CAPTURING THE DYNAMICS OF PROTEIN SEQUENCE EVOLUTION THROUGH SITE-INDEPENDENT STRUCTURALLY CONSTRAINED PHYLOGENETIC MODELS

by

Javier Antonio Alfaro

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
September 2012

DALHOUSIE UNIVERSITY

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "CAPTURING THE DYNAMICS OF PROTEIN SEQUENCE EVOLUTION THROUGH SITE-INDEPENDENT STRUCTURALLY CONSTRAINED PHYLOGENETIC MODELS" by Javier Antonio Alfaro in partial fulfilment of the requirements for the degree of Master of Science.

Dated: September 18, 2012

Supervisor: _____

Readers: _____

_____

_____

Departmental Representative: _____

DALHOUSIE UNIVERSITY

DATE: September 18, 2012

AUTHOR: Javier Antonio Alfaro

TITLE: CAPTURING THE DYNAMICS OF PROTEIN SEQUENCE EVOLUTION THROUGH SITE-INDEPENDENT STRUCTURALLY CONSTRAINED PHYLOGENETIC MODELS

DEPARTMENT OR SCHOOL: Department of Biochemistry and Molecular Biology

DEGREE: MSc CONVOCATION: May YEAR: 2013

_____

Signature of Author

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

ABSTRACT ix

LIST OF ABBREVIATIONS USED x

ACKNOWLEDGEMENTS xi

CHAPTER 1 INTRODUCTION 1

1.1 CONSTRAINTS ON THE EVOLUTION OF BIOLOGICAL ORGANISMS: 1

1.1.1 EVOLUTIONARY CONSTRAINTS ON CODING SEQUENCES 1

1.1.2 BIOPHYSICAL CONSTRAINTS ON PROTEIN CODING GENES: 2

1.2 BUILDING PHYLOGENETIC TREES FROM MOLECULAR | SEQUENCE DATA: 5

1.2.1 PHYLOGENETIC TREES FROM MULTIPLE SEQUENCE ALIGNMENTS: 6

1.3 MARKOV MODELS OF CODING SEQUENCE EVOLUTION: 10

1.3.1 DNA MODELS OF EVOLUTION: 12

1.3.2 CODON MODELS OF PROTEIN EVOLUTION: 12

1.3.3 AMINO ACID MODELS OF PROTEIN EVOLUTION. 13

1.3.4 SPECIALIZED AMINO ACID RATE MATRICES AND THE INNOVATION OF SEMI-EMPIRICAL CODON MODELS: 15

1.3.5 SYNOPSIS OF THE MODELS AND RESULTS INTRODUCED IN THIS WORK: 20

CHAPTER 2 SITE INDEPENDENT STRUCTURALLY CONSTRAINED PHYLOGENETIC MODELS THAT FLEXIBLY FIT DIFFERENT STRUCTURAL ENVIRONMENTS 22

2.1 INTRODUCTION: 22

2.2 MATERIALS AND METHODS: 26

2.2.1 MODEL DESCRIPTION 26

2.2.2 PARAMETER ESTIMATION: 29

2.2.3 DATASETS FOR PERFORMANCE EVALUATION. 30

2.2.4 MODEL COMPARISON BY LIKELIHOOD RATIO TESTING AND BY AIC. 32

2.2.5 EVALUATING GENERAL TRENDS IN SITE-SPECIFIC MATRICES: 36

2.2.6 MEASURING PARTITION-SPECIFIC SIGNIFICANCE TOWARDS OBSERVED LIKELIHOOD GAINS: 36

2.2.7 Identifying sites potentially involved in protein-protein/protein-ligand interactions. 37
2.3 Results and Discussion: 38
2.3.1 Structural constraints significantly improve independence models of protein evolution: 38
2.3.2 A hybrid FOLDX+C_PROSA potential accommodates datasets with | varying evolutionary rates and biophysical constraints: 41
2.3.3 Partitioning for solvent accessibility and secondary structure..... 42
2.3.4 Parameter estimates for a general structurally constrained model of protein evolution. 43
2.3.5 FoldX outperforms PROSA in the hydrophobic protein cores while maintaining a comparable performance on the surface. 53
2.3.6 Highly variable sites in exposed regions of the protein are better modeled by partitioned SCPMs 56
2.3.7 Le and Gascuel matrices: 56
2.3.8 The performance of SCPEs in the vicinity of protein-protein, protein-ligand interactions. 57
2.4 Conclusion: 62
CHAPTER 3 DISCUSSION 64
3.1 Models implemented: 64
3.1.1 Evolutionary constraints on coding sequences: 64
3.1.2 Structurally constrained partition models: 64
3.2 Other constraints on protein evolution ignored by our models. 65
3.2.1 Maintenance of extra-protein interactions represent another selective constraint on sites. 65
3.2.2 Intrinsic constraints on translational accuracy and the fidelity of protein folding: 66
3.2.3 Protein expression level increases the relevance of structural constraints on protein evolution: 66
3.3 Future research directions: 67
3.3.1 A non-stationary structurally constrained model: 67
3.3.2 Applications of non-stationary structurally constrained phylogenetic models: 69
3.3.3 Improving statistical potentials for structurally constrained phylogenetic models. 70

3.3.4 FINAL REMARKS 72
LITERATURE CITED 73
APPENDIX A- SUPPLEMENTARY FIGURES AND TABLES FOR CHAPTER 2 85

LIST OF TABLES

TABLE 2.1: COMPARING LG+ΔG MODELS TO LG. 54
TABLE 2.2: COMPARING 2P+ΔG MODELS TO LG. 55
TABLE 2.3: NUMBER OF TIMES OUT OF 48 THAT A MODEL USING LG BASIS MATRICES WAS PREFERRED BY AIC TO A MODEL USING APPROPRIATE SECONDARY STRUCTURE/SOLVENT ACCESSIBILITY MATRICES FROM LE AND GASCUEL, 2009..... 57

LIST OF FIGURES

FIGURE 1.1: AMINO ACID SEQUENCE CHANGES WITHIN A PROTEIN STRUCTURE MAY HAVE VARYING AFFECTS ON FUNCTION. 3
FIGURE 1.2: TWO ALTERNATIVE PARADIGMS FOR PHYLOGENETIC INFERENCE. 8
FIGURE 1.3: SUMMARY OF THE LIKELIHOOD CALCULATION. 11
FIGURE 1.4: PHYLOGENETIC MODELS THAT ACCOUNT FOR VARYING DEGREES OF STRUCTURAL CONSTRAINT. 18
FIGURE 2.1: ALLOWING STRUCTURAL CONSTRAINTS TO VARY ACROSS STRUCTURAL ENVIRONMENTS OFTEN IMPROVES SCPMS AND ALWAYS IMPROVES NON-STRUCTURALLY CONSTRAINED COUNTER PARTS. 35
FIGURE 2.2: LIKELIHOOD GAINS OVER A STANDARD STRUCTURALLY CONSTRAINED MODEL CAN BE ARRIVED WITH MIXED STATISTICAL-POTENTIAL/EMPIRICAL POTENTIAL APPROACHES AND BY PARTITIONING FOR SECONDARY STRUCTURE AND SOLVENT ACCESSIBILITY. 40
FIGURE 2.3 THE GENERAL STRUCTURALLY CONSTRAINED MODEL: EFFECTS ON STATIONARY FREQUENCIES AND THE MEDIAN Q. 46
FIGURE 2.4: STATIONARY FREQUENCIES DERIVED FOR PARTITION SPECIFIC MODELS. 48
FIGURE 2.5: MEDIAN Q MATRICES FOR MODELS UTILIZING A COMBINATION OF FOLDX AND C_PROSA ENERGIES. 52
FIGURE 2.6: USING STRUCTURALLY CONSTRAINED MODELS OF PROTEIN EVOLUTION TO IDENTIFY INTERACTION SITES. 60
FIGURE 2.7: PSSF FACTORS FOR INTERACTING SITES PLOTTED AGAINST ALIGNMENT INFORMATION. 61

ABSTRACT

Protein function arises from the large scaffold of residue interactions that position critical residues to stabilize the fold and to interact with substrates and other proteins or co-factors. Any accurate model of the evolution of protein sequences should therefore account for the selection pressures to preserve these supporting interactions. It is therefore surprising that the most commonly-used methods for resolving protein sequence phylogenies employ models of the evolutionary process that do not account for these residue-specific constraints. While structurally constrained models of protein evolution have existed for some time, their implementation has been based on complex models that attempt to take into account the effects of multiple substitutions in protein sequences and/or dependence amongst sites in the alignment. Here we propose an alternative approach. We formalize a simple structurally constrained amino acid model of protein evolution that maintains the common phylogenetic inference assumption that sites evolve independently of each other. Our independence energy model adjusts a standard substitution model, such as the Le and Gascuel matrix (LG), on a site-by-site basis in order to incorporate the structural constraint that is based on the change in free energy of folding that arises from introducing single point substitutions at a site in the wild-type protein sequence. We explore the properties of our structurally constrained model as well as two extensions aimed at more accurately incorporating structural constraints into our model and evaluate how well they fit the evolutionary dynamics of a set of protein families.

LIST OF ABBREVIATIONS USED

AIC	Akaike Information Criterion
C_PROSA	Combined PROSA cβ pairwise and surface potentials
DEM	Dependency energy model
F81	Felsenstein-81
FoldX	The FoldX empirical force field
FoldX+C_PROSA	Combination of FoldX empirical energies and PROSA cβ pairwise statistical potentials.
GTR	General time reversible model
IEM	Independence energy model
JTT	Jones Taylor and Thornton
ld	Log(site likelihood) differences
LG	Le and Gascuel matrix
LR	Likelihood ratio test statistic
LRT	Likelihood ratio test
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
NCBI	National Center for Biotechnology Information
P_PROSA	PROSA cβ pairwise statistical potentials
PAM	Accepted point mutation matrix
PISA	Protein interfaces, surfaces and assemblies service at the European Bioinformatics Institute (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html).
PSSF	Partition-specific significance factor
RAS	Rates-across-sites
SCPE	Structurally constrained model of protein evolution
SCPM	Structurally constrained phylogenetic model
SCPM's	A multitude of structurally constrained phylogenetic models
SEPM	Simulation energy phylogenetic models
SSM	Structure based substitution models
WAG	Whelan-and-Goldman model

ACKNOWLEDGEMENTS

It would not have been possible to write this Masters Thesis without the continuous, generous and support of many people.

My principle supervisor Dr. Andrew Roger, whose mentoring, care, and patience was inspiring. His contagious enthusiasum for science guided and helped structure my evolutionary thought. Many fruitful meetings with Dr. Edward Susko will always be remembered as the birthplace of my mathematical mindset and an introduction to many concepts in mathematical modeling. Dr. Christian Blouin, whose insightfulness and ability to translate complex ideas into an understandable form helped greatly to communicate this project to both me and others. Dr. Jan Rainey, I thank for his patience at my many committee meetings.

Finally, I would like to acknowledge Kirsten Kennedy, for her continuous love and support through the completion of both her degree and mine.

CHAPTER 1 INTRODUCTION

1.1 CONSTRAINTS ON THE EVOLUTION OF BIOLOGICAL ORGANISMS:

At the molecular level, the variation that fuels biological evolution of inherited traits results from the fixation of unrepaired errors that occur during genome replication. Over time, variations in an organism's phenotype may arise within a population that are either neutral, beneficial, or detrimental for the survival of an organism in its environment. Ultimately, these variations are caused at the molecular level by mutations and may constitute, for example, partial or complete genome or gene duplications, genome insertions, genome-rearrangements or nucleotide mutations in the form of insertions, deletions, or substitutions (Koonin and Wolf, 2010). Once a mutation has occurred in a member of a population, it may become fixed or removed from the population through either natural selection or genetic drift. Evolutionary 'constraints' manifest because of selection for maintenance of essential functions and such constraints are evident at the level of genome architecture, gene repertoire, and individual gene presence, copy number and sequence. In general, sequences encoding structural RNAs and non-synonymous nucleotide positions in protein coding sequences are among the most strongly constrained when compared to nucleotide positions at synonymous sites in protein coding sequences, and sequences coding for regulatory RNAs and non-coding regulatory sequences (Koonin and Wolf, 2010).

1.1.1 EVOLUTIONARY CONSTRAINTS ON CODING SEQUENCES

Sequences encoding structural RNAs or protein coding genes with important cellular/organismal functions are often subject to purifying selection that 'prunes' molecular phenotypes that drift too distantly from the original gene sequences. This is because, to a first approximation, many of the gene sequences of modern organisms encode molecules whose function is nearly 'optimal', having been honed by selection for function over millions of years. In protein sequences, this purifying selection tends to purge nucleotide mutations from the population that would lead to changes to most of the other amino acids (non-synonymous changes) at the majority of positions in a protein coding gene, compared to those mutations that change the codon specified but do not change the amino acid coded (synonymous changes). For example, an examination of

15,350 pairs of human and mouse orthologous genes has revealed an average non-synonymous to synonymous substitution rate ratio of 0.21 (Nei *et al.*, 2010). Similarly, examining ~10,000 orthologous genes from the human, chimpanzee and macaque genomes demonstrates a similar suppressed rate ratio of 0.25 (Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.*, 2007). Assuming that synonymous substitutions are effectively neutral, these results would suggest that nearly 75-80% of non-synonymous mutations are eliminated by purifying selection (Nei *et al.*, 2010). However, there is still a great deal of variation in the rate of evolution across sites in taxonomically diverse and homologous protein sequence alignments. A more nuanced view of protein evolution is that the majority of sites in the sequence are subject to at least some constraints on the identity/properties of an amino acid that can function in that context while relatively few, if any, sites are completely unconstrained.

1.1.2 BIOPHYSICAL CONSTRAINTS ON PROTEIN CODING GENES:

The observation that most protein sequences are subject to purifying selection suggests that there are some functional constraints on the protein sequence. Some of these functional constraints are related to the biophysical features of the folded protein product. For example, constraints on intra- or inter-molecular contacts and molecular dynamics of proteins may exist such that functional interactions with the correct chemical partners are maintained. While these constraints for function are on the molecule as a whole, changes in protein sequences occur discretely at the amino acid level and each site affected occurs within a particular local structural environment. In the context of the final folded protein, the suitability of a particular amino acid change and the strength of the constraint depends on the functional importance of the structural region within which it is located and to what degree the new residue can interact in the same way and/or have the same dynamic properties (e.g. flexibility) as the original amino acid (Figure 1.1 a-d).

In summary, given sufficient evolutionary time, those sites with amino-acid identities that are critical and uniquely important in light of these functional constraints will be strongly conserved, while sites with amino-acid identities that are not will be observed to vary.

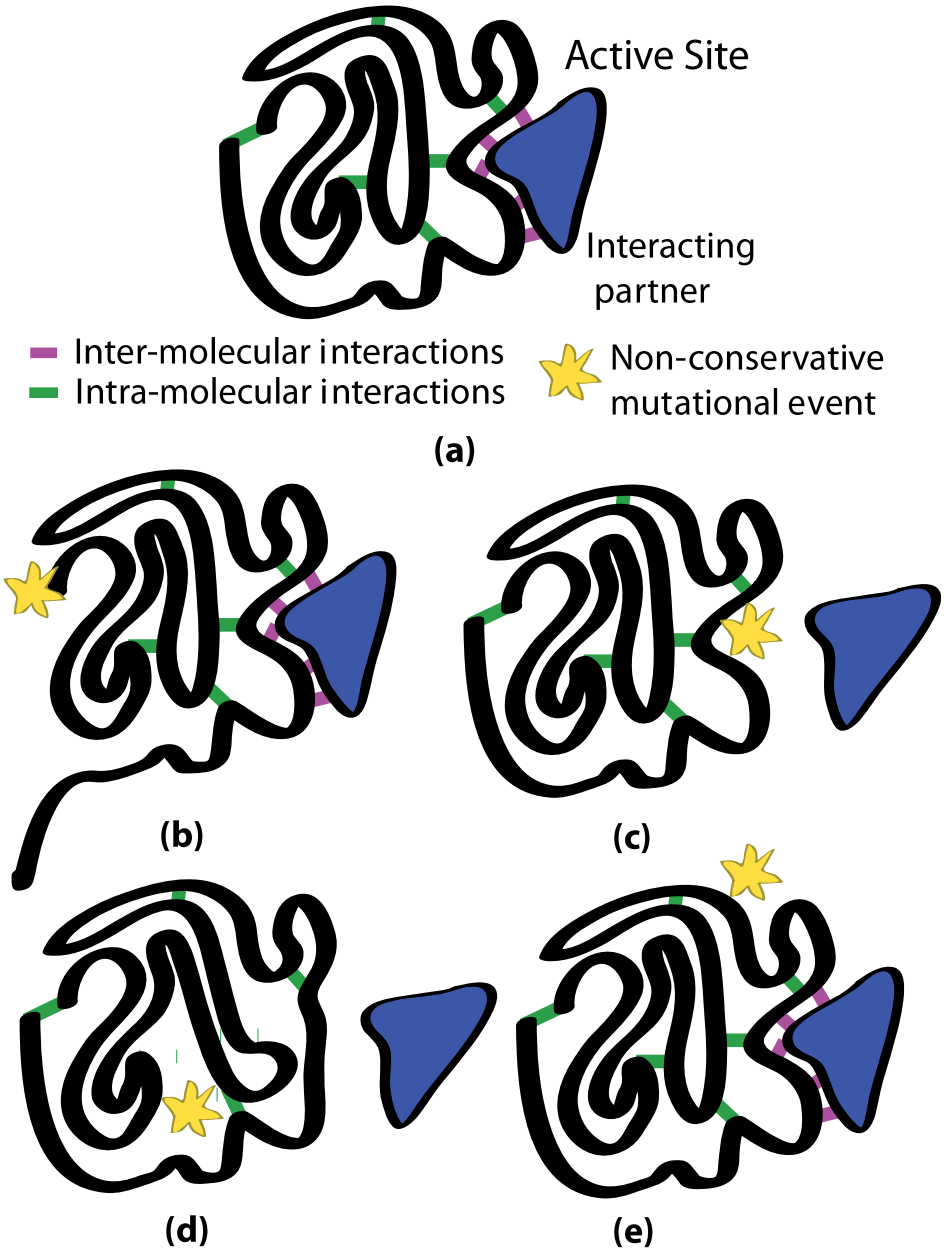


Figure 1.1: Amino acid sequence changes within a protein structure may have varying effects on function.
Amino acids in a protein sequence are under varying selective constraints depending on the relevance of their interactions or local dynamics to the function of the protein. In example (a), we portray a two-dimensional cartoon protein structure exhibiting a varying degree of inter- and intra-molecular contracts. We illustrate the effect of non-conservative mutations to residues that can no longer fulfill the interactions observed in the wildtype protein structure. (b) Mutations located well away from the active site may change the protein structure significantly without affecting function. (c) Mutations affecting residues that directly interact with substrates are usually under a strong purifying constraint due to their direct role in protein function. (d) Mutations within the protein core are much more likely to propagate large changes in protein structure and function than mutations at the surface (e).

The majority of residues in a protein (or homologous sites in an alignment of related proteins with the same function) are not directly involved in interactions with chemical partners. For these sites, the degree of evolutionary constraint is related to the importance of that site for the local or overall structure of the protein. For this reason, solvent accessibility is one of the primary correlates of the rate of evolution at a site (Overington *et al.*, 1992). Residue conservation is much higher in solvent inaccessible regions, where substitutions are likely to affect larger regions of the protein structure, than in solvent accessible regions. These solvent inaccessible regions represent the hydrophobic core of the protein where the satisfaction of hydrogen bonding potential of polar side chains is an important constraint in protein evolution (Worth and Blundell, 2009; Worth and Blundell, 2010). These satisfied polar buried amino acids likely contribute significantly to the overall stability of the protein by holding together distinct secondary structure elements and lending integrity to the overall fold.

In fact, hydrogen bonds between the side chains of amino acids and backbone atoms in the protein are also generally conserved. Amongst these, the amino acid identities of residues involved in interactions between sidechains and main-chain amide groups seem to be more strongly conserved than hydrogen bonds to main chain carbonyl groups (Overington *et al.*, 1990). Unsurprisingly, hydrogen bonds between main chains are also observed, and the sites involved vary a bit more in amino acid composition simply because the side-chain is not directly involved in the bonding interaction (Worth and Blundell, 2010).

These and other biophysical constraints on coding sequences are crucial factors influencing their evolutionary trajectories. However, a complete understanding of protein evolution includes many more factors in addition to the specification of the final protein product. Factors that are relevant include sequence constraints arising from intron splicing (Parmley *et al*., 2007) for reliable gene expression (Drummond *et al.*, 2005; Drummond and Wilke, 2008), correct protein folding (DePristo *et al.*, 2005; James and Tawfik, 2003), protein folding kinetics (Plaxco *et al*, 2000) and the need to avoid opportunistic interactions that detract from, or compete with, the primary protein function (Yang *et al.*, 2012) or protein aggregation (Reumers *et al*., 2009). Protein degradation is also finely controlled, especially with respect to recognizing misfolded proteins that may

arise as a result of mutations (Goldberg, 2003) providing another potential force constraining the evolution of protein sequences.

