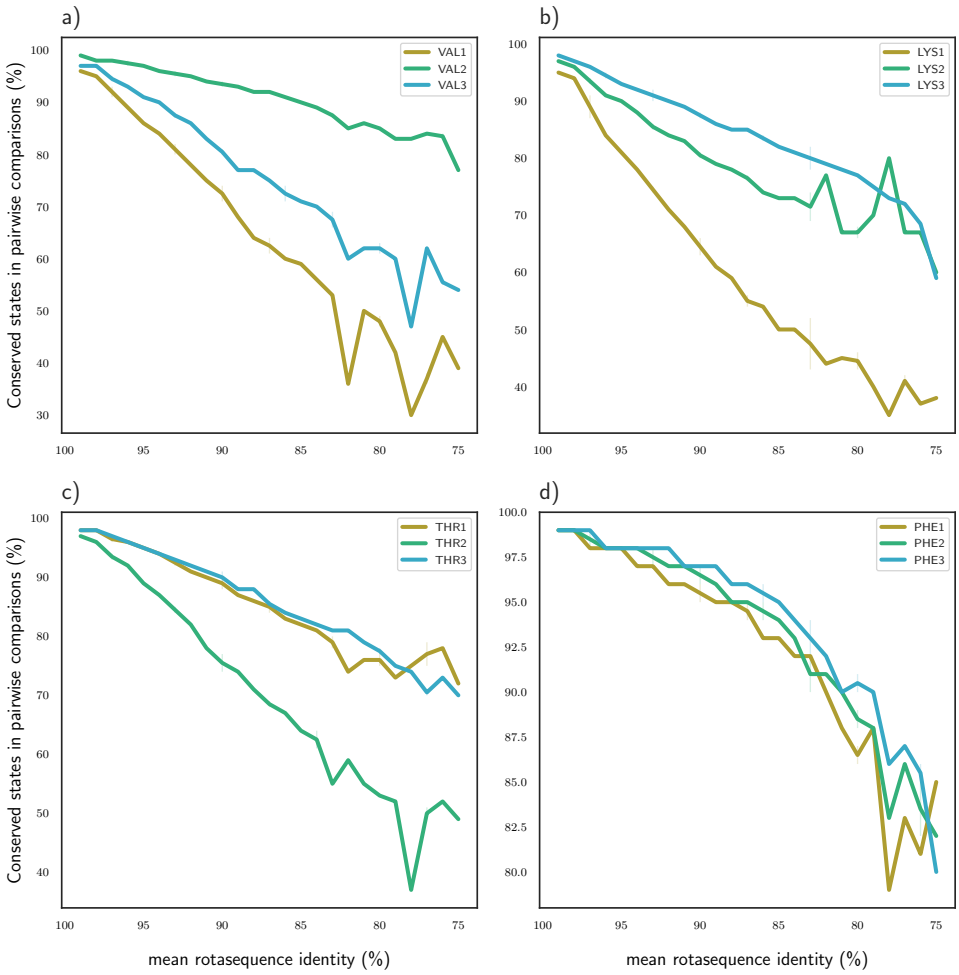

Figure 2.8: Different rotamer states of the same amino acid evolve at different rates. All pairwise site comparisons containing valine, lysine, threonine, or phenylalanine are binned by the overall sequence identity of the two rotasequences that are being compared. The y-axes report the percentage of conserved sites containing a given rotamer state in each bin. The x-axes show the mean rotasequence identity for each of 32 bins. Which of the three rotamer states (AR) is more conserved, and how different their replacement rates are, varies according to the amino acid (A).

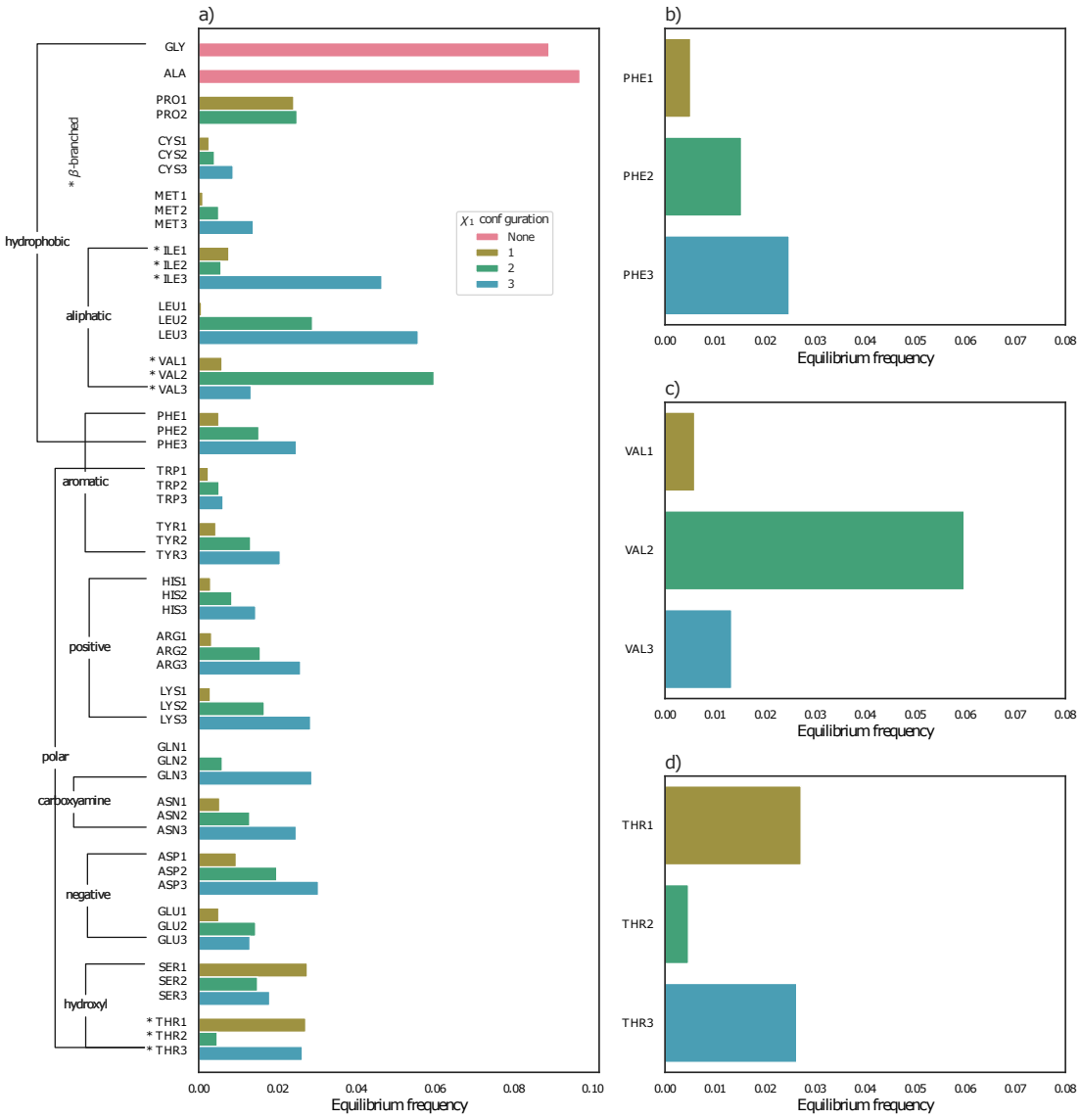

Figure 2.9: Empirical rotamer state equilibrium frequencies from RAM55. **(a)** shows the equilibrium frequency for each of 55 states grouped by biochemical properties. The smaller panels contain close-ups for **(b)** phenylalanine, **(c)** valine, and **(d)** threonine. Rotamer states in configuration 3 are generally more abundant, with some notable exceptions, as highlighted in **(c)** and **(d)**.

It is also interesting to notice the interplay between the overall replacement rate of a given rotamer state and its equilibrium frequency. Fast-evolving residues are less conserved and, depending on how divergent the input alignments are, these states can end up being rarer across the entire dataset. Valine and threonine provide, again, a good illustration of this effect: the replacement trends in Figure 2.8a, c match quite well with the equilibrium frequencies in Figure 2.9c, d with VAL1 and THR2 being replaced more quickly and also having the lowest equilibrium frequencies. For other rotamer states, such as phenylalanine's, this relationship between replacement and frequency is less obvious: PHE3 is the most common of the three (Fig. 2.9b), and yet PHE1, PHE2, and PHE3 are all replaced at very similar rates (Fig. 2.9d), despite PHE1 being $\sim$5 times rarer than PHE3.

What is the impact of this rotamer state "uniqueness" on the substitution process, both in rotamer state and amino acid state space? In Figure 2.10 I report observed rotamer state changes for tryptophan and histidine across the RAM55 dataset with flow thickness representing the number of observed changes between each pair of states. If a tryptophan is in configuration 3 (the most abundant), it will most likely change to PHE3, VAL3, or TYR3 and is very unlikely to change to PHE1 (Fig. 2.10a). On the contrary, if a tryptophan is in configuration 1 (the least abundant), it is very likely to change to PHE1, and less so to PHE3. This illustrates the effect of the χ_1 conservation trends observed in Figures 2.6 and 2.7. These distinct exchange propensities can have an impact on substitutions in amino acid space as well. From HIS1 (Fig. 2.10b), the most frequent replacements lead to tyrosine (TYR2) and threonine (THR3), and from these two to valine (VAL2) and isoleucine (ILE3). From HIS3, however, a change to asparagine (ASN3) is by far the most likely, followed by tyrosine (TYR3), and from there to alanine, glycine, valine (VAL2) and isoleucine (ILE3). These χ_1-dependent replacement processes are invisible to a traditional 20-state amino acid model, which cannot distinguish between, for example, HIS1 and HIS3. In Chapter 4 I further compare RAM55 against standard 20-state models for phylogenetic inference tasks.

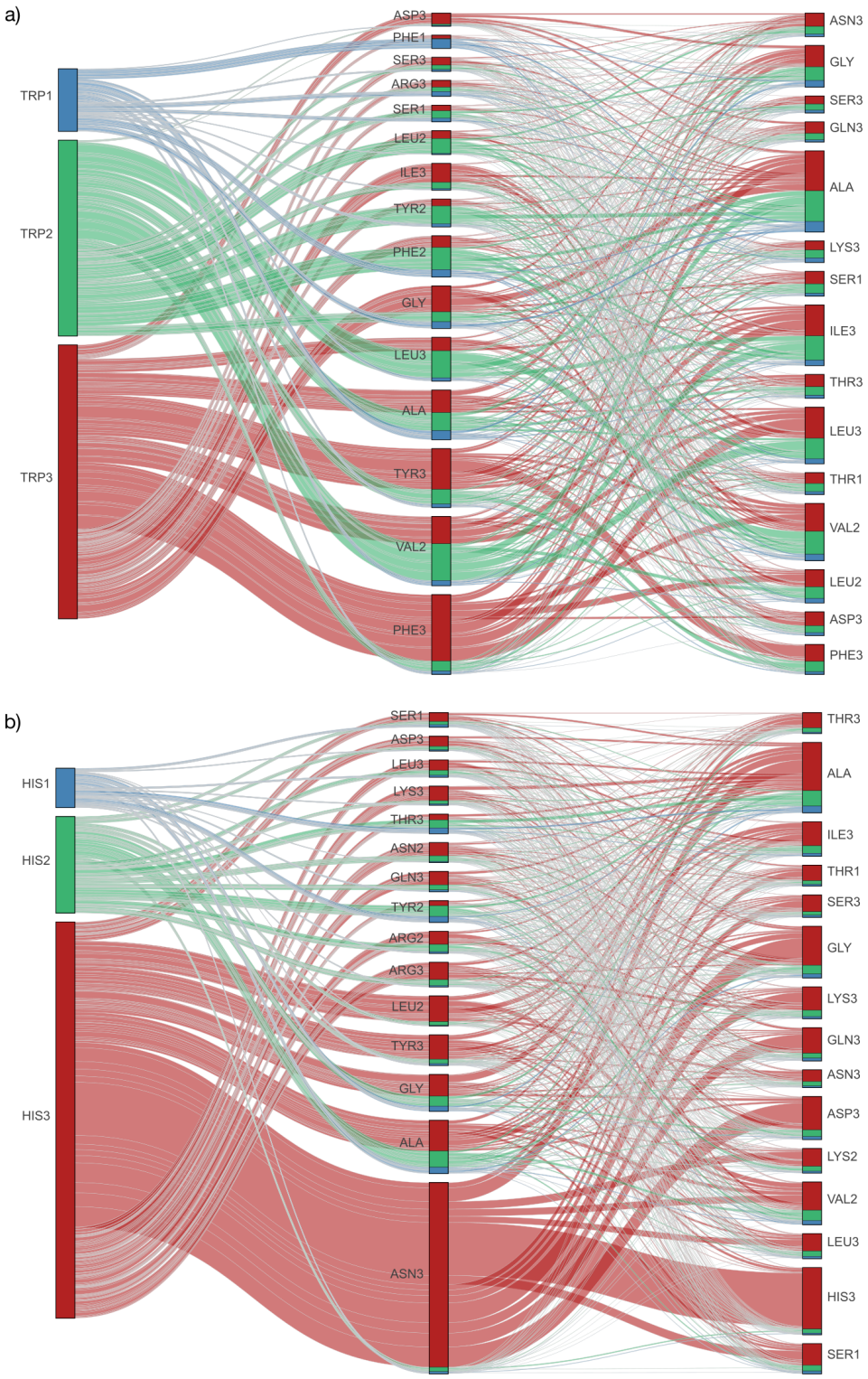

Figure 2.10: Rotamer state replacement patterns impact amino acid substitutions. This paired category plot illustrates the rotamer state substitution process for cases where the amino acid is not conserved (i.e. $AR \rightarrow A'R$ or $AR \rightarrow A'R'$). The left column represents the proportion of pairwise sites comparisons containing each of the three rotamer states for **(a)** tryptophan or **(b)** histidine in the RAM55 dataset. The central column shows the 15 rotamer states that most frequently replace those in the left column, ordered by frequency. The right column contains the 15 states that most frequently replace those in the central column. Line thickness represents the number of observed changes between two states across the RAM55 dataset; line colour indicates the original χ_1 rotamer state (left column). Individual rotamer states show distinct replacement patterns and this influences which substitutions are more likely in amino acid space.

2.6 Rotamer state molecular similarity

As highlighted in section 2.4, one of the likely causes for distinct rotamer state exchange patterns is the necessity to avoid steric clashes with the backbone or with neighbouring residues' side chains. In this context, the backbone configuration, defined by its ϕ and ψ torsional angles, plays an important role by determining the positioning in space of neighbouring residues. Can we investigate the influence that backbone geometry may have on the observed exchangeabilities?

To do so I calculated, for each pair of rotamer states, the overlap between the bivariate joint distributions of their ϕ and ψ backbone dihedral angles obtained from the Dunbrack rotamer library [155, 134]. These overlap values correlate with the corresponding RAM55 exchangeabilities, with a Spearman's ρ of 0.33 (p-value = 1.9×10^{-28}, see Fig. 2.11a), 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 [138]. Indeed 76% (115 out of 153) of the 3×3 and 2×3 sub-matrices corresponding to changes in amino acid have a positive Spearman's ρ between overlap and exchangeability (Fig. 2.11b). Interestingly, some amino acid pairs with a strong diagonal pattern in the 3×3 exchageability sub-matrix show a strong correlation between exchangeabilities and overlap values, as is the case for serine and aspartic acid ($\rho = 0.95$, see Fig. 2.11a,c in red). In other cases, like aspartic acid and leucine, non-diagonal exchangeability patterns also strongly correlate with the overlap values ($\rho = 0.90$, see Fig. 2.11a,c in magenta). This indicates that, for most amino acid pairs, there is a tendency for evolutionary exchanges to be between side chains geometries that accommodate similar backbone geometries and that, for some pairs, this might require a change in χ_1 configuration.

2.7 Conclusion

In this Chapter I have described how, by splitting amino acid states according to their χ_1 configuration, we can uncover distinct exchangeability patterns among rotamer states sharing the same amino acid. I have also shown how individual rotamer states can evolve at different rates and have unique equilibrium frequencies. Many of these patterns might

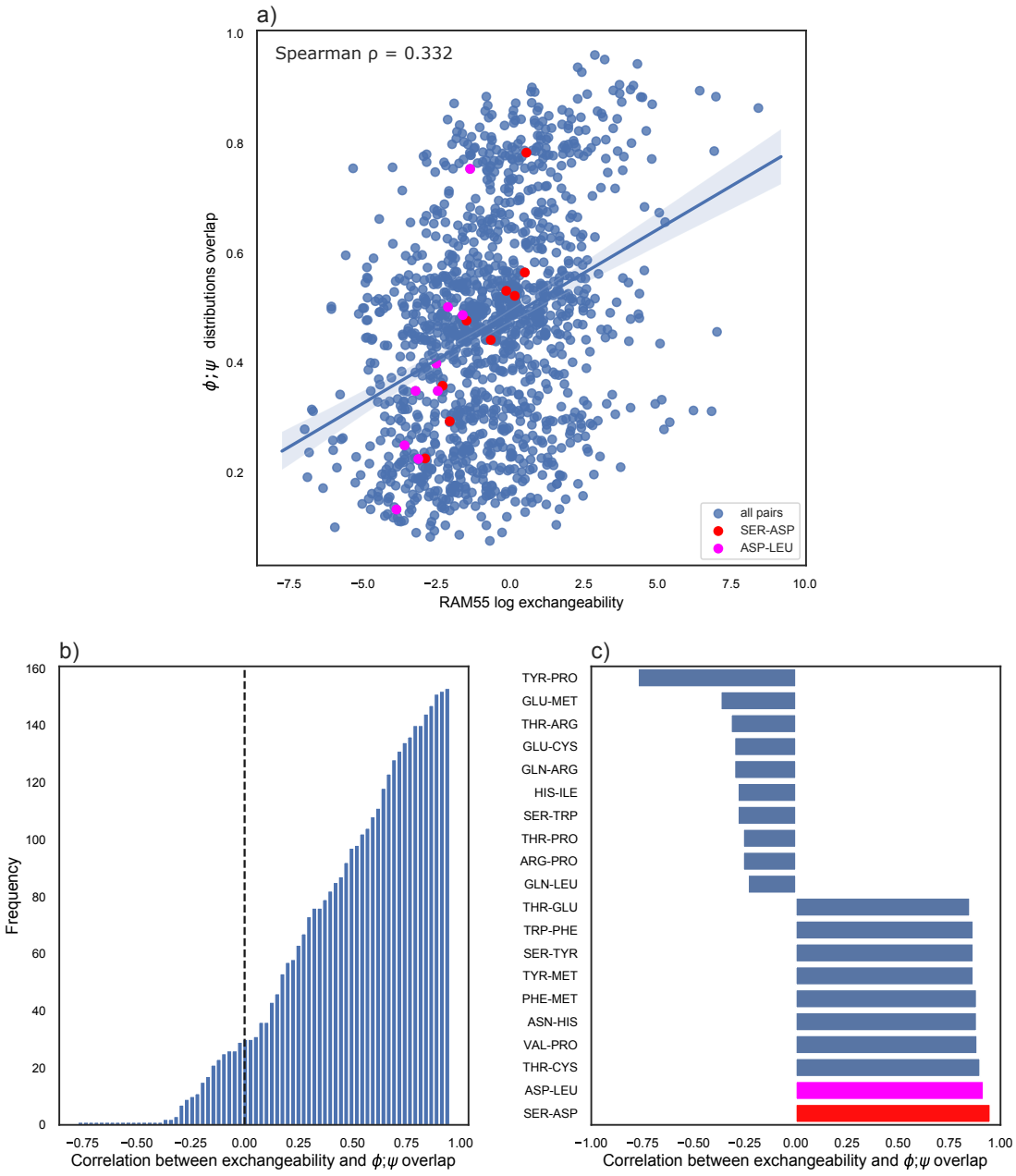

Figure 2.11: Correlation between exchangeability and the overlap between (ϕ, ψ) distributions. For each amino acid pair (excluding alanine and glycine), the Spearman rank-order correlation between the exchangeabilities of their χ_1 configurations in RAM55 and the overlap between their Ramachandran probability distributions [138] in the Dunbrack rotamer library [155] is computed. **(a)** illustrates the correlation across all pairs; **(b)** shows the cumulative histogram of the correlation values for each pair; **(c)** shows the top and bottom 10 pairs according to their Spearman ρ. The two pairs with the highest correlation values are highlighted (see text).

be capturing the effect of local steric constraints on how freely a bulky (or charged) side chains can be positioned without displacing or clashing with those of neighbouring

residues. However, there are still several strong exchange patterns 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 2.6a and 2.7a. Nevertheless, the strength of these associations suggests that RAM55's expanded state set incorporates valuable, biochemically plausible, structural information into the model. RAM55 (Fig. 2.6a) can thus be considered a "high-resolution" version of RUM20 (Fig. 2.6b) generated from the same dataset. In Chapters 4 and 5 I will show how this provides additional inferential power from the ability to distinguish states and state-interchanges according to χ_1 configuration.

3. Rotamer state variability

Parts of the results in this Chapter are based on the RAM55 dataset, previously described in a first-author publication [134]. All results shown here are unpublished, and all data collection and analysis was performed by me.

3.1 Rotamer variability across structures of the same protein

The same amino acid sequence can correspond to multiple PDBe entries which might not be identical to one another. These structures, as well as conserved domains of homologous proteins, are sometimes related through conformational changes. Thermal fluctuations, crystal packing forces and ligand binding are examples of factors that might create such conformational differences and consequently introduce noise in the model's exchange rates. Can we estimate what proportion of rotamer state substitutions we observe is due to conformational changes and thermodynamic instability rather than protein evolution?

In a study of 63 pairs of structures of the same protein, one in the *apo* state and the other in the *holo* state, Zavodszky and Kuhn [192] found that 95% of side-chains did not change rotamer. 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. [125] 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 [193]. These residues are characterised by diffuse electron densities and high thermal B-factors, which for RAM55 [134] were removed with a B-factor filter (see section 2.1). Neither Najmanovich et al. [125] nor Clark et al. [25] found a correlation between side-chain conformational change and backbone conformational change.

These studies strongly suggest that χ_1 rotamer configurations, as defined in RAM55, remain quite stable through conformational changes and thermodynamic fluctuations. I further investigated χ_1 rotamer configuration variability using a larger dataset comprising 284 human thrombin heavy chain (H) structures from PDBe. All of these structures correspond to the same 258-amino acid sequence (Uniprot P00734) which I converted to an alignment of rotasequences following the approach described in Chapter 2.

3.2 Human thrombin structures dataset

In this section I discuss a selection of the preparation, crystallisation, acquisition, and modelling metadata for the aforementioned 284 human thrombin structures: Figure 3.1 summarises the diversity of this dataset of structures of the same protein via a selection of structure metadata. These parameters can have a significant influence on the quality of each structure model and it is important to put them in context before evaluating rotamer state variability across this dataset.

The first prerequisite of structure determination is to produce and purify a sufficient quantity of the protein of interest. Protein production is typically achieved by expressing a recombinant coding gene which is optimised for a host organism [64]. Recombinant human prothrombin can be produced in *E. coli* as a single, non-glycosylated polypeptide chain, since bacteria lack the required post-translational machinery (Fig. 3.1a) [146]. This issue can be particularly problematic for larger, multi-domain eukaryotic proteins expressed in bacteria, which often are non-functional and can be difficult to purify. Expression systems using mammalian cells and organisms (e.g. small rodents) have been developed for applications requiring accurate folding, however, these systems have lower yields than bacteria, as well as other limitations (e.g. time-consuming, toxicity to host cells) [64]. In

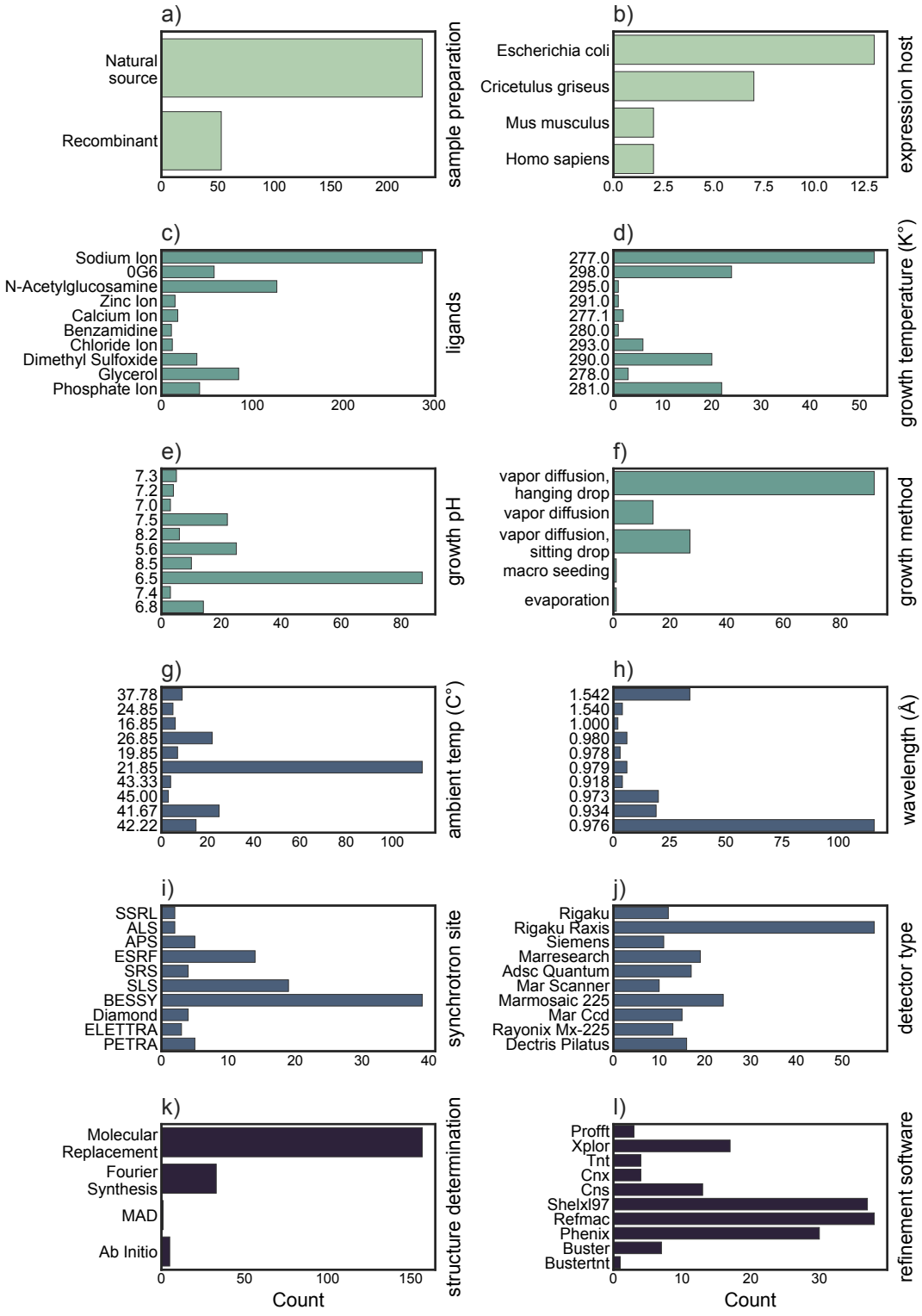

Figure 3.1: Metadata associated with 284 human thrombin structures from PDBe. Each panel shows the count distribution for one "meta-feature". **(a,b)** preparation, **(c-f)** crystallisation, **(g-j)** acquisition, and **(k,l)** modelling metadata. Structures missing any meta-feature annotation are removed from the corresponding analysis.

the case of thrombin, it is also possible to purify the protein directly from human plasma (Fig. 3.1a,b).

The following step in structure determination requires obtaining crystals from the purified protein of interest, with a volume suitable for X-ray analysis (i.e. $0.1-0.4$mm in each dimension) [118]. The extent and resolution of the X-ray diffraction pattern from a crystal depends on its degree of internal order: the more structurally uniform are the molecules in the crystal and the more precise is their periodic arrangement, the larger, more uniform, and more defined is the scattering pattern [118]. In the process of crystallisation, proteins are dissolved in an aqueous solution that is supersaturated in the macromolecule (i.e. the protein is present in excess of its solubility limit) but exhibits conditions that do not significantly perturb the protein's native state [118]. Supersaturation is achieved via addition of precipitating agents (e.g. neutral salts, polymers: see Fig. 3.1c), and by the manipulation of various parameters which influence the number and volume of obtained crystals, including crystal growth temperature (Fig. 3.1d) and pH (Fig. 3.1e). Other factors that can affect the structural state of the macromolecule crystal include metal ions (e.g. Na^{+}, Zn^{+2}, Cl^{-}), inhibitors and activators (e.g. N-Acetylglucosamine, a prothrombotic [50]), cofactors or other small molecules (Fig. 3.1c). Once supersaturation is reached, nucleation and growth of crystals favours the re-establishment of equilibrium [139].

A variety of crystal growth strategies have been developed among which the most common is vapour diffusion (Fig. 3.1f) [139]. In this method, droplets containing purified protein, buffer, and precipitant are allowed to equilibrate with a larger reservoir containing similar buffers and precipitants in higher concentrations. Vapor diffusion can be performed in either hanging-drop or sitting-drop format. Hanging-drop involves placing a drop of protein solution on an inverted cover slip suspended above the reservoir. Sitting-drop places the drop on a pedestal separated from the reservoir. Once grown sufficiently, crystals are extracted from the protein solution and are typically imaged at specific ambient temperatures to preserve protein folding and any protein-ligand interaction of interest (Fig. 3.1g).

Even under the same crystallisation conditions, a marked variation in crystal properties exists [139]. For each individual protein, crystallisation is a matter of finding and optimising a set of parameters that yield the best crystals possible. This parameter set can vary substantially, even when crystallising the same protein, depending on the specifics of the instrumentation and the necessity to co-crystallise various ligands (see Fig. 3.1)).

Once a crystal is obtained, it is placed in an intense beam of X-rays, usually of a single wavelength around 1Å which is on the scale of covalent chemical bonds and the atomic radius (Fig. 3.1h). X-ray beams are generated in synchrotrons, which accelerate charged particles and confine them in a circular loop using magnetic fields. Most synchrotrons are national facilities, such as BESSY II in Germany, the Swiss Light Source (SLS), and the Canadian Light Source (CLS) (Fig. 3.1i). Each has several dedicated beamlines where samples are rotated within the X-ray beam, producing a regular pattern of reflections at each position. The intensities of these reflections can be recorded with traditional image plate detectors (e.g. Rigaku's R-AXIS detectors), or with a highly accurate charge-coupled device (CCD) cameras (e.g. Mars, Rayonix detectors, see Fig. 3.1j) [139].

The recorded dataset of two-dimensional diffraction patterns is converted into a three-dimensional electron density map. To do so both amplitude and phase must be determined [163]. While the former is measured directly during a diffraction experiment, the latter can to be estimated through a variety of approaches including: 1) *ab initio* phasing which exploits known phase relationships between certain groups of reflections, 2) molecular replacement where a known related structure is used to determine the orientation and position of the molecules within the unit cell, and 3) multi-wavelength anomalous diffraction (MAD) phasing [70] in which methionines are replaced with seleno-methionines and reflections are rerecorded at three different wavelengths thus altering the selenium atom scattering in a known way and yielding the position of any methionine residues within the protein (Fig. 3.1k) [139].

For the purposes of the analysis described in the following sections, it is important to notice that the majority of structures in this dataset (157 of 284) has been solved by molecular replacement using a total of 41 distinct starting models. Specifically, 60 of these

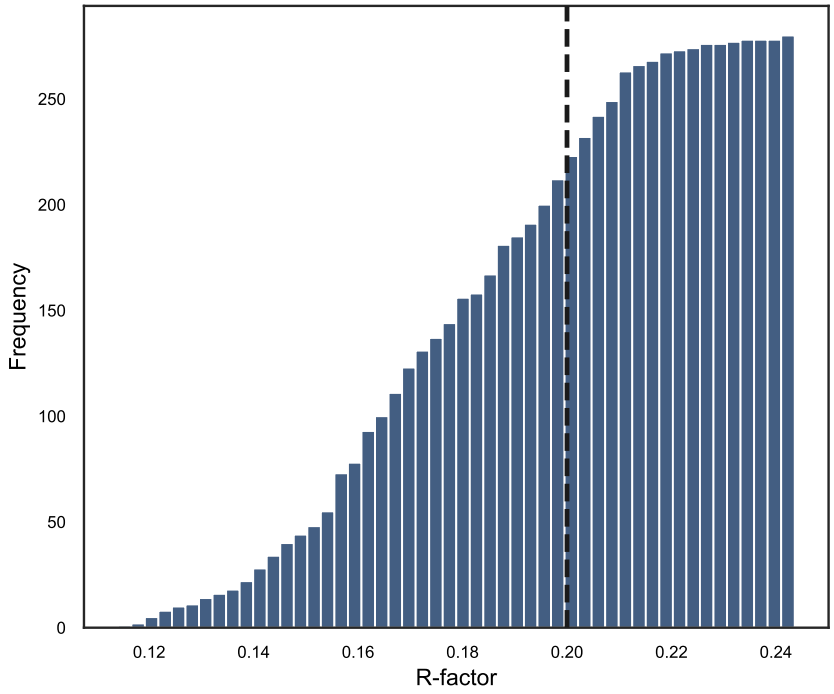

Figure 3.2: R-factor cumulative distribution across 284 human thrombin structures. The R-factor measures how well the refined structure predicts the observed diffraction data. For protein structures with resolution ∼2.5Å, R-factors ≤0.2 (dashed line) are considered indicative of a very accurate model [139]; around 75% of R-factors in this dataset are below this threshold.

structures have been obtained using 1H8D as a model. Although molecular replacement structures are then refined independently of their starting models, this process might introduce some undesirable dependencies in side chain orientation across structures.

From these initial phases, a preliminary electron density can be obtained by Fourier synthesis [163]. The atomic positions in the model and their B-factors can be then refined to fit the observed diffraction data, and successive rounds of refinement are carried out [139]. These steps are performed using computational methods that compare the experimental data to the simulated diffractogram of a model structure, and refine the structural parameters of the model, usually via least squares based minimisation algorithms. Widely-used refinement software includes Shelxl [156], Refmac [124], and Phenix [109] (see Fig. 3.1l). This iterative process continues until the refined model's fit for the observed diffraction data, measured by the R-factor which is closely related to the intensity of the reflection it describes, is maximised [139].

Figure 3.2 shows the R-factor distribution for all 284 thrombin structures in this

dataset: despite the large variety in preparation, crystallisation, acquisition, and modelling conditions, most structure models (∼75%) have R-factors below the 0.2 threshold which is considered indicative of good model fit for the observed data [139]. In the following section I will investigate χ_1 configuration variability across corresponding residues for all 284 thrombin structures, and how it relates to structure quality and similarity.

3.3 Rotamer variability across human thrombins

The 284 structures in the thrombin dataset all share the same 258 amino acid sequence and are all models of the same human protein. Since no relevant span of time separates these proteins (i.e. not enough for protein evolution to occur), any observed change in χ_1 configuration across corresponding residues cam be attributed to the confounding factors mentioned in section 3.1 and represents a source of background noise of a type which might affect the RAM55 replacement rates.

Figure 3.3a shows the rotamer state site-wise identity distribution across the thrombin alignment: at most sites there is strong agreement between the χ_1 configurations of corresponding residues in different thrombin structures. On average 96.76% of structures share the same consensus (i.e. most frequent) rotamer state at each site, a result consistent with previous work by Zavodszky and Kuhn [192] (∼95% identity) and Najmanovich et al. [125] (∼94% identity) on smaller datasets. These studies also highlighted the influence of residue positioning within the structure on side chain configuration variability. Can we observe a similar trend in the thrombin dataset?

I obtained the relative accessible surface area ($RASA$) for all residues as

$$RASA = \frac{ASA}{MaxASA} \tag{3.1}$$

where ASA is a residue's solvent accessible surface area (measured in Å^2) computed with DSSP [170], and $MaxASA$ is the amino acid specific maximum ASA value [168]. Residues with $RASA < 25\%$ are generally classified as buried [108]. Figure 3.3b compares the cumulative identity distributions of exposed and buried residues: the latter tend to have

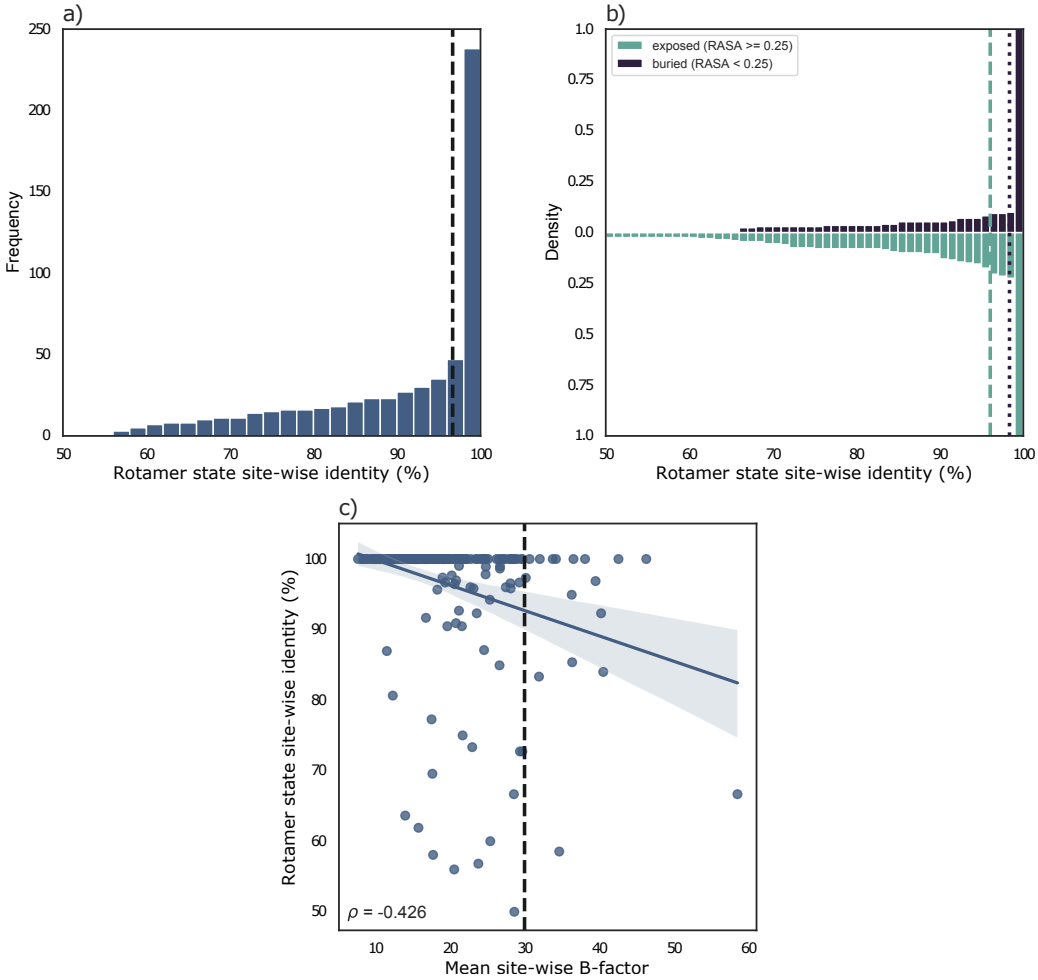

Figure 3.3: Rotamer state identity across the thrombin dataset. For each of 258 sites in the thrombin rotasequence alignment (284 sequences) the percentage of residues that match the consensus rotamer state (i.e. the most frequent) is computed (site-wise identity). **(a)** shows the cumulative site-wise identity distribution across all sites; the mean identity is 96.76% (dashed line). In **b)** the relative accessible surface area ($RASA$) is computed using DSSP [170]: mean site-wise identity is 96.03% for exposed residues (dashed line) and 98.34% for buried residues (dotted line). **(c)** Shows the relation (Spearman's $\rho = -0.41$, p-value $6.8{\times}10^{-10}$) between site-wise identity and mean B-factor for the four atoms that define χ_1 across all 284 residues at a given site (see section 2.1). These results are consistent with previous studies on smaller datasets [192, 125, 193].

slightly higher rotamer state identity (mean 98.34%) than the former (mean 96.03%). This agrees with previous observations [193] suggesting that residues on the protein surface are likely to have "noisier" χ_1 configurations across structures of the same protein, likely because exposed side chains are less constrained by local steric interactions. In line with what was observed by Zhao, Goodsell, and Olson [193], there is also a significant, albeit not strong (Spearman's $\rho = -0.42$, p-value $= 6.8{\times}10^{-10}$), correlation between χ_1 configuration identity and mean B-factor for the four atoms defining χ_1 (see section 2.1).

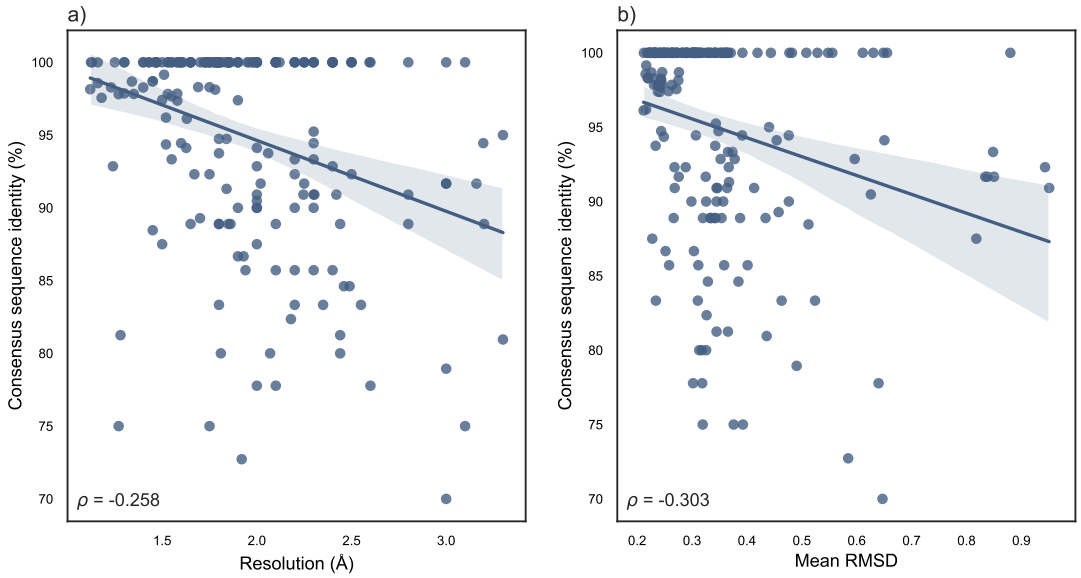

Figure 3.4: The relationship between rotamer state identity, resolution, and root mean square deviation (RMSD). Here rotamer state identity is measured, for each structure, as the percentage sequence identity with the consensus rotasequence. **(a)** shows the correlation between identity and overall structure resolution for all thrombin structures. For each structure I compute the RMSD against all other 257 structures in the dataset using UCSF Chimera [135]. **b)** shows the correlation between identity and the mean RMSD for each structure. High-resolution thrombin structures and those that superimpose well with most others (i.e. have low mean RMSD) tend to show more consistent χ_1 configurations.

The dashed line in Figure 3.3c represents a B-factor $< 30\text{Å}^2$ threshold which removes some of the most variable sites (bottom right), thus mitigating the amount of spurious variability from poorly resolved side chains.

On a whole-structure level, χ_1 configurations from higher-quality models of the same protein, as measured by resolution in Å, tend to agree more frequently with the consensus configuration (Spearman's ρ=-0.26, p-value= 1.92×10^{-4}, see Fig. 3.4a). Both structure resolution's and site B-factors' correlations with rotamer state identity are likely due to the fact that, in poorly-resolved parts of a structure, side-chain conformations are often determined through additional rounds of energy minimisation and refinement which might produce conflicting rotamer state assignments. Finally, for each structure in the thrombin dataset, I computed the root mean square deviation of atomic positions (RMSD), a quantitative measure of similarity between protein structures, against all the remaining 283 structures using UCSF Chimera [135]. A structure's mean RMSD shares a similar relationship with rotamer state identity as its resolution (Fig. 3.4b): this means that 1) most

thrombin structures are very similar to most others (i.e. have a low mean RMSD) and 2) that those structures that superimpose well tend to also share most of the consensus rotasequence states (Spearman's $\rho - 0.3$, p-value 1.08×10^{-4}). This further suggests that, at least for closely-related structures, rotasequence similarity is a good proxy for structure similarity, and that rotasequences that diverge from the consensus sequence correspond to relatively atypical structures.

Collectively, these results and those of previous studies indicate that the amount of "background" (i.e. non-evolutionary) χ_1 configuration changes across structures of the same protein is small ($\sim$3%) and likely more due to imprecise structure determination than to actual side chain conformation change. In fact, most side-chain movement appears 1) to consist of within-rotamer dihedral angle fluctuations [192], 2) to be concentrated in the rotamers further away from the backbone [193], and 3) to involve more frequently solvent-exposed residues (see e.g. Fig. 3.3b and [193]). Thus, after structure and residue quality filtering (see section 2.1, Fig. 3.3b, 3.4a), the effects of thermal motion and variations in crystallisation conditions are negligible with regard to the rotamer state substitution rates described in RAM55.

3.4 Rotamer variability across structural contexts

I have introduced structurally "partitioned" amino acid models in section 1.4. These approaches (e.g., [127, 175, 90, 166, 60, 102, 140]) recognise that the surrounding structural environment influences amino acid replacement rates and amino acid composition across sites and thus explicitly employ a separate 20-state model for each environment category. In particular, solvent exposure (also referred to as solvent accessibility) and secondary structure have been consistently shown to be some of the most important features affecting evolutionary change [127, 60]. For rotamer states, in Perron et al. [134] we have highlighted that, for most amino acid pairs, there is a tendency to exchange between side-chain geometries that accommodate similar backbone geometries (see section 2.6 in this thesis). How do solvent exposure and secondary structure influence rotamer state equilibrium frequencies and replacement patterns?

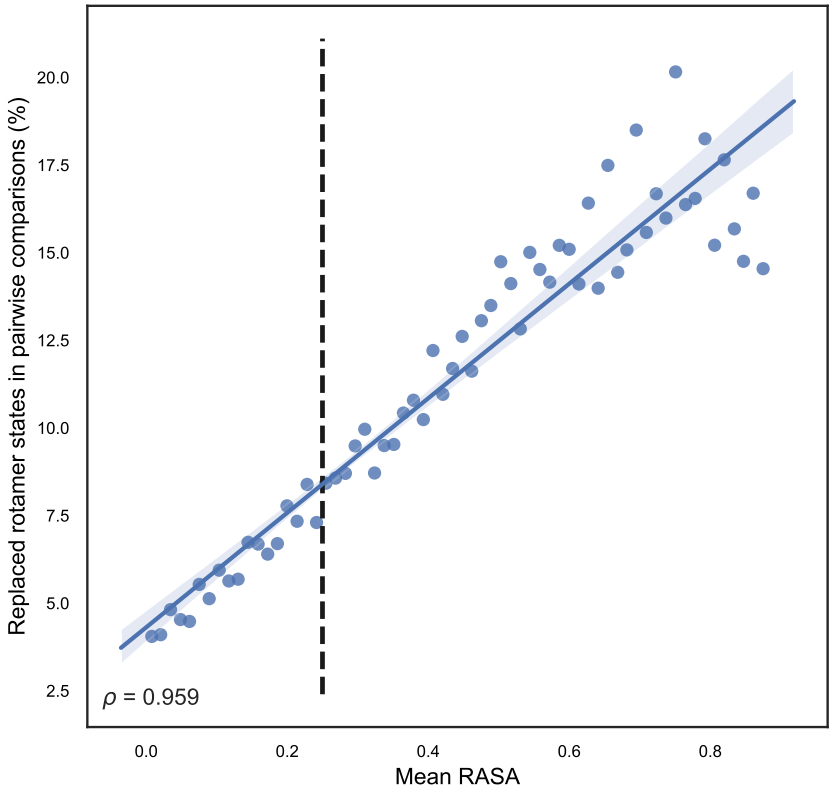

Figure 3.5: Rotamer states at exposed sites are less conserved. Relative accessible surface area (RASA) [168] and rotamer state conservation are computed across 5659 Pfam [46] family alignments and 1.22×10^6 sites using DSSP [170]. RASA is divided in 64 bins and the mean RASA for each bin is shown on the x-axis. The y-axis shows the percentage of sites where a rotamer state change (i.e. $(AR) \rightarrow (A'R)$ or (AR') or $(A'R')$) is detected in pairwise rotasequence comparisons for each RASA bin. The vertical dashed line represents the usual $< 0.25\%$ RASA threshold separating buried and exposed residues [108]. There is a strong tendency for more accessible rotamer states to be less conserved (ρ=0.059, p-value$<<10^{-5}$).

To investigate this, I systematically annotated all rotamer sequences in RAM55 (corresponding to 79,558 unique PDBe entries) using DSSP [170]. The aforementioned 25% RASA threshold is used to partition buried and exposed residues alongside the three more frequent secondary structure classes in the RAM55 dataset, as defined by DSSP: α-helix, β-sheet, and hydrogen-bonded turn. For each of these partitions I then tabulated the raw change counts and from these I computed the exchange rates, equilibrium frequencies, and exchangeabilities in the same way as described in sections 2.2 and 2.3.

In terms of overall replacement rate, similarly to what happens for amino acids [60], there is a very strong tendency (Spearman's $\rho = 0.96$) for rotamer states at exposed sites to change much faster than those at buried sites (Fig. 3.5). This is likely explained, as for

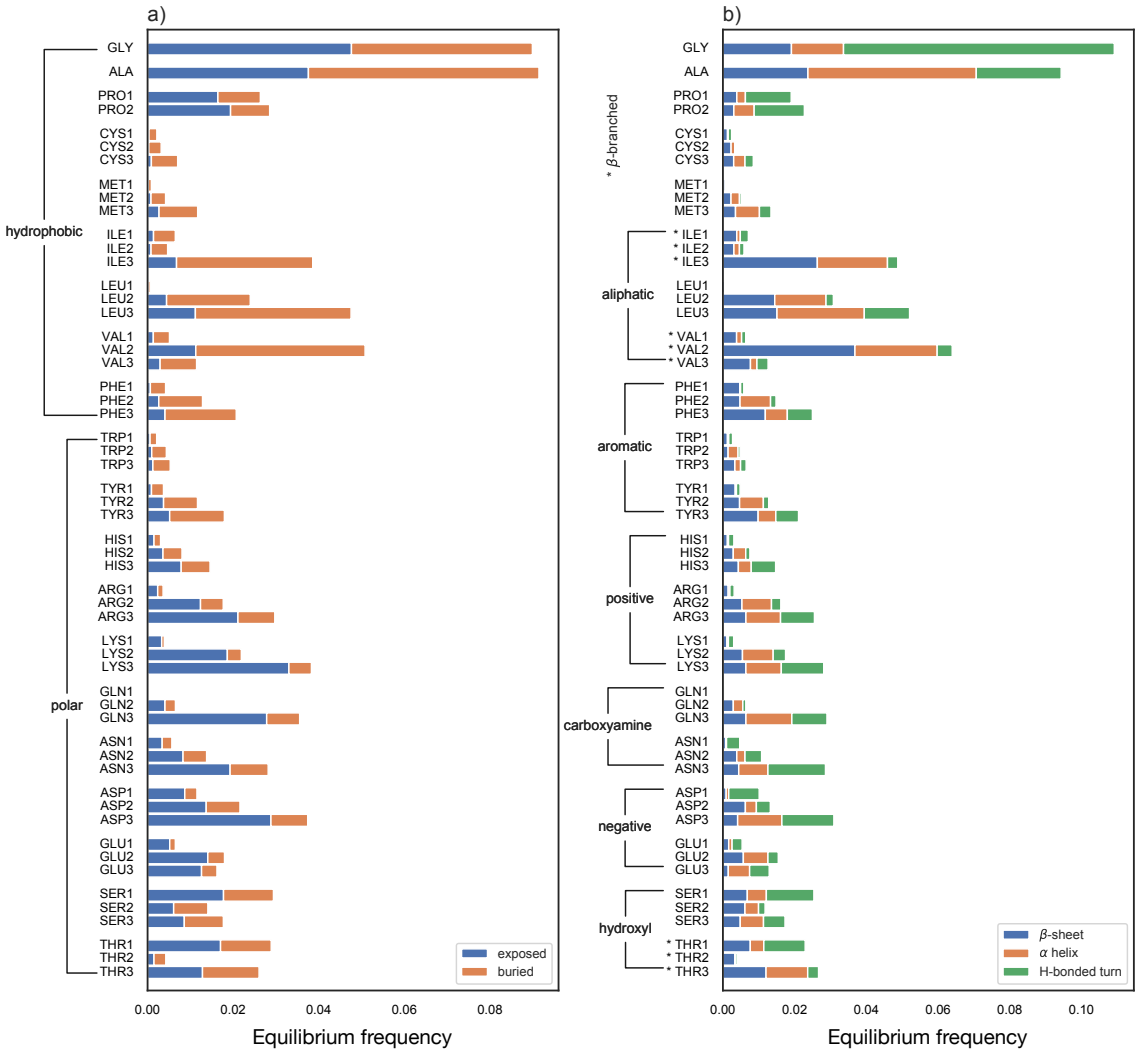

Figure 3.6: Rotamer states equilibrium frequencies differ according to structural context. For each rotamer state, overall bar length indicates its equilibrium frequency in RAM55. In **(a)** the stacked bars represent each residue's frequency at buried or exposed sites across the RAM55 dataset. In **(b)** the stacked bars represent frequencies at α-helix, β-sheet, and H-bonded turn sites, respectively.

amino acid rates, by the fact that residues at buried sites interact with many neighbouring residues and thus amino acid changes (i.e. $(A, R) \to (A', R)$) or χ_1 configuration changes (i.e. $(A, R) \to (A, R')$) are more likely to disrupt the position of other neighbours. For rotamer states the correlation between accessibility and replacement rate appears to be stronger than for amino acids ($\rho = 0.668$) across the RAM55 dataset (see Fig. 1.3), possibly because side chain geometry might be less sterically constrained than side chain type – which determines amino acid identity – at exposed sites, and, conversely, more constrained at buried sites in order to avoid clashes.

Comparing empirical rotamer state frequencies from exposed and buried residues, it is easy to notice a strong composition bias whereby hydrophobic amino acid are much less common at solvent-exposed sites, and vice-versa for polar amino acids (Fig. 3.6a). It is also interesting to note how serine and threonine both have a slightly higher preference for configuration 1 (60°) when exposed, this brings together the side chain hydroxymethyl group and the backbone carboxyl group, facilitating the formation of a hydrogen bond. This might help stabilise the local fold and be preferable to bonding neighbouring solvent molecules in configurations 2 or 3.

Protein secondary structure is a function of backbone geometry and has a very strong influence on which side chain configuration is preferred. For example, the aromatics are more frequently observed in configuration 2 (−180°) when in α-helix regions, and in configuration 3 (−60°) when in turn or β-sheet regions (Fig. 3.6). Leucine, instead, is equally likely to be in configuration 2 or 3 when in a β-sheet but strongly prefers configuration 3 for turns. Contrary to the aromatics, for asparagine and aspartic acid, configuration 3 is more common than 2 in helices and turns while 2 and 3 are equally frequent in β-sheets. Interestingly, glutamic acid, also negatively charged, is equally likely to be in configuration 2 or 3 in helices and turns but, conversely, strongly prefers configuration 2 in β-sheets. Finally, the β-branched residues favour configuration 3 (2 for valine) when in α-helices because of their need to avoid clashes between their side chain γ atoms and the N atom of the following residue. These residue-specific and backbone-dependent χ_1 propensities are consistent with those observed in rotamer libraries (e.g. [40, 115, 155]), despite the fact that the RAM55 dataset is quite varied and much larger, covering ∼50% of all PDBe entries. Further, these findings also align with what has been observed regarding changes in patterns of amino acid substitution in different structural contexts [166, 60, 127, 129].