Here, we propose a framework that incorporates biophysical constraints on the final folded protein product into phylogenetic models that are widely used to infer past evolutionary events. To provide a context for these structurally constrained phylogenetic models and how they fit within the spectrum of available phylogenetic inference methods we introduce some commonly-used methods below.

1.2 BUILDING PHYLOGENETIC TREES FROM MOLECULAR SEQUENCE DATA:

Molecular phylogenetics establishes the historical relationships amongst sequences of DNA, RNA or proteins that reside within organisms or viruses throughout the tree of Life. These phylogenies are often used to reconstruct the evolutionary tree of organisms under the assumption that genetic sequences 'track' organismal history (i.e. that vertical inheritance of genes is the evolutionarily predominant signal). For multicellular eukaryotes, at least, this is probably largely true, although for prokaryotes the assumption is far from proven (Doolittle and Bapteste, 2007). Relationships amongst organisms (taxa) are expressed in terms of a bifurcating tree graph where leaves represent currently existing sequences and internal nodes represent hypothetical last common ancestors. Taxa that group together and share a common ancestral node in the tree that excludes the root of the entire phylogeny are known as 'clades'.

Constructing a phylogenetic tree from molecular sequence data begins with the collection of a set of sequences of interest from the various organisms of interest. These sequences are aligned with homologous sequences that are identified through searching genetic sequence databases such as those housed at the National Center for Biotechnology Information (NCBI at http://www.ncbi.nlm.nih.gov). The phylogenetic reconstruction process depends significantly on correctly determining which sites are positionally homologous in the overall homologous sequences (i.e. the descendant residues/nucleotides at that position in different sequences are positional homologs if they descend from the ancestral sequence at that same position and differ only because of point substitutions that have occurred over the evolutionary tree). However, because homologous sequences will often vary in length because of insertion and deletion events,

sequence alignment programs such as MUSCLE (Edgar, 2004 a,b) or HMMER (Finn *et al*., 2011) are used to estimate a multiple sequence alignment. Each column in a multiple sequence alignment is assumed to include positionally homologous sites. Gaps, corresponding to insertion or deletion events, are added to optimize the score of alignments, although the precise details of how this is done and which methods are most accurate are beyond the scope of this discussion.

1.2.1 PHYLOGENETIC TREES FROM MULTIPLE SEQUENCE ALIGNMENTS:

Interpreting a phylogenetic tree estimated from a multiple sequence alignment requires a deep understanding of the evolutionary process that the molecule in question is undergoing. For example, any misspecification in the model of protein evolution (i.e. large discrepancies between the true process of evolution and the models' process assumptions) may lead to phylogenetic methods to infer incorrect trees.

There are a wealth of different tree estimation methods that vary greatly in their speed and accuracy. Simpler methods rely on low complexity heuristics. For example, Unweighted Pair Group Method with Arithmetic Means (UPGMA) (Sokal and Michener, 1958) and Neighbour joining (NJ) (Saitou and Nei, 1987) methods use an iterative bottom up clustering approach in which the tree is constructed one taxon at a time by choosing the closest sequence according to pairwise 'distances' between sequences in the alignment (Figure 1.2 **a**). Another approach is to explore tree space in search of the tree that best describes the alignment according to some optimality criterion. Simple optimality criteria like 'minimum evolution' (Kidd and Sgaramella-Zonta, 1971) and 'maximum parsimony' (Edwards and Cavalli-Sforza,1963) score alternative trees by computing a simple metric that describes, in some fashion, the fit of the tree to data. The computational simplicity of these metrics comes at the sacrifice of model realism making the trees output by these methods, at best, just good starting points for searching tree space based on more complex optimality criteria.

The current 'gold standard' phylogenetic inference methods are likelihood-based. These methods treat the individual columns in an alignment as independent outcomes of a stochastic process whose probability of occurring under a particular tree topology can be inferred by a parameterized Markov model. Likelihood-based optimality criteria can

be implemented in either maximum likelihood or Bayesian inference paradigms (Figure 1.2 **b**).

LIKELIHOOD-BASED PHYLOGENETICS:

In order to account for the many possible evolutionary paths that could generate the data from a tree, state of the art statistical inference techniques such as maximum likelihood estimation or Bayesian inference are used. When applied to an alignment and given a statistical model of protein evolution, these methods provide estimates for the model's parameters. The 'goal' for both methods is to evaluate various tree topologies in search of the best tree(s) that explain the alignment given a model of sequence evolution. As we have extended the maximum likelihood framework, we focus on this methodology here.

DETERMINING THE MAXIMUM LIKELIHOOD TREE:

Formally, in maximum likelihood phylogenetic estimation, the objective is to search tree space in order to maximize the likelihood of an alignment D containing k sites and N taxa arising from an evolutionary model M, a tree T and a set of model-specific parameters θ.

$$L(\hat{\theta}) = L(D; \hat{T}, \hat{\theta})$$

The calculation of the likelihood is made simpler by the additional assumption that data at each site (D_k) evolves independently (Figure 1.3 a) from the others allowing $L(\hat{\theta})$ to take the form:

$$L(\hat{\theta}) = \prod_k L_k(\hat{\theta}) = \prod_k L_k\left(D_k \middle| \hat{T}, \hat{\theta}\right)$$

The overall probability for a single site in the alignment (the site likelihood) over the tree is then the product of the probability of each substitution event that has occurred over each branch leading to the observed amino acids at the leaves. This calculation is made simpler by the additional assumption that that once two lineages have split along the tree, each branch evolves independently of the others.

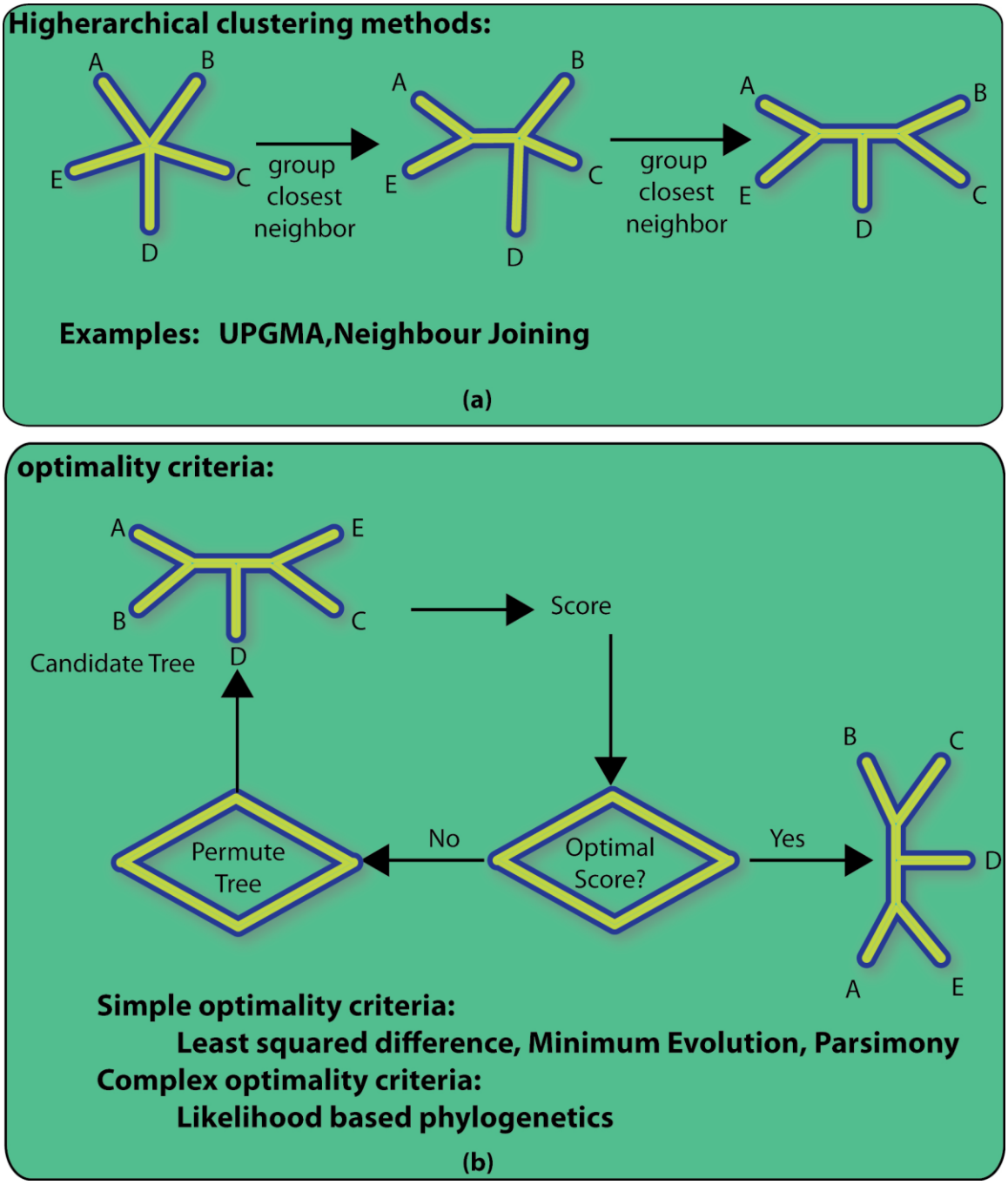


Figure 1.2: Two alternative paradigms for phylogenetic inference.
Phylogenetic trees can be resolved either by an (a) iterative bottom up hierarchical clustering of sequences into groups based on some pairwise distance metric or (b) by exploring several alternative trees in search of the one that best describes the alignment under some optimality criterion.

Since these events and hence states at internal nodes are unknown, all possible amino acid states (*i*) are considered at internal nodes and the overall probability is the sum of the probability of each evolutionary path involving all possible ancestral states at each node. The site likelihood is therefore recursively defined from any internal node v given a set of stationary frequencies π_i for a column *k* in the alignment containing characters D_k:

$$L_k(D_k|M,\theta) = \sum_{i=1}^{20} \pi_i L_k^v(D_k|v=i,\theta)$$

Where $L_k^v(D_k|v=i,T,\theta)$ is recursively defined by:

$$L_k^v(D_k|v=i,\theta) = \left(\sum_{j=1}^{20} P_{ij}(t_1) L_k^{u_1}\left(D_k^{u_1}|u_1=j,\theta\right)\right)\left(\sum_{j=1}^{20} P_{ij}(t_2) L_k^{u_2}\left(D_k^{u_2}|u_2=j,\theta\right)\right)$$

for all nodes v, where u_1 and u_2 are children of v and t_1 and t_2 are the lengths of the branches connecting them to v. $D_k^{u_1}$ and $D_k^{u_2}$ represent the character states (here, amino acids) at the leaves for the sub trees descended from u_1 and u_2 respectively. The recursion terminates at the leaves, which have the amino acid states specified by the data. $P_{ij}(t)$ is the probability of a site being in state *j* after time *t*, given that the process started in state *i* at time 0. For a continuous-time Markov process describing the substitution process, for the *l*'th branch in the tree this probability takes the form $p(j|i,t_l) = [e^{Q|t_l|}]_{ij}$. Therefore, the underlying Markov model (M) is specified by an instantaneous rate matrix **Q** which, varies in size depending on the sequence alphabets being analyzed: DNA models utilize a 4×4 rate matrix, codon models utilize a 61×61 rate matrix and amino acid models utilize a 20×20 rate matrix. An interpretation of the instantaneous rate matrix **Q** is that for some small time interval h, $Q_{ij}h$, is is the approximate probability that residue *i* is substituted with residue *j*.

The complex recursive calculation discussed above, known as the pruning algorithm (Figure 1.3 b), was first introduced by Felsenstein (1981). The location of the root of the tree is typically ignored since widely used phylogenetic models are reversible

and stationary and the direction on the tree in which site probabilities are calculated by the pruning algorithm does not change the final value of the site likelihood.

BAYESIAN PHYLOGENETIC INFERENCE:

Differentiating from the ML approach, in Bayesian phylogenetic inference evidence in favor of certain parameter values θ are considered in light of the posterior distribution $p(\theta|D)$.

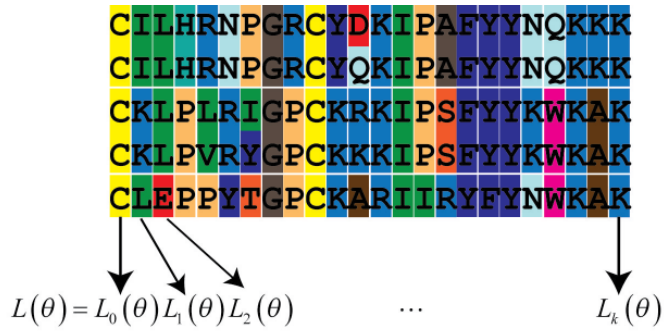
$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

The goal of Bayesian phylogenetic inference is typically to obtain the tree that contains the largest mass in the posterior probability distribution over the space of all possible trees. Obtaining this integral can be difficult, given the high dimensionality of the parameter space and often involves approximations based on Markov Chain Monte Carlo methods. For more details regarding Bayesian phylogenetic inference we refer the reader to Felsenstein (2004).

1.3 MARKOV MODELS OF CODING SEQUENCE EVOLUTION:

Both statistical inference methods discussed above rely on models of protein evolution whose assumptions are supposed to closely match the underlying evolutionary process in order for them to provide accurate phylogenetic estimates. As mentioned in section 1.2.1, the most widely used models of protein evolution assume evolution to be independent across sites and reversible according to a Markov model (M).that relies on an instantaneous rate matrix **Q**.

The reversibility of the underlying Markov model ensures that **Q** can be decomposed into a set of stationary frequencies for the model π as well as a symmetric substitution matrix *S*. π_i represents the stationary or equilibrium amino acid frequency of amino acid *i* that would arise at any site *k* if the Markov process were left evolving for a sufficiently long period of time. Diagonal entries in **Q** are obtained as the minus sum of the off-diagonals for the row and are thus proportional to the rate at which changes leave state *i*. To ensure that the interpretation of an edge length is the expected number of substitutions along that edge, **Q** is then rescaled so that $-\sum \pi_i Q_{ii} = 1$.



(a)

$L_k(\theta)$ as calculated by the pruning algorithm:

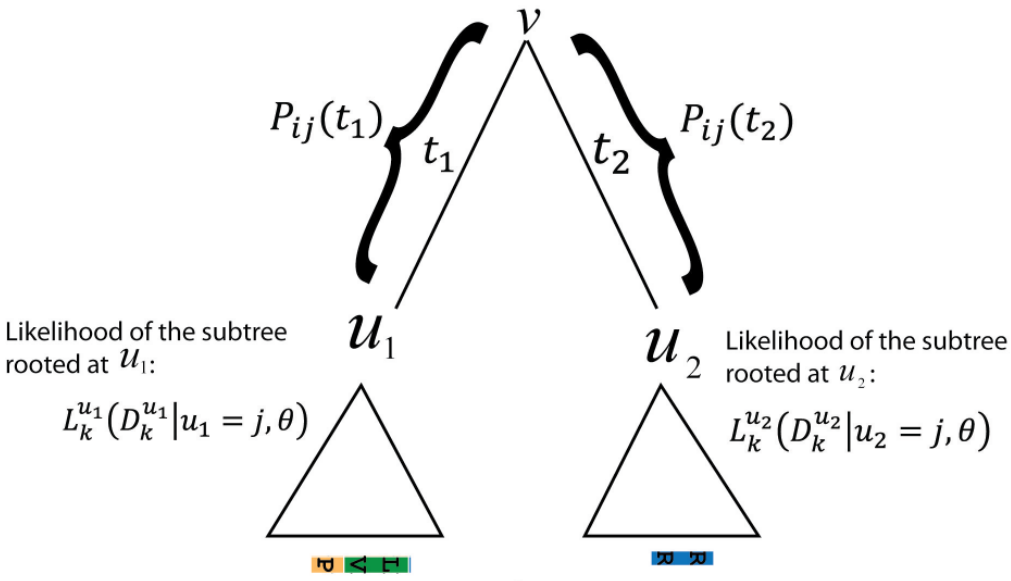
(1) Root node arbetrarily chosen as a starting point for the calculation.
It will be the first node v from which the recursive algorithm is defined.

(2) For each possible aminoacid at the root the likelihood is defined by

$$L_k(\theta) = \sum_{i=1}^{20} \pi_i L_k^v(D_k|v=i,\theta)$$

(3) L_k^v recursively defines the likelihood of the subtree rooted at v as:

$$L_k^v(D_k|v=i,\theta) = \left(\sum_{j=1}^{20} P_{ij}(t_1) L_k^{u_1}(D_k^{u_1}|u_1=j,\theta)\right)\left(\sum_{j=1}^{20} P_{ij}(t_2) L_k^{u_2}(D_k^{u_2}|u_2=j,\theta)\right)$$



(b)

Figure 1.3: Summary of the likelihood calculation.
(a) The likelihood of an alignment under a given phylogenetic model is calculated as the product of the site likelihoods for each column in the alignment. **(b)** The calculation of the site likelihood is recursively defined for all possible evolutionary paths arising from an arbitrarily chosen root. The recursion ends at the leaves, which must assume the observed character state.

Q itself may be empirically derived from sequence alignments or may have parameters to be optimized during phylogenetic inference. Classically, DNA and codon models have typically been highly parameterized while protein models have relied on a more empirical approach. However, both codon and protein models have progressed more recently to semi-empirical approaches that improve common empirical matrices. We briefly examine DNA and codon models before focusing more deeply on amino acid models as this thesis primarily concerns the development of a parameterized amino acid model.

1.3.1 DNA MODELS OF EVOLUTION:

The most general stationary reversible model of DNA, the general time reversible model (GTR), includes 4 stationary nucleotide frequency parameters π_j, 6 exchangeability parameters between nucleotides (lower diagonal of **S**) and a rate heterogeneity parameter in θ. The rate determines the finite set of evolutionary rates under which the sites in the data (columns in the alignment) will be modeled. Early models of nucleotide evolution were special cases of the GTR model and the more complex models, including GTR itself, were proposed later. The simplest model is the Jukes and Cantor model (1969) which assumes stationary frequencies and uniform mutation rates. The Kimura 2-parameter model (1980) includes 2 parameters α and β allowing the transition rate to vary relative to the transversion rate. The Felsenstein-81 (F81) (Felsenstein, 1981) and Hasegawa-Kishino-Yano-85 (HKY85) (Hasegawa *et al.*, 1985) models of DNA evolution extend simpler models by allowing base stationary frequencies to vary. F81 is simply an extension of the Jukes and Cantor model while HKY85 extends the Kimura two parameter model.

1.3.2 CODON MODELS OF PROTEIN EVOLUTION:

When the DNA sequences being analyzed encode proteins it is more realistic to model the data as the codon triplets that each specify one amino acid in the final protein sequence.

INCORPORATING GENERAL BIOPHYSICAL CONSTRAINTS: GOLDMAN AND YANG (1994)

Codon models first proposed by Goldman and Yang (1994) employ a 61×61 rate matrix Q_{ij} corresponding to the exchangeability between codon triplets $i = i_1 i_2 i_3$ and $j = j_1 j_2 j_3$. Starting from the Kimura two parameter model, codon exchangeabilities are adjusted by a factor $e^{cd_{ij}}$ in S_{ij}. Here, d_{ij} accounts for the differences in the physiochemical properties of the resulting substituted amino acid (Grantham, 1974) and c is a parameter to be optimized during statistical inference.

$$Goldman\ and\ Yang\ (1994): \begin{cases} \pi_1 \neq \pi_2 \neq \pi_3 \ldots \neq \pi_{61} \\ S_{ij} = \begin{bmatrix} 0 & more\ than\ 1\ codon\ position\ affected \\ \mu e^{cd_{ij}} & a\ transversion\ has\ occurred \\ \mu\kappa e^{cd_{ij}} & a\ transition\ has\ occurred \end{bmatrix} \end{cases}$$

Codon exchangeabilities can be grouped into those that do not result in an amino acid changing substitution (synonymous) and those that do (non-synonymous) and a relative rate parameter can be associated with each type (Muse and Gaut, 1994). Yang *et al*. (1998) went one step further in a model (Rev0) by including a parameter (ω) that captures the non-synonymous to synonymous rate ratio and κ, from the $HKY85$ model, representing the transition to transversion rate.

$$\text{Yang } et\ al.\ (1998) \begin{cases} \pi_1 \neq \pi_2 \neq \pi_3 \ldots \neq \pi_{61} \\ S_{ij} = \begin{bmatrix} 0 & if\ more\ than\ 1\ codon\ position\ has\ changed \\ 1 & if\ a\ synonymous\ transversion \\ \kappa & if\ a\ synonymous\ transition \\ \omega & if\ a\ non\ synonymous\ transversion \\ \kappa\omega & if\ a\ non\ synonymous\ transition \end{bmatrix} \end{cases}$$

1.3.3 AMINO ACID MODELS OF PROTEIN EVOLUTION.

While parameter-rich Markov models have often been used to study DNA and codon evolution, modeling of amino acid evolution has typically been conducted via an empirical approach. Substitution models have classically been developed by counting observed amino acid changes between closely related sequences in large sequence databases. The first such substitution model was the accepted point mutation (PAM) matrix (Dayhoff and Eck, 1968; Dayhoff *et al.*, 1979; Dayhoff *et al.*, 1983), which was derived from relative numbers of different amino acids aligned to each other. A model

derived using similar a framework includes the widely used Jones-Taylor-Thornton (JTT) model (Jones *et al.*, 1992), a matrix derived from a larger set of protein sequences.