In Figure 3.7 I investigate how context-specific composition biases influence rotamer exchangeabilities. Specifically, I compute the exchange rate matrices, as described in sections 2.2 and 2.3, from rotamer state change counts at exposed or buried sites in the RAM55 dataset. To highlight differences in the exchange patterns, I then cluster these matrices, both along the columns then along the rows, using the UPGMA algorithm [159]

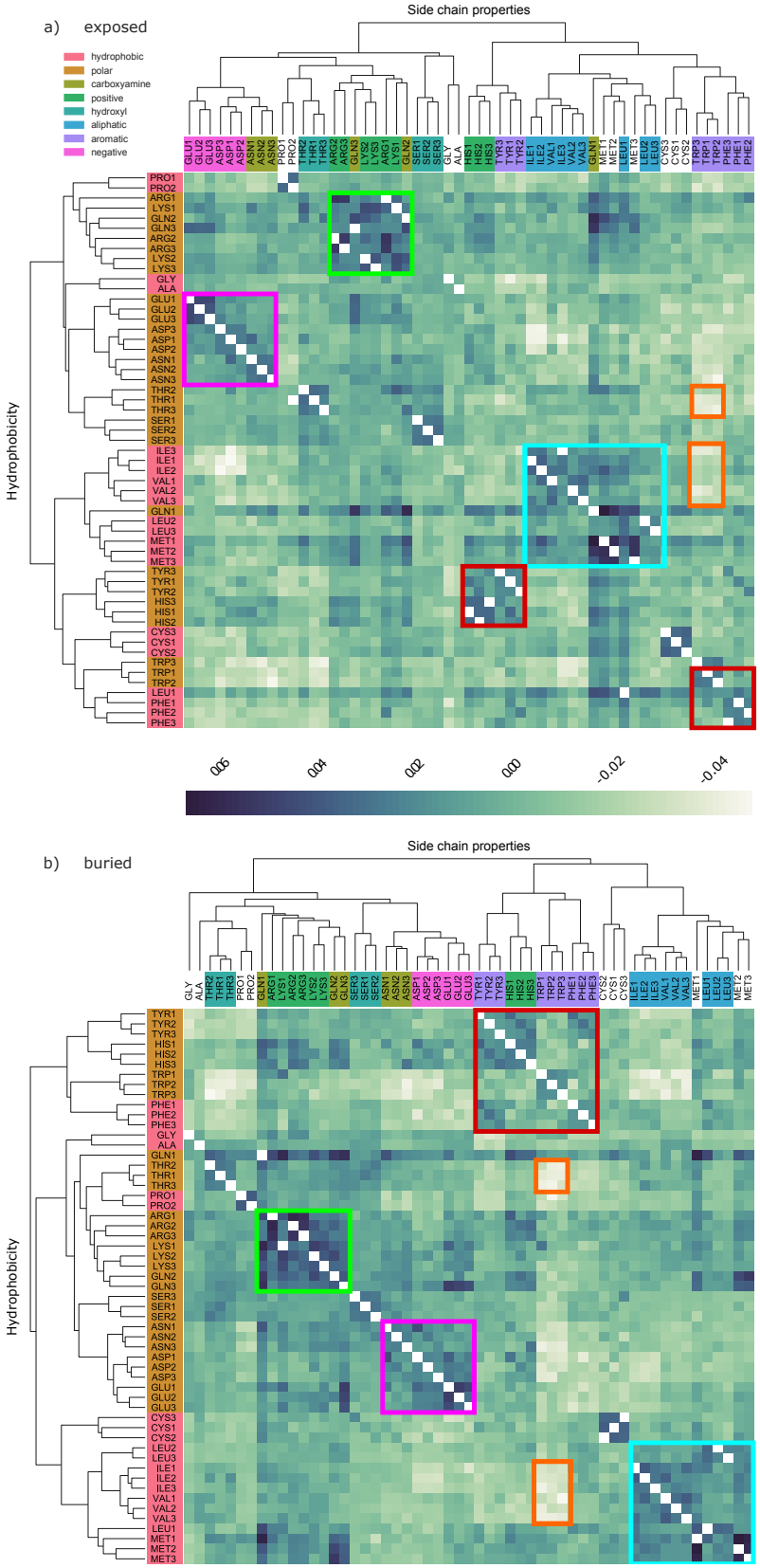

Figure 3.7: Rotamer exchangeabilities differ according to solvent exposure. The RAM55 dataset is partitioned into **a)** a buried subset and **b)** an exposed subset according to RASA. For each an exchangeability matrix is computed (see section 2.3). These matrices are shown as (log-scaled) heat maps and then clustered using the UPGMA algorithm [159] and the Pearson correlation distance between rows and columns. The axis labels are coloured according to the physicochemical properties of residues. Notable heatmap features are highlighted with coloured boxes (see text).

and the Pearson correlation distance. The latter is simply:

$$d_{X,Y} = 1 - \rho_{X,Y} \tag{3.2}$$

where $\rho_{X,Y}$ is the Pearson correlation coefficient between vectors X and Y, which in this analysis correspond to a pair of rows or columns in the exchangeability matrix. In both matrices, residues with similar physicochemical properties (coloured labels) tend to have similar exchange patterns (as measured by Pearson correlation) and cluster together, albeit with some notable differences. These include a stronger segregation of hydrophobic and polar residue for the buried matrix (Fig. 3.7b, y-axis), and a split of the aromatics into two pairs in the exposed matrix (Fig. 3.7a, x-axis). The latter might be due to the divergent replacement patterns of histidine (aromatic and positive) and phenylalanine (aromatic and hydrophobic) which cluster well with tyrosine and tryptophan, respectively (Fig. 3.7a, red squares). Further, polar residues with long side chains like arginine, lysine, and glutamine form a block of rotamer states that exchange quickly with each other (green square). In the buried matrix this block includes glutamine 1 (GLN1), which in the exposed matrix clusters with the aliphatics. Negatively charged glutamic acid and aspartic acid form another fast-exchanging cluster with asparagine (purple square), while also showing a strong tendency to conserve their side chain configuration when exchanging in both matrices. So do the aliphatics and methionine (light blue square). There are also noticeable differences at the levels of individual residues: tryptophan for example is more unlikely to exchange with the β-branched (orange squares) when in a buried context.

Figure 3.8 examines two particular residues, phenylalanine and leucine, and the corresponding subset of rotamer states (i.e. {PHE1, PHE2, PHE3}, {LEU1, LEU2, LEU3}) which show noticeably different behaviours depending on the secondary structure context. In order to quantify how much their exchange patterns differ, for each subset of rotamer states I compute the Bray-Curtis distance [17] between pairs of corresponding 3×64 sub-matrices across all five structural contexts defined above (exposed, buried, α-helix, β-sheet, H-bonded turn) plus the full RAM55 matrix (Fig. 2.6a). The most divergent exchange patters emerge when comparing the β-sheet and α-helix sub-matrices for phenylalanine (Fig. 3.8a,c), and the sub-matrix from the overall RAM55 matrix against

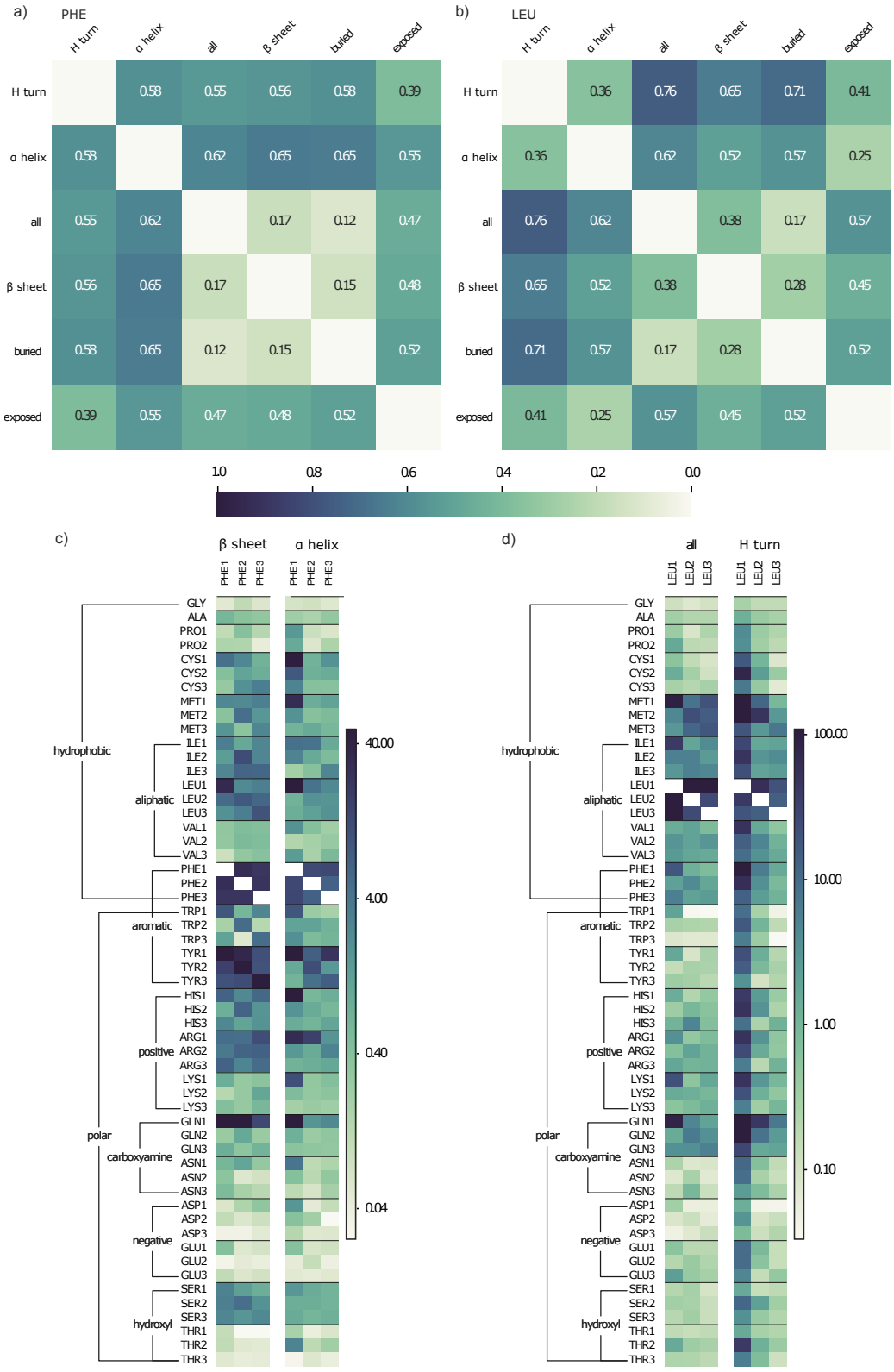

Figure 3.8: Individual rotamer state exchange rates differ according to structural context. The RAM55 dataset is partitioned into five subsets according to the structural context of each site: exposed, buried, α-helix, β-sheet, H-bonded turn. For each subset an exchangeability matrix is computed (see section 2.3) and the Bray-Curtis distance [17] matrix among the rates of **(a)** phenylalanine and **(b)** leucine across these five partitions plus the original RAM55 matrix (all) is shown. **(c)** compares the phenylalanine exchangeabilities in β-sheet and α-helix sites; **(d)** leucine's in turns and the general RAM55 matrix.

the H-turn sub-matrix for leucine (Fig. 3.8b,d). Phenylalanine appears to have weaker signals of χ_1 configuration conservation (i.e. lighter-coloured diagonal patterns) when in a helix, and specifically when exchanging with leucine, tryptophan, and methionine (Fig. 3.8c). Other patterns, for example the ones with arginine, histidine, and serine, are obviously different between the two matrices; however, they are less easy to interpret. For leucine, configuration 1 appears to be replaced more quickly overall in turns (Fig. 3.8d), and diagonal patterns also appear to be less strong in turns when leucine exchanges with aspartic acid and phenylalanine. These results highlight how different secondary structures impose distinct constraints on rotamer states replacements that can be captured by explicitly partitioning the RAM55 models according to local structural context.

3.5 Rotamer variability across protein families

In section 1.3 I have listed a few special-purpose amino acid replacement models which achieve better fit for much data by estimating exchange rates over alignments of specific proteins families (e.g. [1, 112, 30]). For most models, the observed differences across structurally and functionally related families can be largely explained by amino acids composition bias, while the exchangeability of any given amino acid remains for the most part unchanged across related families but might vary between widely different proteins (e.g. transmembrane and globular proteins) [19, 112]. This principle is highly convenient for computational purposes, and thus it is assumed to be true by many general model implementations which simply parameterise the equilibrium frequencies (see equation 1.2) and estimate them for each alignment of interest. This approach is biologically sound, significantly improves inference accuracy, and has thus become standard practice in modern phylogenetics.

Figure 3.9a illustrates the rotamer state composition bias across all family alignments used in computing RAM55. Most rotamer states sharing the same amino acid have both different overall equilibrium frequencies (red lines), and different distributions across families, suggesting that a specific configuration might be preferred over others in certain Pfam domains, similarly to what happens in different structural contexts (see Fig. 3.6).

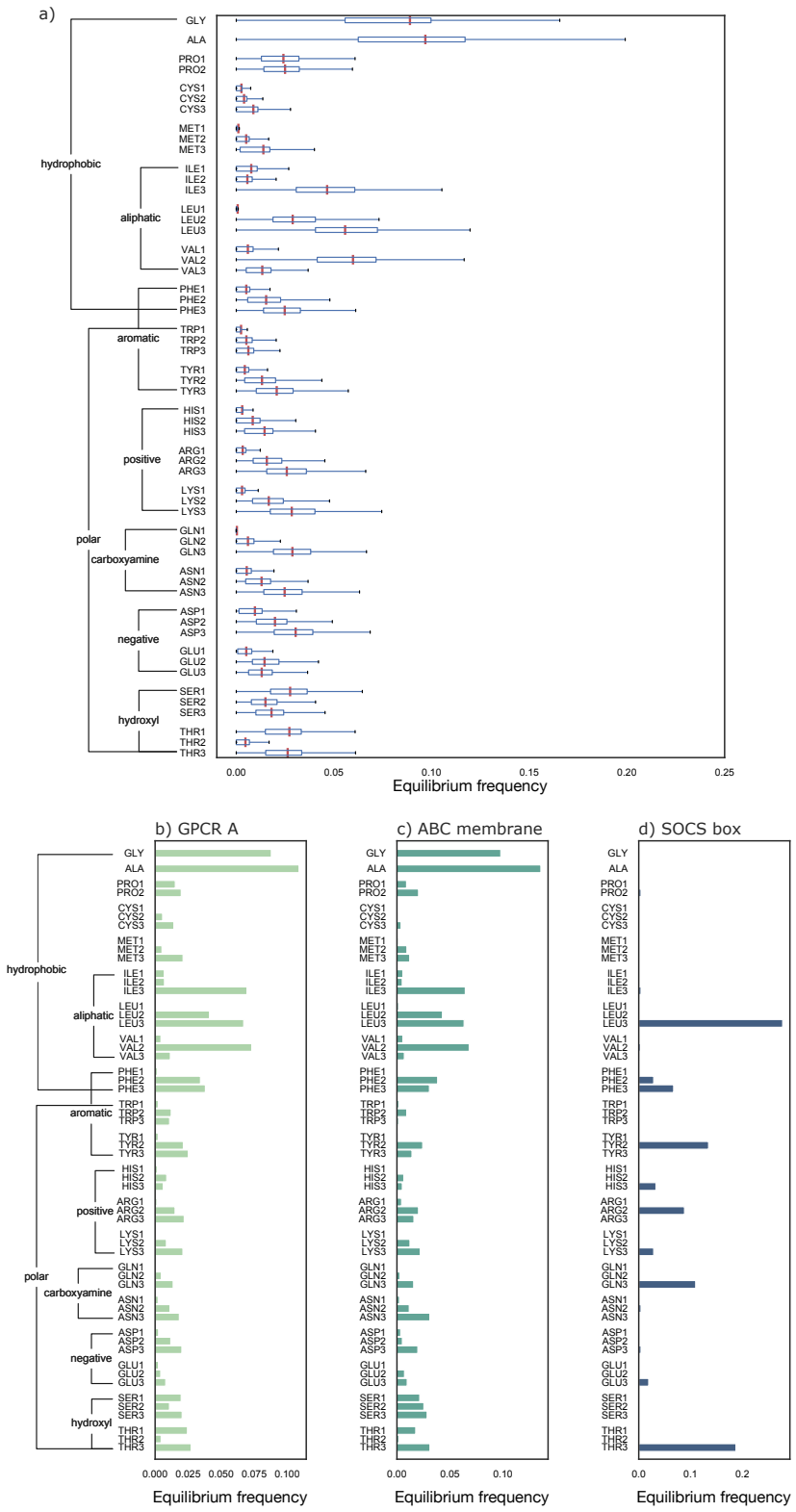

Figure 3.9: Rotamer state composition bias varies across Pfam families. **(a)** Family-specific rotamer state equilibrium frequency distributions across 5659 Pfam alignments are shown as boxplots; red lines indicate the overall state equilibrium frequency. Related Pfam families can be grouped together into clans [46]: barplots represent clan-specific equilibrium frequencies for **b)** GPCR A transmembrane receptors, **c)** membrane components of ABC transporters, and **d)** SOCS proteins.

In Figures 3.9b,c,d I have grouped a subset of Pfam family alignments from the RAM55 dataset according to their Pfam clan, which is meant to aggregate families according to similarity of sequence, structure or profile-HMM [46].

The clans presented here represent three example protein folds with distinct cellular localisations, structural requirements, and functions. Specifically, the G protein-coupled receptor-like A superfamily (CL0192; 354 structures in RAM55) contains various rhodopsin seven-transmembrane receptor families. This clan's main shared motif is the seven transmembrane helix domain, and the function of most of its rhodopsin-like members is light absorption and G-protein activation [165]. Proteins in the ABC transporter membrane domain clan (CL0241; 37 structures in RAM55) instead share a six transmembrane helix region and are responsible for translocation of a variety of compounds across biological membranes [15]. Finally, SOCS proteins (CL0642; 39 structures in RAM55) usually possess substrate-binding domains and, through their C-terminal SOCS box, are able to recruit E3 ubiquitin protein ligases regulating protein turnover by targeting proteins for proteasome-mediated degradation [83].

In GPCR receptors and ABC transporters, the frequency of hydrophobic states is higher, which is consistent with their transmembrane helices being exposed to membrane lipids. There are also a few noticeable state-specific patterns including: 1) the three serine configurations which are almost equally frequent for ABCs while configurations 1 and 3 are much more common for GPCRs, 2) tyrosine in configurations 2 and 3 swapping their frequency order, as well as 3) a similar swap between arginine in configurations 2 and 3.

Rotamer state frequencies across SOCS proteins also show some interesting features, particularly a very strong preference for leucine in configuration 3 and tyrosine in 2. Interestingly, the interaction between the SOCS box and elonginBC, which assists in recruiting E3 ligases, depends upon the first 12 residues of the SOCS box domain and particularly on a deeply buried, conserved leucine [9]. Leucine 4 is the only residue conserved across all human SOCS proteins and its side chain is consistently observed in χ_1 configuration 3 across all SOCS protein structures in the RAM55 dataset. Further, the C-terminal part of the SOCS-box is required for functional interaction with several

receptor motifs, with a single tyrosine residue at position 253 being a critical binding determinant [101]. Similarly to leucine 4, tyrosine 253 is conserved in configuration 3 in all 39 structures in RAM55.

From these results it clearly emerges that, in addition to composition bias at the amino acid level, there are also significant differences in term of χ_1 configuration composition across proteins of different families, as captured by clan-specific rotamer state equilibrium frequencies.

3.6 Rotamer variability across alignment datasets

For assembling the RAM55 alignment dataset (see sections 2.1 and 2.2) [134], we imposed a rotasequence identity threshold of >75% when tabulating changes in sequence pairs (see section 2.2). This corresponds to ∼80% amino acid sequence identity and aims at limiting the chances of observing sites where a multiple changes have occurred (a phenomenon also referred to as multiple hits), which would confound our counts of observed changes. While this is a widely-used rule of thumb for assembling phylogenetic alignment datasets, there are also datasets that are assembled using approaches that emphasise structure similarity over sequence similarity (see section 1.1). One example is the HOMSTRAD dataset [121] which contains high-quality protein structure alignments covering 1032 families and 3454 unique structures. For each family, HOMSTRAD provides a structure-based alignment derived using COMPARER [150] and annotated with JOY [122] according to the local structural context of each residue. Because of this, amino acid sequence identities in HOMSTRAD alignments can be as low as ∼30%, which implies that the proportion of multiple hits should be higher in this dataset. In this section I investigate whether it is possible to detect context-dependent rotamer state exchange patterns, similar to those in RAM55, from counts tabulated from HOMSTRAD alignments that are, on average, more divergent than the ones in the RAM55 dataset. This would confirm that these exchange patterns are robust to different alignment methods and "noisier" datasets.

In order to compute the exchangeability matrix, I followed the same approach described in sections 2.1 and 2.2, but this time relying on HOMSTRAD's structural alignments of amino acid sequences instead of Pfam alignments and removing the $> 75\%$ rotasequence identity threshold. The rotamer state change counts were partitioned according to JOY annotations into buried, exposed, α-helix, and β-sheet subsets which map to the DSSP annotations of the same name used for RAM55. For each subset of sites I then computed an exchangeability matrix as described in section 2.3.

In this thesis, I concentrate on the subset of sites which are annotated as buried since 1) the exposed exchangeability matrix is quite sparse (data not shown), possibly because of lower structure quality in these region as well as smaller dataset size ($\sim$4% of the RAM55 dataset); and since 2) the buried matrix exemplifies most patterns present in the α-helix and β-sheet matrices (data not shown). Figure 3.10a shows the exchangeability matrix from buried HOMSTRAD residues in heatmap form. Many of the most noticeable exchange patterns present in the general RAM55 matrix (Fig. 2.6a) can also be observed here, despite standing out less starkly from the background. In the HOMSTRAD matrix there is strong evidence of χ_1 configuration conservation, particularly among the aromatics (red box), aspartic acid and asparagine, serine and threonine. This is evidenced by the darker cells (i.e. higher exchangeabilities) on the diagonals of the respective 3×3 submatrices.

I then clustered the buried HOMSTRAD exchangeabilities using the UPGMA algorithm and the Pearson correlation distance (see section 3.4). Rotamer states with similar physicochemical properties (including hydrophobicity which is not shown) still tend to cluster together (Fig. 3.10b), but do so in a less orderly pattern than the RAM55 buried rates in Figure 3.7b. Further differences between replacement rates include the hydroxyl (light blue box) and aromatics (red box) clusters which are less distinct from the background and "noisier" in the HOMSTRAD matrix, with LEU1, PHE1, TYR1, TRP1, HIS2, and HIS3 being separated from their respective clusters. Finally, polar residues with long chains (green box) also form a cluster containing some of the arginine, lysine, and glutamine rotamer states. This latter cluster is much closer in composition and intensity to the one in Figure 3.7b. These results suggest that basic rotamer state replacement

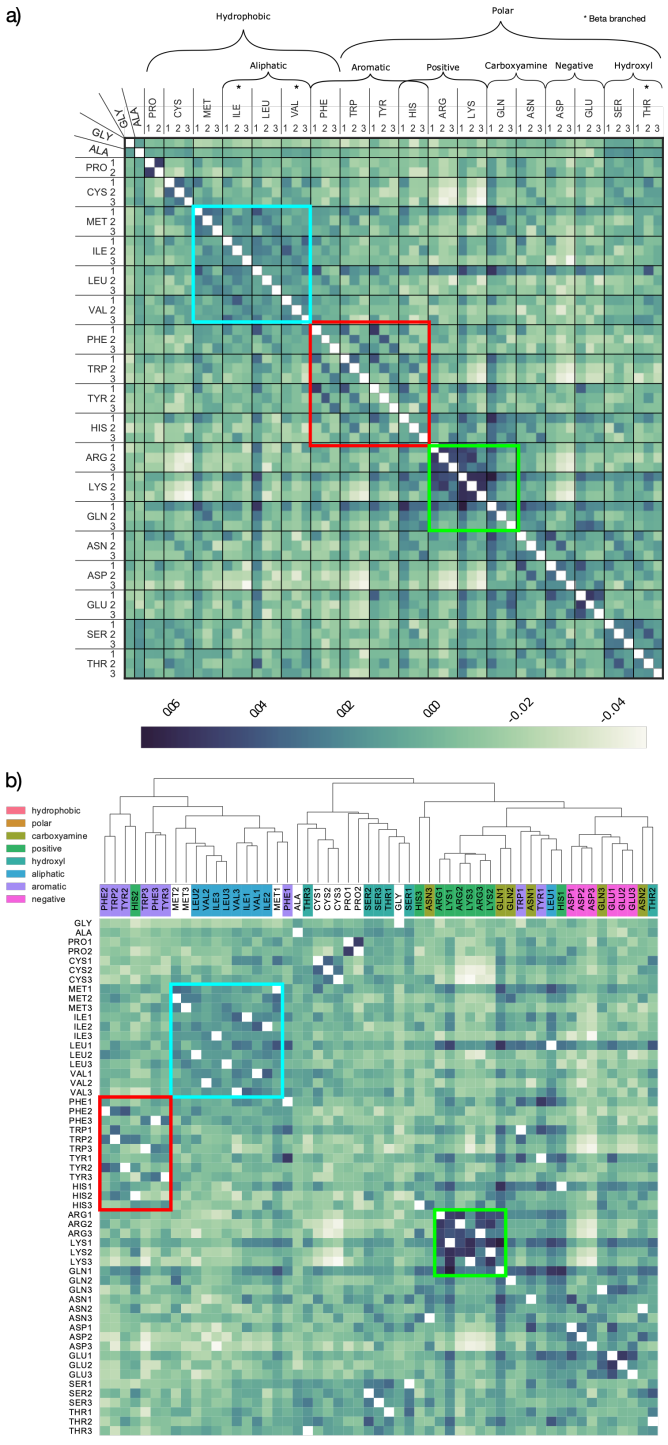

Figure 3.10: Context-dependent rotamer exchange patterns in the HOMSTRAD [121] dataset. Rotamer state change counts are tabulated from HOMSTRAD structure alignments across all buried residues as defined by JOY [122]. An exchangeability matrix is computed from these counts (see section 2.3), **a)** displayed as (log-scaled) heat map, and **b)** clustered using the UPGMA algorithm and the Pearson correlation distance between columns. The axis labels are coloured according to the physicochemical properties of residues. Here the the hydrophobicity labels are not shown for clarity and clustering is only applied to columns.

patters, such as χ_1 configuration conservation among similar residues, can be observed even in datasets containing highly diverged sequences. Other patterns like those within fast-exchanging clusters of related states (see Fig. 3.10a), "blend" into the background, possibly because they are more specific, less strong overall and thus less robust to more diverged alignments and more multiple hits.

3.7 Conclusion

In this Chapter I have first assessed whether χ_1 rotamer configurations are robust to conformational changes and thermodynamic fluctuations brought about by variations in the preparation, crystallisation, acquisition, and modelling conditions of protein structures of the same protein. A study of 284 human thrombin structures confirms results from previous studies and shows 1) that the amount of non-evolutionary χ_1 configuration changes, which might confound the RAM55 exchange rates, is quite small and 2) that structure quality filtering likely removes some of the spurious changes.

I then investigated how rotamer state exchange patterns might vary depending on local structural contexts. Results show state-specific differences in replacement rates and equilibrium frequencies depending on solvent exposure and secondary structure of sites. These differences are likely explained by a combination of the side chain physicochemical properties and the side chain geometry, which might cause specific rotamer states to be more favourable in a given context. Examples include SER1 which facilitates the formation of an hydrogen bond with the backbone at exposed sites that might reduce destabilising interaction with solvent molecules, and β-branched residues which prefer configuration 3 in α-helices to avoid clashes between their side chain γ atoms and the N atom of the following residue. There are also differences in how strictly χ_1 configuration is conserved when exchanging between similar residues, with constraints varying across secondary structure classes, primarily. These results share strong similarities with, and expand upon, what has been observed for amino acid replacements in different structural environments [127, 129, 166, 60].

Protein family alignments in RAM55 show strong rotamer state composition bias. As

for structural context, some of this variation can be attributed to the physicochemical properties of the side chain (i.e. the amino acid identity), as is the case for hydrophobic states abundance in transmembrane regions. However, there is clear evidence of specific χ_1 configurations being strictly favoured within a motif, as in SOCS proteins alignments. These findings too mirror similar observations for amino acid composition bias [19, 112, 2, 37] and improve our understanding of the interplay between protein sequence and structure evolution.

Finally, I computed and analysed rotamer states exchangeabilities using the HOMSTRAD structural alignment dataset. Some of the general exchange patterns observed in RAM55, such as χ_1 conservation among the aromatics and fast exchanges among long-chained states, appear to be robust to structural alignment methods and higher sequence divergence as found in HOMSTRAD. Regardless, it is probably advisable to rely on phylogenetic alignments and more closely related sequences when building an empirical evolutionary model as excessive divergence is likely to introduce unnecessary noise to exchange rate and "average out" informative patterns. However, in future studies it might be interesting to investigate strategies for teasing apart the effects of increased sequence divergence and increased structural divergence on rotamer state exchanges, as observed between RAM55 and HOMSTRAD. This could be accomplished through IRM estimation methods which allow for variable divergence levels and don't filter out more diverged sequences and structures, along the lines of those developed for amino acid IRM estimation (e.g. [181, 75, 104, 29]).

All these results suggest that RAM55 might benefit from optimising (or empirically estimating) the rotamer state equilibrium frequencies from the alignment dataset of interese, as I will investigate in Chapters 4 and 5. Further, partitioning RAM55 according to the five structural categories defined here (buried, exposed, α-helix, β-sheet, H-turn) might also improve inference accuracy. However, implementing and validating this latter strategy represents a substantial project and goes beyond the scope of my PhD.

Both these approaches have been successfully applied to amino acid replacement models in the past (e.g., [127, 175, 90, 166, 60, 102, 140]) and our ability to detect context-dependant variations in rotamer state replacement patterns suggests that these techniques have the potential to similarly improve model fit for the RAM55 model. Of the two, a "partitioning" strategy is, however, more difficult to implement within existing phylogenetic software and would require either 1) a known structure in order to appropriately categorise each site, or 2) a more complex modelling strategy which could assign the optimal partition matrix to each site from sequence data alone. HMM-based approaches have been successfully demonstrated for structurally-partitioned amino acid models (see section 1.4), their application to the RAM55 model is for future work.

4. RAM55 for phylogenetic inference

This chapter has been adapted and expanded from a first-author publication [134]. All data collection and analysis was performed by me, the original Perron et al. [134] manuscript was written by me, edited collectively, and agreed upon by all co-authors.

Chapter 2 described distinct replacement patterns for residues according to their χ_1 rotamer configuration, and introduced the rotamer-aware RAM55 model [134]. In Chapter 3, I have also shown that χ_1 configurations have low "background" variability across structures of the same protein. These findings suggest that the RAM55 model provides additional, robust, and biochemically plausible information about the evolutionary process. This Chapter aims to quantify the contribution of χ_1 configurations in RAM55, as well as investigating whether RAM55 enables us to use this information to improve our ability to infer phylogenies.

4.1 Kullback-Leibler divergence

In Perron et al. [134], we measured the amount of information lost, regarding the amino acid sequence, when a 20-state model is used to approximate RAM55 by computing the Kullback-Leibler (KL) divergence in *bits* [97] for each rotamer state (A, R) and its corresponding amino acid state (A).

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 a 20-state model. For instance, a histidine residue is more likely to be substituted with an asparigine when in configuration 3 than when in any of the other χ_1 configurations (see Fig.

2.10). In the RUM20 model there is no distinction between the three χ_1 configurations: 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 to that obtained 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 loss can be quantified using the KL divergence which measures the divergence between the amino acid probability distribution at time t, when starting with rotamer state (A, R) for 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:

$$D_{\mathrm{KL}}(P_{\mathrm{RAM55}}(t,(A,R))||P_{\mathrm{RUM20}}(t,A)) = \sum_{a\in S_{20}}\left(\sum_{r\in R_a} P_{\mathrm{RAM55}}(t,(A,R),(a,r))\log_2\frac{\sum_{r\in R_a} P_{\mathrm{RAM55}}(t,(A,R),(a,r))}{P_{\mathrm{RUM20}}(t,A,a)}\right) \quad (4.1)$$

with S_{20} being the 20-state space, 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 eqn. (4.2) below), and $P_{\mathrm{RAM55}}(t,(A,R),(a,r))$ the $((A,R),(a,r))$ element of $P_{\mathrm{RAM55}}(t)$.

At $t = 0$, no loss occurs due to the amino acid sequence being fully known in both models. As $t \to \infty$, both models tend towards the equilibrium amino acid frequencies and the loss tends towards zero. The differences between the two models manifest in between these extremes. Figure 4.1 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 (e.g. VAL1, ASP2, ASP3, SER2), 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 $\sim$20% amino acid sequence identity.

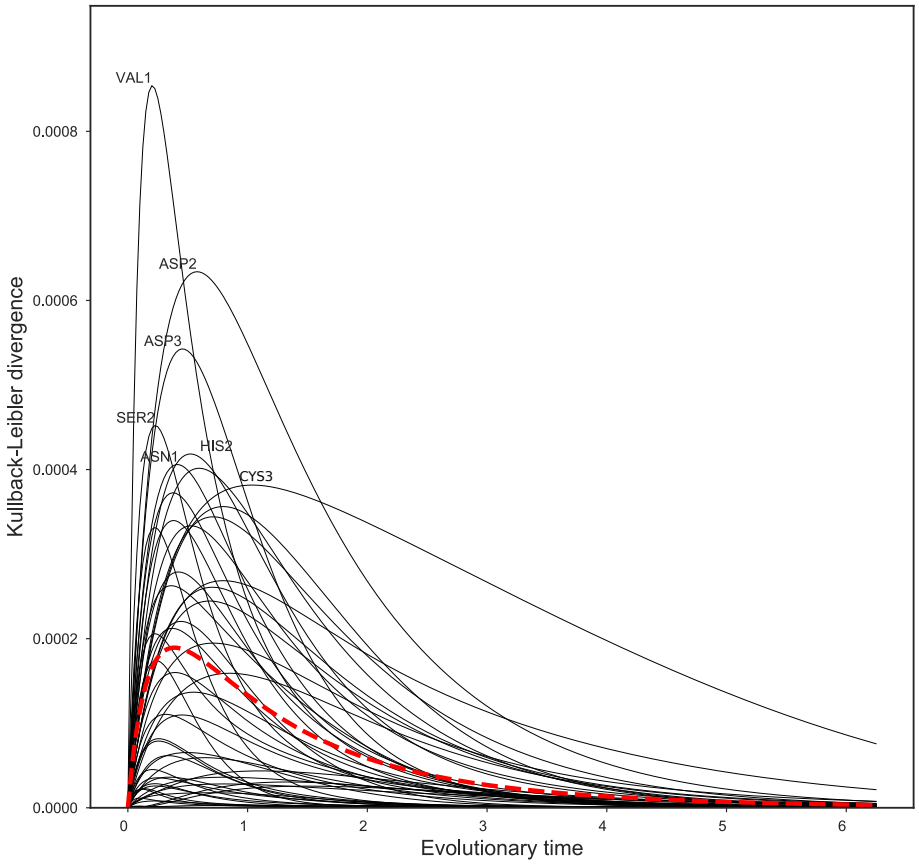

Figure 4.1: Kullback-Leibler (KL) divergence [97] measures the amount of information about the amino acid state that is lost when the structure-free RUM20 model is used to approximate the 55-state model, RAM55. KL divergence is computed for every pair of amino acid state and corresponding rotamer state 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.

4.2 Likelihood calculation and maximisation over phylogenies

To investigate the effect of RAM55's expanded state space on phylogenetic inference, in Perron et al. [134] we implemented various replacement models using a ML framework for phylogenetic inference applied to multiple sequence alignments [47]. This standard approach searches for the tree T that maximises the likelihood function with substitutions, occurring as independent events on the branches of T, modelled by a Markov process. Markovian state substitutions over time t are described by a probability matrix defined by

$$P(t) = e^{tQ} \tag{4.2}$$

where Q is the IRM of the Markov process, scaled so that, at equilibrium, the model produces on average one amino acid substitution per unit of time (see section 2.3), and with Q's on-diagonal elements set to 0 for of mathematical and computational convenience.

The likelihood of T, which describes both tree topology and branch lengths, given data (i.e. the alignment) X and IRM Q can then be computed as:

$$L(T|Q, X) = \prod_i L(T|Q, X_i) \tag{4.3}$$

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 eqn. (4.2) to each tree branch and using Felsenstein's pruning algorithm [49]. Maximising L over T provides estimates for tree shape and branch lengths (i.e. $\widehat{T}$), and thus the most-likely phylogeny given the observed data and the current substitution model. This framework also allows to match the model's state frequencies to the observed data, an approach usually designated as "+F" in phylogenetic software, by simultaneously maximising L over these frequencies. If the model contains any other free parameters, they can be estimated using this same strategy.

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 [188]. The shape parameter of the gamma distribution, α, is usually estimated by ML along with T as it is considered specific to each dataset. Models using the gamma distribution to model rate heterogeneity are denoted "+G".

All ML tree inferences were performed using RAxML-NG [94]. Following the needs of RAM55, particularly the necessity to accommodate a custom state space with character coding that went beyond the traditional amino acid and codon alphabets, I collaborated with Kozlov et al. [94] to include these functionalities into RAxML-NG's development. This allows RAM55 (and any other model with custom state spaces and rate matrices of any size) to leverage a highly optimised likelihood computations library [53], as well as the flexibility and user-friendliness of one of the most widely-used phylogenetic packages. By implementing RAM55 in RAxML-NG I can infer tree topology, branch lengths and likelihoods that can be used for model fitting and comparisons. RAxML-NG also permits the +F and +G variants of substitution models through its "+FO" and "+G" options: these variants further improve RAM55's fit for the data as I show in the following sections.

RAM55's expanded state-space has some inevitable repercussions for CPU time, not least because 20-state models benefit fully from bespoke likelihood computation algorithms in RAxML-NG, whereas the RAM55 model currently works with general likelihood algorithms that are less efficient. Nevertheless, computation times remain acceptable, tending to be 5–10 times longer than using 20-state models (Fig. 4.2).

4.3 Log Likelihood comparison across models

Phylogenetic models are usually compared in terms of fit for a specific alignment dataset with the assumption that the highest likelihood indicates the model that better approximates the observed evolutionary process [47, 189]. The Akaike Information Criterion (AIC), an information-theoretic score, is frequently used for model selection [3, 161] and fits well with the models described in this Chapter, most of which are non-nested (i.e. they do not share any term) and, in cases, have different state spaces (e.g. the amino acid

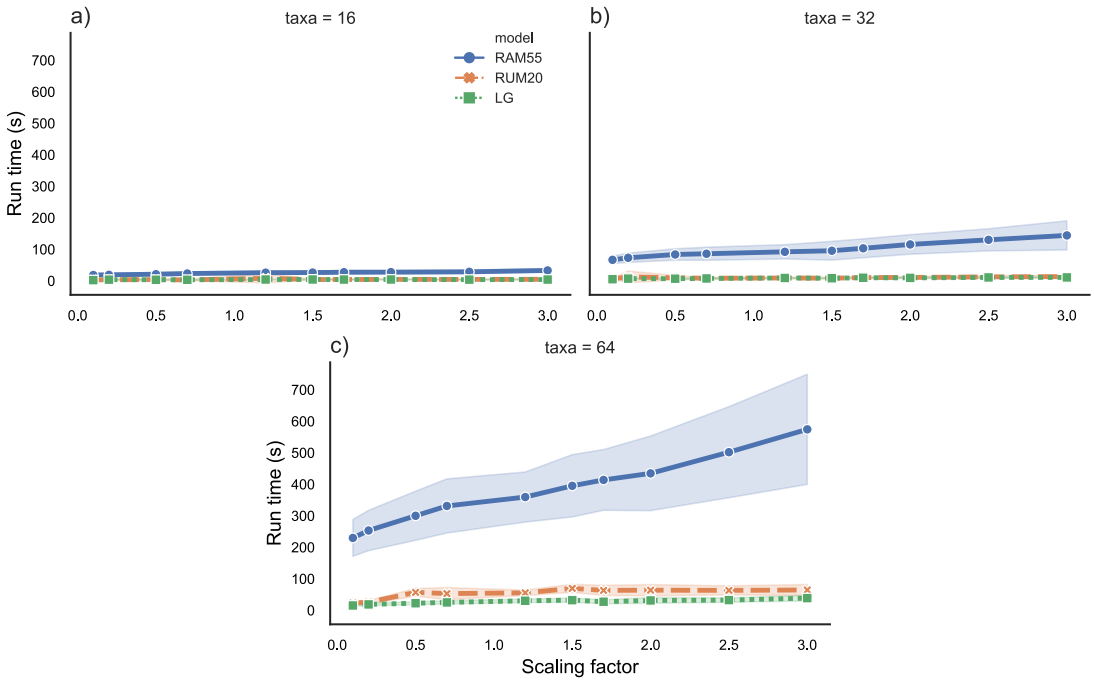

Figure 4.2: Run time comparisons. Rotasequence alignments (200 sites, 100 replicates per scaling factor) were simulated using RAM55 and the 16-, 32- and 64-taxon trees shown in Figure 4.3), scaled according to the factors shown on the x-axis (see section 4.4). The plots report mean run times in seconds and standard deviation (shaded ares) for ML tree inference from these rotasequence alignment data under the RUM20, LG and RAM55 models in RAxML-NG. Masked amino acid alignments (see section 4.4) are used for the RUM20 and LG analyses).

state space and the rotamer state space). The AIC score is defined as

$$AIC = 2k - 2\log \widehat{L} \tag{4.4}$$

where k is the number of estimated parameters in the model and $\widehat{L}$ is the maximized value of the likelihood function of eqn. (4.3). However, the likelihood function depends upon a model's state-space [5]: 20-state models of amino acid substitution cannot be directly compared with RAM55 as they exist in different state-spaces.

Whelan et al. [182] have developed a generalized "correction" allowing the comparison of likelihoods between state-spaces. This strategy is applicable to any two state-spaces (D,C), where multiple states of D are treated equivalently in C, providing that 1) each state in D maps to a single state in C and that 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:

$$L(X^D|\theta^C) = L(X^C|\theta^C) \sum_{\text{taxon } p} \sum_{\text{site } q} \frac{\pi^D_{d(p,q)}}{\pi^C_{c(p,q)}} \tag{4.5}$$

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. For the purpose of this analysis, model D corresponds to RAM55, whose distinct states can be uniquely compounded into amino acid states (e.g. TRP3 $\rightarrow$ TRP), and model C corresponds to a 20-state amino acid model (e.g. WAG, LG or RUM20) whose states can be mapped to a unique set of rotamer states (e.g. TRP $\rightarrow$ {TRP1, TRP2, TRP3}).

As an independent approach to test the contribution of knowledge of rotamer state substitutions, in Perron et al. [134] we also generated 55-state models that were expanded versions of the 20-state LG model [105]. Likelihoods then are directly comparable to RAM55's, since they share the same state-space. This model expansion operation was performed without the introduction of any information about the additional states (LGexp model) or, alternatively, by accounting for just the observed frequencies (but not the relative exchange rates) of these additional states in the RAM55 dataset (LGbyfreq-exp). In each case, I started from LG's exchangeabilities and reconstructed an artificial raw substitution count matrix by reversing 20-state versions of eqns. (2.7) and (2.2). For LGexp, this reconstructed counts matrix N was then expanded into a 55-state counts matrix ($\bar{N}$) according to

$$\bar{n}_{(A,R),(A',R')} = \frac{n_{A,A'}}{|R_A|.|R_{A'}|} \tag{4.6}$$

where $|R_A|.|R_{A'}|$ is the product of the dimensions of a sub-matrix in $\bar{N}$ corresponding to a single cell of N. Eqns. (2.2) and (2.7) 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

$$\bar{n}_{(A,R),(A',R')} = \pi_{(A,R)}\pi_{(A',R')}n_{A,A'} \tag{4.7}$$

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.

Thus, all models can be compared according to their fit for a specific dataset using eqn. (4.4) with the likelihood term corresponding either simply to $\widehat{L}$ for 55-state models (RAM55, LGexp, LGbyfreq-exp) or to the state-corrected likelihood obtained from eqn. (4.5) for 20-state models (RUM20, LG, WAG). The latter is referred to as a "state-corrected AIC score".

4.4 Tree generation and alignment simulation

The first step in evaluating a new model consists in detecting whether its use can produce an improvement in model fit and inference accuracy under ideal circumstances. This is accomplished by applying the model to an alignment dataset which has been obtained through simulation under the model itself. Here, simulated sequence alignments were obtained under RAM55 using four trees generated through a Yule process (8, 16, 32 or 64 taxa; branch lengths $\in [0.01, 0.5]$; see Fig. 4.3a), as well as a pruned (mammals) and scaled version of the Ensembl-compara species tree [72] (see Fig. 4.3b).

The substitution simulation approach is based on the concept outlined by Method 1 of Fletcher and Yang [52], which I appropriately implemented for RAM55's expanded state-set. An ancestral sequence is generated by sampling rotamer states according to their equilibrium frequencies in RAM55, and is assigned to the root node of the guide tree (Fig. 4.4). The tree is then traversed using a preorder strategy in which each node is visited before its descendants. For each node, a sequence is generated starting from its parent node's sequence. The replacement process is simulated independently at each

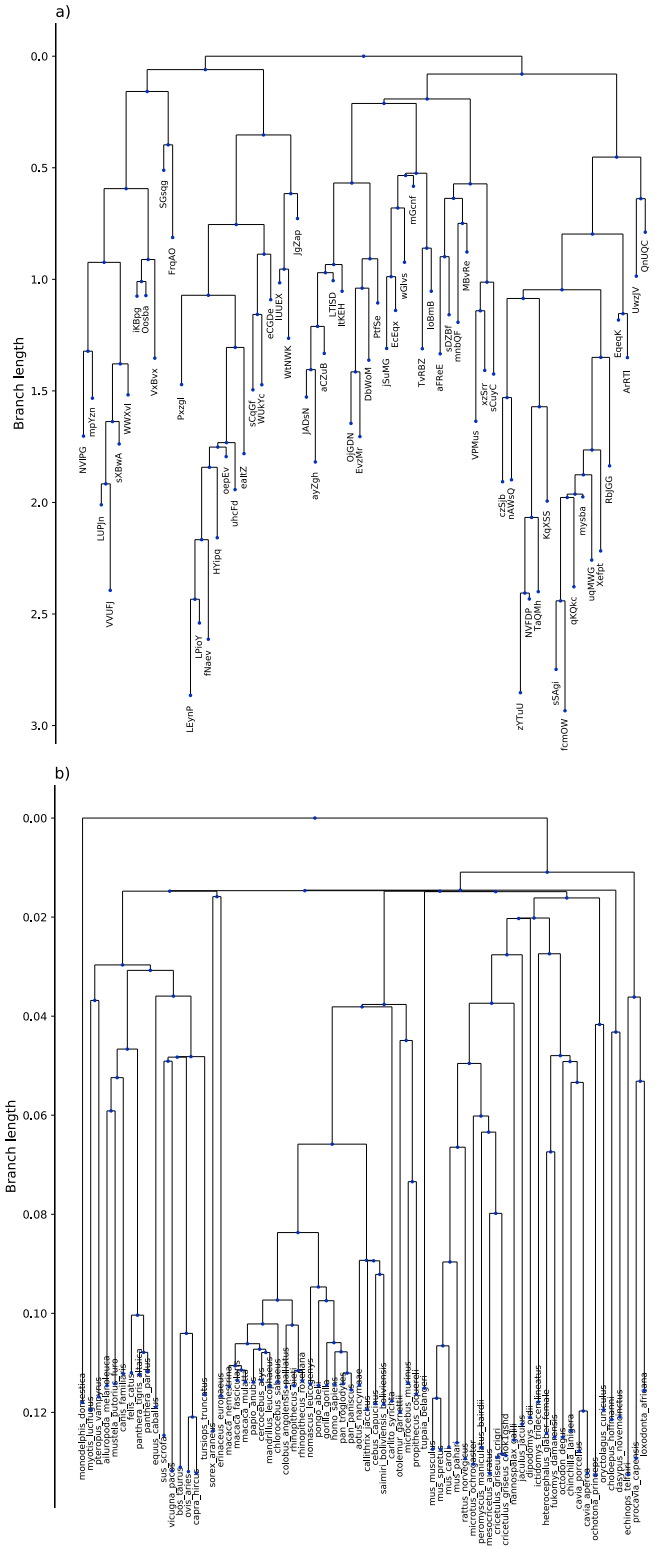

Figure 4.3: Phylogenetic trees used for simulation. **(a)** 64-taxon tree generated by a Yule process; **(b)** a pruned (mammals only) and scaled version of the Ensembl-compara species tree [72]. Branch length height is shown on the y-axis, for both tree the average branch length is ∼0.26.

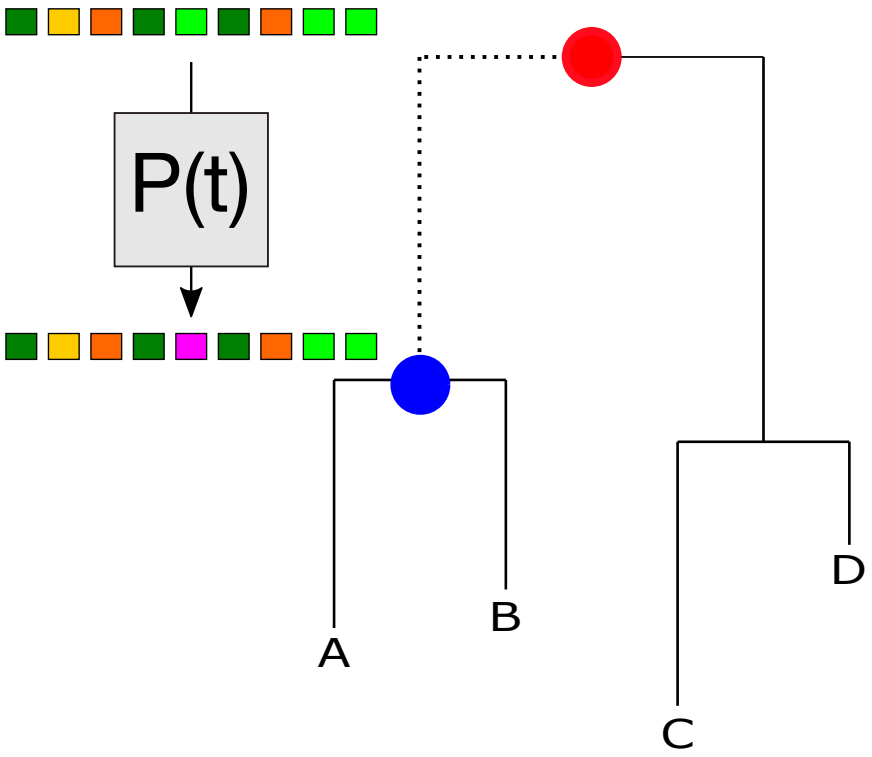

Figure 4.4: Alignment simulation along a tree under RAM55. The root node (red dot) sequence is generated by sampling rotamer states according to their equilibrium frequencies in RAM55. The tree is then traversed from root to tips using a preorder strategy. For each descendent node (blue dot) evolution is simulated independently for each descendent at each site in the parent's sequence (coloured boxes) according to the replacement probabilities in $P(t)$ (see eqn. 4.2) where t is the length of the edge connecting the descendent node to its parent node (dotted line). The final output is an alignment of simulated sequences, one for each terminal node (i.e. A, B, C, D).

site by replacing the rotamer states according to a probability matrix $P(t)$ (see eqn. 4.2) where t is the length of the edge connecting the descendent node to its parent node.

To allow investigation of a realistic range of sequence divergences (around 10–85%) while maintaining consistent tree topologies, all branches 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. We generated 100 rotasequence alignments for each scaled tree, each 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 analysed in this form; their constituent rotasequences were converted

into amino acid sequences for inference under 20-state models by masking (i.e. removing) the χ_1 configuration component of their rotamer states (i.e. $(A, R) \rightarrow A$).

4.5 Model benchmarking: Simulation

In Chapter 2 I illustrated how RAM55 adds biologically meaningful information that is not found in traditional amino acid models. Here, I investigate whether a typical protein sequence alignment contains enough information to permit identification of the true generating model, in a case where the data were generated under RAM55. Further, I also consider whether RAM55 affects our ability to infer phylogenies compared to models using the traditional 20-state space.

To examine our ability to detect χ_1 configuration-influenced evolution, I assessed RAM55's goodness-of-fit when analysing alignments simulated using the model itself. As described above, 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. From these simulated alignments I then infer the corresponding phylogenies by maximum likelihood using RAM55 or other models that are widely used for phylogenetic analysis of amino acid sequences (see section 4.2). The popular LG model [105] in particular represents a good benchmark for RAM55 as it is a general amino acid replacement model (along the lines of earlier ones such as JTT and WAG) whose rates are estimated by ML from a large and varied protein alignment dataset, similar to RAM55's collection of Pfam alignments. LG and its more recent variants (e.g. LG4X [103]) are also widely used for phylogenetic inference purposes ($\sim$ 2000 citations) and are already implemented in RAxML-NG. Figure 4.5 compares AIC scores [3] or state-corrected AIC scores (see section 4.3) 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 Tab. 4.1). 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, a version of the 20-state LG model "uniformly

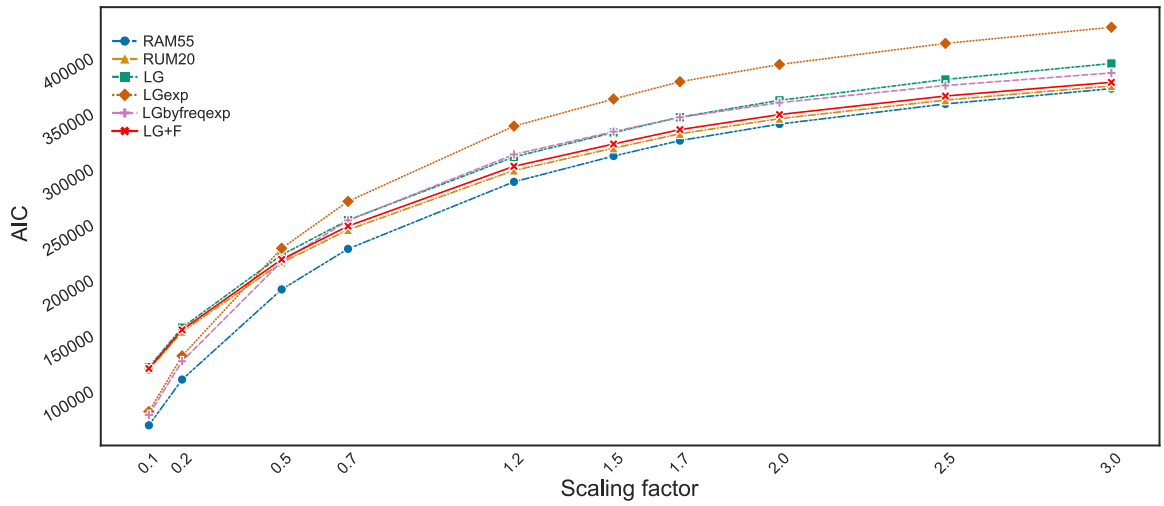

Figure 4.5: Improvement in model fit when using RAM55 under ideal circumstances. The y-axis shows AIC values for competing models; each data point corresponds to the mean AIC value of trees inferred from 100 alignments of 1000 sites each, simulated under RAM55 using a 64-taxa tree (Fig 4.3a).

expanded" to 55 states but incorporating no structural information (see section 4.3), 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 non-informative complexity to a 20-state model penalises its performance, whilst 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, I 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 I considered both 1) the Euclidean distance [47], 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, on average, at least as close or closer to the reference phylogenies than those inferred by amino acid replacement models such as the RUM20 model – computed from the same dataset as RAM55 – or LG. Figure 4.6a compares the distributions of Euclidean distances between inferred and reference trees, estimated using the RAM55, RUM20 and LG models in simulations of 1000 sites on a

64-taxon phylogeny. Shifts towards 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 (see Fig. 4.7), as well as when simulating over a larger, empirical tree (Fig. 4.6b).

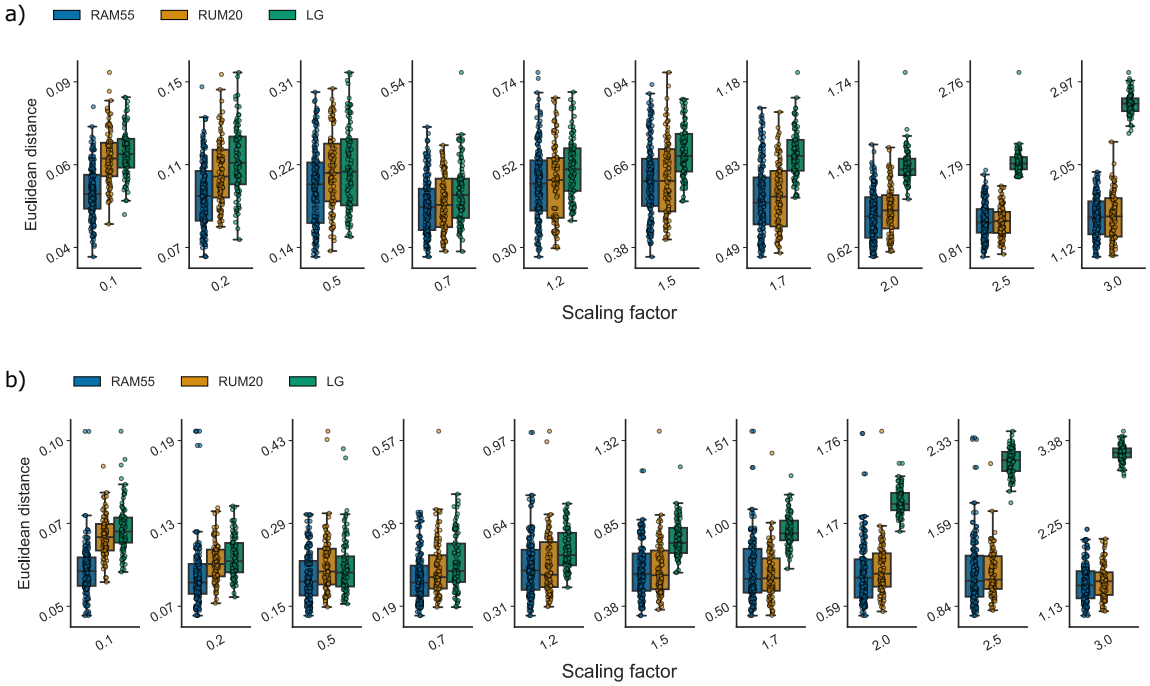

Figure 4.6: Improvement in phylogenetic inference accuracy when using RAM55 under ideal circumstances. Comparison of the RAM55 model (blue bars) against LG (green bars) and RUM20 (orange bars) in terms of Euclidean distance of inferred trees from the reference phylogeny used to simulate the alignment. Boxplots illustrate distance distribution and median (horizontal lines); scatterplot points represent individual distance values. In **(a)** tree inference is performed on alignment data sets (1000 sites, 64 taxa, 100 replicates per scaling) simulated using RAM55 and a Yule process reference phylogeny (Fig. 4.3a), scaled according to the factors on the x-axis. In **(b)** a pruned and scaled version of the Ensembl-compara species tree Fig. 4.3b) is used to simulate the alignments (1000 sites, 70 mammal taxa, 100 replicates per scaling). 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 4.8 — highlighting one internal branch for each of four topologies — illustrates this with branch length estimates from RAM55 having smaller variances,

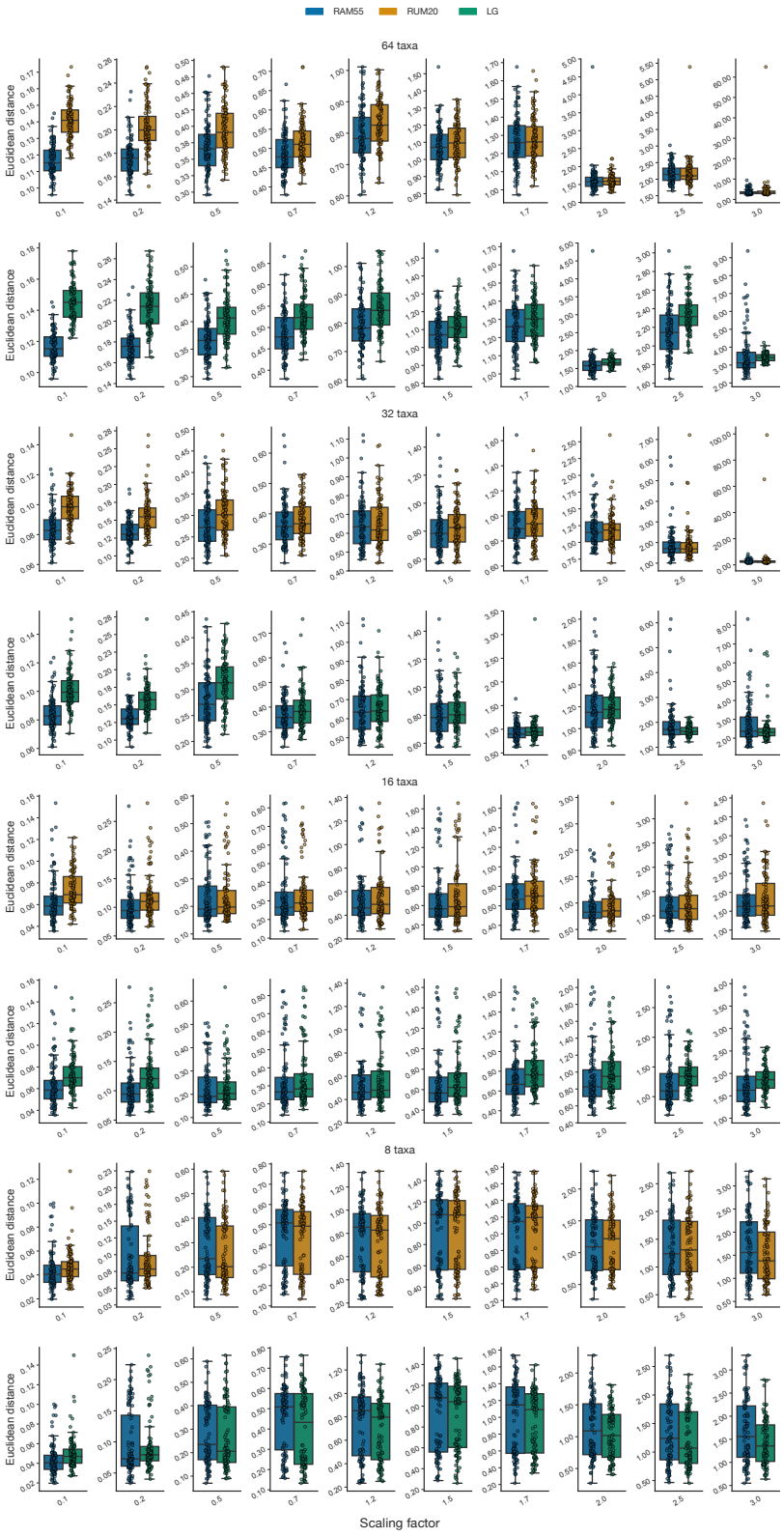

Figure 4.7: Additional inference accuracy analysis using RAM55 under ideal circumstances. Rotasequence alignments (200 sites, 100 replicates per scaling) are simulated under RAM55 and multiple trees, scaled according to the factors on the x-axis. RUM20, LG and RAM55 itself are then used to perform inference over the simulated alignments and the resulting trees are compared to the original phylogeny in terms of Euclidean distance. RAM55 can infer trees that are closer to the original as shown by distance distributions and medians shifted towards lower values.