In contrast to substitution models generated by counting the observed frequencies of interchanges between residues in an alignment, more recent methods to estimate empirical substitution matrices utilize a general time reversible model of protein evolution where all parameters in the model (including S_{ij} entries, $\pi's$, the tree topology and the branch lengths) are estimated from the database of aligned protein families. The goal is to obtain a general **S** matrix and π vector through maximum likelihood estimation in a GTR model framework considering all of the alignments and pre-estimated fixed trees in the database jointly. The resulting substitution model, **Q** (where **Q** = **S** π), is then used as a general fixed model of amino acid interchange that can be applied to potentially any protein family to estimate a phylogeny (Figure 1.4 a).

Due to the complexity of the GTR models, simplifying approximations are often necessary in the calculations that utilize these models. By relaxing the requirement for the estimation of an optimal tree topology and using instead approximate phylogenies derived by neighbor joining (with branch lengths approximated under the JTT model), a more accurate substitution model – the Whelan-and-Goldman model (WAG) – has been described (Whelan and Goldman, 2001). The current 'state-of-the-art' general model of protein evolution, the Le and Gascuel matrix (LG) (Le and Gascuel, 2008), was estimated using a method similar to that of Whelan and Goldman but also accounting for the variability of evolutionary rates across sites during the parameter estimation. This improvement, along with a much larger dataset comprising 50000 sequences and 6.5 million sites utilized by these authors, resulted in a significant improvements in model fit.

MODIFYING GENERALIZED AMINO ACID SUBSTITUTION MODELS:

Markov models such as JTT, LG or WAG are often implemented in the maximum likelihood framework along with a few additional model features that are known to improve the fit to data. The likelihood for a site is usually evaluated as a mixture model averaging over the site probability evaluated under multiple rate categories (Yang, 1994) derived from a discretized gamma distribution whose shape parameter $\hat{\alpha}$ is optimized by ML for all sites in the alignment. This mixture may also include an 'invariable sites'

component that accounts for complete conservation of a proportion of sites in the alignment; this proportion is also estimated by ML (Fitch and Margoliash 1967; Fitch 1986; Shoemaker and Fitch 1989). The **Q** matrix of the Markov model is also often adjusted be a product of the empirical exchangability matrix (**S**) from JTT, LG or WAG and a $\hat{\pi}$ vector approximated by the observed proportion of each amino acid in the alignment under analysis (Cao *et al*., 1994).

1.3.4 SPECIALIZED AMINO ACID RATE MATRICES AND THE INNOVATION OF SEMI-EMPIRICAL CODON MODELS:

Models of both amino acid and codon substitution have undergone significant increases in complexity in recent years. Generalized amino acid substitution models have been estimated for proteins encoded by genes on mitochondrial and plastid genomes (Adachi and Hasegawa, 1996; Adachi *et al.*, 2000) as well as for viral proteins (Dimmic *et al.*, 2002; Dang *et al.,* 2010). Others have attempted to develop mixture models incorporating information about physico-chemical similarity (Koshi and Goldstein 1995; Dimmic *et al.*, 2000). Newer 'general' codon models, which were traditionally parameterized like the simpler DNA models, have recently been estimated using an empirical approach. For example, Schneider *et al.* (2005) estimated a 'general codon model' based on a dataset of 8.3 million aligned codons from vertebrates.. This work has spurred the development of a number of semi-empirical codon models that combine these empirically derived substitution models with parameters derived from classic parametric codon models (Doron-Faigenboim and Pupko, 2007; Kosiol *et al.*, 2007; Zoller and Schneider, 2012).

STRUCTURALLY CONSTRAINED MODELS OF CODING SEQUENCE EVOLUTION

Ideally, protein evolutionary models should take into account the selection pressures that preserve the large scaffold of residue interactions that position critical residues to interact with substrates, other proteins or co-factors. Indeed, a multitude of structurally constrained models of protein evolution (SCPE's) have been proposed that attempt to incorporate some form of a protein structure-based constraint into various phylogenetic models of protein evolution (Fornasari *et al.*, 2002; Parisi and Echave, 2004; Parisi and Echave, 2005; Fornasari *et al*., 2007; Juritz *et al.*, 2012; Robinson *et al.*,

2003; Rodrigue *et al.*, 2005; Rodrigue *et al*., 2006; Kleinman *et al*., 2006; Rodrigue *et al.*, 2009; Bonnard *et al.*, 2009; Kleinman *et al.*, 2010). There are several flavours of structurally constrained models that differ in how they account for site-wise interdependence.

A variety of amino acid substitution matrices have been produced for different secondary structure and solvent accessibility characters (Lüthy *et al.* 1991; Overington *et al.*, 1992; Koshi and Goldstein, 1995; Goldman *et al.*, 1998; Le and Gascuel, 2010; Liò and Goldman, 2002) and even for transmembrane proteins (Jones *et al.*, 1994 a,b*).* Recently, structure-based substitution models (SSM) have been incorporated into partitioned models and parametric mixture models (Le and Gascuel, 2010), and appear to lead to a dramatic improvement in the fit to real data relative to the simpler empirical substitution models such as LG (Figure 1.4 b).

Simulation energy phylogenetic models (SEPM) attempt to incorporate structural constraints into popular phylogenetic methods by the estimation of site-specific substitution models intended to capture the effects of the structural environment found at each site (Fornasari *et al.*, 2002; Parisi and Echave, 2004; Parisi and Echave, 2005; Fornasari *et al*., 2007; Juritz *et al.*, 2012). These methods are based on simulating protein evolution under a structural constraint and are used to generate large alignments that are then used in the estimation of a substitution model (Figure 1.4 c). The structural constraint typically enters the model through the use of energy potentials measuring the free energy of particular sequence-structure fits that is incorporated into a function for accepting or rejecting proposed amino acid changes during the simulation.

Dependence energy models (DEM) models attempt to explicitly account for the interdependencies amongst sites due to a structural constraint by replacing the 20×20 amino acid rate matrix by a rate matrix meant to denote the exchangeability amongst sequences of length n from the set of all possible sequences of dimension $4^n \times 4^n$ for DNA models or $20^n \times 20^n$ for amino acid models (Robinson *et al.,* 2003; Kleinman *et al.*, 2010; Bonnard *et al.*, 2009; Rodrigue *et al.*, 2005; Rodrigue *et al.*, 2005). While these DEM models can explicitly account for the underlying dependence between amino acids, they are computationally complex when compared to either SSM or IEM models (Figure 1.4 d).

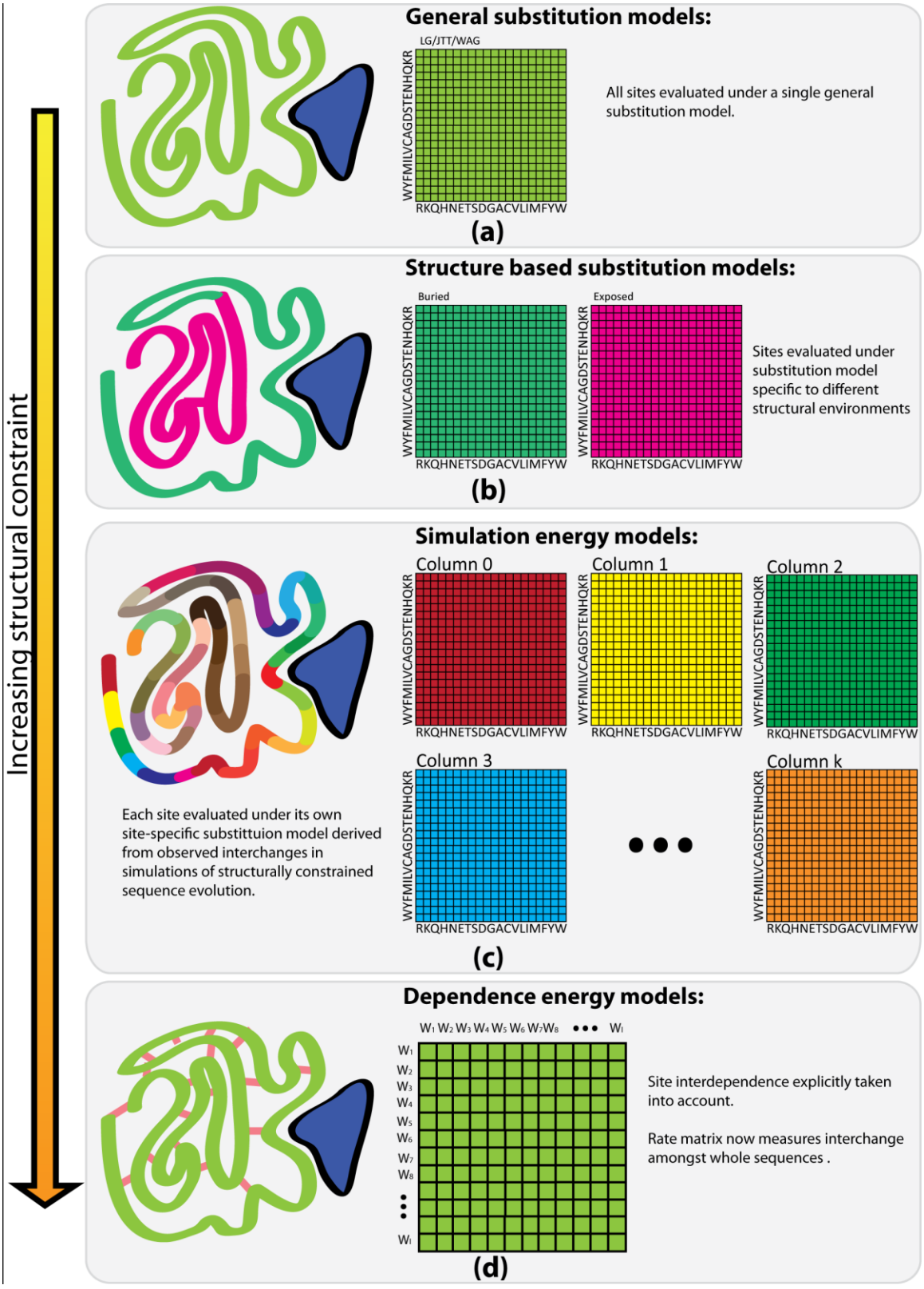

Increasing structural constraint
General substitution models:
LG/JTT/WAG
WYFMILVCAGDSTENHQKR
RKQHNETSDGACVLIMFYW
All sites evaluated under a single general substitution model.
(a)
Structure based substitution models:
Buried
Exposed
WYFMILVCAGDSTENHQKR
RKQHNETSDGACVLIMFYW
Sites evaluated under substitution model specific to different structural environments
(b)
Simulation energy models:
Column 0
Column 1
Column 2
Column 3
Column k
Each site evaluated under its own site-specific substittuion model derived from observed interchanges in simulations of structurally constrained sequence evolution.
(c)
Dependence energy models:
W1 W2 W3 W4 W5 W6 W7 W8 ••• Wl
Site interdependence explicitly taken into account.
Rate matrix now measures interchange amongst whole sequences .
(d)

Figure 1.4: Phylogenetic models that account for varying degrees of structural constraint.
Phylogenetic models come in varying degrees of structural constraint. (a) phylogenetic models incorporating general substitution models such as JTT, LG and WAG assume that all sites are modeled by a general set of physico-chemical constraints. (b) Structure-based substitution models (i.e. Le and Gascuel, 2010) may provide substitution models specific to structural environments. For example, separate substitution models could be made for exposed and buried sites. (c) Simulated protein evolution under a structural constraint can be used to create site-specific substitution models (i.e. Parisi and Echave, 2001). (d) Dependency energy models explicitly model site-wise interdependence by using a rate matrix that models the interchange between entire sequences $W_1, W_2, W_3, W_4 \ldots W_l$.

INCORPORATING BIOPHYSICAL CONSTRAINTS INTO PHYLOGENETIC MODELS:

A detailed introduction to, and discussion of, structurally constrained models will be provided in chapter 2 where our own structurally constrained framework is introduced. However, given that all structurally constrained phylogenetic models incorporate biophysical constraints using existing *in silico* methods for estimating energy potentials, what follows is a brief introduction to how these potentials are derived. There are three general types of potentials that could be used to measure the effect of single to multiple amino acid substitutions at sites in protein structures: statistical potentials, empirical force fields and purely physical force fields.

STATISTICAL POTENTIALS:

A statistical, or knowledge-based, potential is an energy function that traditionally uses the Boltzmann law to convert observed frequencies of interactions in protein structure databases into potentials (Sippl, 1993). These potentials are obtained as a function of the ratio of observed to expected frequencies, where expected frequencies are derived from a hypothetical reference state where no interactions occur. Single body potentials, such as the solvation potential rely on the distance of a residue to some external field. Pairwise or multi-body potentials are based on the frequency of occurrence of pairs of amino acids or groups of amino acids. As an example, the interaction potential of amino acids *i,j* distance *r* apart can be calculated as:

$$\Delta E^{ij}(r) = -kTln[f^{ij}(r)/f^{s}(r)]$$

Where k is the Boltzmann constant, T is the temperature $f^{ab}(r)$ is the frequency of observing amino acid i and j distance r apart and $f^{s}(r)$ is the average frequency of the reference state.

An excellent summary of the theory backing statistical potentials and a comparison amongst various statistical potentials can be found in Sippl (1995) and Rykunov and Fiser, (2010). Some popular statistical potentials that have been employed in phylogenetic models are Prosa pairwise and solvation potentials (Sippl, 1993), Bastolla contact potentials (Bastolla *et al.*, 2001) and a series of potentials derived in Bonnard *et al.* (2009).

EMPIRICAL AND PHYSICAL FORCE FIELDS:

A more robust methodology for predicting the stability of a protein structure is through the use of physical potentials available in a variety of molecular dynamics packages (Bash *et al.*, 1987; Pitera and Kollman, 2000; Prevost *et al.,* 1991). However, these potentials have, so far, been avoided by the phylogenetics community largely due to the extreme computational complexity required to determine ΔG. Alternatively, empirical force fields such as FoldX (Guerois *et al.*, 2002) serve as a kind of compromise by combining a physical description of the interactions with experimental data on how mutations in specific structural environments have affected ΔG. The FoldX empirical force field has a complex physical description that incorporates energy terms associated with van der Waals interactions, solvation energy, hydrogen bond formation, water bridges, electrostatic contribution of charge group interactions, entropy costs for fixing the backbone and entropy costs for fixing the side chain (Guerois *et al.*, 2002). In producing the final potential, the weights for many of these terms were optimized by utilizing the ProTherm database that contains thermodynamic information for proteins and their mutants (Kumar *et al.*, 2006). As a result FoldX is optimized to predict the mutational effect of single or multiple substitutions on a wild type protein, making it ideal for approximating the relative exchangeabilities of amino acids at a site. It has since been validated through an analysis of rhodopsin in which a highly significant correlation between FoldX energy changes and the average age of night blindness and daytime vision loss onset was found (Rakoczy *et al.*, 2011).

1.3.5 SYNOPSIS OF THE MODELS AND RESULTS INTRODUCED IN THIS WORK:

Here we formalize and extend the independence energy model framework. In the second chapter, we introduce our model and evaluate its performance across 48 datasets containing sites located in a wide variety of structural environments. By implementing a series of structurally constrained partition models, we determine if model fit can be improved significantly by allowing the structural constraint to vary across different secondary structure and solvent exposure categories. Then, we evaluate whether a standard general amino acid model of protein evolution, such as LG or JTT, is preferred

for each secondary-structure/solvent-exposure category or if perhaps an appropriate structurally constrained substitution model performs better (Le and Gascuel, 2010).

CHAPTER 2 SITE INDEPENDENT STRUCTURALLY CONSTRAINED PHYLOGENETIC MODELS THAT FLEXIBLY FIT DIFFERENT STRUCTURAL ENVIRONMENTS.

2.1 INTRODUCTION:

An examination of any set of homologous protein sequences reveals that the evolutionary process leads to a diverse distribution of sequence patterns across sites. The observed exchangeability between amino-acids at a site in an alignment is influenced by several factors including: (1) The ease with which one codon can be converted to another arising from the number and type of substitutions required to move between them; (2) Codon usage biases that vary from species to species (Miyata *et al.,* 1979, Grantham *et al.*, 1980) and arise from differential protein expression or differences in the availability of translational machinery components (Andersson and Kurland 1990; Sharp *et al.*, 1993; Akashi and Eyre-Walker, 1998; Willie and Majewski, 2004; Sharp *et al.*, 2005); and (3) The exposure of a mutation to purifying or positive selective forces that preserve or alter protein structure or function.

It is possible to generalize the effects of codon exchangeabilities and codon usage biases to the rates of amino-acid interchange across sites in an alignment. However the physicochemical constraints on protein sequences visible to selection are site-specific and frequently ignored in phylogenetic Markov models of amino acid replacement. Probabilistic phylogenetic methods evaluate the likelihood of an alignment given a Markov model with parameters that include the topology and branch lengths of the tree, and the shape parameter of a gamma distribution that accounts for differing rates across sites (Yang, 1994). In these models, sites are usually treated as independent and identically distributed, with observed exchangeabilities amongst amino acids modeled by general rate matrices (e.g. the Jones-Taylor and Thornton (JTT), Whelan and Goldman (WAG) and Le and Gascuel (LG) models), meant to reflect average rates of amino acid replacement, derived from large databases of aligned protein families (Jones *et al.,* 1992; Le and Gascuel, 2008; Whelan and Goldman, 2001).

The foregoing Markov models are specified by an 20X20 instantaneous rate matrix $\mathbf{Q}$, whose off-diagonal entries $Q_{ij}h$, are interpretable as the approximate probability that amino acid i is substituted with j in a small interval of time h. A

conditional probability matrix, with entries P_{ij} , over a branch of length l can then be calculated by $P_{ij}(l) = [e^{Ql}]_{ij}$. The overall probability for a single site in the alignment (herein referred to as the site likelihood) over the tree is then the product of the probability of each possible substitution event that could have occurred over each branch leading to the observed amino acids at the leaves. Because these events and hence states at internal nodes are unknown, all possible states are considered at internal nodes and the overall probability is the sum of the probability of each evolutionary path involving all possible ancestral states at each node. This complex calculation is accomplished via an efficient 'pruning algorithm' that was first introduced by Felsenstein (1981).

Since evolution is treated as independent across the sites in the sequence the overall likelihood for an alignment and tree, $L(\theta)$, is the product of the foregoing sitewise probabilities (Equation 1).

(1) $L(\theta) = \prod_k L_k(\theta)$

Where θ denotes all unknown parameters and $L_k(\theta)$ is the k'th site likelihood or, equivalently, the probability of the data at the k'th site.

Although these models are very useful approximations, particular amino-acid substitutions at different sites are known to have different impacts on protein activity and/or stability. Therefore model realism demands rate matrices that accurately describe the fitness effects of mutations at particular sites. As homologous proteins are often structurally very similar (Sander and Schneider, 1991) it is possible to estimate the impact of substitutions at individual sites that simply perturb the stability of protein folds and to generalize this information to all sequences in an alignment to generate 'structurally constrained' phylogenetic models.

Structurally constrained phylogenetic models alter Q in a site-specific way, employing a function $F(\Delta G_{S^x}, \Delta G_{S^y})$ that expresses a relationship between the folding energy of two sequences S^x and S^y and the instantaneous rate of exchange between them. Approximation of ΔG for a particular protein of known 3-dimensional structure can be arrived upon either through the use of statistical potentials (Sippl, 1993; Hamelryck *et al.*, 2010), empirical effective energy potentials (Guerois *et al.,* 2002; Yin

*et al.,*2007; Johnston *et al.*, 2011) or physical potentials (Brooks *et al.*, 1983; Van Der Spoel *et al.*, 2005; Bueno *et al.*, 2007; Benedix *et al.*, 2009).

A multitude of structurally constrained phylogenetic models (SCPM's) have been proposed that incorporate some form of a protein structure constraint into their framework. These fall into three general classes that differ in their approach to site-wise interdependence: (1) Structure-based substitution models (SSMs) with distinct amino-acid exchangeability matrices for residues located in different structural environments estimated from a database of alignments with assigned structural classes (Le and Gascuel, 2010), (2) Simulation energy phylogenetic models (SEPMs) where sequences are simulated under a conventional independence model but those drifting too far from wildtype folding energy are ignored (Parisi and Echave, 2001), giving rise to site-specific rate matrices based on the observed accepted substitutions that occurred during simulation (Fornasari *et al.,* 2002; Parisi and Echave, 2004; Parisi and Echave, 2005; Fornasari *et al.,* 2007; Juritz *et al.*, 2012) and (3) dependence energy models (DEM) models that attempt to explicitly account for the interdependencies amongst sites due to a structural constraints by estimating a rate matrix representing the exchangeability amongst entire sequences of length N (for nucleotides this matrix is $4^N x 4^N$, and for amino acids $20^N x 20^N$) from the set of all possible sequences (Robinson *et al.*, 2003; Rodrigue *et al.*, 2005; Rodrigue *et al*., 2006; Kleinman *et al.*, 2006; Rodrigue *et al.*, 2009; Bonnard *et al.*, 2009; Kleinman *et al.*, 2010). While these DEM models can explicitly account for the underlying dependence between amino-acids in tertiary structures, they are extremely computationally complex when compared to either SSM or SEPM models and require the application of methods that approximate the rate matrices. Here, we propose a novel and computationally tractable SCPM framework that maintains the computational simplicity of SEPM but that incorporates some of the advantages of SSM and DEMs.

The simplest structurally constrained phylogenetic models of protein evolution are those based on structure-based substitution models. Here, alignments containing at least one sequence of known structure have been used to build different substitution models specific to unique structural environments. The latter models have been incorporated into a several phylogenetic estimation programs (Goldman *et al.*, 1998; Le and Gascuel, 2010). In a recent example (Le and Gascuel, 2010), 11 structural

environment specific substitution models were created from a dataset of alignments collected from the HSSP database (Schneider and Sander, 1996). Three sets of models were derived from this dataset including: i) a solvent-exposed and a buried pair of matrices, ii) a matrix for each of the three secondary structure categories (α-helix, β-sheet and other) and, iii) six substitution matrices representing each of the possible combinations of solvent exposure and secondary-structure categories (Exposed-α-helix, Exposed-β-sheet, Exposed-other, Buried-α-helix, Buried-β-sheet, Buried-other). Their work demonstrated statistically significant improvement in model fit when using a variety of partition and mixture models based on these categorical substitution models.

Simulation energy phylogenetic models (SEPMs), are a type of SCPM that incorporate a site-specific structural constraint by tracking the energies of sequences simulated under an independence model (Parisi and Echave, 2001). SEPMs (Fornasari *et al.,* 2002; Parisi and Echave, 2004; Parisi and Echave, 2005; Fornasari *et al.,* 2007; Juritz *et al.*, 2012) employ site-specific rate matrices constructed from proportions of observed substitutions in simulations of structurally constrained protein evolution. Starting from a protein of known structure, each evolutionary time step of the simulation first mutates the sequence, and then assigns a distance score S to the sequence that is used to modify the probability of accepting the sequence into an alignment that is then used to infer a site-specific $Q_{ij}^{(k)}$ (Parisi and Echave, 2004). Since not all substitutions are observed, the site-specific rate matrix implied by the energy constraint alone (${Q_{ij}}^{E^{(k)}}$) is corrected by an established substitution model such as JTT.

The dependency energy model (DEM) approach is more computationally complex than either the SSM or IEM frameworks (Robinson *et al.,* 2003; Rodrigue *et al.*, 2005; Rodrigue *et al.*, 2006; Kleinman *et al.,* 2006; Rodrigue *et al.*, 2009; Bonnard *et al.*, 2009; Kleinman *et al.*, 2010). Rodrigue and colleagues (2005) developed an amino-acid based DEM where the Markov generator is $20^N X 20^N$. In their Bayesian implementation (equation 2), the rate matrix takes the form:

$$(2)\ R_{s^x s^y} = \begin{cases} -\sum_{s^y \neq s^x} R_{s^x s^y}, & if\ s^y = s^x \\ Q_{xy} e^{\beta F(\Delta G_{s^x}, \Delta G_{s^y})}, & if\ s^y \neq s^x \\ 0, & Otherwise \end{cases}$$

Here, each column of the rate matrix corresponds to a particular sequence and the columns range over all possible sequences. As is usually the case with rate matrices, $R_{s^x s^y} t$ is interpretable as the approximate probability that sequence s^x will evolve into sequence s^y in a small interval of time t. Sequences differing by multiple substitutions are ignored leading to large numbers of entries in the rate matrix that are 0. With such large rate matrices, calculation of substitution probabilities based on eigen-value decomposition techniques are infeasible. Instead approximations to entries in this matrix are made possible by employing Markov chain Monte Carlo (MCMC) techniques. For these studies a simple pairwise 'contact' energy potential matrix was used to approximate changes in free energy associated with amino acid substitutions for a given fixed 3-dimensional structure of a single representative sequence.

2.2 MATERIALS AND METHODS:

2.2.1 MODEL DESCRIPTION

The foregoing SSM, SEPM and DEM frameworks rely on the idea that amino-acid exchangeabilities at a site are directly dependent on the change in the functional 'activity' associated with substituting some amino acid *i* to another amino acid *j* at that site. While both the SEPM and DEM approaches account for dependencies amongst sites, here, instead, we introduce a computationally tractable independence energy model (IEM) that requires only a measure of amino-acid suitability at each site. Similar to the frameworks discussed above, our model assumes that the exchangeability between amino-acids *i* and *j* is related to the change in free energy of the transition by a function proportional to $e^{-\tau \Delta\Delta G_{ij}}$ where τ is a weighting parameter. Our model differs from SSM by utilizing predicted free energy changes associated with substitutions at individual sites in the specific structures under consideration and does not rely on simulation, instead incorporating free-energy change directly into a Markov model of sequence change. Finally, it differs from the DEM approach by keeping the simplifying assumption of independence between sites, while utilizing more sophisticated energy-potential calculations for predicting site-specific free energy changes.

We consider four structurally constrained partitioned Markov models of protein evolution that preserve the independence across sites assumption. As the model describes

amino acid interchange, rate matrices are 20X20 and, although these matrices vary between sites, direct calculation of substitution probabilities through eigen-decomposition is possible. These four models differ in complexity with regards to the number of different structural environments that are taken into account.

The first model (equations 3-4) pools all sites together in a single class. Similar to the Rodrigue and colleagues (2005) model (equation 2) we use a general substitution model, in this case LG or JTT, to form the basis of the model $Q_{ij}^{(B)}$ for each site. This matrix is adjusted by multiplying its entries by $Q_{ij}^{E^{(k)}}$, a rate matrix encompassing the energy contribution to exchangeabilities that depends on the particular site *k* under consideration. The structural constraint first enters the model through a function $E_j^{(k)}(\Delta G_j, \boldsymbol{\tau}) = \tau_1 \Delta G_{j_1} + \tau_2 \Delta G_{j_2} \ldots + \tau_K \Delta G_{j_K}$ where $\boldsymbol{\tau} = [\tau_1, \tau_2 \ldots \tau_K]$ is a parameter vector governing the weight associated with various alternative approximations of the change in folding energy terms $(\Delta G_{j_1}, \Delta G_{j_2} \ldots \Delta G_{j_K})$ at a site. We abbreviate $E_j^{(k)}(\Delta G_j, \boldsymbol{\tau})$ as E_j and introduce the term $E_{ij} = E_j - E_i$ to represent the overall energetic consequences of substituting amino acid *i* to amino acid *j*. We have experimented with various alternative definitions for ΔG_j using statistical potentials coming from the PROSA package ($K = 1$), empirical force fields from FoldX ($K = 1$) or a combination of both ($K = 2$). The parameter f_j is included to allow variation in the average site-specific stationary frequencies.

Model 1: a general structurally constrained model of protein evolution (LG+ΔG; JTT+ΔG)

(3) $Q_{ij}^{(k)} \propto Q_{ij}^{E^{(k)}} Q_{ij}^{(B)}$

(4) $Q_{ij}^{E^{(k)}} \propto f_j e^{-E_{ij}^{(k)}}$

An examination of $\Delta G_j{}^{(k)}$ over sites having different structural environments reveals that the median value of $\Delta G_j{}^{(k)}$ varies amongst the various structural environment partitions for both PROSA and FoldX energies (supplementary Figure A1 for PROSA energies and supplementary Figure A2 for FoldX energies). It is clear from both of these potentials that substitutions in solvent exposed regions result in lower energies for hydrophilic amino acids while substitutions in buried regions result in lower energies for

hydrophobic amino acids. Furthermore, substitutions in the hydrophobic core of the protein structure are more likely to increase ΔG_j. Similarly, an examination of ΔG_j values across different secondary structures demonstrates that some residues, such as glycine and proline, have much lower ΔG_j values in loops than in beta-sheets and helices.

To take advantage of changes in the behavior of our energy potentials amongst protein structure categories, we further developed our model by allowing structural constraints to vary amongst these different environments (equations 5-6). We partition datasets into P different structural classes and estimate parameters separately for each class. The three structurally constrained partitioned models we develop are: (i) a 2 partition model 2P/2P_LG, which segregates residues into exposed and buried categories based on their relative solvent accessibility, (ii) 3P/3P_LG, which segregates residues by their secondary structure into extended/helix/other categories; and (iii) 6P/6P_LG, which segregates residues by solvent accessibility and secondary structure (see methods for a more detailed description). The partitioning here is the same as that employed by Le and Gascuel (2010) and the nomenclature here reflects models that incorporate the appropriate structure-based substitution model (2P/3P/6P) in $Q_{ij}^{(B)}$ as opposed to models that use LG for each partition (2P_LG/3P_LG/6P_LG).

Model 2/3/4: Structurally constrained partition models of protein evolution

(5) $Q_{ij}^{(k)} \propto Q_{ij}^{E^{(k,P)}} Q_{ij}^{(B)}$

(6) $Q_{ij}^{E^{(k,P)}} \propto {f_j}^P e^{-E_{ij}^{(k,P)}}$

Here, P indicates the structural partition that site k belongs to and $E^{(k,P)}$ depends on P because the $\boldsymbol{\tau}$ weights in are allowed to vary across partitions. The partitions are defined as follows:

$Partition\ (2P) = [\text{Exposed}, \text{Buried}]$

$Partition\ (3P) = [\text{Alpha Helix}, \text{Beta Sheet}, \text{Other}]$

$Partition\ (6P)$

$$= \begin{bmatrix} \text{Exposed and Alpha helix}, \text{Exposed and Beta sheet}, \text{Exposed and Other} \\ \text{Buried and Alpha helix}, \text{Buried and Beta sheet}, \text{Buried and Other} \end{bmatrix}$$

All of the models considered here are time-reversible. Defining $\pi_j^{(B)}$ as the stationary frequency of j for the base model, this can be seen by considering the ratio of rates of amino acid exchange:

$$(7)\ \frac{Q_{ij}}{Q_{ji}} = \frac{Q_{ij}^{E^{(k)}} Q_{ij}^{(B)}}{Q_{ji}^{E^{(k)}} Q_{ji}^{(B)}} = \left[\frac{f_j e^{-2E_j^{(k)}}}{f_i e^{-2E_i^{(k)}}}\right]\left[\frac{\pi_j^{(B)}}{\pi_i^{(B)}}\right]$$

Since the ratio of the entries in the rate matrices can be expressed in the form $\frac{\alpha_j}{\alpha_i}$, it follows that the model is time reversible. Moreover the site-specific stationary frequencies $\pi_j^{(k)}$ for the model are equal to α_j, allowing them to take the form:

$$(8)\ \pi_j^{(k)} \propto f_j e^{-2E_j^{(k)}} \pi_j^{(B)}$$

for non-partitioned models and for partitioned models:

$$(9)\ \pi_j^{(k)} \propto {f_j}^P e^{-2E_j^{(k,P)}} \pi_j^{(B)};$$

Finally, each model is implemented with a single rates-across-sites (RAS) mixture model for which the relative rates of evolution for sites are assumed to be independent and identically distributed from a discretized version of a Γ distribution with 8 rate categories defined by a single shape parameter α as in Yang (1994).

2.2.2 PARAMETER ESTIMATION:

Most of the parameters in the model including the edge lengths, the α parameter from the Γ distribution and **τ** are estimated through maximum likelihood. Multidimensional parameter optimization was performed using the nonlinear programming routine E04UCF in the FORTRAN77 libraries of the Numerical Algorithms Group. Tree topologies for each data set were estimated using RaxML under an LG+4Γ model and were fixed for model testing.

The f_j parameters are estimated from the observed amino acid frequencies across *k* sites, $\hat{\pi}_j^{(k)}$, as well as the average energy observed by the substitution of amino acid *j* at a given site. From (8) it follows that

$$(10)\ f_j \propto \frac{{\pi_j}^{(k)} e^{2E_j^{(k)}}}{\pi_j^{(B)}}$$

Substituting the observed site-specific frequencies, $\hat{\pi}_j^{(k)}$ into (10) and taking averages across all n sites we obtain an estimate for $\hat{f}_j$:

$$(11)\hat{f}_j \propto \sum_{k=1}^{n} \frac{\hat{\pi}_j^{(k)} e^{2E_j^{(k)}}}{\pi_j^{(B)} n}$$

Because of potential sparseness issues associated with small partitions, our estimation of $\hat{f}_j$ for a partition P with m sites is calculated as a mixture of that obtained for averages across sites in the whole alignment and averages obtained across a particular partition.

$$(12)\ \hat{f}_j^{\,P} \propto \hat{\omega} \sum_{k \in P} \frac{\hat{\pi}_j^{(k)} e^{2E_j^{(k,P)}}}{\pi_j^{(B)} m} + (1-\hat{\omega}) \sum_{k=1}^{n} \frac{\hat{\pi}_j^{(k)} e^{2E_j^{(k)}}}{\pi_j^{(B)} n}$$

Here, $\hat{\omega}$ is a small positive constant optimized in the range [0,1] and typically estimated in the range [0.5,1], chosen to ensure that no $\hat{f}_j^{\,P}$ is 0. The proportionality constant $\hat{f}_j^{\,P}$ and $\hat{f}_j$ never requires explicit calculation due to the final **Q** matrix being rescaled so that $-\sum \pi_j Q_{jj} = 1$. Note that the estimate of f_j depends on $\boldsymbol{\tau}$ that is embedded in the $E_j^{(k)}$ term. As maximum likelihood estimation of these parameters proceeds, every new value of the parameters in $\boldsymbol{\tau}$ leads to a new $\hat{f}_j$.

2.2.3 DATASETS FOR PERFORMANCE EVALUATION.

As a starting dataset we used the 300 test alignments plus structures described by Le and Gascuel (2010). These 300 datasets were selected randomly from 1771 non-redundant datasets obtained from the HSSP database (Le and Gascuel, 2010). The majority of datasets chosen were made up of enzymes involved in anabolic and/or catabolic metabolic pathways where loss of function would likely affect organismal fitness. The intensive tests we conducted using our model required that we reduce the number of datasets further so we took a representative sample of 48 alignments from the supplied test dataset (Fig A3). Due to the difficulty assigning an energy score to gaps in alignments we ignored all gap-containing sites. The resulting 48 datasets varied in length from 103 to 897 non-gap containing sites and contained between 11-and 93 sequences each (Table S1). For each dataset a tree was estimated using RAXML (Stamatakis, 2006) under the LG+4Γ model. As a measure of the amount of information in each dataset we define the 'information content' as the number of non-gap sites x the number of

sequences x the normalized tree branch length (the total tree length divided by the number of branches in the tree).

Solvent accessibility and secondary structure classifications were obtained using DSSP (Kabsch and Sander, 1983; Joosten *et al.*, 2011) with the protein structure for the seed sequence in order to determine the solvent accessibility and secondary structure classification of each site in the alignment. As in Le and Gascuel (2010) and Goldman *et al.,* (1998), we used a 10% relative accessibility threshold in order to assign sites to exposed versus buried categories, which leads to roughly equal numbers of buried and exposed residue assignments. Each dataset was partitioned in three different ways for each partition model tested in this paper. For the two partition model (2P/2P_LG) we used two solvent accessibility categories exposed and buried and the average dataset had 61% exposed sites and 39% buried sites. For the three partition model (3P/3P_LG) we partitioned the dataset into β-sheet(extended)/α-helix/Other categories with dataset wide average occupancies 22%, 28% and 47% sites respectively. Finally for the six partition model (6P/6P_LG) we combined the two previous partitions into six secondary structure/solvent exposure categories which on average contained around 16% of sites per partition. The number of sites found in each category did vary from alignment to alignment and has been summarized in supplementary table S1. The protein structures considered here ranged in resolution from 1.30Å to 5.00Å. We subjected each of the 48 structures to several rounds of energy minimization using the FoldX force-field in order to correct minor errors that may have resulted from the mis-positioning of side chains due to low x-ray crystallographic resolvability. Furthermore, since we were using FoldX potentials to evaluate the free energy of each amino acid substituted at each site, and this software tool involves a local minimization of amino-acid positions after a mutant is generated, this initial minimization maximized the accuracy and consistency of calculated changes in the free energy of folding of the protein structure.

The FoldX empirical physical potential (FoldX), PROSA pairwise β-carbon-based side-chain pairwise interaction potentials (P_PROSA), and PROSA combined (C_PROSA) pairwise interaction and surface potentials (the combination is performed by PROSA using default parameter settings) were used to introduce structural constraints into our model. For each site in the alignment, an approximation to ΔG_{fold} was obtained

for substituting each of the 20 amino acids into the wild type protein sequence at the site. Creating a protein mutant in FoldX involves first mutating the residue in question to itself followed by an adjustment to neighbouring residues in order to find the energy minimum for the wiltype folding energy (ΔG_{wt}). We correct for a resulting slight variation in ΔG_{wt} across sites by scaling each ΔG_j values such that $\Delta G_j = \Delta G_j - \Delta G_{wt}$. PROSA does not have the same variations in ΔG_{wt} across sites, but the same transformation was applied to P_PROSA and C_PROSA energy values. The resulting distribution for ΔG_j across sites for P_PROSA β-pairwise potentials (Figure A1) and FoldX (Figure A2) has been depicted for an example PDB (1XG2).

The FoldX, P_PROSA and C_PROSA energies derived above were incorporated into our models utilizing ($K = 1$) in the calculation of E_j and E_{ij} above. To determine if a mixed statistical potential/empirical potential based approach would perform better than either potential on its own, for the FoldX+C_PROSA mixed potential we associated a different τ parameter with each kind of potential ($K = 2$).

2.2.4 MODEL COMPARISON BY LIKELIHOOD RATIO TESTING AND BY AIC.

Most models presented thus far are nested with the LG + Γ as the simplest basis model and parameter estimates required for model nesting are shown in Figure 2.1a for $K = 1$. Partition models generalize to the single partition model when ω=0 and energy models generalize to their equivalent non-energy models when $\boldsymbol{\tau} = \mathbf{0}$. Model nesting is possible as the more complex partitioned models will become equivalent to the simpler non-partition (or simpler partitioned models) when the relevant partition-specific $\boldsymbol{\tau}$ parameters of the more complex models are constrained to have the same values. Furthermore, under the simpler null model, the relevant partition-specific $\hat{f}_j$'s will tend to equality as the number of sites in each partition increases. Nested models permit likelihood ratio tests (LRTs) to be used to compare model fits. The likelihood ratio test (equation 12) expresses, for two nested models, the increase in log-likelihood expected for data under a more complex model.

(12) **Likelihood ratio test statistic** $(\mathrm{LR}) = -2(\ln(L_S) - \ln(L_C))$

$L_S = likelihood\ simple\ model;\ L_C = likelihood\ more\ complex\ model$

The significance of an observed improvement by a more complex model is measured by calculating a p-value associated with the probability, under the simple model, that the observed improvement in likelihood occurred by chance. For standard LRTs, the p-value obtained using that LR is known to be chi-square distributed with degrees of freedom equal to the difference in the number of adjustable parameters between the models in question (the number of parameters are summarized in supplementary table S2). For this particular setting, the simpler models are on the boundary of parameter space of the more complex model and as a result the true null distribution is usually a mixture of chi-squared distributions with different numbers of degrees of freedom (Self and Liang, 1987). It follows that using the standard chi-squared distribution as described above for our application, the LRTs will be conservative: the probability of false rejection for an α-level test is less than α. For the purpose of this analysis the α-level threshold of 0.01 was used to denote a significant LRT.

For the comparisons of non-nested models, we used the Akaike Information Criterion (AIC) (equation 13) to get an absolute ranking of model fit between non-nested models. Below, δ refers to the number of free parameters in the model and L is the maximized value of the likelihood function for the estimated model.