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 me to infer more-reliable phylogenies from alignments simulated under the 55-state model itself than does any of the 20-state models investigated.

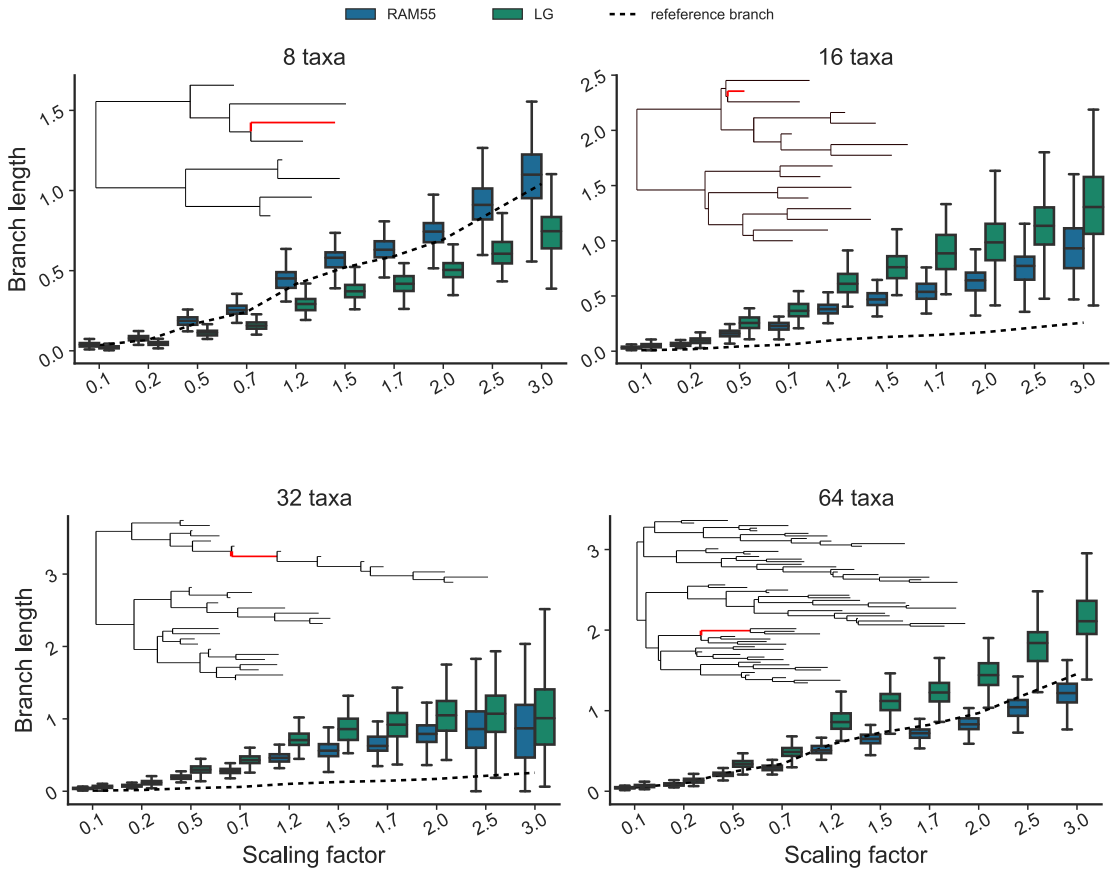

Figure 4.8: Examples of individual branch length inferences illustrate the tendency for RAM55 to give estimates closer to the reference value. Trees inferred using RAM55 (blue) or LG (green), analysing alignments (200 sites, 100 replicates per scaling) simulated using RAM55 and reference phylogenies 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).

4.6 Model benchmarking: Case studies

Having established that RAM55 can result in detectable improvements in model fit and inference accuracy in a simulation scenario, representing an idealised situation of a correctly

Taxa	Scaling factor	RAM55	RUM20	LG	LGexp	LGbyfreq-exp	LG+F
4	0.1	1.00	0.00	0.0	0.0	0.0	0.0
4	0.2	1.00	0.00	0.0	0.0	0.0	0.0
4	0.5	1.00	0.00	0.0	0.0	0.0	0.0
4	0.7	1.00	0.00	0.0	0.0	0.0	0.0
4	1.2	1.00	0.00	0.0	0.0	0.0	0.0
4	1.5	0.96	0.04	0.0	0.0	0.0	0.0
4	1.7	0.99	0.01	0.0	0.0	0.0	0.0
4	2.0	0.96	0.04	0.0	0.0	0.0	0.0
4	2.5	0.93	0.07	0.0	0.0	0.0	0.0
4	3.0	0.77	0.23	0.0	0.0	0.0	0.0
8	0.1	1.00	0.00	0.0	0.0	0.0	0.0
8	0.2	1.00	0.00	0.0	0.0	0.0	0.0
8	0.5	1.00	0.00	0.0	0.0	0.0	0.0
8	0.7	1.00	0.00	0.0	0.0	0.0	0.0
8	1.2	0.99	0.01	0.0	0.0	0.0	0.0
8	1.5	0.99	0.01	0.0	0.0	0.0	0.0
8	1.7	0.99	0.01	0.0	0.0	0.0	0.0
8	2.0	0.96	0.04	0.0	0.0	0.0	0.0
8	2.5	0.86	0.14	0.0	0.0	0.0	0.0
8	3.0	0.89	0.11	0.0	0.0	0.0	0.0
16	0.1	1.00	0.00	0.0	0.0	0.0	0.0
16	0.2	1.00	0.00	0.0	0.0	0.0	0.0
16	0.5	1.00	0.00	0.0	0.0	0.0	0.0
16	0.7	1.00	0.00	0.0	0.0	0.0	0.0
16	1.2	1.00	0.00	0.0	0.0	0.0	0.0
16	1.5	1.00	0.00	0.0	0.0	0.0	0.0
16	1.7	0.95	0.05	0.0	0.0	0.0	0.0
16	2.0	0.95	0.05	0.0	0.0	0.0	0.0
16	2.5	0.84	0.16	0.0	0.0	0.0	0.0
16	3.0	0.80	0.20	0.0	0.0	0.0	0.0
32	0.1	1.00	0.00	0.0	0.0	0.0	0.0
32	0.2	1.00	0.00	0.0	0.0	0.0	0.0
32	0.5	1.00	0.00	0.0	0.0	0.0	0.0
32	0.7	1.00	0.00	0.0	0.0	0.0	0.0
32	1.2	1.00	0.00	0.0	0.0	0.0	0.0
32	1.5	1.00	0.00	0.0	0.0	0.0	0.0
32	1.7	1.00	0.00	0.0	0.0	0.0	0.0
32	2.0	1.00	0.00	0.0	0.0	0.0	0.0
32	2.5	1.00	0.00	0.0	0.0	0.0	0.0
32	3.0	0.98	0.02	0.0	0.0	0.0	0.0
64	0.1	1.00	0.00	0.0	0.0	0.0	0.0
64	0.2	1.00	0.00	0.0	0.0	0.0	0.0
64	0.5	1.00	0.00	0.0	0.0	0.0	0.0
64	0.7	1.00	0.00	0.0	0.0	0.0	0.0
64	1.2	1.00	0.00	0.0	0.0	0.0	0.0
64	1.5	1.00	0.00	0.0	0.0	0.0	0.0
64	1.7	1.00	0.00	0.0	0.0	0.0	0.0
64	2.0	1.00	0.00	0.0	0.0	0.0	0.0
64	2.5	1.00	0.00	0.0	0.0	0.0	0.0
64	3.0	1.00	0.00	0.0	0.0	0.0	0.0

Table 4.1: Best model (AIC) for each category of simulated alignment. 1000-site alignments are simulated under the RAM55 model and various randomly-generated reference phylogenies (4,8,16,32 and 64 taxa; see Fig. 4.3) scaled according to a set of scaling factors. For each phylogeny and scaling factor pair the table reports the proportion of 100 replicates where each model achieves the lowest AIC (or state-corrected AIC, see section 4.3) when compared against all other models.

specified model, I further assessed RAM55's performance on three empirical amino acid alignments for which I can obtain corresponding structural information, and compared goodness-of-fit of inferred phylogenies across models.

Alignments PF00514 and PF07714 correspond to two Pfam family alignments and their structural information from PDBe: β-catenin-like repeat and tyrosine kinase, respectively. Here I followed the procedure outlined in section 2.1 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 recognises as part of that family's domain. The third alignment was obtained by querying Uniprot [173] with the term "rubisco" and manually filtering 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 I do not follow Pfam's definition of a family domain and this results in a longer alignment (46 taxa, 681 sites). While these datasets represent three diverse protein families, and cover a range of sequence lengths and alignment sizes, the application of RAM55 in its current form to empirical analyses is limited by the need to obtain structural information for all proteins in the dataset of interest. In Chapter 5 I will discuss a strategy that mitigates this issue.

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 RUM20, and the 55-state LGbyfreq-exp model which recognises the frequencies of the 55 states but not their structural information content (see section 4.3, 4.2).

Figure 4.9 shows the goodness-of-fit (measured by AIC values, see section 4.3) 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 RAM55 is implemented in RAxML-NG [94], it was also straightforward to incorporate a discrete gamma model of rate heterogeneity [185], maximum likelihood estimation of equilibrium frequencies from the observed data, or both in combination (see section 4.2). 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 datasets, and thus appears to be a valuable model of protein sequence evolution. It is interesting to note that RAM55+G shows slightly better fit than RAM55 across all three datasets, this might indicate that modelling χ_1 configurations captures some, but not all sources of rate variations among sites.

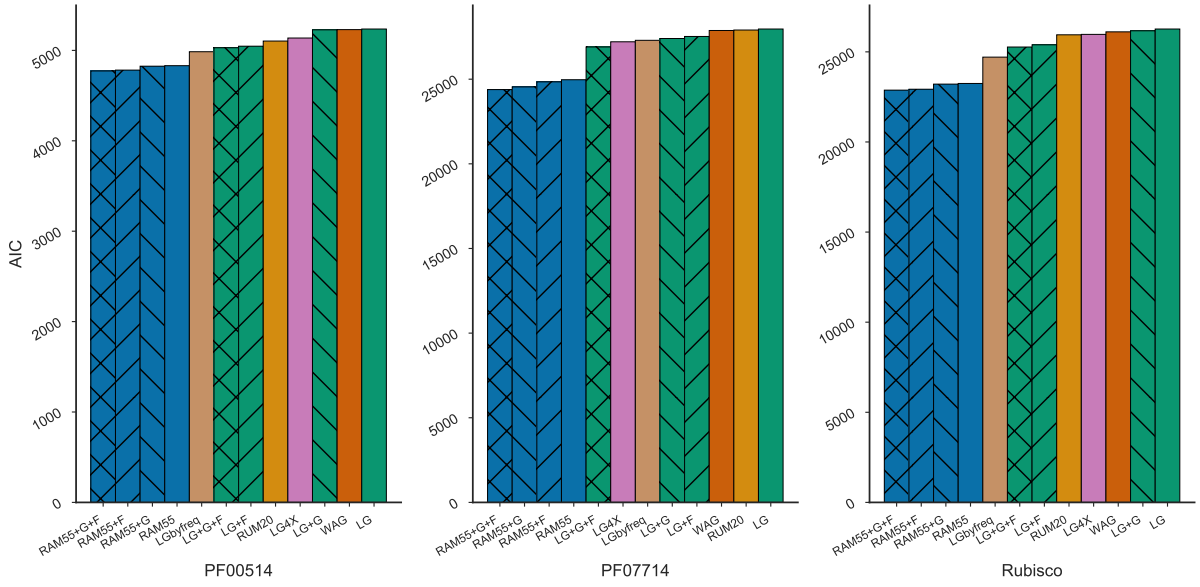

Figure 4.9: A comparison of RAM55 variants against other models in terms of Akaike information criterion (AIC) on three empirical rotasequence alignments: PF00514 (β-catenin-like repeat), PF07714 (tyrosine kinase) and rubisco. '+G' models use a discrete gamma model of rate heterogeneity with 4 categories; '+F' models use ML-estimated state frequencies obtained from the observed data.

4.7 Conclusion

The results shown in this Chapter indicate that RAM55 can produce detectable improvements in model fit and phylogenetic inference accuracy, both on simulated data and on empirical protein sequence alignments which can be mapped to known structures. Interestingly, incorporating a discrete gamma model of rate heterogeneity into RAM55 results in improved model fit for empirical alignments. Here, the 4 replacement rate categories likely capture some of the context-dependent constraints on the replacement process at

a given site; this, in turn, suggests that further improvements might be obtained by replacing the general RAM55 IRM with a small number of context-specific rate matrices, similarly to partitioned models (see section 1.4) and mixture models (see section 1.5). As mentioned in section 3.7, an obvious way of doing so would rely on splitting RAM55 according to solvent exposure (buried, exposed) and secondary structure (α-helix, β-sheet, and H-turn). However, implementing this approach requires either a known structure in order to assign each site to the correct category, or further model complexity and some additional uncertainty the form of implicit (or explicit) structure prediction through an HMM which would infer each site's category from sequence data alone. While similar strategies have been successfully demonstrated for amino acid models (e.g. [166, 60]), an implementation for RAM55 would only suitable for an extended study of its own.

In future studies, it might also be interesting to systematically investigate whether the use of RAM55 for phylogenetic inference benefits mostly 1) analyses of closely-related proteins, where little information is found across many similar (or identical) amino acid sequences, or whether RAM55 proves most useful to 2) analyses of highly-diverged sequences, where conserved χ_1 configurations might retain additional information, even when amino acid sequence identity is low.

Because of its ease of implementation within traditional ML inference frameworks and software, RAM55 has immediate applications beyond inferring phylogenies: these include ancestral sequence reconstruction and side chain configuration prediction for homology modelling as described in the following Chapter. However, one apparent limitation of RAM55 as described so far is its reliance on the availability of high quality structures for all proteins in a dataset of interest in order for the χ_1 configurations to be assigned. In the following Chapter I will outline an approach that mitigates this issue and increases RAM55's real-world viability by encoding missing structural information as ambiguous rotamer states.

5. Other RAM55 applications

Sections 5.1 and 5.2 have been adapted and expanded from a first-author publication [134]: all data collection and analysis for the results shown these sections were performed by me, the original Perron et al. [134] manuscript was written by me, edited collectively, and agreed upon by all co-authors. Sections 5.3 and 5.4 have been adapted and expanded from a co-authored publication [179]: all data collection and analysis for the results shown in sections 5.3, 5.4 was performed by me, the corresponding sections of the Weber et al. [179] manuscript were written by me, edited collectively, and agreed upon by all co-authors.

RAM55 contains useful, biochemically plausible information (see Chapters 2 and 3) and is a valuable model of protein sequence evolution which can improve phylogenetic inference accuracy and model fit over traditional 20-state models (see Chapter 4). In this Chapter I assess 1) whether RAM55 can be applied to ancestral sequence reconstruction (ASR), another standard phylogenetic problem; 2) whether RAM55 can help with side chain configuration inference for homology modelling; and 3) whether each of these approaches is still viable when some of the structural information is missing.

5.1 Ancestral amino acid reconstruction

ASR is a technique that employs sequence (and structure) information of extant proteins and a model of the evolutionary process to infer ancestral protein sequences and investigate their structural and functional properties [167, 119, 66]. Having established RAM55's ability to infer reliable phylogenies when structural information is available (see Chapter 4), here I describe the approach we followed in Perron et al. [134] to evaluate whether

it can be used for further inference tasks. First, I illustrate the reconstruction of ancestral amino acid states, which is a task that can also be achieved under standard 20-state substitution models in which side-chain configurations are not modelled. These models provides a reliable benchmark against which I can evaluate RAM55's ASR performance.

Phylogenetic inference algorithms compute the likelihood of a given tree by integrating over the ancestral states at internal nodes (see section 4.2), however, in a ML framework it is also possible to infer what these ancestral (and thus un-observed) states are, given the observed states at terminal nodes, a tree topology, and a model of the evolutionary process. There are two broad categories of approaches to the problem of ASR. 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 maximises the likelihood of the observed data at that site [186, 187]. In Perron et al. [134] and Weber et al. [179], we used both the marginal reconstruction algorithm [186] and Pupko *et al.*'s version of the joint reconstruction algorithm [137] to infer ancestral rotamer and amino acid sequences. For both these algorithms, I developed efficient and scalable python implementations, opportunely adapted to fit RAM55's expanded state-space, which I deployed on EMBL-EBI's HPC infrastructure with up to 10^4 simulation and inference processes running in parallel (see Code and Data Availability).

To test whether RAM55 allowed us to correctly infer unobserved ancestral amino acid states starting from data simulated under the model itself, I simulated rotasequence alignments under RAM55 (see section 4.4) using trees with fixed topology (see Fig. 4.3) 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 sequences were compared position-by-position against the simulated reference sequences and the results reported in terms of percent sequence

identity (percent correct inference). When RAM55 is used to reconstruct ancestral amino acids the model outputs a rotasequence which is then masked and compared to the reference, which also consists of a masked rotasequence and is similarly compared to the amino acid sequence inferred under LG.

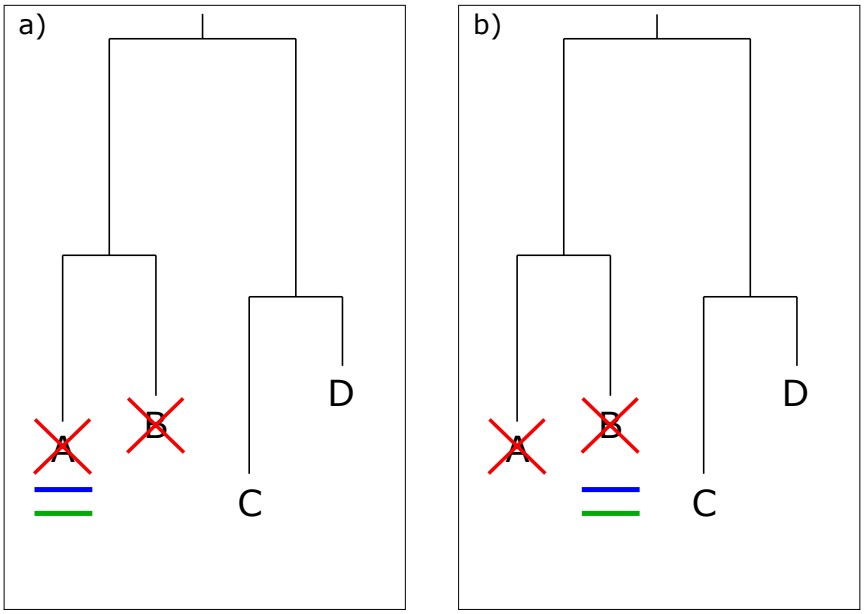

Figure 5.1: Illustration of the "leave-leaves-out" (LLO) procedure. LLO is used to evaluate a model's ability to reconstruct terminal nodes' states. **(a)** A pair of sibling terminal nodes (A, B) is selected, their sequences are removed from the alignment (red ×) and then all internal node sequences are reconstructed along with A's sequence (in blue). This can be compared to its original sequence (in green). **(b)** The analogous procedure is then followed to reconstruct B's sequence and compare it to the original.

In Perron et al. [134] we also investigated RAM55's performance when reconstructing ancestral amino acid sequences from empirical rotasequences. Ideally this would be performed by comparison of inferred ancestral rotasequences with known ancestral structures. While an increasing number of resurrected ancestral protein structures have been resolved (e.g. [88, 78, 69, 141, 26, 55]), 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 the reconstructed rotasequences with reference ancestral rotasequences obtained from deposited structures. To overcome this, I employed a "leave-leaves-out" (LLO) approach (Fig. 5.1) in which I 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 neighbour (B) 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: the assumption is that the ability to infer terminal sequences (with no information from the nearest sibling) implies a similar ability to reconstruct internal sequences. Here, I first validate LLO on simulated data, to verify that model performance when reconstructing under LLO does not differ drastically from the accuracy achieved when reconstructing internal nodes.

Figure 5.2 shows that RAM55 performs equally or better than LG in terms of of amino acid state reconstruction accuracy using the joint algorithm, particularly at longer evolutionary distances (see also Fig. 5.3). Very similar results are achieved using the marginal reconstruction approach (see Fig. 5.4,5.3). Both terminal amino acid sequences (node A) and internal sequences (nodes B, C) are reconstructed, to evaluate model performance under LLO. Figure 5.2 illustrates that estimating terminal sequences can be as hard, if not harder, than it is reconstructing internal sequences. This is likely due to removing the nearest neighbour sequence, which "isolates" these nodes on the tree more than many internal nodes, thus depriving the model of the most closely-related, and likely most useful, sequence. Further, sequence reconstruction accuracy varies significantly across nodes with a node's distance from the observed sequences at the tips (or its distance from the root node) being possibly among the primary causes of this variance, as illustrated in Figure 5.3. However, the topological context of each node is also likely to play a role in determining reconstruction accuracy and future studies might develop more sophisticated metrics for assessing ASR complexity at any given node. These results suggest that LLO is a viable approach to evaluate reconstruction accuracy on empirical alignments where no reference sequence is available for internal nodes.

In Table 5.1 I reconstructed ancestral amino acid sequences under RAM55 or LG starting from extant empirical rotasequences (masked for LG) from the β-catenin-like repeat alignment (PF00514, as in Fig. 4.9). The LLO approach was used for each of the 13 terminal sequences, which were reconstructed both with the marginal and joint

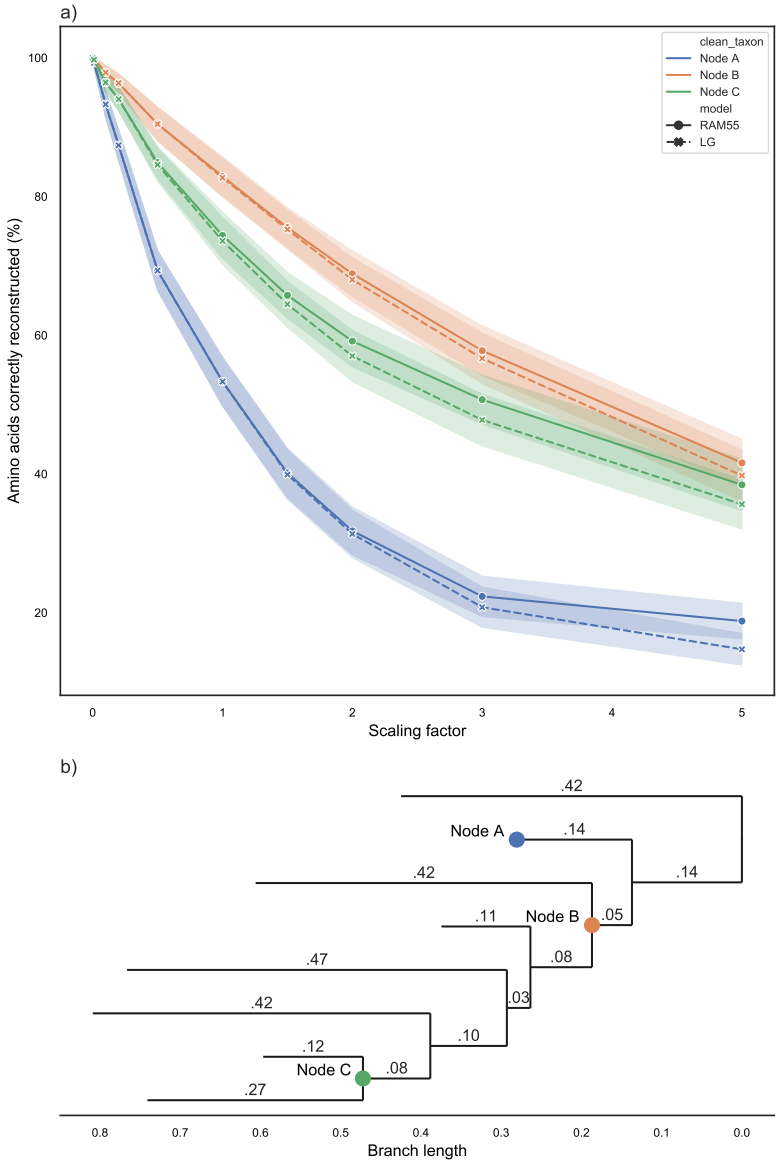

Figure 5.2: Ancestral amino acid joint reconstruction accuracy. Amino acid states are inferred using joint reconstruction from rotasequence alignments (200 sites) simulated under RAM55 using an 8-taxon reference phylogeny which is scaled according to the factors reported on the x-axis. The joint reconstruction algorithm is then employed along with RAM55 (continuous lines) and the true phylogeny to reconstruct rotasequences at various internal (B, C) or terminal (A) nodes, the latter using the LLO procedure (see Fig. 5.1). The same procedure is repeated using LG on masked alignments (dashed lines). The y-axis indicates the percentage of amino acid states correctly reconstructed for each inferred sequence (accuracy). Each data point corresponds to the mean accuracy of 100 simulation replicates for a given taxon. Shaded areas indicate the accuracy standard deviation.

algorithms using the phylogeny inferred using RAxML-NG and RAM55 to guide the reconstruction process. Similarly to what happens with simulated data, RAM55 and LG

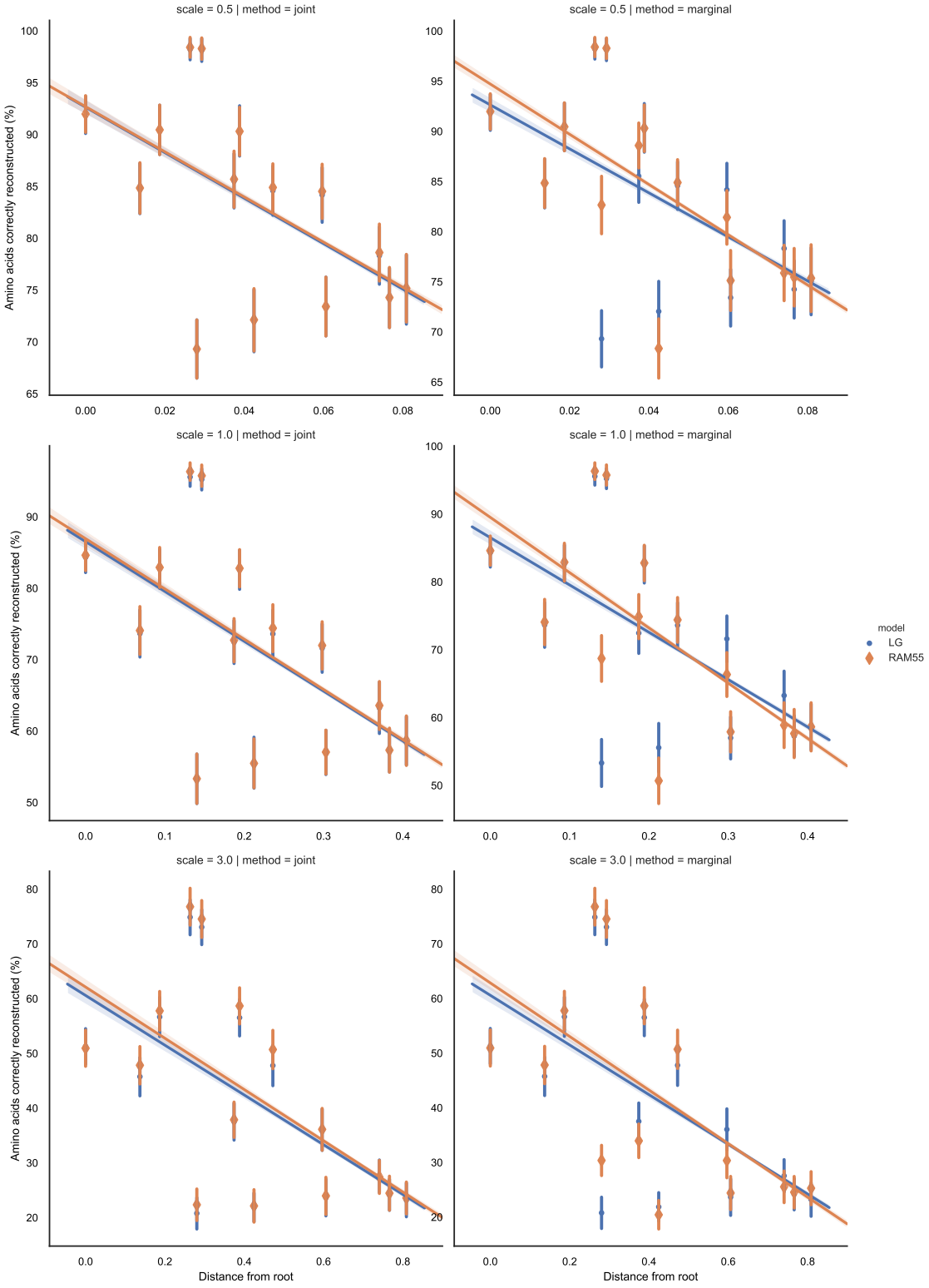

Figure 5.3: Amino acid reconstruction accuracy and distance from root. Amino acid states are reconstructed using both joint and marginal algorithms from simulated data (same as in Fig. 5.2) under RAM55 and LG for all internal and terminal nodes, the latter using the LLO procedure (see Fig. 5.1). Each data point indicates (on the y-axis) the mean percentage of amino acid states correctly reconstructed (accuracy) from 100 replicates for a given node. The x-axis report each node's distance from the root node; the error bars indicate the accuracy standard deviation.