(13) **Akaike Information Criterion** $= 2\delta - 2L$

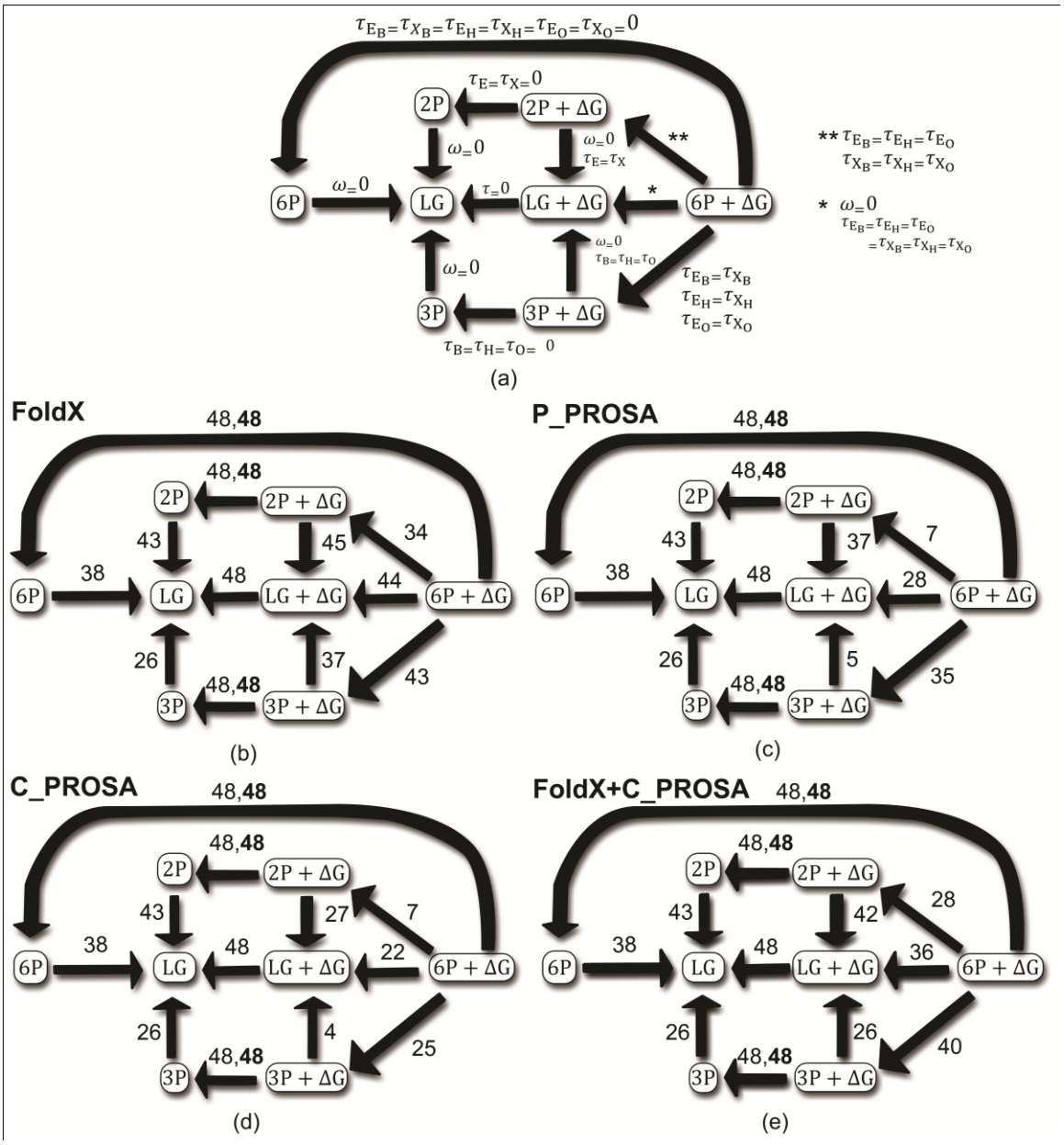

(a)
(b)
(c)
(d)
(e)
FoldX
P_PROSA
C_PROSA
FoldX+C_PROSA
6P
LG
6P + ΔG
2P
2P + ΔG
3P
3P + ΔG
LG + ΔG
** $\tau_{E_B}=\tau_{E_H}=\tau_{E_O}$ $\tau_{X_B}=\tau_{X_H}=\tau_{X_O}$
* $\omega_{=}0$ $\tau_{E_B}=\tau_{E_H}=\tau_{E_O}=\tau_{X_B}=\tau_{X_H}=\tau_{X_O}$

Figure 2.1: Allowing structural constraints to vary across structural environments often improves SCPMs and always improves non-structurally constrained counter parts.

Likelihood ratio test (LRT) results for nested models amongst the 48 datasets using the FoldX force field or P_PROSA or C_PROSA statistical potentials or both in the calculation of ${Q_{ij}}^{E^{(k)}}$. **(a)** Conditions for model nesting between the models implemented in this paper. Arrows indicate that nesting from a more complicated model to a simpler model along with the parameter conditions that satisfy nesting. The number of LRTs rejecting the null hypothesis using a p-value of 0.01 along with FoldX **(b),** P_PROSA**(c)**, C_PROSA**(d)**, and FOLDX+C_PROSA **(e)** in the calculation of ${Q_{ij}}^{E^{(k)}}$. For Partition models, those using partition specific exchangeability matrices as (${Q_{ij}}^{B}$) are in bold (Le and Gascuel, 2010) while those using LG are not.

2.2.5 EVALUATING GENERAL TRENDS IN SITE-SPECIFIC MATRICES:

To understand general trends in site-specific rate matrices produced by our four models, we have subsampled 1092 **Q** matrices from sites in our dataset for each model/partition combination. Putting these matrices together, each Q_{ij} entry in the matrix consists of a distribution of 1092 amino acid exchangeabilities. We obtained the ratio of the exchangeability observed for that site compared to the exchangeability observed in a general substitution model using LG+8Γ. In this way we obtained a matrix where each *i,j* entry contains 1092 Q_{ij}^{P}: Q_{ij}^{LG} ratios. To understand general trends in how the energy model differed from standard phylogenetic models, we obtained a measure of central tendency (the median) and variance (the mean average absolute value of the deviation from the mean) for each entry in this matrix of ratios. This calculation was completed after removing the top 10% and bottom 10% of extreme values to remove any outliers that might obscure the general trends we were trying to portray. Heat maps of the median Q ratio matrices can be found in Figures 2.3,2.5 A15-A18. Heat maps showing the mean absolute deviation from the median Q ratio matrices can be found in figures A19-A22. The size of the subsample was chosen to correspond to the size of our least populated partition, solvent inaccessible loop regions.

We also examined stationary frequencies coming from the **Q** matrices for the same subsamples of sites described above. Median stationary frequencies from our subsamples have been presented in figure 2.4, and figures A7-A10. An examination of all the stationary frequencies coming from these same sites/models/partitions for all 1092 sites has been presented as boxplots in figures A11-A14.

2.2.6 MEASURING PARTITION-SPECIFIC SIGNIFICANCE TOWARDS OBSERVED LIKELIHOOD GAINS:

To assess the performance of our nested models across various structural categories we define a partition-specific significance factor (PSSF) (equation 13 below) that expresses the ratio of the average likelihood difference within a partition C ($\overline{ld}_c$) between the complex model and a simpler model to the same quantity observed across the rest of the sites in the dataset D ($\overline{ld}_D$).

(13) partition specific significance factor $(PSSF) = \overline{ld}_c / \overline{ld}_D$

Consider model M1 nested within model M2. A PSSF>0 implies M2 fits a site category best while PSSF<0 suggests the opposite. PSSF≈1 indicates that the likelihood gain within a site category is on par with the average likelihood gain across the remaining sites. PSSF>1 indicates that a site category is particularly well modeled under the partition being evaluated while 0<PSSF<1 indicates a smaller contribution to the likelihood gain.

2.2.7 IDENTIFYING SITES POTENTIALLY INVOLVED IN PROTEIN-PROTEIN/PROTEIN-LIGAND INTERACTIONS.

To assess the performance of our model at sites potentially involved in protein-protein and protein-ligand interactions, we searched the inferred biomolecular interaction server (http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi) for sites amongst our 48 datasets known to be involved in interactions. We constrained our search to only include sites with at least partial conservation of binding sites amongst non-redundant homologous members of binding site clusters. For protein-protein interactions, we ensured that sites returned were validated by the Protein interfaces, surfaces and assemblies service (PISA) at the European Bioinformatics Institute (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html). For protein-ligand or protein-ion interactions we constrained our search to those flagged as biochemically relevant.

2.3 RESULTS AND DISCUSSION:

2.3.1 STRUCTURAL CONSTRAINTS SIGNIFICANTLY IMPROVE INDEPENDENCE MODELS OF PROTEIN EVOLUTION:

Our structurally constrained models of protein evolution were applied to the 48 test datasets and the results are summarized in Figure 2.1 as the number of likelihood ratio tests rejecting the simpler models in favour of the more complex structurally constrained models. Models incorporating structural constraints are always significantly better than models that do not regardless of the potential used to generate the energies, the number of partitions utilized, or the backing substitution matrix chosen for each partition. In every case the single partition LG+ΔG model was significantly better than the non-energy alternative with a p-value that decreased as more sites and sequences were added to the alignment (Figure 2.2a). Similarly, our partitioned energy models utilizing LG for each partition outperformed a partitioned model that was not based on energy but only had partition-specific frequencies. The support for our partition energy model was high with $\widehat{\omega}$ values (the parameter controlling the mixture of $f_j^{Partition}$ and $f_j^{Dataset}$) greater than 0.5 for the great majority of partition energy models (Figure A4).

The benefit of partitioning is clear for FoldX (Figure 2.1b), P_PROSA (Figure 2.1c), and FoldX+C_PROSA (Figure 2.1e) energies where 45, 37 and 42 of the datasets were improved by partitioning sites into the 2 partition solvent exposure categories. Combining Prosa β pairwise potentials with surface potentials in the C_Prosa potential (Figure 2.1d) reduced the requirement for partitioning but 27 datasets still demonstrated significant likelihood gains upon partitioning for solvent exposure. The 6 partition model showed a similar trend with 44 datasets performing significantly better than the single partition energy model using FoldX energies and 28 performing significantly better using

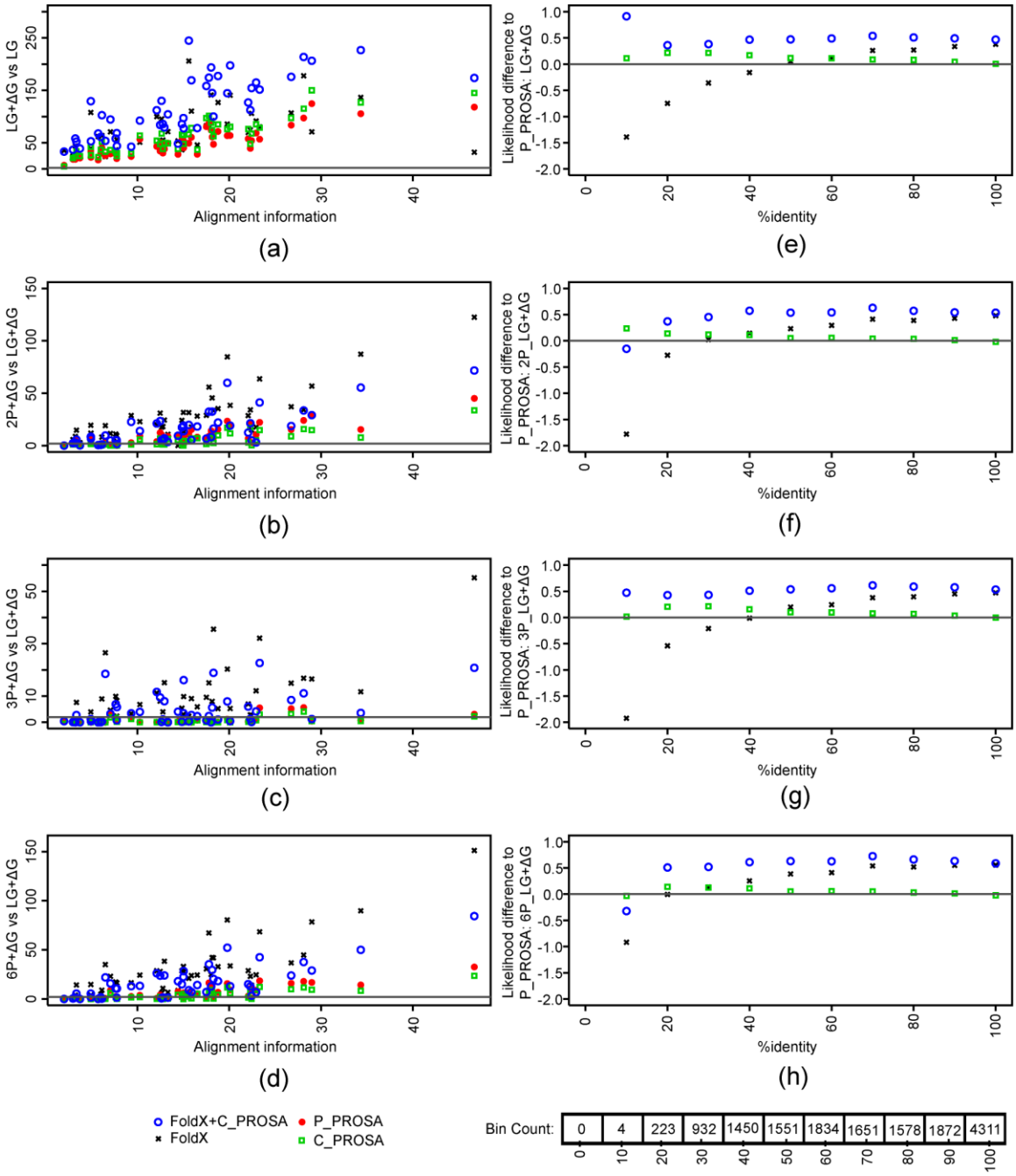

LG+ΔG vs LG
Alignment information
(a)
Likelihood difference to P_PROSA: LG+ΔG
%identity
(e)
2P+ΔG vs LG+ΔG
Alignment information
(b)
Likelihood difference to P_PROSA: 2P_LG+ΔG
%identity
(f)
3P+ΔG vs LG+ΔG
Alignment information
(c)
Likelihood difference to P_PROSA: 3P_LG+ΔG
%identity
(g)
6P+ΔG vs LG+ΔG
Alignment information
(d)
Likelihood difference to P_PROSA: 6P_LG+ΔG
%identity
(h)
FoldX+C_PROSA
FoldX
P_PROSA
C_PROSA
Bin Count:
0
4
223
932
1450
1551
1834
1651
1578
1872
4311

Figure 2.2: Likelihood gains over a standard structurally constrained model can be arrived upon with mixed statistical-potential/empirical potential approaches and by partitioning for secondary structure and solvent accessibility.

LHS: $-log(p\ value)$ obtained for the likelihood ratio test of the **(a)** 1-partition (LG+ΔG), **(b)** 2-partition 2P+ΔG, **(c)** 3-Partition (3P+ΔG) and **(d)** 6-Partition (6P+ΔG) structurally constrained models when compared to LG (LG+ΔG) or LG+ΔG (2P+ΔG, 3P+ΔG, 6P+ΔG). The gray horizontal bar corresponds to a p-value of 0.01. (Alignment information = [number of taxa] × [number of sequences] × [normalized branch length]) **RHS:** Mean likelihood differences for sites binned into 10% identity categories when compared to an equivalent P_PROSA model, which tended to be worst performing. Plots are shown for the 1 partition **(e)**, 2 partition **(f)**, 3-partition**(g)** and 6 partition**(h)** model.

P_PROSA energies, 22 with C_PROSA and 36 performing better when combining the two. While substantially fewer partitioned models appear to be preferred to simpler unpartitioned (or less partitioned) models when P_PROSA and C_PROSA energies are used, it should be noted that with so many partitions some categories were sparse (Table S1). 9/10 and 8/10 datasets that had more than 30 sequences and at least 20 sites per partition class preferred the 6 partition model over the single partition model using P_PROSA or C_PROSA potentials. Indeed a clear trend can be observed demonstrating that datasets containing more information (Figure 2.2b,c,d) were more likely to be associated with smaller p-values for LRTs. This trend continued with the 3 partition model only for FoldX and FoldX+C_PROSA where 37 and 26 of the datasets were better modeled respectively by the 3 partition secondary structure model. For most models tested, the estimates of the rates-across-sites parameters, the α (Figure A5) and normalized tree branch lengths (Figure A6) remain relatively unperturbed by the introduction of the energy potentials into the site-specific models.

It should be noted that there is a fundamental difference between utilizing FoldX and Prosa pairwise potentials. The calculation of $\hat{f}_j$ (equation 11) relies in part on the averaging of the term $e^{2E_j^{(k)}}$ that reflects the average effect of mutating to an amino acid j over the entire dataset. P_PROSA and C_PROSA potentials are a simpler approximation to $\Delta G_{folding}$ that benefit somewhat from partitioning for solvent exposure but can not easily generalize trends across secondary structures. FoldX on the other hand is much more heterogeneous amongst these differing structural environments. It makes sense then that models using FoldX are greatly assisted by the calculation of $\hat{f}_j^P$ across specific structural categories where a more accurate value for the average effect of mutating to an amino acid j can be obtained.

2.3.2 A HYBRID FOLDX+C_PROSA POTENTIAL ACCOMMODATES DATASETS WITH VARYING EVOLUTIONARY RATES AND BIOPHYSICAL CONSTRAINTS:

Force fields have typically been avoided in the construction of SCPMs due to their computational complexity and the assumption that they might bias sequences too strongly towards the reference structure used to generate them. Here, we more closely

examine the effects of adding more realism into the SCPM through the use of empirical physical potentials as well as our mixed physical/statistical potential approach.

The LRT p-values obtained across the 48 datasets (Figure 2.2a) indicate that models based on the FoldX potential and the mixed FoldX+C_PROSA potentials are most often preferred to models using P_PROSA or C_PROSA Energies. AIC test score results show a broadly similar pattern (Supplementary document 1). To better understand how the two methods of obtaining $Q_{ij}^{E(k)}$ differ in their performance, we examined the distribution of average site-wise likelihood differences between our various ΔG approximations compared to P_PROSA (our worst performing model) binned across varying degrees of conservation (Figure 2.2e). For the general LG+ΔG model, it is clear that PROSA is better able to model highly variable sites where the same amino acid appeared in <50% of sequences, which represents 17% of the sites over all the 48 datasets. The remaining 83% of sites displaying >50% identity were better modeled by FoldX energies. This suggests that the FoldX empirical force field tends to bias the model towards the reference structure (and sites that are identical to this structure in homologs) while P_PROSA and C_PROSA more flexibly model variable sites. The hybrid statistical potential/empirical force field was generally observed to perform at least as well as P_PROSA and C_PROSA on these highly variable sites while surpassing all models for more conserved sites. Thus, the hybrid statistical potential/empirical force field approach appears to strike a balance between structural specificity and flexibility across sequences in the alignment.

2.3.3 PARTITIONING FOR SOLVENT ACCESSIBILITY AND SECONDARY STRUCTURE.

We investigated if allowing biophysical constraints to vary across different secondary structure and solvent accessibility categories would improve model fit when assessed over different levels of sequence conservation. Our partitioned models performed better on the more variable sites as can be observed by average site-likelihood differences tending to become more positive, with respect to P_PROSA, across both variable and invariable sites (Figure 2.2f,g,h). Notably, partitioning for solvent exposure improved the performance FoldX-based analyses across highly variable sites. Again, the rationale for this behaviour is likely that partitioned structurally constrained phylogenetic

models utilizing FoldX take into account environment-specific properties of the energy function leading to a model that better accounts for environment-specific sources of amino-acid variability.

2.3.4 PARAMETER ESTIMATES FOR A GENERAL STRUCTURALLY CONSTRAINED MODEL OF PROTEIN EVOLUTION.

STATIONARY FREQUENCIES:

Parameter estimates in the absence of partitioning have been summarized in Figure 2.3. Median stationary frequencies reported for a subsample of 1092 sites (Figure 2.3a-d) are in general comparable to those of the standard LG model. However, a sign test reveals that many of the differences although slight are statistically significant ($p = 0.01$). Proline, for example, displays a depressed stationary frequency that is most likely the result of its generally unfavourable impact on protein stability. Boxplots for the stationary frequencies calculated for these 1092 sites display a high degree of variation about the median Figure A7-A10 (LG+ΔG). In general P_PROSA and C_PROSA stationary frequencies tended to be much less variable than FoldX or FoldX+C_PROSA stationary frequencies. The distribution of stationary frequencies amongst different amino acid types shows a clear bias to stationary frequencies between 0 and 0.2. Nearly all residue types were represented in at least some sites with stationary frequencies greater than 0.2, which is significantly greater than the median stationary frequency of a residue in the standard LG model (Also depicted in Figure 2.3). Residues showing the highest degree of variation in stationary frequency were small aliphatic amino acids glycine, alanine and the branched-chain amino acids leucine and valine. The wider range of amino acid stationary frequencies observed when using FoldX and FoldX+C_PROSA energies demonstrates the improved plasticity of these energy potentials to model the diversity of site patterns in an alignment. Interestingly, glycine appeared to be the most versatile amino acid in terms of stationary frequencies having, for a minority of sites, stationary frequencies approaching 1. On average, sites containing at least one amino-acid stationary frequency >0.2 were well modeled by our structurally constrained model of protein evolution as the median value for the likelihood difference between the energy model at these sites and the no-energy model was always positive (data not shown). Again glycine is unique in exhibiting a large tail away from the median stationary

frequency when using FoldX energies with one site even assuming a stationary frequency near 1. A mutation to a glycine can be quite a significant change as exemplified by the broad range of energy values assigned by FoldX to these mutations in Figure A2 although these large stationary frequencies for glycine were mostly associated with significant likelihood gains at the site.

The median stationary frequencies of amino acids for our structurally constrained partitioned models are plotted against those obtained for a single partition energy model across 1092 randomly selected sites for models using FoldX+C_PROSA in Figure 2.4 (this model shows the same general patterns as P_PROSA (Figure A11), C_PROSA (Figure A12), FoldX (Figure A13)). The single partition model displayed a range of median frequencies with the lowest frequencies corresponding to rare amino-acids such as "W" or "C" having median frequencies near 0.01 while the most frequent amino acids are small hydrophobic "LVA" having median frequencies near 0.1. The 2P_LG+ΔG and 6P_LG+ΔG models show a clear trend where the larger hydrophilic amino-acids "RKQHNEDY" have much smaller stationary frequencies in partition models evaluated over buried categories (X) than exposed categories (E). Conversely, though perhaps not as clear in the case of 6P_LG, the hydrophobic amino-acids "AMVIL" display higher frequencies in buried categories than when exposed to solvent. We limit our discussion of the 3 partition model to those incorporating FoldX energies, which were the analyses that showed the clearest likelihood gains. Here, 3P_LG displayed several trends in agreement with the literature (Costantini *et al.*, 2006; Jiang *et al.*, 1998). Residues typically associated with β-sheets (B) displayed median stationary frequencies similar "TMFYCW" or larger "VIL" than their equivalent calculated in the general energy model. Residues typically observed in α-helices (H) displayed a similar trend ("MALEK"). Furthermore residues often observed to be destructive to secondary structure formation ("GP") displayed lower median stationary frequencies in both the α-helix category and β-sheet category while larger stationary frequencies were observed in the Other category (O). Similar to the single partition models, a larger variation in the stationary frequencies was observed for models incorporating FoldX energies over models utilizing P_PROSA or C_PROSA energies (Figure A7-A10).

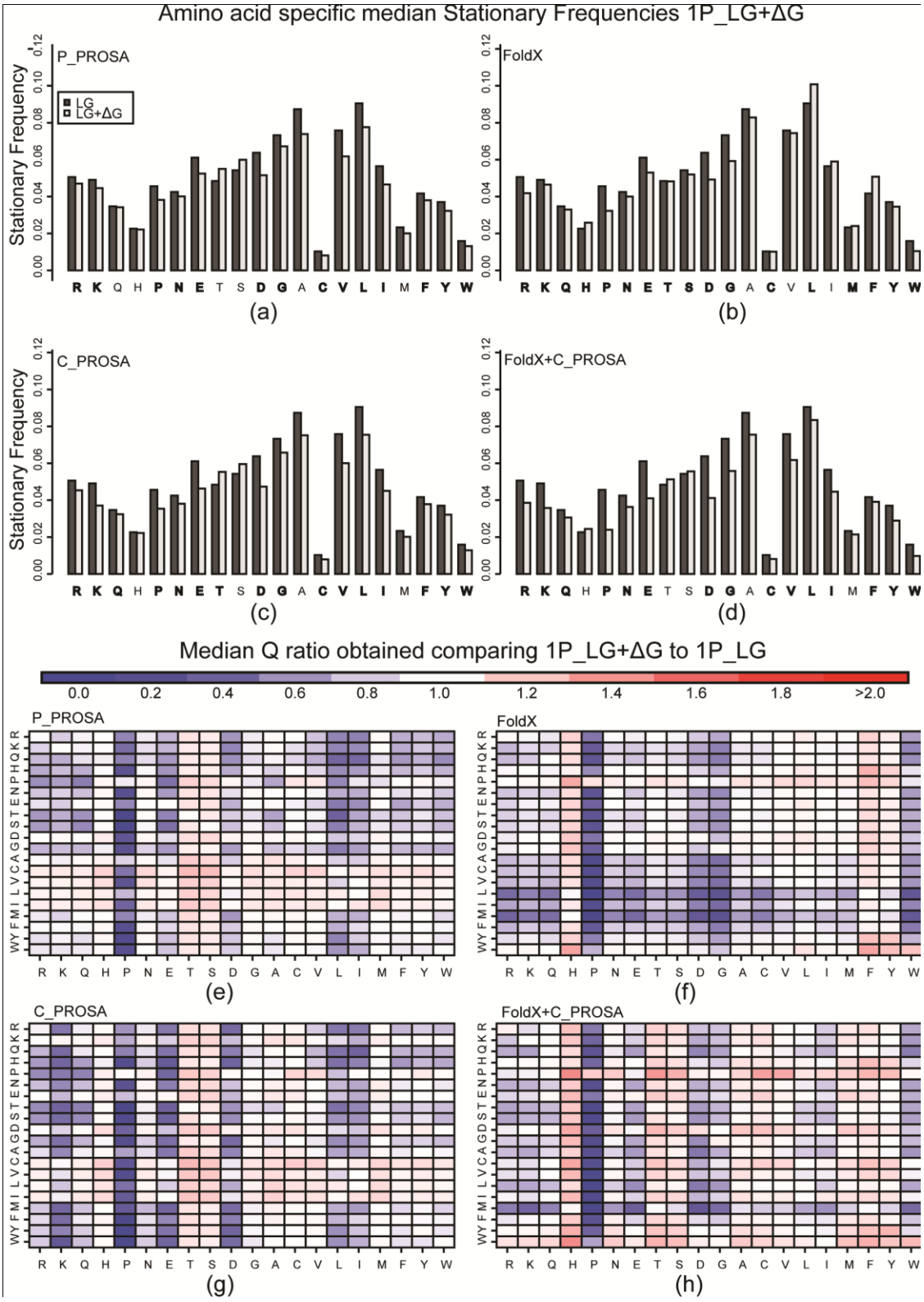

Amino acid specific median Stationary Frequencies 1P_LG+ΔG
P_PROSA
LG
LG+ΔG
Stationary Frequency
0.00 0.02 0.04 0.06 0.08 0.10 0.12
R K Q H P N E T S D G A C V L I M F Y W
(a)
FoldX
(b)
C_PROSA
(c)
FoldX+C_PROSA
(d)
Median Q ratio obtained comparing 1P_LG+ΔG to 1P_LG
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 >2.0
P_PROSA
W Y F M I L V C A G D S T E N P H Q K R
(e)
FoldX
(f)
C_PROSA
(g)
FoldX+C_PROSA
(h)

Figure 2.3 The general structurally constrained model: Effects on stationary frequencies and the median Q.

Stationary frequencies sampled from single partition models utilizing P_PROSA potentials **(a)**, C_PROSA potentials **(b)**, FoldX potentials **(c),** FoldX+C_PROSA potentials **(d)**. Columns with bold lettering represent statistically significant differences by the sign test at a p-value threshold of 0.01. Median stationary frequencies were obtained from randomly sampling 1092 sites from amongst 15028 sites in our dataset. We calculate the median ratio of $Q_{ij}^{(k)}$ entries found using the LG+ΔG to those found when using the standard LG model.

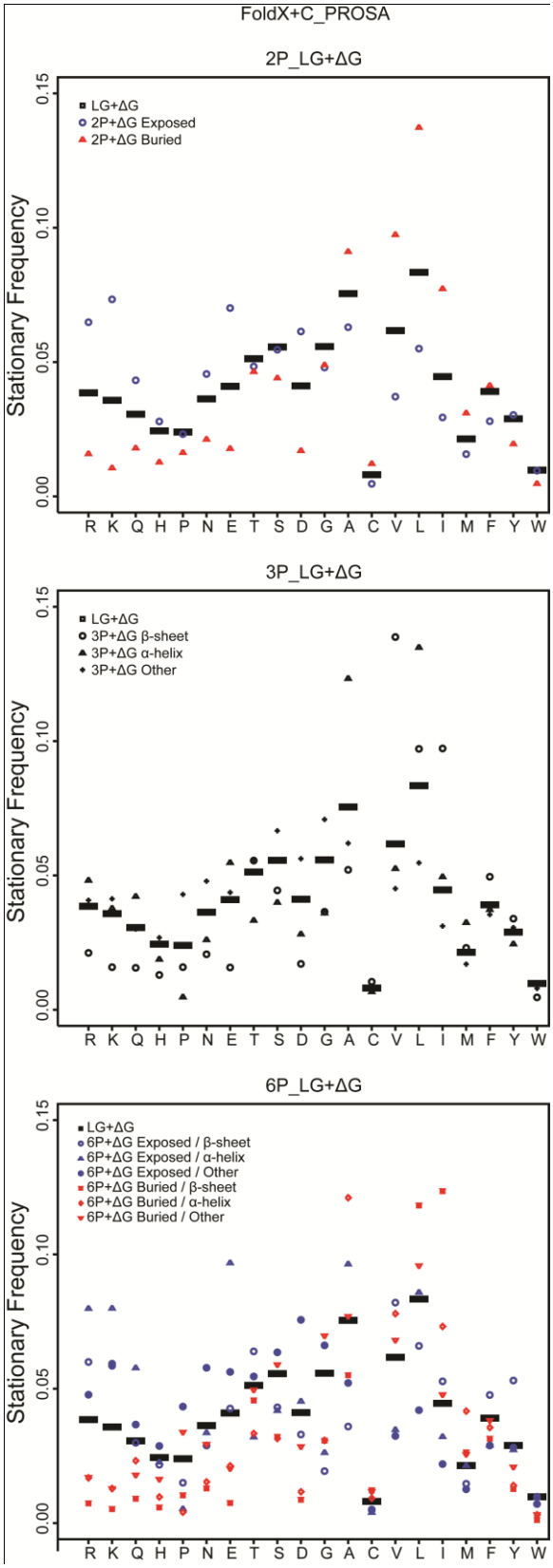

FoldX+C_PROSA
2P_LG+ΔG
LG+ΔG
2P+ΔG Exposed
2P+ΔG Buried
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
3P_LG+ΔG
LG+ΔG
3P+ΔG β-sheet
3P+ΔG α-helix
3P+ΔG Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
6P_LG+ΔG
LG+ΔG
6P+ΔG Exposed / β-sheet
6P+ΔG Exposed / α-helix
6P+ΔG Exposed / Other
6P+ΔG Buried / β-sheet
6P+ΔG Buried / α-helix
6P+ΔG Buried / Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W

Figure 2.4: General trends in the stationary frequencies derived for partition-specific models.

We plot the median stationary frequencies derived from 1092 sites randomly sampled from the 48 datasets in this paper using both FoldX+C_PROSA energies when compared to the stationary frequencies observed under the LG+ΔG model for the 2 partition **(a)**, 3 partition **(b)** and 6 partition **(c)** model. Section 2.2.5 details the methodology used to produce this figure.

THE SUBSTITUTION MODEL:

The median ratio of amino-acid exchangeabilities $Q_{ij}^{(k)}: Q_{ij}^{LG}$ for subsamples of 1092 sites are shown for the four single partition models in this dataset (Figure 2.3e-h). Models utilizing P_PROSA and C_PROSA potentials displayed differing patterns with respect to LG than FoldX. In particular, P_PROSA and C_PROSA tend to show an increased probability of exchange away from smaller hydrophobic amino-acids such as "AMVILF" to a variety of hydrophilic and hydrophobic amino acids while FoldX tends to favour transitions away from the hydrophilic residues "RNDQEHKS" to hydrophobic amino acids. However, many similarities also exist, such as a uniform depression in exchangeabilities to both aspartate and proline. Our mixed statistical potential/empirical potential approach (Figure 2.3h) has median ratios that significantly resemble those observed for FoldX (Figure 2.3f). However, close examination reveals a clear influence of C_PROSA (Figure 2.3g) on **Q**. For example, exchangeabilities to hydroxylated amino acids ("ST") are enhanced in both C_PROSA (Figure 2.3g) and FoldX+C_PROSA (Figure 2.3h) relative to FoldX alone (Figure 2.3f).

In order to better understand the site-to-site variability (plasticity) of the entries in the **Q** matrices presented in Figure 2.3, we examined the mean absolute deviation away from the median $Q_{ij}^{(k)}: Q_{ij}^{LG}$ in the LG+ΔG panel of supplementary figures A19-A22. These 4 panels show that residues near the diagonal, which are similar in terms of hydrophilicity and size, tend to display large variances in their exchangeability from site to site in the alignment. This variance quickly drops as residues become more and more physiochemically distinct.

The effect of median stationary frequency variation amongst the various structural categories is evident upon inspecting the resulting median $Q_{ij}^{P}: Q_{ij}^{LG}$ matrices for the partitioned FoldX+C_PROSA models (Figure 2.5). In general, median Q matrices become more heterogeneous moving from the 2-partition and 3-partition models to the 6 partition model. The two partition model and the 6 partition models clearly show that the exchangeabilities to hydrophilic amino acids has increased in the exposed categories while the opposite is true for buried categories. Amino-acid propensities for various secondary structure classes is known to vary amongst fold types (Costantini *et al.*, 2006;

Jiang *et al.*, 1998). Despite the pooling of sites from different fold classes here, the 3-partition model does display some expected trends, such as an increased propensity of hydrophobic amino-acids such as "CVLIMFY" in beta-sheets, a suppression of prolines and glycines in both sheets and helices, and their more probable incorporation in loops. Amino-acid propensities have been shown to vary amongst secondary structure classes in different solvent exposures (Momen-Roknabadi *et al.*, 2008), thus it is interesting to see here that while the 6 partition model has some similarities the 2 partition model, there are differences due to the additional incorporation of partitioning by secondary structure. These differences increase the ability of the models to better fit a wide range of exchangeabilities observed in these specific structural environments. The mean average deviation about these median ratios reveals that the same general trends are observed as in the single partition model with similar amino-acids in general showing more variability in $Q_{ij}^{(k)}$ than dissimilar amino acids (figures A19-A22).

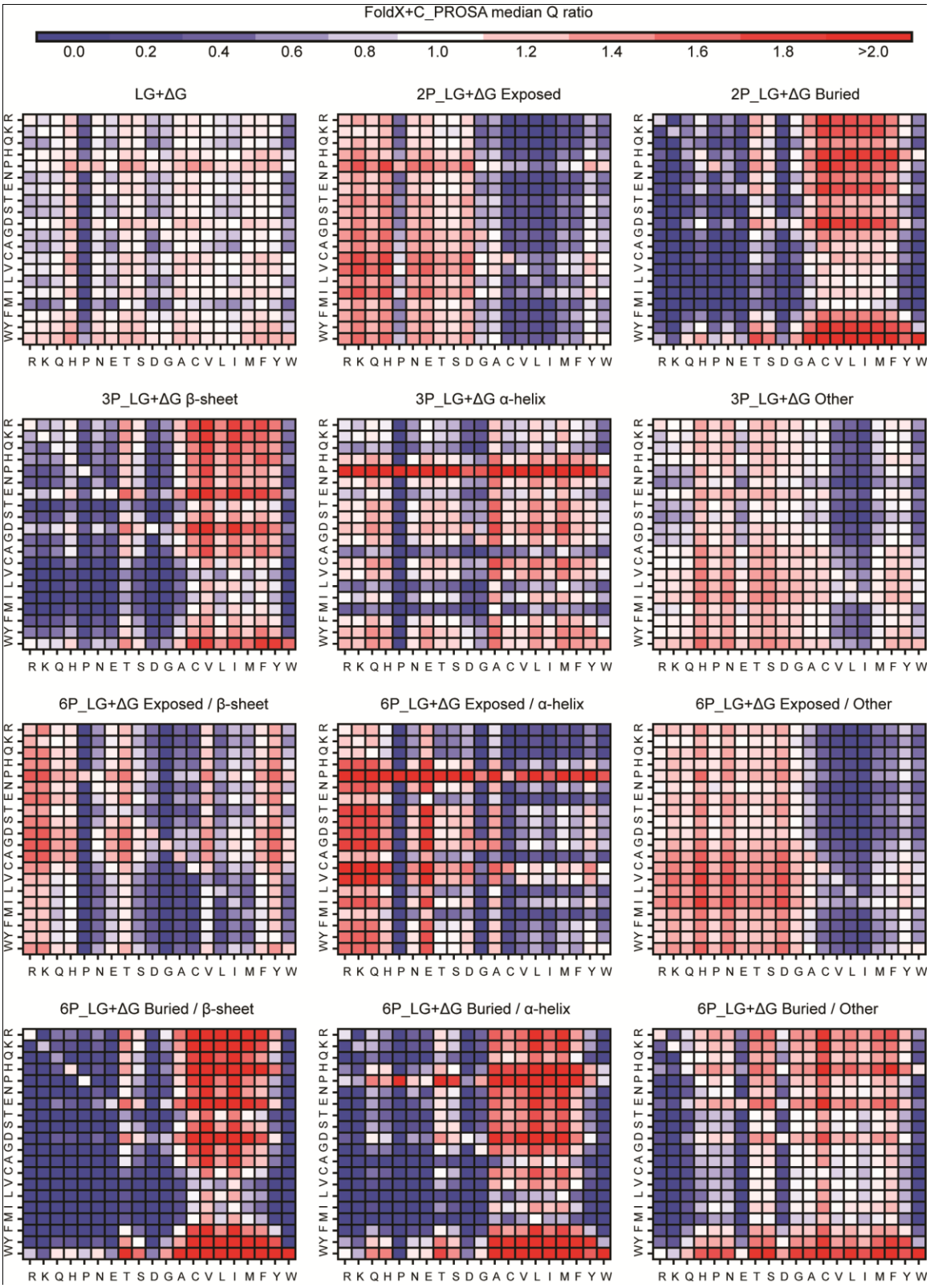

FoldX+C_PROSA median Q ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
>2.0
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure 2.5: Median Q matrices for models utilizing a combination of FoldX and C_PROSA energies.

After running our structurally constrained partition models in all 48 datasets we randomly sampled 1092 sites being analyzed under a particular partition category/model combination. We calculate the median ratio $Q_{ij}^{E\,(k,P)}$: Q_{ij}^{LG} entries found using the various models (M) analyzed in this paper.

2.3.5 FoldX outperforms PROSA in the hydrophobic protein cores while maintaining a comparable performance on the surface.

The 48 datasets analyzed here constitute 15028 sites without gaps. Each site can be categorized into the various structural environments analyzed in this paper. The PSSF and percent of sites with likelihood differences greater than 0 obtained between the various energy models compared to LG and LG+ΔG or back to the same partition model but without energies has been summarized in Table 2.1 for models tested. A main difference between PROSA-based and FoldX-based models appears to be with regards to the enhanced ability for FoldX to model the hydrophobic core of the protein structure. The comparison between LG+ΔG and LG reveals that 89% of buried sites are modeled better by LG+ΔG:FoldX than LG while only 64% and 66% of sites are better modeled when using the P_PROSA or the C_PROSA in LG+ΔG. On the other hand, C_PROSA performs best on 64% of solvent exposed sites while P_PROSA and FoldX have similar performance for these sites. The best performing model combines both FoldX and PROSA energies. While its performance in the hydrophobic core of the protein structure is not quite as good as FoldX, it is still significantly improved with respect to C_PROSA with 82% of sites being modeled better than LG. Exposed sites are generally improved for both FoldX and C_PROSA energies with 70% being better modeled by the FoldX+C_PROSA combined potential over LG which is greater than the 62-64% arrived at from the other energy potentials on their own.

This bias towards modeling the protein core induced by utilizing more complex potentials like FoldX is what we strive to correct for by allowing the parameter $\hat{f}_j^P$ to vary between significant structural environments by partitioning. Since the bulk of the likelihood gains of LG+ΔG models over the standard LG model come from sites in the hydrophobic core, parameter estimates will favour increasing the likelihood of these sites to the possible detriment of sites exposed to the surface. While C_PROSA energies do not suffer from this imbalance, modeling approximately 63-64% of sites better than LG for both exposed and buried sites, it is important to note that likelihood gains associated with C_PROSA energies are significantly smaller regardless (Figure 2.2a).

Table 2.1: Comparing LG+ΔG models to LG.
Partition-specific significance factors and the % of likelihood differences greater than 0.

		All sites		Exposed		Buried	
Energy type	**Residue Hydrophobicity Class[a]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**
P_PROSA	All Sites	1.00	63%	1.02	62%	0.97	64%
	Hydrophillic	1.11	66%	1.31	68%	0.59	61%
	Hydrophobic	0.77	60%	0.11	49%	1.08	65%
	Other	1.53	63%	1.58	64%	1.39	62%
C_PROSA	All Sites	1	65%	1.02	64%	0.98	66%
	Hydrophillic	1.02	66%	1.33	70%	0.24	55%
	Hydrophobic	0.87	63%	0.06	50%	1.25	70%
	Other	1.46	66%	1.53	66%	1.26	65%
FoldX	All Sites	1	74%	0.5	62%	1.59	89%
	Hydrophillic	0.5	63%	0.28	55%	1.08	82%
	Hydrophobic	1.48	87%	0.87	75%	1.77	93%
	Other	1	69%	0.69	63%	1.84	84%
FoldX+C_PROSA	All Sites	1	76%	0.76	70%	1.28	82%
	Hydrophillic	0.77	72%	0.79	72%	0.71	72%
	Hydrophobic	1.18	81%	0.54	67%	1.48	87%
	Other	1.19	72%	1.06	70%	1.53	78%
Counts		All sites		Exposed		Buried	
	All Sites	15028		8106		6922	
	Hydrophillic	6605		4757		1848	
	Hydrophobic	6826		2181		4645	
	Other	1597		1168		429	

a: Sites containing > 80% hydrophobic/hydrophilic residues labeled as such. All other sites pooled into an Other category.

b: partition-specific significance factor =$\overline{\boldsymbol{ld}}_{\boldsymbol{C}}/\overline{\boldsymbol{ld}}_{\boldsymbol{D}}$.

c: % of sites in the category with likelihood differences greater than 0.

Table 2.2: Comparing 2P+ΔG models to LG.
Partition-specific significance factors and the % of likelihood differences greater than 0.