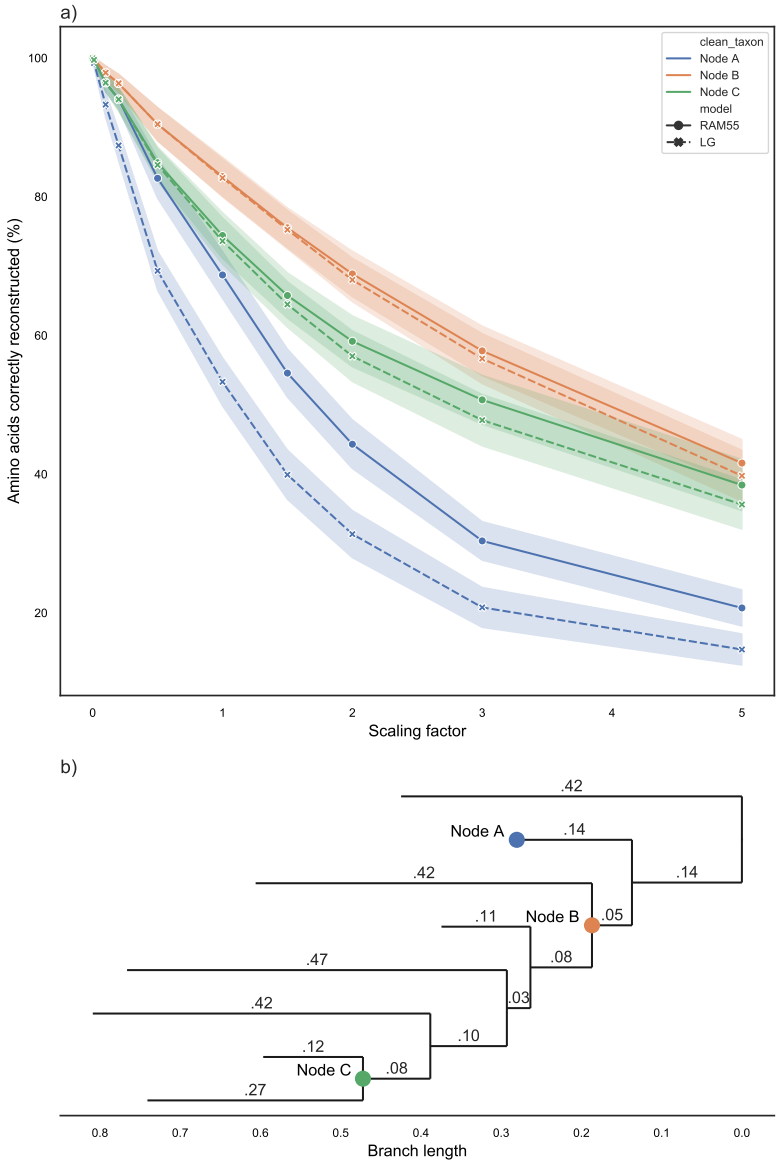

Figure 5.4: Ancestral amino acid marginal reconstruction accuracy. Amino acid states are inferred using marginal reconstruction from simulated rotasequence (same as in Fig. 5.2) under RAM55 (continuous lines) and LG (dashed lines) for various internal (B, C) or terminal (A) nodes, the latter using the LLO procedure (see Fig. 5.1). The y-axis indicates the percentage of amino acid states correctly reconstructed for each inferred sequence (accuracy). Each data point corresponds to the mean accuracy of 100 simulation replicates for a given taxon. Shaded areas indicate the accuracy standard deviation.

are closely matched in amino acid reconstruction accuracy, and appear to perform equally well under both joint and marginal algorithms. These results indicate 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 sequences as well

as a 20 state model, provided that the extant proteins can be mapped to known structures. While accurate ancestral amino acid reconstructions can benefit protein resurrection studies, RAM55 can also infer information about ancestral side-chain conformation by reconstructing ancestral rotamer states, which is not possible by any other method.

Taxa	RAM55 joint	LG joint	RAM55 marginal	LG marginal
Q71VM4/147-186	75.15%	76.05%	72.75%	72.16%
P52292/410-447	76.35%	75.75%	73.05%	73.05%
P52294/158-197	80.84%	80.24%	76.95%	76.35%
P52293/251-279	81.14%	80.84%	78.44%	77.55%
Q9C2K9/158-194	89.22%	89.82%	81.74%	81.44%
O60684/113-154	88.92%	88.32%	85.03%	85.03%
P35222/229-262	89.22%	89.22%	89.52%	89.22%
P25054/649-689	88.92%	88.32%	89.22%	89.52%
Q02821/372-411	97.01%	97.01%	97.31%	96.71%
Q02248/229-262	96.41%	96.41%	97.01%	96.41%
Q99959/386-423	95.51%	95.51%	95.81%	95.81%
O60716/440-481	96.41%	96.41%	97.31%	97.31%
O44326/361-402	99.40%	99.10%	99.40%	99.40%

Table 5.1: Empirical amino acid reconstruction accuracy. Amino acid state reconstruction from the PF00514/β-catenin-like repeat alignment using RAM55 or LG and either the joint or marginal reconstruction algorithms. Scores represent the percentage of sites correctly reconstructed. Reconstruction was limited to terminal nodes using the LLO approach, due to lack of a reference for internal sequences.

5.2 Ancestral rotamer state reconstruction

The ASR approach described in the previous section allows RAM55 to reconstruct ancestral rotamer states starting from an alignment of extant proteins of known structures. When assessing RAM55's rotamer state reconstruction accuracy, it is harder to find an adequate benchmark. No other model of rotamer state replacement is currently available and, as described above, resurrected structures are of little help for this purpose due to their rarity and difficulty in mapping to datasets of extant sequences and structures. I thus evaluate RAM55's accuracy when jointly reconstructing ancestral rotamer states from data simulated under the model itself and an 8-taxa phylogeny. Figure 5.5 shows that RAM55 is able to infer the correct ancestral rotamer state for residues belonging to internal and terminal sequences with an accuracy that closely matches its amino acid reconstruction performance in Figure 5.2. This suggests that the model is able to infer

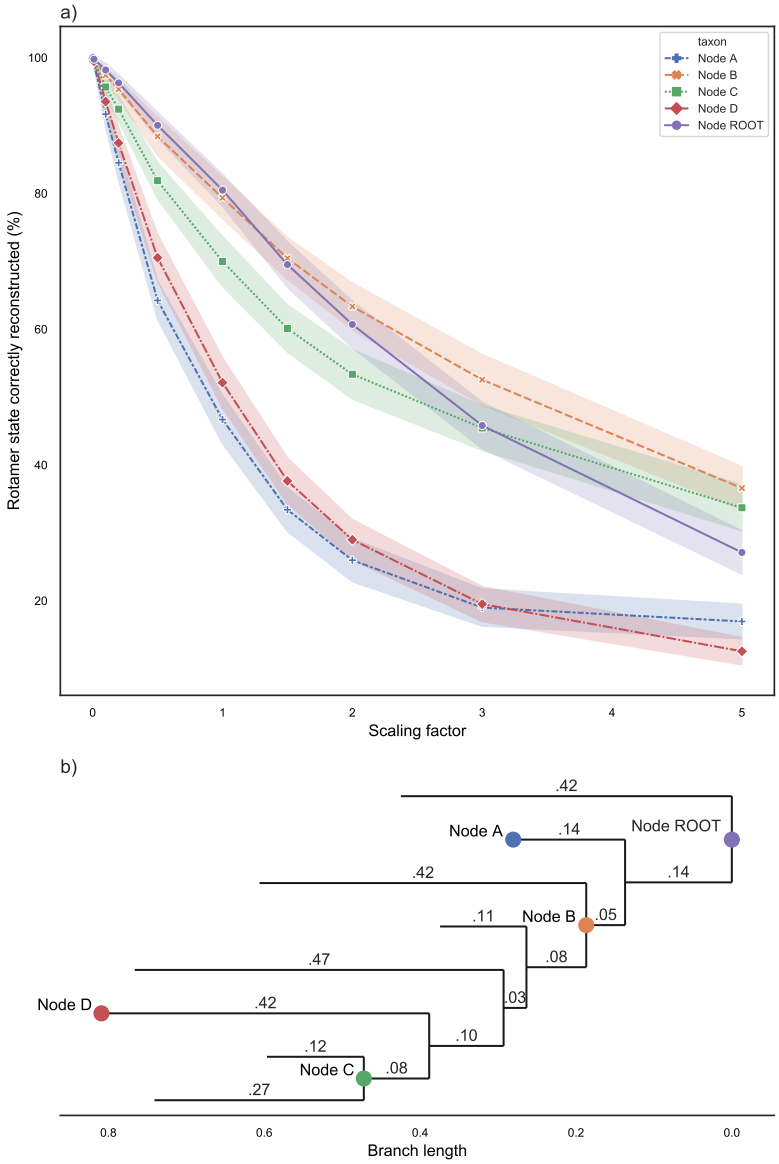

Figure 5.5: Rotamer state reconstruction accuracy. Rotamer states inferred using joint reconstruction from rotasequences (200 sites) simulated under RAM55 using an 8-taxon reference phylogeny and scaling its branches according to the factors reported on the x-axis The joint reconstruction algorithm is then employed along with RAM55 and the true phylogeny to reconstruct rotasequences at various internal (ROOT, B, C) or terminal (A, D) nodes, the latter using the LLO procedure (see Fig. 5.1). The y-axis indicates the percentage of rotamer states correctly reconstructed for each inferred sequence. Each data point corresponds to the mean accuracy of 100 simulation replicates for a given taxon. Shaded areas indicate the accuracy standard deviation.

the correct χ_1 configuration in almost all cases when the amino acid is accurately reconstructed. As in Figure 5.2, terminal node sequences are reconstructed using the LLO approach: this appears to pose a comparable, if not harder, challenge than reconstructing

internal sequences. Similar results are obtained for other simulated alignment lengths and numbers of taxa in reference phylogenies, as well as the margin reconstruction algorithm (data not shown). These reconstructed ancestral rotamer states could be used to predict side-chain geometry for homology modelling of ancestral proteins, to assess which configuration better fits the evolutionary data. I further explore this application in section 5.4.

5.3 Using ambiguity on mixed data

In a real-world scenario, we could be interested in a specific set of protein sequences, some of which might not map to any known structure. Applying RAM55 as implemented so far in this context would require discarding those amino acid sequences lacking structural information, as these cannot be expressed in the 55-state alphabet RAM55 currently uses. Is there a way to utilise a model with a larger state-space on sequences belonging to a smaller state space contained within it?

In Weber et al. [179], we considered cases where the characters at the tips of a phylogenetic tree are only available in an aggregated state-space A with m states, rather than a larger state space S that separates characters into n distinct states, where $m < n$. Each state a_i in $A = \{a_1, ..., a_m\}$ corresponds to one or more states s_j in $S = \{s_1, ..., s_n\}$, while each state in S maps to a single state in A. This scenario corresponds, for example, to the relationship between the amino acid state space and the codon state space where each amino acid corresponds to a unique set of codon states and each codon maps to a unique amino acid. For RAM55, each amino acid state (A) can represent multiple rotamer states (S) and, conversely, each rotamer state corresponds to a single amino acid. If only amino acid sequences, rather than rotamer sequences, are accessible and modelling the data in S might be informative, we can take advantage of these mappings.

Treating data observed in A as ambiguous states in S is similar to the “covariotide" models, conceptualised by Fitch and Markowitz [51], where each nucleotide may be in an “on” or “off” state that cannot be directly observed [77, 172, 56]. Ambiguity has also

been used to encode population allele frequencies using small samples as input [35], and to handle sequence uncertainty [95].

In order to estimate phylogenetic models under maximum likelihood when the data do not match the model state-space, in Weber et al. [179] we modified an established method for handling alignment gaps and ambiguous characters. Where a column of sequence data is missing for one or more tip nodes k, the convention is to set the entries for each possible state x_j in the conditional probability vector $L_k(x_j)$ to 1 (see [47] p. 225). This correctly represents the case where all states have equal probabilities of being observed at this node in the alignment given the true state of the model: $L_k(x_j)$ holds the probability of observing the data observed at the terminal nodes descendant from node k, conditional on the model and the state at node x_j.

In a case where the observed data are in the aggregated state space A, and we are interested in modelling in the separate state space S, we can proceed in a similar manner and perform inference using a model in S since there is an unique mapping from each state in A onto a set of states in S. For ambiguity at terminal node k under RAM55, for any rotamer state (A, R), the probability of observing x_j given (true) state (A, R) is 1 if x_j=A and 0 otherwise, independent of R. Hence, our observation in space A (e.g. TRP) is ambiguous with respect to the character in state space S (e.g. {TRP1, TRP2, TRP3}). In the following, the term "ambiguous" is used to refer to instances where incomplete information about the state at a given site is available but the character is not missing, as opposed to less strict uses of this term in the past where its definition covered missing data as well. Here, where data are completely absent for an alignment position, the same vector is encoded by $L_k(x_j) = \{1, ..., 1\}$. Once L_k has been set according to this modification, the calculation of the likelihood proceeds as normal following Felsenstein's pruning algorithm [49] as described in section 4.2.

To generate sequences in rotamer space with known phylogenies and ancestral states, in Weber et al. [179] I used a set-up similar to the one described in Chapter 4. Briefly, I randomly generated trees (32-taxon) using a Yule process and scaled the branches according to a scaling factor ($[0.1, 1]$). I then performed a continuous-time Markov chain

simulation along the branches for 1000 replicates under RAM55. To emulate cases where structural information in not available for some of the terminal nodes, I generated mixed alignments by "masking" (i.e. removing the χ_1 configuration information) a specific proportion of the terminal sequences, leaving only amino acids. The choice of which terminal sequences are to be removed requires some consideration as well, since some might be more informative than others because of their sequence and their topological context. For this analysis I randomly and independently select a set of terminal sequences for each simulation replicate. The size of this set varies according to a predetermined masking factor which represents the proportion of the tips which is to be masked. Amino acid states are then treated as ambiguous in the inference step (see below). Further, sequences can be "discarded" from the simulated alignment to illustrate the loss of information caused by removing all sequences where no structure is available. This is done by replacing a specific proportion of the alignment sequences with gap characters using the same randomisation strategy described above.

To reconstruct ancestral rotamer states (see section 5.2) I followed the approach described earlier in this section, encoding each amino acid (state-space A) observed in the alignment as ambiguous in rotamer space (state-space S) in the conditional probability vector. This procedure allows us to use the RAM55 model to obtain rotamer sequences at internal nodes using a mixture of amino acid and rotamer sequences, and the tree used for simulation as input. To be able to obtain posterior probabilities for reconstructions, I then apply my own implementation of the marginal reconstruction algorithm of Koshi and Goldstein [91] (see Code and Data Availability). Joint reconstruction gives qualitatively similar results [137]. To assess the accuracy of the reconstructions, I examined the proportion of sites with matching characters. For example, if both the simulated and reconstructed state at a given site is R1 (Arginine in configuration 1), a match is recorded. Meanwhile, if any other amino acid or any other configuration of R (R2 or R3) is reconstructed, this is recorded as a mismatch.

In Weber et al. [179], we opted to benchmark the model using reconstruction accuracy rather than branch lengths, as ancestral side-chain configurations represent an output that would be otherwise unobtainable. Varying the proportion of masked sites in the alignment

allows us to compare scenarios where structures are available for some of the sequences of interest, or none at all, similar to what would be observed for real empirical data. However, it is equally feasible to input sequences that are masked on a per-site basis into our ASR approach. This might correspond to a scenario where a protein structure has only been partially resolved –because of flexible loops or subunits that are particularly hard to crystallise– or one where the available structure only partially satisfies the quality filter. The reconstruction accuracy for the data where rotamer information is available for all of the tips provides a benchmark for the performance of ambiguity coding.

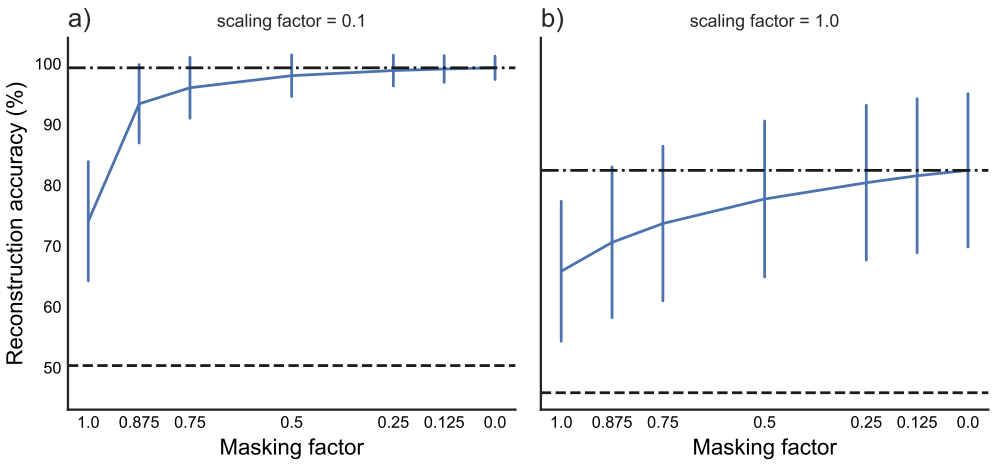

Figure 5.6: Accuracy of ancestral rotamer sequence reconstruction from mixed data under RAM55 increases and shows lower variance when less rotamer configuration data is removed. The x-axis shows the fraction of χ_1 configuration data removed (i.e. masked). The vertical bars show the standard deviation of the reconstruction accuracies, centred around the median. The black dash-dot lines represent the maximum accuracy reached on full (unmasked) alignments; black dashed lines show the accuracy achieved by reconstructing the amino acids under RUM20 and randomly assigning ('guessing') the χ_1 configuration. **(a)** results when all branches of the tree are multiplied by 0.1, showing greater overall accuracy in this case. **(b)** results for the standard simulation tree.

Figure 5.6 illustrates a relatively modest reduction in overall rotamer state reconstruction accuracy between simulations where χ_1 configurations are known for all taxa, and simulations where this information is not available for any of the taxa (∼15% difference for the unscaled tree, Figure 5.6b). Reconstruction under RAM55 using only amino acid sequences is markedly more accurate than the only alternative approaches of using a conventional empirical amino acid model to reconstruct the protein sequence and randomly assigning ("guessing") rotamer states (Fig. 5.6, dashed line), or assigning them based on

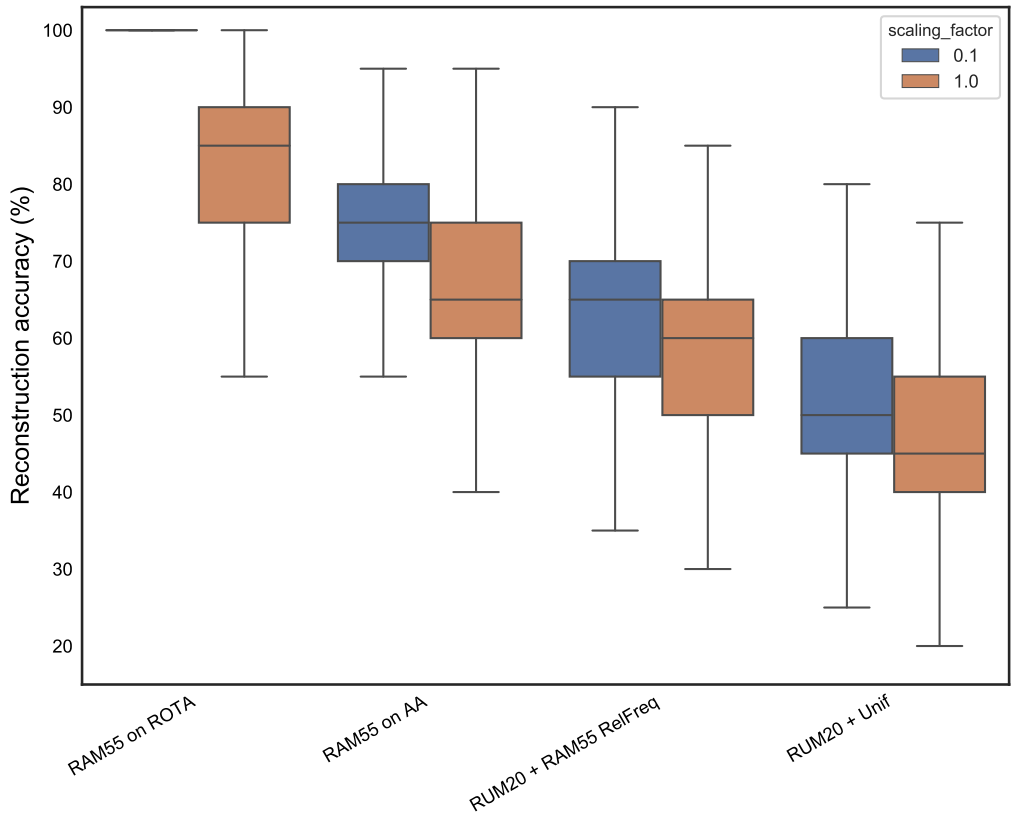

Figure 5.7: Comparison of ancestral rotamer state reconstruction under RAM55 against random assignment. Ancestral rotamer states are reconstructed from simulated data (same as in Fig. 5.6). "RUM + Unif" refers to amino acid sequence reconstructions where χ_1 configurations were assigned at random (based on the assumption of equal frequencies), while "RUM20 + RAM55 RelFreq" were assigned χ_1 configurations based on the equilibrium frequencies of RAM55. These assignments all fare worse than reconstruction under RAM55 from either amino acid ("RAM55 on AA") or rotamer states ("RAM55 on ROTA"), indicating that the information retained is attributable to the model. For the first panel the "RAM55 on ROTA" reconstructions are correct across all sites and replicates.

the equilibrium frequencies of the RAM55 model (see Fig. 5.7). Hence, it is advantageous to reconstruct under the rotamer-aware model, even when the input data are only available in the aggregated state-space.

As expected, the overall accuracy depends on how difficult the ancestral sequence reconstruction problem is, with a shallower tree showing higher sequence identity overall between simulated and reconstructed characters (Fig. 5.6a). For this tree, the ancestral sequences are correctly inferred in almost all replicates for smaller masking factors ($[0-0.5]$), and, remarkably, average reconstruction accuracy remains above 75% even when all χ_1 information is masked. The taller tree (Fig. 5.6b) represents a harder problem with

an average maximum accuracy of ∼80%, however, accuracy appears to decrease more slowly as χ_1 configuration data is removed. This might be explained by the fact that this taller tree contains more substitutions, which might be contribute more information to the model and counter the loss of χ_1 information to some extent. It is also interesting to notice that the greatest increase in performance appears between alignments with no rotamer configuration information present and those with 12.5% of sequences containing that information (Fig. 5.6). This suggests that little structural information is required in order to achieve ancestral reconstruction of rotamer states with acceptable accuracy. Intuitively, the fraction of correctly inferred states declines with increasing distance from the tips of the tree (Spearman's rank correlation coefficient -0.537, $p < 0.001$; details not shown).

Here, I also considered how the certainty with which the model assigns the correct ancestral state responds to χ_1 configuration data being masked at the tips of the tree. When using the marginal reconstruction approach, this certainty is described by the posterior probability value of the correct prediction. Unsurprisingly, the posterior probability for the correct state declines as data is removed (see Fig. 5.8, orange markers); thus we observe a drop in the certainty of the reconstruction preceding the drop in accuracy. Interestingly, when a corresponding proportion of the sequences is discarded (i.e. all characters are replaced by gaps) rather than masked (see Fig. 5.8, red markers), model certainty appears to drop even faster. This suggests that a similar relationship might exist for reconstruction accuracy as well (see below).

To examine the robustness of rotamer state reconstruction from ambiguous data, I also assessed ASR accuracy under a simple model violation scenario, simulating data under RAM55 with gamma-distributed rates [188] and reconstructing under RAM55 without rate heterogeneity. When rotamer configuration is masked, we observe a larger decline in accuracy compared to a scenario with no violation, which is expected given that the amino acid sequence contains less signal (see Fig. 5.9).

Inferring ancestral rotamer states from ambiguous amino acid sequences appears to be a viable, robust strategy resulting in surprisingly good accuracy even when little or no

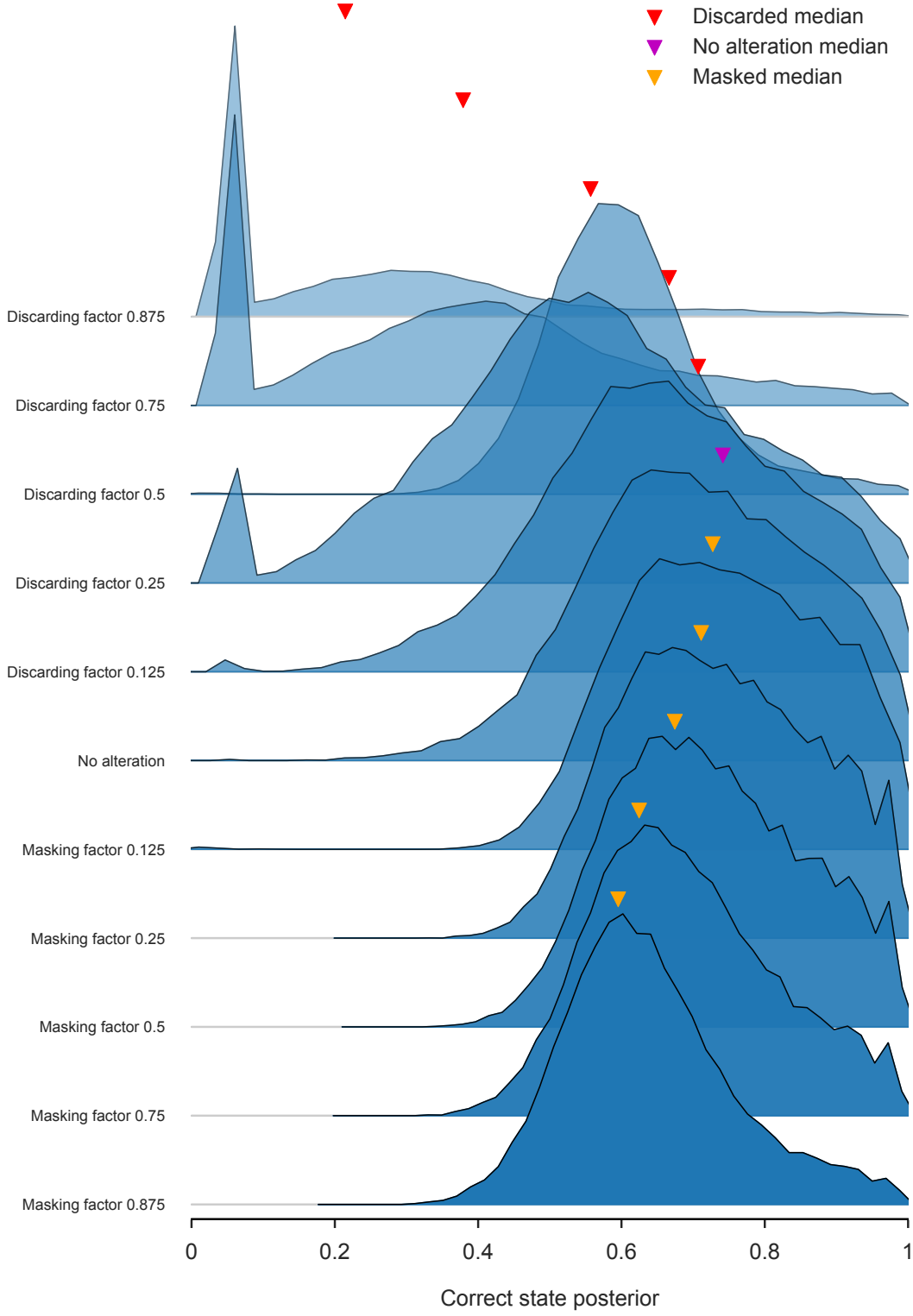

Figure 5.8: Model certainty for ancestral rotamer state reconstruction on masked or discarded sequences. When performing ASR (same as in Fig. 5.6b), masking any proportion (x-axis) of rotasequences (orange markers), rather than discarding them (red markers) results in higher posterior probabilities assigned by RAM55 to the correct ancestral rotamer state. As more data is removed, decrease in certainty precedes the decrease in accuracy (see Fig. 5.6, 5.10). Each subplot reports the kernel density estimation for the posterior probability of the correct ancestral state across all sites in the ancestral sequence for 1000 replicates. Medians for each distribution are highlighted by triangle markers. "No alteration" refers to the condition where no rotasequence is masked or removed.

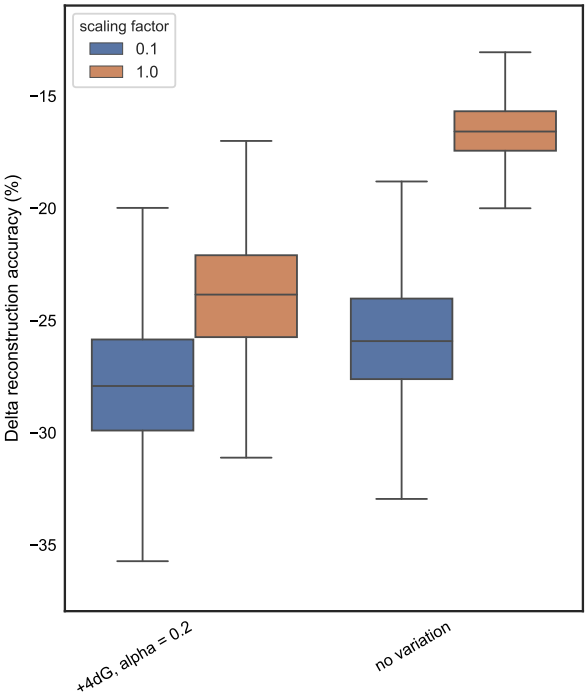

Figure 5.9: Ancestral rotamer state reconstruction on masked data under RAM55 given model violation. Delta reconstruction accuracy is computed by subtracting the accuracy obtained when reconstructing on full (i.e. unmasked) rotasequences from that obtained on masked amino acid sequences (i.e. fully masked), under RAM55. Results labelled "no variation' are computed from the accuracies reported in Fig. 5.6. Results labelled "+4dG, alpha = 0.2" correspond to the delta accuracy obtained simulating under RAM55 plus a 4-category gamma distribution with $\alpha = 0.2$, and reconstructing under RAM55 without rate variation. Results where all branches of the tree are multiplied by 0.1 (in blue) show a modest decline in accuracy ($\sim$2% larger) with rate heterogeneity. Results for the standard simulation tree (in orange) show a larger decline ($\sim$7% larger) in the model violation scenario. Model violations produce noisier estimates and exacerbate the drop in accuracy when reconstructing on masked sequences.

structural data is available. In a realistic scenario this approach might allow us to include proteins that cannot be mapped to a known structure (or one of sufficient quality) into our analysis, rather than discarding them. Can we quantify the improvement in reconstruction accuracy that might be produced by including this additional ambiguous data?

I compared the accuracy of rotamer state reconstructions under two scenarios: 1) all rotamer state information (amino acid and χ_1 configuration) is discarded from a proportion of sequences, and 2) masking is used so that amino acid, but not rotamer, sequences are available for a proportion of the alignment. This provides a measure of the advantage gained by considering additional amino acid sequences where no structural information

is available. For the unscaled tree, masking 50% of the χ_1 configurations produces ancestral reconstructions that are comparable in accuracy to trees where 12.5% of terminal sequences are replaced by gaps (Fig. 5.10b), indicating a noticeable advantage for including amino acid sequences where full rotamer state information is unavailable, rather than discarding them. In other words, augmenting half of the amino acid sequences with χ_1 configuration data is approximately as informative as having 87.5% of the full rotamer state information and completely discarding the remaining 12.5% of the data. Further, removing all rotamer information and reconstructing with ambiguity is equivalent to retaining 50% of the original information.

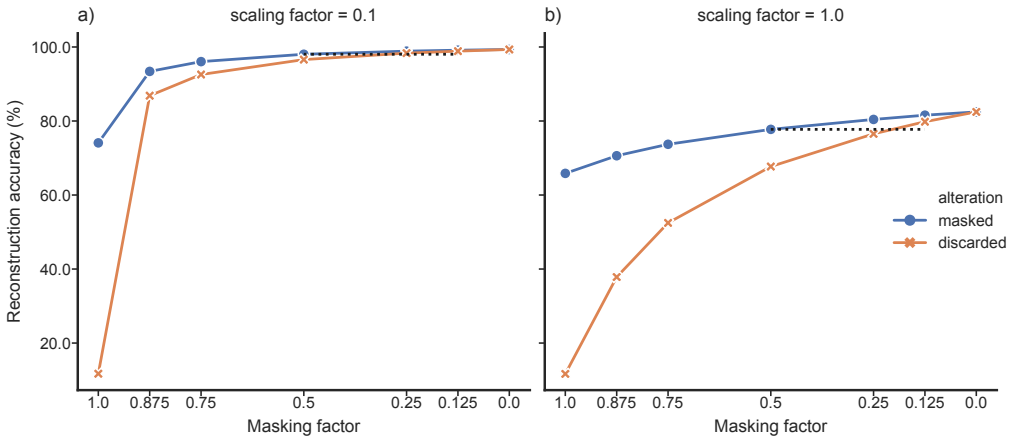

Figure 5.10: Ancestral rotamer reconstruction accuracy on masked or discarded sequences. The x-axis shows the fraction of sequence data removed under two scenarios. The blue circles reflect the the amount of χ_1 configuration data that has been masked (i.e. replaced with amino acids), and the orange crosses represent the amount of sequences that has been replaced with gaps (discarded). Masking half of the χ_1 configurations produces rotamer state reconstruction accuracies comparable with those obtained by replacing 1/8−1/4 of sequences in the alignments with gaps (see black dashed lines). As in Figure 5.6, the shallower tree, **(a)**, shows higher overall accuracy. Masking rotasequences retains useful information for ASR that is lost by discarding them.

These results suggest that the ambiguity coding approach described in this section is a novel and valuable strategy that allows us to perform accurate inference from mixed input data. Specifically, it is surprisingly useful for extending the potential applications of RAM55 to many real-world scenarios (such as the one I illustrate below) where can it be very beneficial to consider amino acid sequences that lack structures.

5.4 Side-chain predictions for homology modelling

Prediction of side-chain conformations is an important part of protein structure and interaction modelling. When possible, configuration predictions for a target protein are based on the side chain orientation observed in a closely related template protein of known structure. Thus, inferring side-chain configurations for a protein where no such template is available or for target residues that are not conserved in the template, represents an additional challenge and computational burden [178]. In previous sections I have described how ASR algorithms under RAM55 can be modified to infer terminal node sequences, and how they can be augmented by ambiguity coding to accurately infer χ_1 configurations even when some (or much) of the structural information is missing from the alignment of interest. By combining these approaches, the ASR strategy can be put to practical use in the context of modelling side-chain configurations for extant target proteins. Its evolutionarily-backed predictions could benefit homology modelling strategies by providing a sensible starting point which might speed up the side-chain configuration refinement process.

The task of inferring a target protein's χ_1 configurations from a dataset of homologous sequences and structures is quite similar to the LLO procedure described in section 5.2. One important difference is that here the target's amino acid sequence is known, and thus the rotamer state inference is constrained to the observed amino acid at any given position. To investigate whether RAM55 can accurately predict χ_1 configurations in a realistic scenario, in Weber et al. [179] I consider two manually curated empirical datasets consisting of 16 ADK structures and 30 RuBisCO structures from PDBe [28], respectively. For each dataset, a multiple amino acid sequence alignment is generated using MAFFT [82]. A rotamer configuration is then assigned to each amino acid in the alignment based on structural information in PDBe, creating a rotasequence alignment (see section 2.1). The tree for the reconstruction is estimated from the rotamer sequence alignment using RAxML-NG under RAM55 ([94], see section 4.2). I then mask, in turn, each terminal rotamer sequence in the alignment and predict each amino acid's χ_1 configuration using RAM55 and the marginal reconstruction algorithm as discussed previously, constrained to the observed amino acid. Prediction accuracy can then be computed against the original

rotamer sequence in terms of the percentage of rotamer states correctly reconstructed. In order to benchmark RAM55's accuracy I first establish a baseline accuracy by assigning the χ_1 configuration according to a uniform probability distribution (rand) and the relative equilibrium frequency of each possible configuration according to RAM55 (rel freq).

A widely used strategy to predict side-chain configurations in unresolved structures consists in assigning to each amino acid the same configuration found at the corresponding site in the nearest neighbour's structure [162, 178]; here I refer to this approach as "Nearest Neighbour Configuration" (NNC). NNC is only applicable to sites where the amino acid is conserved in the template, and so my implementation of NNC falls back to a "rel freq" strategy for non-conserved sites. I also evaluated a scenario where no structural information is available for the nearest homolog sequence (Masked Nearest Neighbour, MNN). Here, RAM55 can process all mixed data using the ambiguity coding. As a baseline for this scenario, I apply the NNC approach as described previously, here relying on the second-closest structure instead.

As a further baseline, I also predict the side chain configurations starting from the target protein's backbone structure using SCWR4 [96]. This software uses the same backbone-dependent rotamer library as RAM55 [155], a simple energy function based on the library rotamer frequencies, short-range soft van der Waals interaction potentials, and a graph decomposition algorithm to solve the combinatorial side chain packing problem. The SCWR4 approach differs significantly from all other methods in this analysis since it 1) requires the target protein's backbone structure, rather than its amino acid sequence, to be known, and since 2) it does not rely on any evolutionary information from closely related sequences or structures (as opposed to RAM55, NNC, and MNN).

Predicting χ_1 side chain configurations using RAM55 is more accurate ($\sim$11% median improvement for both datasets) than NNC when the nearest neighbour's structure is available (Fig. 5.11). Further, RAM55 can make use of all the available rotamer sequence information, as well as the nearest neighbour's amino acid information, even when the nearest neighbour's structure is not available. Meanwhile, the traditional approach would

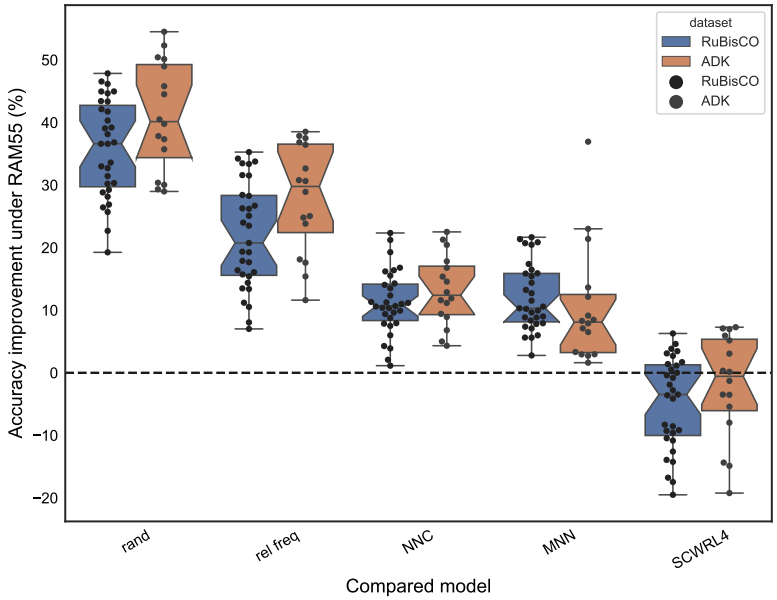

Figure 5.11: Improvement in χ_1 configuration prediction accuracy using sequence and structural information. RAM55 predictions make use of all the information available in two mixed empirical alignment (amino acid and rotasequences) datasets where the nearest neighbour of each target sequence is available as a rotasequence (for NNC) or is masked (for MNN). Accuracy improvement (y-axis) is obtained by computing the difference between RAM55's prediction accuracy and that of five other strategies: random configuration assignment (rand), assignment based on the RAM55 equilibrium frequencies (rel freq), based on the nearest neighbour's configuration (NNC), the second-nearest configuration (MNN), or only the target's backbone structure using SCWRL4. Each datapoint corresponds to the χ_1 configuration prediction accuracy improvement for one target protein; boxplot notches indicate the 95% confidence interval for the median accuracy (horizontal lines). Use of RAM55 results in a higher prediction accuracy than rand, rel freq, NNC, and MNN in both the ADK (16 taxa) and RuBisCO (30 taxa) dataset (Wilcoxon one-sided test P-values < 0.001). RAM55's accuracy is not significantly worse than SCWRL4's for the ADK dataset (Wilcoxon one-sided test P-values$= 0.07$) but it is so for the RuBisCO dataset$= 0.01$

instead rely on the second-nearest structure (MNN). This results in improved reconstruction accuracy for RAM55 (∼9% and ∼12% median improvement, respectively) over MNN (Fig. 5.11). For both NNC and MNN analyses, the improvements with RAM55 are driven by strongly increased accuracy at non-conserved sites (results not shown).

However, SCWRL4 is the most accurate predictor of χ_1 configurations out of all examined here. RAM55's median prediction accuracy is quite close to SCWRL4's on the ADK dataset, albeit noticeably worse for some taxa, and is slightly worse ($\sim$ 5%) for the RuBisCO dataset. When assessing these results it is important, however, to consider that SCWRL4, unlike RAM55, requires the target protein's backbone to be solved. In the case of a target protein of unknown structure, it would be necessary to proceed by *ab initio* modelling or homologous modelling if sufficiently similar template structure were available. This could introduce structure modelling errors which might negatively affect SCWRL4's side chain prediction accuracy that, in the best-case scenario analysis performed here, are instead based on the true backbone structure.

The strategy based on RAM55 provides plausible predictions of χ_1 configurations using an approach that, as opposed to NNC, MNN, or SCWRL4 explicitly models the evolutionary process along the branches of the phylogeny and can make use of amino acid information when structures are not available. Further, evolutionarily-backed side chain predictions might provide orthogonal information to that of predictors which rely on structure alone such as SCWRL4. RAM55-based predictions could thus benefit the side chain homology modelling process by creating an informed prior to constrain the search space, particularly where close homologs with unresolved structures might otherwise be discarded by traditional strategies.

5.5 Conclusion

In this Chapter I have discussed how RAM55 can be applied to perform structurally-aware reconstruction of ancestral sequences and side chain configuration prediction for extant proteins. Both amino acid and structural configuration states can be quite reliably inferred. Although there is little improvement in amino acid sequence reconstruction

over traditional 20-state models, use of RAM55 could benefit ASR methods by providing better phylogenies (see Chapter 4), which are valuable in themselves but also help towards obtaining reconstructions of structural information, i.e. χ_1 configurations, that are simply not possible by any other method.

Many real-world applications of RAM55, such as side chain configuration prediction, might often rely on data with a mixture of amino acid sequences and rotasequences, or amino acid sequences alone. RAM55 can be applied to this type of data, using a novel approach we described in Weber et al. [179], by treating ambiguity regarding the rotamer state in the same manner in which sequencing errors and other forms of ambiguity are handled thus allowing the information gained by accounting for the distinct evolutionary signatures of the rotamer states to be applied to many bioinformatics tasks relying on evolutionary modelling. The results in this Chapter highlight the utility of using amino acid sequences alone, and "mixed" inputs where even a limited amount of structural information leads to considerable improvements in the accuracy of χ_1 configuration prediction. Being able to include data from amino acid sequences improves prediction accuracy over modelling side-chains based on the nearest available structure alone. This is because evolutionary conserved amino acid sequences hold information about structure as well and RAM55, thanks to its internal representation of sequence and structure evolution, can benefit from this.

In particular, side chain configuration prediction under RAM55 could assist homology modelling strategies, specifically the steps involving modelling both conserved side-chains based on a known template structure, and non-conserved side-chain modelling achieved by searching a rotamer library and minimising an energy function [184, 96, 155]. In this context, using RAM55 and the marginal ancestral reconstruction algorithm makes it possible to obtain evolutionarily-backed posterior probabilities for each of the possible configurations at a given site. This distribution might provide a more-robust prior for further refinement and might introduce information orthogonal to what can be provided by structure-based methods such as SCWRL4. In the Rosetta modelling process [107], during the conformational search stage, rotamer states are sampled, using an energy function, and are then accepted or rejected using Monte Carlo methods [107]. These states

could be preferentially sampled according to their posterior probabilities according to the RAM55 model. Alternatively, the scoring function could be adjusted by replacement of the rotamer probability term – which currently favours generally more-prevalent rotamer states independent of evolutionary context – by a site-specific rotamer probability that also depends on the evolutionary context of that amino acid [4]. Such an approach could also complement artificial intelligence-based developments in the exploitation of residue co-evolution in modelling the protein backbone [177, 113, 153, 183].

Both here and in Weber et al. [179] we did not explore the use of RAM55 for inferring phylogeny from mixed data or plain amino acid sequences. This analysis would have required substantial additional work and we instead prioritised different, simpler problem to assist the development of our novel approach and illustrate its potential. A systematic exploration of tree inference accuracy from mixed data under RAM55 and other models is for future studies.

6. Conclusions

The primary goal of my PhD project was to create a structurally-aware amino acid replacement model, evaluate its contribution to our understanding of protein evolution, and assess its usefulness for inference purposes. As I have highlighted in Chapter 1, the influence of structural context in protein sequence evolution was recognised early on and models incorporating structural information are both more biologically plausible and more valuable for inference tasks than those relying on sequence information alone. However, many structurally-aware models are limited by computational complexity, reliance on MCMC algorithms, and their need for known, high quality structures. This is particularly true for models attempting to model sequence and structure change simultaneously in continuous spaces, as well as for those which constrain sequence evolution using stability or misfolding propensity. To me, these observations suggest that it might be preferable to prioritise model viability, by introducing biologically meaningful improvements within the general framework of practical phylogenetic inference approaches, rather than developing a comprehensive model that can only be applied to a minimal amount of data and a restricted set of tasks.

The RAM55 model is thus as simple as possible and as close as possible to the "Dayhoff-Eck" empirical approach. This ensures its compatibility with well-established phylogenetic inference strategies and has allowed for its implementation in popular phylognetics software, maximising its utility for the molecular evolution community. RAM55 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. As I have confirmed in Chapter 3, these configurations are robust to conformational

changes and thermodynamic fluctuations brought about by variations in the crystallisation and modelling conditions of protein structures. Further, as I described in Chapter 2, χ_1 configurations are conserved over evolutionarily meaningful time spans. The exchange patterns among RAM55's states introduce valuable, biochemically sound, information into the model by capturing effects of local steric constraints, for example those dictating how a side chain can be positioned without displacing or clashing with neighbouring residues (see Chapter 2). Further, some rotamer state exchange patterns appear to vary according to the structural context (i.e. solvent exposure, secondary structure) and functional context (i.e. Pfam domain, motif) context of a site, while others such as those indicating χ_1 configuration conservation appear to be more general and are observed in almost all contexts (see Chapter 3). These patterns deserve further exploration, perhaps by relating 3D-QSAR molecular descriptors to the exchange rates as has been attempted for amino acid exchange rates and 1-D biochemical properties [63, 34, 194].

Even if a model accurately captures some additional aspect of protein evolution, its measurable effects might be so small as to be undetectable in the amount of data available for a typical analysis, as opposed to the much larger amount used to build the model itself. RAM55 captures enough information to detect the χ_1 configuration-aware expanded state space from a standard alignment of protein sequences of known structures. It consistently offers detectably better fit to data compared to models that use the traditional 20-state space while inferring equally or more reliable phylogenies than any of the 20-state models. This argues in favour of its consideration for phylogenetic analysis of protein sequences. Moreover, when applied to empirical data, the model provides a better fit than any of the traditional models evaluated (see Chapter 4).

In Chapters 4 and 5 I focused on ancestral sequence reconstruction as another application for RAM55. ASR is a useful evolutionary biology technique with a wide variety of applications including studying how the physical properties of proteins shaped their evolutionary history, vaccine development and resurrection of proteins from extinct organisms. I have shown that RAM55 can be applied to perform structurally-aware reconstruction of ancestral amino acid and rotamer states. RAM55 could thus be a valuable tool for ancestral protein resurrection and might improve both phylogeny estimation, and help in

obtaining reconstructions of side chain orientation, which represents a unique output and cannot be replicated using traditional 20-state models.

I have also discussed how properly adapted ASR algorithms and RAM55 can be applied to data with a mixture of amino acid sequences and rotasequences, or amino acid sequences alone using a novel approach that repurposes statistical techniques for handling sequencing errors and other forms of ambiguity. This method significantly increases RAM55's viability for many realistic evolutionary modelling scenarios, including side chain configuration prediction from a known template structure. The approach I describe in Chapter 5 explicitly models the rotamer state evolutionary process along a phylogeny for a protein alignment of interest: its output is a posterior distribution of χ_1 configurations for each residue of the target protein. These results could provide an informative prior for rotamer states sampling and help reduce the number of side chain energy minimisation and refinement rounds in an homology modelling algorithm. More generally, evolutionarily-backed χ_1 configuration predictions could complement residue co-evolution information in modelling the protein backbone as well.

Aside from the applications I explored here, rotamer state replacement rates might also prove useful in the context of sequence alignment, by providing additional statistical power through the introduction of structural information. This could be of particular relevance for harder alignment problems, such as highly diverged protein sequences, and might improve sequence similarity estimates as well as protein classification into functionally meaningful families, along the lines of Pfam [46], CATH [31], and HOMSTRAD [121].

From assessing the development of models of protein evolution over the past 50 years, as well as the results from RAM55, I believe it clearly emerges that future modelling approaches 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, even though the challenge of building tractable models increases the more factors are considered. In this thesis I have shown that it is possible to take valuable steps in the direction of more plausible models and build on decades of research into protein sequence, structure, and evolution. The results shown here

strongly suggest χ_1 configuration, with its implementation and computational advantages, as another candidate structural feature for future models. The process of exploring the best combinations, and indeed devising practical algorithms and computational strategies to implement them, is for subsequent studies.

Acknowledgements

I would like to express my sincerest gratitude to my supervisor at EMBL-EBI, Dr. Nick Goldman, for his guidance throughout my PhD and specifically for his example as an excellent scientist, leader, and scientific writer.

I would similarly like to thank my co-supervisor at EMBL-EBI, Dr. Iain H. Moal, for his mentoring during the first half of my PhD, for his help in shaping my career and life choices, and his enthusiasm in sharing the dark arts of brewing and PDB file parsing.

I would also like to thank all other past and present members of the Goldman group at EMBL-EBI, in particular Dr. Nicola De Maio, Dr. Claudia C. Weber, and Conor Walker, for their help, advice, and for creating a wonderful team environment.

I would like to acknowledge Dr. Alexey M. Kozlov and Prof. Alexandros Stamatakis at the Heidelberg Institute for Theoretical Studies for their help with RAM55's implementation, for many helpful discussions, and for hosting me during our collaboration.

I would also like to thank Prof. John P. Overington for allowing me to join his team at Medicines Discovery Catapult for a few months, which helped me gain a better perspective on research beyond academia, and for his invaluable career advice.

I would like to show my appreciation for my Thesis Advisory Committee members: Dr. Evangelia Petsalaki, Dr. Simone Weyand, and Dr. Toby Gibson for their advice and encouragement throughout my PhD.

Finally, I would like to thank EMBL and EMBL-EBI for their support and for creating a positive, diverse, and welcoming environment for research and scientific training.

Code and Data Availability

All analyses described in this thesis, other than where explicitly referenced in the main text or in this section, were performed using code written by me which is available on bitbucket.org/uperron, all repositories mentioned below can be found in my bitbucket home page. More specifically:

- https://bitbucket.org/uperron/ram55_annot is a Snakemake workflow which generates the RAM55 dataset starting from Pfam protein alignments (available at ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release) and Pfam-PDBe mappings (available at ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/pdbmap.gz), applies structure quality filters, computes the χ_1 dihedral angles from atomic coordinates retrieved from the corresponding PDBe structure using Biopython's PDB utilities [27] (code available at https://github.com/biopython/biopython/blob/3c6f4e0c8e43cf8cf999737e195cedfe0a49d843/Bio/PDB), assigns χ_1 configurations according to the Dunbrack rotamer library [155] (available at http://dunbrack.fccc.edu/bbdep2010/), and cross-references these assignments using PDBe's own rotamer assignments which are available through the PDBe validation API (available at https://www.ebi.ac.uk/pdbe/api/doc/validation.html). This corresponds to the workflow described in section 2.1. Each residue in the dataset is further annotated using DSSP and other tools. This repository also contains code that computes the raw rotamer state replacement counts and from these the IRM and exchangeabilities (see section 2.3) for the analysis in Chapters 2 and 3.

- The annotated RAM55 dataset itself is available on 10.5281/zenodo.3820056 in .hd5 format.

- https://bitbucket.org/uperron/ram55_annot_eda/ contains jupyter notebooks that replicate the RAM55 dataset analysis described in sections 2.1, 2.4, 2.5, 2.6, 3.4, 3.5, and 3.6.

- https://bitbucket.org/uperron/rotastate_variability contains jupyter notebooks that replicate the human thrombin dataset analysis shown in sections 3.2 and

3.3.

- https://bitbucket.org/uperron/mdc_homstrad is a Snakemake workflow which computes RAM55-style rotamer state exchangeabilities using the HOMSTRAD alignment dataset and its pre-computed JOY structural annotations.

- https://bitbucket.org/uperron/ram55 contains code used to generate trees and simulate sequence evolution along their branches as described in section 4.4, as well as all trees used in this analysis (in Newick format), all the empirical alignments referenced in section 4.6, frequencies and exchange rate matrices for RAM55 and RUM20, and examples of how to implement these models in RAxML-NG.

- https://bitbucket.org/uperron/ram55_paper_fig contains jupyter notebooks that replicate the analysis and benchmarking of the RAM55 model for phylogenetic inference and ancestral sequence reconstruction. This corresponds to sections 4.1, 4.2, 4.5, 4.6, 5.1, 5.2.

- https://bitbucket.org/uperron/rotamer_ambiguity is a Snakemake workflow which performs the RAM55 inference benchmark on mixed data shown in sections 5.3 and 5.4. This repository includes my implementations of the marginal and joint ASR algorithms.

- https://bitbucket.org/uperron/ambiguity_coding_fig contains jupyter notebooks that replicate the figures in sections 5.3 and 5.4.

Bibliography

[1] J. Adachi and M. Hasegawa. "Model of amino acid substitution in proteins encoded by mitochondrial DNA". *J. Mol. Evol.* 42 (1996), pp. 459–468.

[2] J. Adachi, P. J. Waddell, W. Martin, and M. Hasegawa. "Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA". *J. Mol. Evol.* 50 (2000), pp. 348–358.

[3] H. Akaike. "A new look at the statistical model identification". *IEEE Trans. Automat. Contr.* 19.6 (1974), pp. 716–723.

[4] R. F. Alford et al. "The Rosetta all-atom energy function for macromolecular modeling and design". *J. Chem. Theory Comput.* 13.6 (2017), pp. 3031–3048.

[5] D. R. Anderson and K. P. Burnham. "Avoiding pitfalls when using information-theoretic methods". *J. Wildl. Manage.* 66.3 (2002), pp. 912–918.

[6] M. Arenas and D. Posada. "Computational design of centralized HIV-1 genes". *Curr. HIV Res.* 8.8 (2010), pp. 613–621.

[7] M. Arenas, A. Sánchez-Cobos, and U. Bastolla. "Maximum-likelihood phylogenetic inference with selection on protein folding stability". *Mol. Biol. Evol.* 32.8 (2015), pp. 2195–2207.

[8] M. Arenas, C. C. Weber, D. A. Liberles, and U. Bastolla. "ProtASR: an evolutionary framework for ancestral protein reconstruction with selection on folding stability". *Syst. Biol.* 66.6 (2017), 1054–1064.

[9] J. J. Babon et al. “The SOCS box domain of SOCS3: structure and interaction with the elonginBC-cullin5 ubiquitin ligase”. *J. Mol. Biol.* 381.4 (2008), pp. 928–940.

[10] I. Bahar, A. R. Atilgan, and B. Erman. “Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential”. *Fold. Des.* 2.3 (1997), pp. 173–181.

[11] I. Bahar and A. Rader. “Coarse-grained normal mode analysis in structural biology”. *Curr. Opin. Struct. Biol.* 15.5 (2005), pp. 586–592.

[12] U. Bastolla and M. Arenas. “The Influence of Protein Stability on Sequence Evolution: Applications to Phylogenetic Inference”. In: Computational Methods in Protein Evolution. Humana Press, New York, NY, 2019, pp. 215–231.

[13] U. Bastolla, M. Vendruscolo, and E.-W. Knapp. “A statistical mechanical method to optimize energy functions for protein folding”. *Proc. Natl. Acad. Sci.* 97.8 (2000), pp. 3977–3981.

[14] W. Bergsma. “A bias-correction for Cramér’s V and Tschuprow’s T”. *J. Korean Stat. Soc.* 42.3 (2013), pp. 323 –328.

[15] M. Blight and I. Holland. “Structure and function of haemolysin B, P-glycoprotein and other members of a novel family of membrane translocators”. *Mol. Microbiol.* 4.6 (1990), pp. 873–880.