		All sites		Exposed		Buried	
Energy type	**Residue Hydrophobicity Class[a]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**
P_PROSA	All Sites	1	67%	1.02	65%	0.98	70%
	Hydrophillic	1.15	70%	1.49	75%	0.28	59%
	Hydrophobic	0.8	65%	-0.18	44%	1.26	75%
	Other	1.23	64%	1.34	62%	0.96	68%
C_PROSA	All Sites	1	68%	1.04	66%	0.95	70%
	Hydrophillic	1.08	69%	1.48	75%	0.06	53%
	Hydrophobic	0.85	67%	-0.11	46%	1.31	77%
	Other	1.29	66%	1.42	65%	0.93	68%
FoldX	All Sites	1	78%	0.9	73%	1.12	84%
	Hydrophillic	0.82	74%	0.96	77%	0.48	68%
	Hydrophobic	1.18	85%	0.77	69%	1.38	92%
	Other	0.95	69%	0.9	67%	1.08	74%
FoldX+C_PROSA	All Sites	1	78%	0.95	74%	1.06	82%
	Hydrophillic	0.88	75%	1.04	78%	0.48	68%
	Hydrophobic	1.08	82%	0.65	68%	1.29	89%
	Other	1.14	72%	1.12	71%	1.19	75%
			All sites		Exposed		Buried
Counts	All Sites		15028		8106		6922
	Hydrophillic		6605		4757		1848
	Hydrophobic		6826		2181		4645
	Other		1597		1168		429

a: Sites containing > 80% hydrophobic/hydrophilic residues labeled as such. All other sites pooled into an Other category.

b: Partition-specific significance factor =$\overline{\boldsymbol{ld}}_{\boldsymbol{C}}/\overline{\boldsymbol{ld}}_{\boldsymbol{D}}$.

c: % of sites in the category with likelihood differences greater than 0.

2.3.6 HIGHLY VARIABLE SITES IN EXPOSED REGIONS OF THE PROTEIN ARE BETTER MODELED BY PARTITIONED SCPMS

For models using FoldX or FoldX+C_PROSA energies, partitioning the dataset into solvent-exposure and secondary structure based categories improves the performance of exposed residues to the mild detriment of buried residues. While only 62%/70% (FoldX/FoldX+C_PROSA) of residues exposed to solvent were modeled well by LG+ΔG when compared to LG (Table 2.1), 73%/74% and 74%/76% of exposed residues were better modeled by the 2P_LG+ΔG and 6P_LG+ΔG models (Table 2.2, Table A4). While partitioning on secondary structure only showed mild gains, all three partitions showed approximately a 5% improvement in residues having a positive likelihood difference (see Table A3).

2.3.7 LE AND GASCUEL MATRICES:

Despite the clear likelihood gain reported by Le and Gascuel (2010) when using the Le and Gascuel partition-specific substitution matrices without a structural constraint, there was still a significant amount more to gain from incorporating structural constraints into the model. Indeed all 48 datasets were significantly improved by adding either PROSA, C_PROSA, FoldX or FoldX+C_PROSA structural constraints to the 2P, 3P and 6P Le and Gascuel partition-specific basis matrices (Figure 2.1). Of further interest, comparing the AIC of models that used partition-specific basis matrices (Le and Gascuel, 2010) and models that used the single LG basis matrix showed that partition-specific basis matrices were more often preferred by SCPMs utilizing P_PROSA and C_PROSA energies but less often preferred by models incorporating FoldX or mixed FoldX+C_PROSA energies (Table 2.3). Here, 2P_LG+ΔG was favoured over 2P +ΔG only 18/19 times out of 48 for P_PROSA/C_PROSA energies respectively and 27/28 times out of 48 for FoldX/FoldX+C_PROSA energies. The same was not true for 3P+ΔG vs 3P_LG+ΔG where, in general, almost all models preferred utilising LG as Q_{ij}^{basis} instead of the partition-specific basis matrices. Finally, following a similar pattern as the 2 partition model, for the 48 datasets 6P_LG+ΔG was favoured over 6P+ΔG 18/19 times for P_PROSA/C_PROSA and 28/31 times for FoldX/FoldX+C_PROSA. Even the single partition energy model LG+ΔG was often preferred by AIC over any of the Le and Gascuel partition-specific basis matrices without structural constraints.

While the foregoing results were unexpected, they can be rationalized as follows. The Le and Gascuel partition-specific matrices capture generalized substitution rates observed amongst secondary structure classes at different solvent accessibilities and it seems that the FoldX and FoldX+C_PROSA models, in most cases, are able to capture the similar information in a partition-specific context. Moreover, using PROSA or FoldX energy constraints provide additional model fit gains by taking into account site-specific features beyond secondary structure and solvent accessibility such as residue packing and side chain neighbouring interactions.

Table 2.3: Number of times out of 48 that a model using LG basis matrices was preferred by AIC to a model using appropriate secondary structure/solvent accessibility matrices from Le and Gascuel, 2009.

	2P	3P	6P
P_PROSA	19	40	18
C_PROSA	18	39	19
FoldX	27	44	28
FoldX+C_PROSA	29	43	31

2.3.8 THE PERFORMANCE OF SCPEs IN THE VICINITY OF PROTEIN-PROTEIN, PROTEIN-LIGAND INTERACTIONS.

Important protein-protein and protein-ligand interaction sites as well as sites that were not structurally constrained at all were expected to be poorly modeled by SCPEs compared to sites involved in interactions that stabilized the overall protein fold. We assessed whether our structurally constrained partition models could be used to highlight areas where structural constraints are not sufficient to model the evolutionary dynamics at sites of interaction.

Figure 2.6 illustrates how our model could be used in identifying sites involved in functional interactions to other molecules using the example of pectin methylesterase. Pectin, one of the main components of the plant cell wall, is secreted in a methyl esterified form and later de-esterified by pectin methylesterases. We used the crystal structure of pectin methyl-esterase (PDBID 1XG2) from our dataset and examined the differences in the $\log(site\ likelihood)$ between LG+ΔG and LG (Figure 2.6a). This initial mapping demonstrates that our general 1 partition energy model may not be sufficient to highlight regions involved in extra-protein interactions as there are many

poorly modeled regions well away from the active site. The additional improvement observed when comparing 6P_LG+ΔG to LG (Figure 2.6b) reveals that residues involved in the binding to substrate (Figure 2.6c), product (Figure 2.6d) or a protein inhibitor (Figure 2.6e) are generally worse modeled compared to the rest of the protein surface (Figure 2.6f). Interestingly, this type of analysis also revealed some evidence of problems with the structural constraints implied by our energy models. Sites involved in interactions with the inhibitor (Figure 2.6e) are modeled quite well by our general structural constraint model (Figure 2.6a) despite the fact that the conservation in the sequence here arises from interactions with the inhibitor and not a constraint for the overall protein fold. Figure A23 that compares 6P_LG+ΔG to LG+ΔG reveals that the active site and inhibitor binding site are much more poorly modeled for the 6 partition model relative to the single partition model. Although the 6 partition model fits better than LG at these sites, many residues on the fringes of the active site are poorly modeled with respect to LG+ΔG. This indicates that the additional structural specificity of information harvested by 6 partition models will lead to this model being 'positively misled' if there are important structural/functional features constraining sites that are not included in the energy calculations.

Similar analyses of the distributions of the PSSF for datasets containing interacting sites meeting our criteria inferred from the biomolecular interaction server (see methods) are summarized in Figure 2.7. We inspected all interacting sites in our protein structures and often found that at least one of our four models would highlight some of the residues involved in the binding sites defined by PISA. However, it is most often the case that these binding sites are relatively conserved and show at least some improvement in fit ($0<PSSF<1$) for all our structurally constrained models relative to non-energy models. Thus, we caution that, while our method can be used to illuminate sites involved extra-protein interactions, some sites may still be missed if negative site likelihood differences relative to non-energy models are the only criteria used to determine them.

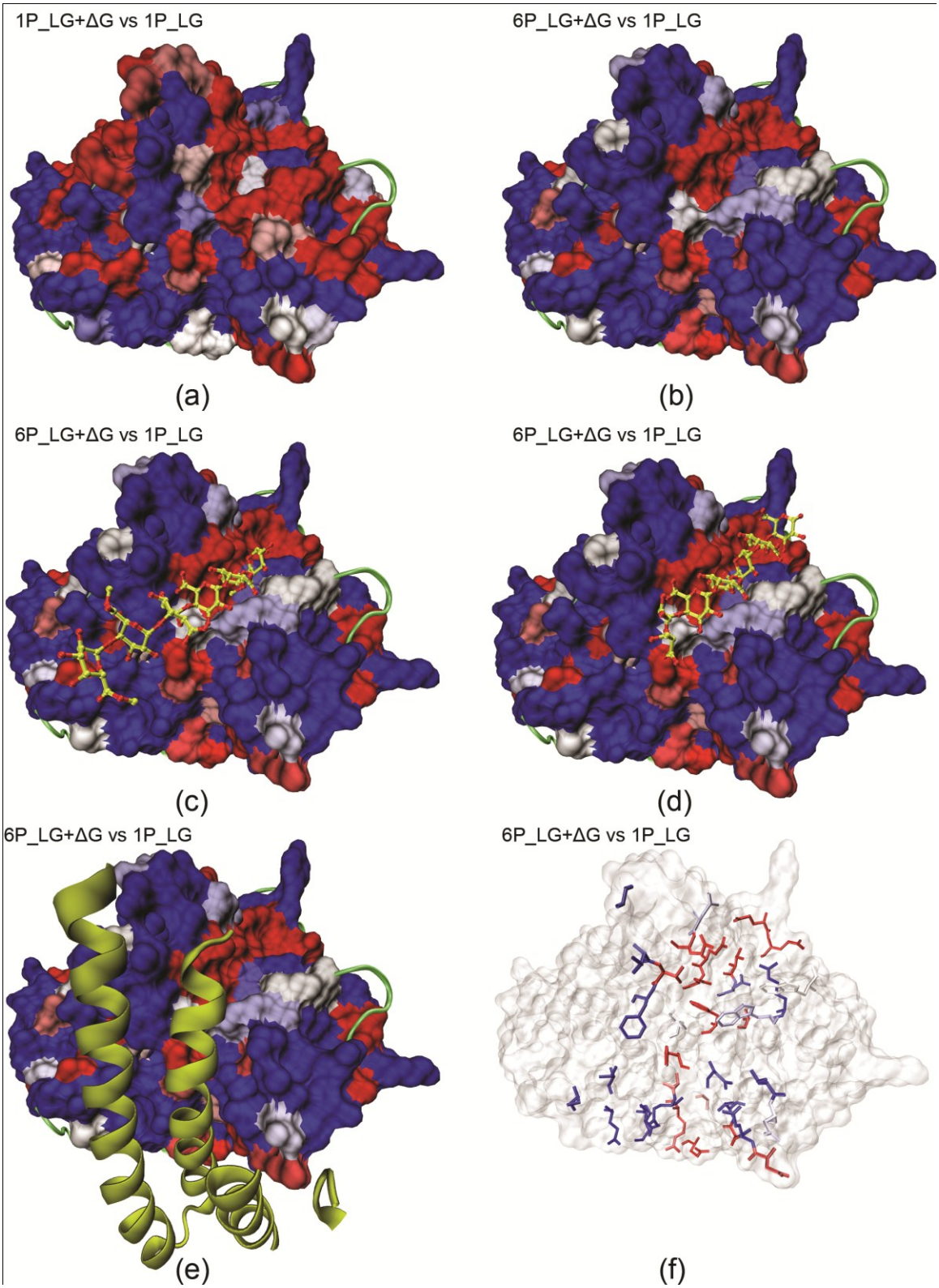

1P_LG+ΔG vs 1P_LG
(a)
6P_LG+ΔG vs 1P_LG
(b)
6P_LG+ΔG vs 1P_LG
(c)
6P_LG+ΔG vs 1P_LG
(d)
6P_LG+ΔG vs 1P_LG
(e)
6P_LG+ΔG vs 1P_LG
(f)

Figure 2.6: Using structurally constrained models of protein evolution to identify interaction sites.

Using FoldX + C_PROSA energies, we plot log(Site likelihood) differences (ld) between the same site modeled under two models and map the result onto the structure of pectin methyl-transferase (1XG2). **(a)** Compares LG to LG+ΔG to examine the effect of incorporating energies into the model. **(b)** Compares LG to 6P_LG + ΔG to establish the effect of partitioning on the model. Blue regions have ld>1 implying the more complex model gave a greater site-likelihood. Red regions indicate ld<0 LG fit that site best. Colors in between have 0<ld<1. Partitioning reveals a poorly modeled patch on the surface that corresponds well to substrate and product binding sites **[(c) ,(d)]** inferred by superposition from related protein structures (PDBID: 2NTB and 2NSP) as well as an inhibitor binding site present in the original crystal structure **(e)**. **(f)** Residues potentially engaged in hydrogen bonds (cut-off: 3.4Å), hydrophobic interactions (cut-off: 5Å), or simply making space for these interaction partners (cut-off: small residues "GASPVTCILN" within 5Å) seem to be poorly modeled with respect to the rest of the protein.

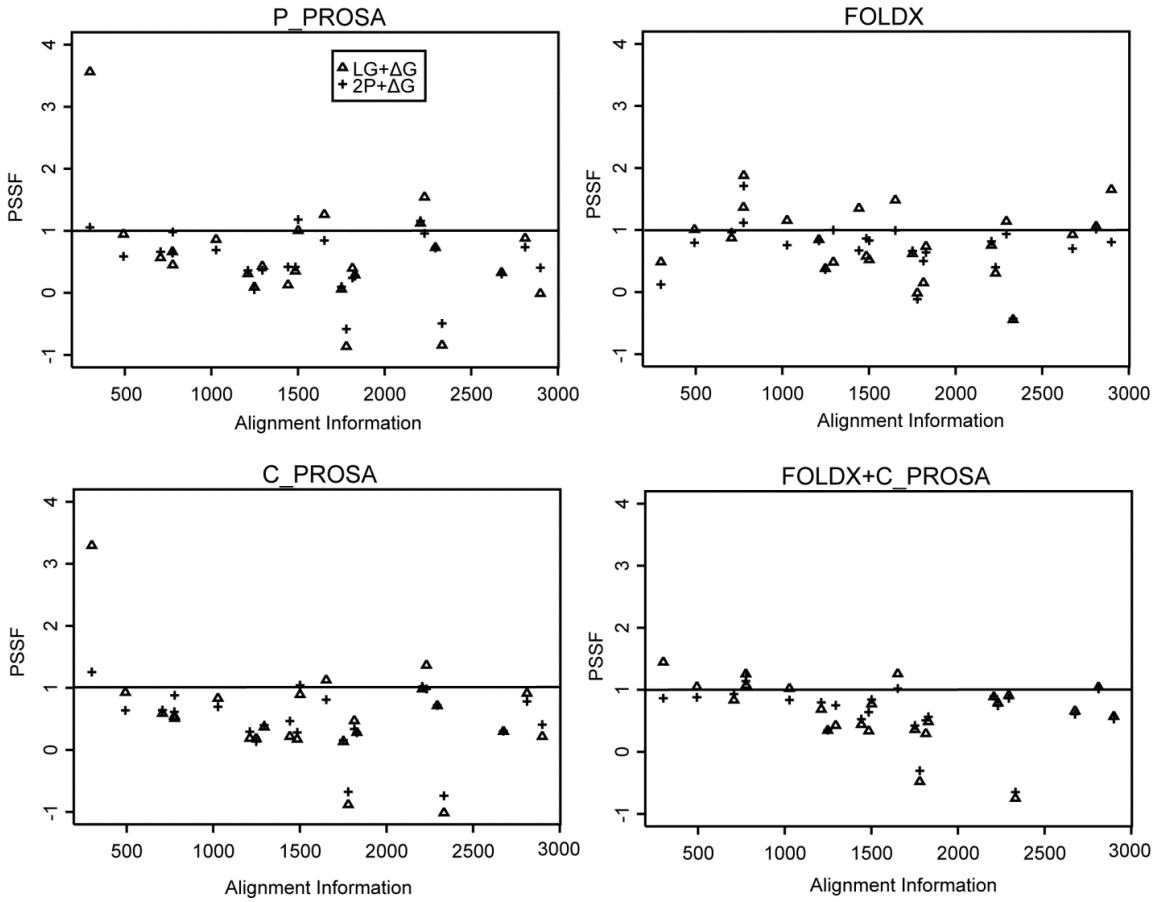


Figure 2.7: PSSF factors for interacting sites plotted against alignment information.

By treating sites involved in interactions as a structural partition, we plot the PSSF (see methods) for sites identified to have interacting partners to ligands or other proteins in 23 alignments from our dataset. We plot the PSSF against alignment information ([number of taxa] × [number of sequences] × [normalized branch length]) to illustrate a slight tendency for the PSSF to decrease with increasing alignment information. Plots are presented for LG+ΔG (LG+ΔG) and 2P_LG (2P) models utilizing P_ PROSA, C_PROSA, FoldX and FoldX+C_PROSA energies.

2.4 CONCLUSION:

We have developed a novel structurally constrained model of protein evolution that formalizes the independence energy model framework. Our model is implemented as a standard maximum likelihood phylogenetic model in which sites are assumed to evolve independently. In four models, we allow parameters to be estimated over the entire dataset or separately by solvent accessibility, secondary structure or both. In 16 variants of our model we tested each of these models using estimates for ΔG obtained from PROSA pairwise β-carbon-based pairwise statistical potentials (P_PROSA), combined PROSA pairwise and surface potentials (C_PROSA), the FoldX physics-based empirical force field (FoldX) and a combination of both FoldX and C_PROSA potentials (FoldX+C_PROSA).

Regardless of the potential used, we find statistically significant likelihood gains over LG from our simple energy model. Models using a combination of physical and statistical potentials (FoldX+C_Prosa) tended to outperform models using either potential on its own. The FoldX potential better modeled conserved residues and for residues in the hydrophobic core of the protein structure, while PROSA performed better at capturing more variable sites and on the hydrophilic exterior. Our mixture of potentials approach provides the best fit to both the hydrophobic core and solvent accessible region. Allowing parameters to be estimated separately for residues on the surface also helps to resolve biases in parameter estimates found within the protein core.

We implemented three additional variants of our models incorporating environment-specific substitution models (Le and Gascuel, 2010) as the base substitution model instead of LG. While the latter models do not generally perform as well as 'backing matrices' for the energy model as models utilizing LG we find that there still significant likelihood gains to be made by incorporating our structurally constrained model of protein evolution over these already structurally constrained general substitution models.

Amongst our models, we believe that the 2P_LG model is likely the best candidate model to extend in future implementations if the proteins of interest do not contain many sites and are very highly conserved. On the other hand, 6P_LG model heavily partitions the alignment (Table SI) and while it is generally the best performing

model, it requires more data (i.e. variable sites and sequences) to justify the additional adjustable parameters that must be estimated from the data. Of all the partitioned models, the 3P_LG model improves model fit over the non-partitioned energy model the least.

The structurally constrained model we have implemented is a comparatively simple SCPM because it is based on single point substitutions at a site and it cannot handle multiple substitution scenarios that are accounted for by both SEPMs and DEMs. This is not necessarily a negative feature however, as the accuracy of the free energy calculations is likely to become inaccurate as the sequence in question drifts away from that of the template structure used (Khatun et al., 2004). This is especially likely for empirical effective energy potentials which are trained to approximate experimental free energy changes in databases, such a ProTherm, of proteins with single to just a few substitutions (Kumar et al., 2006). Furthermore, by simplifying the process of generating rate matrices, we have the ability to explore the development of non-stationary structurally constrained models that allow drift in the protein structure across the tree. Such models could be invaluable in the development of more accurate methods for ancestral protein reconstruction and to serve as a null model upon which to detect instances of positive evolution.

CHAPTER 3 DISCUSSION

3.1 MODELS IMPLEMENTED:

3.1.1 EVOLUTIONARY CONSTRAINTS ON CODING SEQUENCES:

We have implemented and explored a variety of structurally constrained phylogenetic models that attempt to accommodate different biophysical constraints. We first introduced a structural constraint using a single parameterization across the entire dataset measuring the strength of the structural constraint (Equations 3-4). In this model, each site in an alignment is fit using a general substitution model adapted on a site-by-site basis to reflect structure-specific constraints on amino-acid replacement. Although this specification performs quite well with respect to a model not incorporating structural constraints at all, it has a problematic underlying assumption: that constraints can be uniformly applied to all sites in protein substructures in the same way.

It is not entirely clear that the relationship between structure and function, and more specifically the relationship between ΔG and protein fitness should be uniform across all structural environments. On one hand, the argument could be made that the relationship is negligible only at sites where substitutions could have small effects on ΔG, as is the case for many residues on the surface of the protein. However, destabilizing mutations do arise in our potentials at the protein surface that, if they are reliably calculated, may not affect the overall fold but instead induce only local structural rearrangements. If these substitutions are located distantly from the functional centers of a protein, they may have relatively insignificant effects on protein function and be considered neutral. In contrast, similarly destabilizing substitutions in the protein core, are more likely to affect not only the final protein product but also the kinetics of the folding process itself.