[16] C. Bonnard, C. L. Kleinman, N. Rodrigue, and N. Lartillot. “Fast optimization of statistical potentials for structurally constrained phylogenetic models”. *BMC Evol. Biol.* 9.1 (2009), p. 227.

[17] J. R. Bray and J. T. Curtis. “An ordination of the upland forest communities of southern Wisconsin”. *Ecol. Monogr.* 27.4 (1957), pp. 325–349.

[18] W. J. Bruno. “Modeling residue usage in aligned protein sequences via maximum likelihood”. *Mol. Biol. Evol.* 13 (1996), pp. 1368–1374.

[19] Y. Cao, J. Adachi, A. Janke, S. Pääbo, and M. Hasegawa. "Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene". *J. Mol. Evol.* 39 (1994), pp. 519–527.

[20] M. Carroni and H. R. Saibil. "Cryo electron microscopy to determine the structure of macromolecular complexes". *Methods* 95 (2016), pp. 78–85.

[21] C. J. Challis and S. C. Schmidler. "A stochastic evolutionary model for protein structure alignment and phylogeny". *Mol. Biol. Evol.* 29 (2012), pp. 3575–3587.

[22] V. B. Chen et al. "MolProbity: all-atom structure validation for macromolecular crystallography". *Acta Cryst. D* 66.1 (2010), pp. 12–21.

[23] S. C. Choi, A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne. "Quantifying the impact of protein tertiary structure on molecular evolution". *Mol. Biol. Evol.* 24 (2007), pp. 1769–1782.

[24] C. Chothia and A. M. Lesk. "The relation between the divergence of sequence and structure in proteins". *EMBO J.* 5 (1986), pp. 823–826.

[25] J. J. Clark, M. L. Benson, R. D. Smith, and H. A. Carlson. "Inherent versus induced protein flexibility: comparisons within and between apo and holo structures". *PLoS Comput. Biol.* 15.1 (Jan. 2019), e1006705.

[26] B. E. Clifton et al. "Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein". *Nat. Chem. Biol.* 14.6 (2018), pp. 542–547.

[27] P. J. Cock et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics". *Bioinformatics* 25.11 (2009), pp. 1422–1423.

[28] wwPDB consortium. "Protein Data Bank: the single global archive for 3D macromolecular structure data". *Nucleic Acids Res.* 47.D1 (Oct. 2018), pp. D520–D528.

[29] C. C. Dang, V. S. Le, O. Gascuel, B. Hazes, and Q. S. Le. "FastMG: a simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets". *BMC Bioinform.* 15 (2014), p. 341.

[30] C. C. Dang, Q. S. Le, O. Gascuel, and V. S. Le. "FLU, an amino acid substitution model for influenza proteins". *BMC Evol. Biol.* 10 (2010), p. 99.

[31] N. L. Dawson et al. "CATH: an expanded resource to predict protein function through structure and sequence". *Nucleic Acids Res.* 45.D1 (2017), pp. D289–D295.

[32] M. O. Dayhoff and R. V. Eck. "A model of evolutionary change in proteins". In: Atlas of Protein Sequence and Structure. Ed. by M. O. Dayhoff. Washington D.C: National Biomedical Research Foundation, 1968, pp. 33–41.

[33] M. O. Dayhoff, R. V. Eck, and C. M. Park. "A model of evolutionary change in proteins." In: Atlas of Protein Sequence and Structure. Ed. by M. O. Dayhoff. Vol. 5. Washington D.C: National Biomedical Research Foundation, 1972, pp. 89–99.

[34] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. "A model of evolutionary change in proteins." In: Atlas of Protein Sequence and Structure, ed. by M. O. Dayhoff. Vol. 5. Washington D.C: National Biomedical Research Foundation, 1978. Chap. 22, pp. 345–352.

[35] N. De Maio, D. Schrempf, and C. Kosiol. "PoMo: an allele frequency-based approach for species tree estimation". *Syst. Biol.* 64.6 (2015), pp. 1018–1031.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". *J. R. Stat. Soc. Series B* 39.1 (1977), pp. 1–22.

[37] M. W. Dimmic, J. S. Rest, D. P. Mindell, and R. A. Goldstein. "rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny". *J. Mol. Evol.* 55 (2002), pp. 65–73.

[38] R. L. Dunbrack and M. Karplus. "Backbone-dependent rotamer library for proteins. Application to side-chain prediction". *J. Mol. Biol.* 230.2 (1993), pp. 543–574.

[39] R. L. Dunbrack. "Rotamer libraries in the 21st century". *Curr. Opin. Struct. Biol.* 12.4 (2002), pp. 431 –440.

[40] R. L. Dunbrack and F. E. Cohen. "Bayesian statistical analysis of protein side-chain rotamer preferences". *Protein Sci.* 6.8 (1997), pp. 1661–1681.

[41] K. A. Dunn, W. Jiang, C. Field, and J. P. Bielawski. "Improving evolutionary models for mitochondrial protein data with site-class specific amino acid exchangeability matrices". *PLoS ONE* 8 (2013), e55816.

[42] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press, 1998.

[43] J. Echave, S. J. Spielman, and C. O. Wilke. "Causes of evolutionary rate variation among protein sites". *Nat. Rev. Genet.* 17 (2016), pp. 109–121.

[44] J. Echave. "Evolutionary divergence of protein structure: the linearly forced elastic network model". *Chem. Phys. Lett.* 457.4-6 (2008), pp. 413–416.

[45] R. V. Eck and M. O. Dayhoff. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Spring, 1966.

[46] S. El-Gebali et al. "The Pfam protein families database in 2019". *Nucleic Acids Res.* 47.D1 (2019), pp. D427–D432.

[47] J. Felsenstein. *Inferring Phylogenies*. 1st edition. Sinauer Associates, Sunderland, MA, 2004.

[48] J. Felsenstein. "Phylogenies and the comparative method". *Am. Nat* 125 (1985), pp. 1–15.

[49] J. Felsenstein. "Evolutionary trees from DNA sequences: a maximum likelihood approach". *J. Mol. Evol.* 17.6 (1981), pp. 368–376.

[50] T. H. Fischer, A. P. Bode, M. Demcheva, and J. N. Vournakis. "Hemostatic properties of glucosamine-based materials". *J. Biomed. Mater. Res.* 80.1 (2007), pp. 167–174.

[51] W. M. Fitch and E. Markowitz. "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution". *Biochem. Genet.* 4.5 (1970), pp. 579–593.

[52] W. Fletcher and Z. Yang. "INDELible: a flexible simulator of biological sequence evolution". *Mol. Biol. Evol.* 26.8 (2009), pp. 1879–1888.

[53] T. Flouri et al. “The phylogenetic likelihood library”. *Sys. Bio.* 64.2 (2015), 356–362.

[54] M. S. Fornasari, G. Parisi, and J. Echave. “Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations”. *Mol. Biol. Evol.* 19 (2002), pp. 352–356.

[55] F. Galaway, R. Yu, A. Constantinou, F. Prugnolle, and G. J. Wright. “Resurrection of the ancestral RH5 invasion ligand provides a molecular explanation for the origin of P. falciparum malaria in humans”. *PLOS Biol.* 17.10 (2019).

[56] N. Galtier. “Maximum-likelihood phylogenetic analysis under a covarion-like model”. *Mol. Biol. Evol.* 18 (2001), pp. 866–873.

[57] E. García-Portugués et al. “Toroidal Diffusions and Protein Structure Evolution”. In: Appl. Dir. Stat. Mod. Methods Case Stud. 2019, pp. 61–94.

[58] E. A. Gaucher, S. Govindarajan, and O. K. Ganesh. “Palaeotemperature trend for Precambrian life inferred from resurrected proteins”. *Nature* 451.7179 (2008), pp. 704–707.

[59] M. Golden et al. “A generative angular model of protein structure evolution”. *Mol. Biol. Evol.* 34.8 (2017), pp. 2085–2100.

[60] N. Goldman, J. L. Thorne, and D. T. Jones. “Assessing the impact of secondary structure and solvent accessibility on protein evolution”. *Genetics* 149 (1998), pp. 445–458.

[61] N. Goldman and S. Whelan. “A novel use of equilibrium frequencies in models of sequence evolution”. *Mol. Biol. Evol.* 19 (2002), pp. 1821–1831.

[62] G. H. Gonnet, M. A. Cohen, and S. A. Benner. “Exhaustive matching of the entire protein sequence database”. *Science* 256 (1992), pp. 1443–1445.

[63] R. Grantham. “Amino acid difference formula to help explain protein evolution”. *Science* 185.4154 (1974), pp. 862–864.

[64] S. Gräslund et al. “Protein production and purification”. *Nat. Methods* 5.2 (2008), p. 135.

[65] S. Guindon et al. “New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0”. *Syst. Biol.* 59.3 (2010), pp. 307–321.

[66] Y. Gumulya and E. M. Gillam. “Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the ‘retro’approach to protein engineering”. *Biochem. J.* 474.1 (2017), pp. 1–19.

[67] A. Halpern and W. J. Bruno. “Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies”. *Mol. Biol. Evol.* 15 (1998), pp. 910–917.

[68] M. J. Harms and J. W. Thornton. “Evolutionary biochemistry: revealing the historical and physical causes of protein properties”. *Nat. Rev. Genet.* 14.8 (2013), pp. 559–571.

[69] K. M. Hart et al. “Thermodynamic system drift in protein evolution”. *PLoS Biol.* 12.11 (2014), e1001994.

[70] W. A. Hendrickson. “Analysis of protein structure from diffraction measurement at multiple wavelengths”. *Trans. Am. Crystallogr. Assoc* 21.11 (1985).

[71] J. L. Herman, C. J. Challis, Á. Novák, J. Hein, and S. C. Schmidler. “Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure”. *Mol. Biol. Evol.* 31 (2014), pp. 2251–2266.

[72] J. Herrero et al. “Ensembl comparative genomics resources”. *Database* 2016 (2016).

[73] J. K. Hobbs et al. “On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus”. *Mol. Biol. Evol.* 29.2 (2012), pp. 825–835.

[74] I Holmes and G. Rubin. “An expectation maximization algorithm for training hidden substitution models”. *J. Mol. Biol.* 317.5 (2002), pp. 753–764.

[75] I. Holmes and J. P. Rubin. “An expectation maximization algorithm for training hidden substitution models”. *J. Mol. Biol* 317 (2002), pp. 753–764.

[76] Y.-F. Huang and G. B. Golding. “Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures”. *PLoS Comput. Biol.* 10 (2014).

[77] J. P. Huelsenbeck. "Testing a covariotide model of DNA substitution". *Mol. Biol. Evol.* 19.5 (2002), pp. 698–707.

[78] A. Ingles-Prieto et al. "Conservation of protein structure over four billion years". *Structure* 21.9 (2013), pp. 1690–1697.

[79] J. L. Jensen and A.-M. K. Pedersen. "Probabilistic models of DNA sequence evolution with context dependent rates of substitution". *Adv. Appl. Prob* 32 (2000), pp. 499–517.

[80] D. T. Jones, W. R. Taylor, and J. M. Thornton. "A new approach to protein fold recognition". *Nature* 358 (1992), pp. 86–89.

[81] K. Katoh and D. M. Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". *Mol. Biol. Evol.* 30.4 (2013), pp. 772–780.

[82] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform". *Nucleic Acids Res.* 30.14 (2002), pp. 3059–3066.

[83] B. T. Kile et al. "The SOCS box: a tale of destruction and degradation". *Trends Biochem. Sci.* 27.5 (2002), pp. 235–241.

[84] C. L. Kleinman, N. Rodrigue, N. Lartillot, and H. Philippe. "Statistical potentials for improved structurally constrained evolutionary models". *Mol. Biol. Evol.* 27 (2010), pp. 1546–1560.

[85] C. L. Kleinman, N. Rodrigue, C. Bonnard, H. Philippe, and N. Lartillot. "A maximum likelihood framework for protein design". *BMC Bioinform.* 7.1 (2006), p. 326.

[86] P. S. Klosterman et al. "XRate: a fast prototyping, training and annotation tool for phylo-grammars". *BMC Bioinform.* 7.1 (2006), p. 428.

[87] P. Koehl and M. Levitt. "Sequence variations within protein families are linearly related to structural variations". *J. Mol. Biol.* 323.3 (2002), pp. 551–562.

[88] A. Konno, A. Kitagawa, M. Watanabe, T. Ogawa, and T. Shirai. "Tracing protein evolution through ancestral structures of fish galectin". *Structure* 19.5 (2011), pp. 711–721.

[89] C. Korostensky and G. H. Gonnet. “Using traveling salesman problem algorithms for evolutionary tree construction”. *Bioinformatics* 16.7 (2000), pp. 619–627.

[90] J. M. Koshi and R. A. Goldstein. “Mutation matrices and physical-chemical properties: correlations and implications”. *Proteins* 27 (1996), pp. 336–344.

[91] J. M. Koshi and R. A. Goldstein. “Probabilistic reconstruction of ancestral protein sequences”. *J. Mol. Evo.* 42.2 (1996), pp. 313–320.

[92] C. Kosiol and N. Goldman. “Different versions of the dayhoff rate matrix”. *Mol. Biol. Evol.* 22.2 (2005), pp. 193–199.

[93] D. L. Kothe et al. “Ancestral and consensus envelope immunogens for HIV-1 subtype C”. *Virology* 352.2 (2006), pp. 438–449.

[94] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis. “RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference”. *Bioinformatics* 35.21 (2019), pp. 4453–4455.

[95] O. Kozlov. “Models, Optimizations, and Tools for Large-Scale Phylogenetic Inference, Handling Sequence Uncertainty, and Taxonomic Validation”. PhD thesis. Karlsruhe, Germany: Karlsruhe Institute of Technology, 2018.

[96] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack Jr. “Improved prediction of protein side-chain conformations with SCWRL4”. *Proteins* 77.4 (2009), pp. 778–795.

[97] S. Kullback and R. A. Leibler. “On Information and Sufficiency”. *Ann. Math. Statist.* 22.1 (1951), pp. 79–86.

[98] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura. “MEGA X: molecular evolutionary genetics analysis across computing platforms”. *Mol. Biol. Evol.* 35.6 (2018), pp. 1547–1549.

[99] J.-S. Lai, B. Rost, B. Kobe, and M. Boden. “Evolutionary model of protein secondary structure capable of revealing new biological relationships”. *bioRxiv* (2019), p. 563452.

[100] N. Lartillot and H. Philippe. "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process". *Mol. Biol. Evol.* 21 (2004), 1095–1109.

[101] D. Lavens et al. "The C-terminus of CIS defines its interaction pattern". *Biochem. J.* 401.1 (2007), pp. 257–267.

[102] S. Q. Le and O. Gascuel. "Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial". *Syst. Biol.* 59 (2010), pp. 277–287.

[103] S. Q. Le, C. C. Dang, and O. Gascuel. "Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates". *Mol. Biol. Evol.* 29.10 (2012), pp. 2921–2936.

[104] S. Q. Le and O. Gascuel. "An improved general amino acid replacement matrix". *Mol. Biol. Evol.* 25.7 (2008), pp. 1307–1320.

[105] S. Q. Le and O. Gascuel. "An improved general amino acid replacement matrix". *Mol. Biol. Evol.* 25.7 (2008), pp. 1307–1320.

[106] S. Q. Le, N. Lartillot, and O. Gascuel. "Phylogenetic mixture models for proteins". *Philos. Trans. Royal Soc. B* 363.1512 (2008), pp. 3965–3976.

[107] A. Leaver-Fay et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules". *Meth. Enzymol.* 487 (2011), pp. 545–574.

[108] E. D. Levy. "A simple definition of structural regions in proteins and its use in analyzing interface evolution". *J. Mol. Biol.* 403.4 (2010), pp. 660–670.

[109] D. Liebschner et al. "Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix". *Acta Cryst. D* 75.10 (2019), pp. 861–877.

[110] P. Liò and N. Goldman. "Modeling mitochondrial protein evolution using structural information". *J. Mol. Evol.* 54 (2002), pp. 519–529.

[111] P. Lio and N. Goldman. "Models of molecular evolution and phylogeny". *Genome Res.* 8.12 (1998), pp. 1233–1244.

[112] P. Liò and N. Goldman. "Using protein structural information in evolutionary inference: transmembrane proteins". *Mol. Biol. Evol.* 16 (1999), pp. 1696–1710.

[113] Y. Liu, P. Palmedo, Q. Ye, B. Berger, and J. Peng. "Enhancing evolutionary couplings with deep convolutional neural networks". *Cell Syst.* 6.1 (2018), 65 –74.e3.

[114] P. Lopez, D. Casane, and H. Philippe. "Heterotachy, an important process of protein evolution". *Mol. Biol. Evol.* 19.1 (2002), pp. 1–7.

[115] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. "The penultimate rotamer library". *Proteins* 40.3 (2000), pp. 389–408.

[116] A. Löytynoja. "Phylogeny-aware alignment with PRANK". In: Multiple Sequence Alignment Methods. Humana Press, Totowa, NJ, 2014, pp. 155–170.

[117] D. S. Marks et al. "Protein 3D structure computed from evolutionary sequence variation". *PLoS One* 6.12 (2011), e28766.

[118] A. McPherson. "Introduction to protein crystallization". *Methods* 34.3 (2004), pp. 254–265.

[119] R. Merkl and R. Sterner. "Ancestral protein reconstruction: techniques and applications". *Biol. Chem.* 397.1 (2016), pp. 1–21.

[120] J. L. Milne et al. "Cryo-electron microscopy–a primer for the non-microscopist". *FEBS J.* 280.1 (2013), pp. 28–45.

[121] K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington. "HOMSTRAD: a database of protein structure alignments for homologous families". *Protein Sci.* 7.11 (1998), pp. 2469–2471.

[122] K. Mizuguchi, C. M. Deane, T. L. Blundell, M. S. Johnson, and J. P. Overington. "JOY: protein sequence-structure representation and analysis." *Bioinformatics* 14.7 (1998), pp. 617–623.

[123] D. W. Mount. *Bioinformatics: sequence and genome analysis*. Vol. 1. Cold Spring Harbor Laboratory Press, New York, NY, 2001.

[124] G. N. Murshudov et al. "REFMAC5 for the refinement of macromolecular crystal structures". 67.4 (2011), pp. 355–367.

[125] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman. "Side-chain flexibility in proteins upon ligand binding". *Proteins* 39.3 (2000), pp. 261–268.

[126] L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies". *Mol. Biol. Evol.* 32.1 (2015), pp. 268–274.

[127] J. Overington, M. S. Johnson, A. Sali, and T. L. Blundell. "Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction". *Proc. R. Soc. London B* 241 (1990), pp. 132–145.

[128] J. Overington, M. S. Johnson, A. Sali, and T. L. Blundell. "Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction". *Proc. Royal Soc. Lond. B* 241.1301 (1990), pp. 132–145.

[129] J. Overington, D. A. N. Donnelly, M. S. Johnson, A. Sali, and T. O. M. L. Blundell. "Environment-specific amino-acid substitution tables – tertiary templates and prediction of protein folds." *Protein Sci.* 1 (1992), pp. 216–226.

[130] G. Parisi and J. Echave. "Structural constrains and the emergence of sequence patterns in protein evolution". *Mol. Biol. Evol.* 18 (2001), pp. 750–756.

[131] F. Pazos, M. Helmer-Citterich, G. Ansiello, and A. Valencia. "Correlated mutations contain information about protein-protein interactions". *J. Mol. Biol* 271 (1997), pp. 511–523.

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

[133] U. Perron, I. H. Moal, J. L. Thorne, and N. Goldman. "Probabilistic Models for the Study of Protein Evolution". In: Handbook of Statistical Genomics. John Wiley Sons, Ltd, 2019. Chap. 12, pp. 347–30.

[134] U. Perron, A. M. Kozlov, A. Stamatakis, N. Goldman, and I. H. Moal. "Modeling Structural Constraints on Protein Evolution via Side-Chain Conformational States". *Mol. Biol. Evol.* 36.9 (2019), pp. 2086–2103.

[135] E. F. Pettersen et al. "UCSF Chimera—a visualization system for exploratory research and analysis". *J. Comput. Chem.* 25.13 (2004), pp. 1605–1612.

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

[137] T. Pupko, I Pe'er, R. Shamir, and D. Graur. "A fast algorithm for joint reconstruction of ancestral amino acid sequences." *Mol. Biol. Evol.* 17.6 (2000), pp. 890–896.

[138] C Ramakrishnan and G. Ramachandran. "Stereochemical criteria for polypeptide and protein chain conformations: II. Allowed conformations for a pair of peptide units". *Biophys. J.* 5.6 (1965), pp. 909–933.

[139] G. Rhodes. *Crystallography made crystal clear: a guide for users of macromolecular models*. Elsevier Academic Press, 2010.

[140] S. Rios et al. "GPCRtm: An amino acid substitution matrix for the transmembrane region of class A G Protein-Coupled Receptors." *BMC Bioinform.* 16.1 (2015), p. 206.

[141] V. A. Risso, J. A. Gavira, E. A. Gaucher, and J. M. Sanchez-Ruiz. "Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins". *Proteins* 82.6 (2014), pp. 887–896.

[142] D. M. Robinson, D. Jones, H. Kishino, N. Goldman, and J. L. Thorne. "Protein evolution with dependence among codons due to tertiary structure". *Mol. Biol. Evol.* 20 (2003), pp. 1692–1704.

[143] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. "Protein evolution with dependence among codons due to tertiary structure". *Mol. Biol. Evol.* 20.10 (2003), pp. 1692–1704.

[144] N. Rodrigue, H. Philippe, and N. Lartillot. "Assessing site-interdependent phylogenetic models of sequence evolution". *Mol. Biol. Evol.* 23 (2006), pp. 1762–1775.

[145] N. Rodrigue, C. L. Kleinman, H. Philippe, and N. Lartillot. "Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons". *Mol. Biol. Evol.* 26.7 (2009), pp. 1663–1676.

[146] G. L. Rosano and E. A. Ceccarelli. "Recombinant protein expression in Escherichia coli: advances and challenges". *Front Microbiol.* 5 (2014), p. 172.

[147] B. Rost. "Protein structures sustain evolutionary drift". *Fold. Des.* 2 (1997), S19–S24.

[148] R. B. Russell, M. A. S. Saqi, R. A. Sayle, P. A. Bates, and M. J. E. Sternberg. "Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation". *J. Mol. Biol* 269 (1997), pp. 423–439.

[149] N Saitou and M Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Mol. Biol. Evol.* 4.4 (1987), pp. 406–425.

[150] A. Sali and T. L. Blundell. "Definition of general topological equivalence in protein structures". *J. Mol. Biol.* 212.2 (1990), pp. 403–428.

[151] D. Sankoff. "Minimal mutation trees of sequences". *SIAM J. Appl. Math.* 28.1 (1975), pp. 35–42.

[152] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, and A. M. J. J. Bonvin. "Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age". *Proteins* 86 (2018), pp. 51–66.

[153] R. Service. "Google's DeepMind aces protein folding". *Science* (Dec. 2018), doi:10.1126/science.aaw2747.

[154] E. Shakhnovich, V. Abkevich, and O. Ptitsyn. "Conserved residues and the mechanism of protein folding". *Nature* 379.6560 (1996), pp. 96–98.

[155] M. V. Shapovalov and R. L. Dunbrack. "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions". *Structure* 19.6 (2011), pp. 844–858.

[156] G. M. Sheldrick. "Crystal structure refinement with SHELXL". *Acta Cryst. C* 71.1 (2015), pp. 3–8.

[157] Z. Sheng et al. "Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation". *Front. Immunol.* 8.MAY (2017).

[158] L. Si Quang, O. Gascuel, and N. Lartillot. "Empirical profile mixture models for phylogenetic reconstruction". *Bioinformatics* 24.20 (2008), pp. 2317–2323.

[159] R. R. Sokal. "A statistical method for evaluating systematic relationships." *Univ. Kansas, Sci. Bull.* 38 (1958), pp. 1409–1438.

[160] A. Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". *Bioinformatics* 30.9 (2014), pp. 1312–1313.

[161] J. Sullivan and P. Joyce. "Model selection in phylogenetics". *Annu. Rev. Ecol. Evol. Syst.* 36.1 (2005), pp. 445–466.

[162] M. Sutcliffe, F. Hayes, and T. Blundell. "Knowledge based modelling of homologous proteins, part II: rules for the conformations of substituted sidechains". *Protein Eng. Des. Sel.* 1.5 (Oct. 1987), pp. 385–392.

[163] G. Taylor. "The phase problem". *Acta Cryst. D* 59.11 (2003), pp. 1881–1890.

[164] W. R. Taylor and D. T. Jones. "Deriving an amino acid distance matrix". *J. Theor. Biol* 164 (1993), pp. 65–83.

[165] A. Terakita. "The opsins". *Genome Biol.* 6.3 (2005), p. 213.

[166] J. L. Thorne, N. Goldman, and D. T. Jones. "Combining protein evolution and secondary structure". *Mol. Biol. Evol.* 13 (1996), pp. 666–673.

[167] J. W. Thornton. "Resurrecting ancient genes: experimental analysis of extinct molecules". *Nat. Rev. Genet.* 5.5 (2004), pp. 366–375.

[168] M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, and C. O. Wilke. "Maximum allowed solvent accessibilites of residues in proteins". *PloS one* 8.11 (2013).

[169] E. R. M. Tillier and T. W. H. Lui. "Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments". *Bioinformatics* 19 (2003), pp. 750–755.

[170] W. G. Touw et al. "A series of PDB-related databanks for everyday needs". *Nucleic Acids Res.* 43.D1 (2015), pp. D364–D368.

[171] K. Trueblood et al. “Atomic dispacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature”. *Acta Cryst. A* 52.5 (1996), pp. 770–781.

[172] C. Tuffley and M. Steel. “Modeling the covarion hypothesis of nucleotide substitution”. *Math. Biosci.* 147.1 (1998), pp. 63–91.

[173] UniProt Consortium. “UniProt: the universal protein knowledgebase”. *Nucleic Acids Res.* 45.D1 (2017), pp. D158–D169.

[174] C. Venien-Bryan, Z. Li, L. Vuillard, and J. A. Boutin. “Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery”. *Acta Cryst. F* 73.Pt 4 (Apr. 2017), pp. 174–183.

[175] H. Wako and T. L. Blundell. “Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures”. *J. Mol. Biol* 238 (1994), pp. 693–708.

[176] H. C. Wang, K. Li, E. Susko, and A. J. Roger. “A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny”. *BMC Evol. Biol.* 8 (2008), p. 331.

[177] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu. “Accurate de novo prediction of protein contact map by ultra-deep learning model”. *PLOS Comput. Biol.* 13.1 (Jan. 2017), pp. 1–34.

[178] A. Waterhouse et al. “SWISS-MODEL: homology modelling of protein structures and complexes”. *Nucleic Acids Res.* 46.W1 (May 2018), W296–W303.

[179] C. C. Weber, U. Perron, D. Casey, Z. Yang, and N. Goldman. “Ambiguity coding allows accurate inference of evolutionary parameters from alignments in an aggregated state-space”. *Sys. Bio.* (2020). in press.

[180] L. C. Wheeler, S. A. Lim, S. Marqusee, and M. J. Harms. “The thermostability and specificity of ancient proteins”. *Curr. Opin. Struct. Biol.* 38 (June 2016), pp. 37–43.

[181] S. Whelan and N. Goldman. "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach". *Mol. Biol. Evol.* 18 (2001), pp. 691–699.

[182] S. Whelan, J. E. Allen, B. P. Blackburne, and D. Talavera. "ModelOMatic: fast and automated model selection between RY, nucleotide, amino acid, and codon substitution models". *Syst. Biol.* 64.1 (2015), pp. 42–55.

[183] J. Xu. "Distance-based protein folding powered by deep learning". *Proc. Natl. Acad. Sci. U.S.A.* 116.34 (2019), pp. 16856–16865.

[184] J. Xu. "Rapid protein side-chain packing via tree decomposition". In: RECOMB. Springer. 2005, pp. 423–439.

[185] Z. Yang. "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites". *Mol. Biol. Evol.* 10.6 (1993), pp. 1396–1401.

[186] Z Yang, S Kumar, and M Nei. "A new method of inference of ancestral nucleotide and amino acid sequences." *Genetics* 141.4 (1995), pp. 1641–1650.

[187] Z. Yang and R. Nielsen. "Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage". *Mol. Biol. Evol.* 25 (2008), pp. 568–579.

[188] Z. Yang. "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods". *J. Mol. Evo.* 39.3 (1994), pp. 306–314.

[189] Z. Yang. *Molecular Evolution: a Statistical Approach*. Oxford University Press, 2014.

[190] Z. Yang. "PAML 4: phylogenetic analysis by maximum likelihood". *Mol. Biol. Evol.* 24.8 (2007), pp. 1586–1591.

[191] Z. Yang, R. Nielsen, and M. Hasegawa. "Models of Amino Acid Substitution and Applications to Mitochondrial Protein Evolution". *Mol. Biol. Evol.* 15.12 (1998), pp. 1600–1611.

[192] M. I. Zavodszky and L. A. Kuhn. "Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis". *Protein Sci.* 14.4 (2005), pp. 1104–1114.

[193] S. Zhao, D. S. Goodsell, and A. J. Olson. "Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation". *Proteins* 43.3 (2001), pp. 271–279.

[194] S. Zoller and A. Schneider. "Improving phylogenetic inference with a semiempirical amino acid substitution model". *Mol. Biol. Evol.* 30.2 (2013), pp. 469–479.

[195] Z. Zou and J. Zhang. "Amino acid exchangeabilities vary across the tree of life". *Sci. Adv.* 5.12 (2019), eaax3124.