3.1.2 STRUCTURALLY CONSTRAINED PARTITION MODELS:

In Chapter 2 we challenged an underlying assumption made by all currently existing structurally constrained phylogenetic models that structural constraints can be applied uniformly across sites. We focused on the development of a series of structurally constrained partition models that allow for alternative parameterizations of our general model in differing secondary structure and solvent accessibility categories (Equations 5-

6). The improvements in model by these partitioned models is consistent with the notion that the relationship between ΔG and fitness differs across structural environments including solvent exposure and secondary structure. Additionally, while each site may be subject to its own substitution process, sites located within specific structural environments may share some similarities in their energetic preferences for (or against) certain amino acids. For example, sites located within alpha helices will probably exhibit low preference for transitions to proline while sites located within the core of a protein structure will exhibit elevated transitions between hydrophobic amino-acids and potentially reduced transitions to hydrophilic amino-acids. Therefore, another contributing factor towards the likelihood improvements we observe in partitioned structurally constrained models is in correcting potential errors in potentials by factoring in these general trends.

Understanding that the relationship between folding energy and function differs across structural environments opens the possibility for future research directions where mixture models that allow several alternative weightings on the structural constraints are used to model the evolution of sites in an unsupervised way.

3.2 OTHER CONSTRAINTS ON PROTEIN EVOLUTION IGNORED BY OUR MODELS.

3.2.1 MAINTENANCE OF EXTRA-PROTEIN INTERACTIONS REPRESENT ANOTHER SELECTIVE CONSTRAINT ON SITES.

Proteins must interact with other molecules such as ligands, metal co-factors, prosthetic-groups or other proteins in order to carry out their function. These interactions may be obligate or transient. Examples of obligate interactions include coordinated metal co-factors, prosthetic groups and the alternative domains of oligomeric-proteins, whereas transient interactions are exemplified by substrate (ligand) binding and many types of protein-protein interactions. Because they are usually directly related to the primary function of a protein, interacting sites are constrained to a degree that likely varies according to the importance of the interaction (Franzosa and Xia, 2009; Valdar WS and Thornton, 2001; Kim *et al.*, 2006; Eames and Kortemme, 2007). While maintaining functional interactions to other molecules and proteins is important, some have speculated that there is a related constraint to avoid potentially toxic interactions and

secondary activities with other molecules or other proteins in the cell (Zhang *et al.*, 2008; Deeds *et al.,* 2007).

3.2.2 INTRINSIC CONSTRAINTS ON TRANSLATIONAL ACCURACY AND THE FIDELITY OF PROTEIN FOLDING:

Translational accuracy and the fidelity of protein folding may also influence protein sequence evolution (Gingold and Pilpel, 2011; Yang *et al.*, 2010). Errors can occur during translation whereby the ribosome erroneously incorporates an amino acid into the growing polypeptide from a charged tRNA that has a one base mismatch relative to the correct codon on the mRNA. Such errors have been postulated to occur in yeast in the order of 10^{-5} errors per codon (Stanfields *et al*., 1998) and 10^{-2} errors per codon in *B. subtilis* (Meyerovich *et al*., 2010). Mistranslation errors that occur when tRNAs are incorrectly charged with the wrong amino acid are also postulated to occur at a frequency of 10^{-4} errors per amino-acyl tRNA synthesis reaction (Ibba and Sol, 2000). This potential for mistranslation error, along with the observation that, in *Drosophila,* sites that are strongly evolutionarily conserved tend to utilize optimal codons, has led to the proposal that selection favors optimal codons at sites where mistranslation errors could most greatly compromise protein function (Akashi, 1994). Mistranslation errors may not always be significant but can lead to missense substitutions or premature termination that can disrupt protein folding, leading to selection for coding sequences that translate with reduced translational error rates (Drummond *et al.*, 2005; Drummond and Wilke, 2008; Drummond and Wilke, 2009). Even error-free proteins can have a propensity to misfold (Dobson, 2003), and thus the fidelity of protein folding has also been suggested as a significant constraint on protein evolution (Drummond *et al.*, 2005; Drummond and Wilke, 2008; Drummond and Wilke, 2009; Yang *et al.*, 2010).

3.2.3 PROTEIN EXPRESSION LEVEL INCREASES THE RELEVANCE OF STRUCTURAL CONSTRAINTS ON PROTEIN EVOLUTION:

The rate at which different proteins accumulate substitutions over time varies. The most significant predictor of overall evolutionary rate appears to be the expression level of a protein; the more highly expressed a gene is, the slower it tends to evolve (Pal *et al.,* 2001; Krylov *et al.*, 2003; Drummond *et al.*, 2005; Lemos *et al.*, 2005). It is thought that

the strength of the various constraints imposed on protein sequences discussed above are amplified with increased expression level (Wolf *et al.*, 2010).

3.3 FUTURE RESEARCH DIRECTIONS:

The SSM, IEM and DEM structurally constrained protein evolution frameworks presented in the literature thus far rely on the idea that amino-acid exchangeabilities at a site remain constant across the tree. However, this is a problematic assumption because protein structures 'drift' over evolutionary time. Therefore, as these models are extended to problems involving more and more diverse sequences, the substitution model will fit less well to those sequences that are evolutionarily distant to the sequence of the reference structure. One simple way to address this problem of structural drift that should be explored would be to average substitution models generated from different structures in the alignment corresponding to relatively distantly-related sequences. However, there will still be a point at which sequences will have drifted too far to be fit well by even these averaged models. An alternative, more sophisticated, solution is to construct non-stationary models that 'drift'over the phylogenetic tree, allowing the substitutions in that part of the tree to be preferentially modeled by the nearest structure. Fortunately, the fact that we only work with 20 substitutions at each site in our model (in contrast to the DEM models) makes the generation of these non-stationary structural constraint models feasible.

3.3.1 A NON-STATIONARY STRUCTURALLY CONSTRAINED MODEL:

The premise behind the non-stationary structurally constrained model described above is that the model implied by a given structure will best 'fit' the sequences that are closest in sequence to the structure sequence in the phylogenetic tree. Furthermore, multiple sequences whose structures are available could be included in an analysis and each of these structures can be used to generate a substitution model. Then the resulting substitution models could be mixed in some way proportional to their distance away from the leaf structure used to generate the model for a given branch in the phylogenetic tree.

In such a bottom-up inference of the substitution process, the matrix **Q** would have to change across the tree as the likelihood is evaluated from the leaves to the root

node v_0. Consider a branch vu in a rooted tree with node v ancestral to u and having branch length t. In our classic structurally constrained model we calculate $P_{vu}(t)$ across the tree with the same $\mathbf{Q}$ used to model each branch for a given site. However, in a non-stationary model we wish to specify $\mathbf{Q}_{(vu)}$ at v as a mixture of *m* structurally constrained models originating from the subset of leaf taxa that have structures available.

One possibility is to model each branch as a mixture of alternative structurally constrained models in a way similar to that used to deal with rate categories across sites (Yang, 1994). The likelihood for a given site at a given branch would then be computed as the weighted average of the likelihoods calculated under each of *m* distinct models. The weights on each alternative model would have to be optimized as separate parameters for each branch. However, this is a very computationally complex solution and requires the estimation of (2*T*-3)(*m*-1) additional parameters for *T* taxa, requiring a large amount of data to obtain reliable parameter estimates.

Another alternative is to instead optimize a single parameter per structurally constrained model that determines a 'decay rate' of the structural constraint associated with a given model away from the leaf its sequence occupies. Below we demonstrate how our structurally constrained model could be used to determine the best weighted average of a set of structural constraints emanating from the sequences at the leaf taxa $[l_0, l_1, l_2 \ldots l_m]$ and we then increase the realism, hence complexity, of this model. We start with the model specification:

$$Q_{ij} \propto Q_{ij}^{E} Q_{ij}^{LG}$$

Where Q_{ij}^{E} is defined identically to Chapter 2. To obtain the best average of *m* structural constraints we could optimize a parameter set $P = [p_0, p_1, p_2 \ldots p_m]$ with the calculated values of ΔG_j for each alternative structural constraint such that $E_j = -p_0 \Delta G_j^{l_0} - p_1 \Delta G_j^{l_1} - p_2 \Delta G_j^{l_2} \ldots - p_k \Delta G_j^{l_m}$. Optimizing P over the entire alignment would lead to the best weighted average of several structural constraints across the tree, but this model would be stationary across the tree and not vary across branches.

We now convert the weighted average model described above to a model accounting for structural divergence from one sequence in an alignment to another across the tree. Modifying this weighted average involves making the strength of the structural

constraint P depend on the distance D=$[d_0, d_1, d_2 \ldots d_m]$ from leaves containing a structural constraint to the most ancestral node (v) in the branch vu being evaluated. As an example, each structural constraint could take one additional parameter ($s_0, s_1, s_2 \ldots s_m$) to allow for an exponential decay rate for the m'th energy term in E_j. The resulting non-stationary model would have E_j at v decay with distance and therefore be of the form:

$$E_j(D) = -p_0 e^{-s_0 d_0} \Delta G_j^{l_0} - p_1 e^{-s_1 d_1} \Delta G_j^{l_1} - p_2 e^{-s_2 d_2} \Delta G_j^{l_2} \ldots - p_m e^{-s_m d_m} \Delta G_j^{l_m}$$

3.3.2 APPLICATIONS OF NON-STATIONARY STRUCTURALLY CONSTRAINED PHYLOGENETIC MODELS:

ANCESTRAL PROTEIN SEQUENCE RECONSTRUCTION AND PROTEIN STRUCTURE PREDICTION:

A non-stationary model of this sort would represent a landmark advance in phylogenetic modeling capable of more accurately inferring ancestral protein structure constraints from a set of present-day known protein structures. Here, ancestral states at a node can be estimated by finding the state that maximizes the partial likelihood at that node. It would be interesting to see how our non-stationary structurally constrained model would perform in ancestral sequence reconstruction when compared to a variety of available methods.

A non-stationary structurally constrained phylogenetic model represents the first attempt to make a structural constraint leaf-specific. This affords an opportunity to treat the actual protein structure as partially unknown and optimized by the ML. Homology modeling, a technique used to infer a protein structure from closely-related sequences, typically produces a set of closely scoring protein structures that are similar to correct protein structure. This methodology could be extended by first resolving a set of protein structure decoys at a taxon and then ranking these decoys according to the likelihood that the structural constraints that they imply would lead to the substitution patterns observed in the sequence alignment. This could be applied to infer the structure of the protein at leaves where no structure is known but where a set of known structures is available for the alignment.

BENEFITS TO THE STRUCTURAL BIOLOGY COMMUNITY:

Non-stationary structurally constrained models benefit the structural biology community as well. From a structural biology perspective, arguments about biological function arise as a discussion of the critical intermolecular contacts that give rise to the dynamics and activity of a protein. There are readily available tools incorporated in molecular dynamics packages as well as a plethora of experimental methods that allow experimenters to test these biophysical hypotheses for any given protein structure. However, there is no similar package that assesses whether these biophysical hypotheses translate into structural constraints that leave an evolutionary 'footprint' in an alignment.

Non-stationary structurally constrained models could also be used to better capture the shifts in substitution patterns frequently observed to occur at functionally divergent sites. One of the greatest unfulfilled promises of the structurally constrained model framework is to find a way to take into account varying constraints for function across the tree. Known functional constraints could be also defined at each leaf and allowed to decay with increasing distance from the leaves only to eventually be replaced by functional constraints known to exist at other structurally constrained leafs.

Perhaps the most helpful feature of our model is the ability to test the structural constraint hypothesis on a site by site basis. As discussed in Chapter 2, sites with larger site likelihoods under our models when compared to non-structurally constrained models are likely to be under a structural constraint. However, sites where the improvement in model fit is small, or where non-structurally constrained models fit best are either evolving under a differing evolutionary process or are poorly modeled due to errors in the potentials used to postulate free energy changes. A non-stationary structurally constrained model would improve the performance of these models by helping to identify sites where a change in the substitution model is observed simply because of the change in a structural constraint.

3.3.3 IMPROVING STATISTICAL POTENTIALS FOR STRUCTURALLY CONSTRAINED PHYLOGENETIC MODELS.

A complete understanding of how a protein's overall three dimensional fold constrains its sequence depends on the accuracy with which biophysical characteristics

can be translated from known protein structures into evolutionary constraints. However, it is not always clear how a single point mutation might affect the function of a protein and the impact of multiple substitutions is even less predictable. Outside of a few model systems, very little is known about the mechanistic details of protein folding, or the dynamics required for function. This is true even without considering amino-acid substitutions, and the error associated with these kinds of predictions increases as more and more substitutions are taken into account at once. For this reason, the results obtained by phylogenetic energy models (including our own described in Chapter 2) should be treated with a degree of caution. In our own research, we have limited our models to incorporating biophysical constraints arising only from single point mutations to try to limit this source of error.

The focus of the research presented in this thesis has been on advancing the structurally constrained model framework as opposed to improving the accuracy with which biophysical traits of proteins can be expressed as a structural constraint. However, a reader interested in this area should consult Kleinman *et al.* (2010) where a series of statistical potentials were developed with terms related to a variety of biophysical constraints such as pairwise distance interactions, torsion angles, solvent accessibility, and flexibility of the residues.

Perhaps the most practical yet unaccounted biophysical constraint in structurally constrained models is a constraint to preserve critical contact between a protein and other ligands, co-factors or prosthetic groups. The reason that this has not been accounted for to date is that there are very few residues involved in explicit functional contacts relative to the number of residues involved in supporting interactions; Therefore, including a structural constraint for function would lead to minimal likelihood gains after exhausting attempts to create a generalized framework to deal with these molecular interactions. Perhaps a more interesting constraint to explore would be a constraint to avoid potentially toxic protein-protein or protein-ligand interactions. Such a constraint has been discussed in a review by Liberles (2012) but has yet to be incorporated in to a structurally constrained phylogenetic model due to the difficulty in determining how a constraint with so many unknowns might be visible to selection via single amino-acid replacements.

3.3.4 FINAL REMARKS

As a final word of advice to future researchers in the field of structural constrained phylogenetic modeling. Advancements in this field come from two types of innovation. The first is from improvements in the potentials used to infer relative fitness profiles for amino-acids on a site by site basis. As described above, there are many aspects outside of simply maintaining the protein fold to be explored when improving existing site-independent/dependent structurally constrained frameworks. The second is through innovations in the way these fitness profiles are included into existing phylogenetic models. Structurally constrained models have yet to adopt a philosophy of structural drift across the derived phylogeny. Success in this field requires a unique person with a clear understanding of mathematical modeling, numerical methods, protein structure and evolution. I highly recommend both these research directions to patient, enthusiastic and skilled bio-physicists, mathematicians or mathematically competent computer scientists.

LITERATURE CITED

Adachi J, Hasegawa M. 1996. Model of amino acid substitution and applications to mitochondrial protein evolution. J Mol Evol. 42:459-68.

Adachi J, Waddell PJ, Martin W, Hasagawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded in chloroplast DNA. J Mol Evol. 50:348-58.

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev. 8:688-93.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics. 136(3):927-35.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. Microbiol Rev. 54:198-210.

Bash PA, Singh UC, Langridge R, Kollman PA. 1987. Free energy calculations by computer simulation. Science. 236:564-8.

Bastolla U, Farwer J, Knapp EW, Vendruscolo M. 2001. How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins. 44:79-96.

Benedix A, Becker CM, de Groot BL, Caflisch A, Böckmann RA. 2009. Predicting free energy changes using structural ensembles. Nat Methods. 6(1):3-4.

Bonnard C, Kleinman CL, Rodrigue N, Lartillot N. 2009. Fast optimization of statistical potentials for structurally constrained phylogenetic models. BMC Evol Biol. 9:227.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem. 4: 187–217.

Bueno M, Camacho CJ, Sancho J. 2007. SIMPLE estimate of the free energy change due to aliphatic mutations: superior predictions based on first principles. Proteins. 68:850-62.

Cao Y, Adachi J, Janke A, Paabo S, Hasegawa M. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. J Mol Evol. 39:519-27.

Costantini S, Colonna G, Facchiano AM. 2006. Amino acid propensities for secondary structures are influenced by the protein structural class. Biochem Biophys Res Commun. 342:441-51.

Dang CC, Le QS, Gascuel O, Le VS. 2010. FLU, an amino acid substitution model for influenza proteins. BMC Evol Biol. 10:99.

Dayhoff MO, Barker WC, Hunt LT. 1983. Establishing homologies in protein sequences. Methods Enzymol. 91:524-45.

Dayhoff MO, Eck RV. 1968. Atlas of Protein Sequence and Structure. 1967-1968. National Biomedical Research Foundation, Silver Spring, Maryland.

Dayhoff MO, Schwartz RM, Orcutt BC. 1979. A model of evolutionary change in proteins. Pp. 345-352 in Atlas of Protein Sequence and Structure, Volume 5, Supplement 3, 1978, ed. MO Dayhoff. National Biomedical Research Foundation, Washington, D.C.

Deeds EJ, Ashenberg O, Gerardin J, Shakhnovich EI. 2007. Robust protein-protein interactions in crowded cellular environments. Proc Natl Acad Sci U S A. 104(38):14952-7.

DePristo MA, Weinreich DM, Hartl DL. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet. 6:678-87.

Dimmic M, Rest J, Mindell D, Goldstein R. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol. 55:65-73.

Dimmic MW, Mindell DP, Goldstein RA. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac Symp Biocomput. 2000:18-29.

Dobson CM. 2003. Protein folding and misfolding. Nature. 426(6968):884-90.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A. 104(7):2043-9.

Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. Mol Biol Evol. 24:388-97.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102(40):14338-43.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 134(2):341-52.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10(10):715-24.

Eames M, Kortemme T. 2007. Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. Structure. 15(11):1442-51.

Edgar RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity BMC Bioinformatics. 5:113.

Edgar RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-7.

Edwards AWF and Cavali-Svorza LL. 1963. The reconstruction of evolution. Annals of Human Genetics. 27: 105-106.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol. 17:368-76.

Felsenstein J. 2004. Inferring Phylogenies, Sinauer Associates

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39:W29-W37.

Fitch WM, Margoliash E. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet. 1:65-71.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. Science. 155:279 284

Fitch WM. 1986. An estimation of the number of invariable sites is necessary for the accurate estimation of the number of nucleotide substitutions since a common ancestor. Prog Clin Biol Res. 218:149–59.

Fornasari MS, Parisi G, Echave J. 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. Mol Biol Evol. 19:352-6.

Fornasari MS, Parisi G, Echave J. 2007. Quaternary structure constraints on evolutionary sequence divergence. Mol Biol and Evol. 24:349-51.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol. 26(10):2387-95.

Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. Mol Syst Biol. 7:481.

Goldberg AL. 2003. Protein degradation and protection against misfolded or damaged proteins. Nature. 426(6968):895-9.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 149:445-58.

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

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8:r49-r62.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science. 185:862-4.

Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol. 320:369-87

Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, Andreetta C, Boomsma W, Bottaro S, Ferkinghoff-Borg J. 2010. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PLoS One. 5:e13714.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the Human-Ape splitting by a molecular clock of mitochondrial-DNA. J Mol Evol. 22:160-74.

Ibba M, Soll D. 2000. Aminoacyl-tRNA synthesis. Annu Rev Biochem. 69:617-50.

James LC, Tawfik DS. 2003. Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. Trends Biochem Sci. 28:361-8.

Jiang B, Guo T, Peng L-W, Sun Z-R. 1998. Folding Type-Specific Secondary Structure Propensities of Amino Acids, Derived from α-Helical, β-Sheet, α/β, and α+β Proteins of Known Structures. Biopolymers. 45:35-49.

Johnston MA, Søndergaard CR, Nielsen JE. 2011. Integrated prediction of the effect of mutations on multiple protein characteristics. Proteins. 79:165-78.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS. 8:275-82.

Jones DT, Taylor WR, Thornton JM. 1994a. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry. 33:3038-49.

Jones DT, Taylor WR, Thornton JM. 1994b. A mutation data matrix for transmembrane proteins. FEBS Letters. 339:269-75.

Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C, Vriend G. 2011. A series of PDB related databases for everyday needs. Nucleic Acids Res. 39(Database issue):D411-9.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. pp. 21-132 in Mamalian Protein Metabolism, Vol. III, ed. M.N Munro. Academic Press, New York.

Juritz E, Palopoli N, Fornasari MS, Fernandez Alberti S, Parisi G. 2012. Protein conformational diversity modulates sequence divergence. Mol Biol Evol. Epub ahead of print, doi:10.1093/molbev/mss080.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22:2577-2637.

Khatun J, Khare SD, Dokholyan NV. 2004. Can contact potentials reliably predict stability of proteins? J Mol Biol. 336:1223-38.

Kidd KK and LA Sgaramella-Zonta. 1971. Phylogenetic analysis: Concepts and methods. American Journal of Human Genetics. 23: 235-252.

Kim PM, Lu LJ, Xia Y, Gerstein M.B. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. Science. 314:1938-41.

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

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

Kleinman CL, Rodrigue N, Lartillot N, Philippe H. 2010. Statistical potentials for improved structurally constrained evolutionary models. Mol Biol Evol. 27(7):1546-60.

Kleinman CL, Rodrigue N, Lartillot N, Philippe H. 2010. Statistical potentials for improved structurally constrained evolutionary models. Mol Biol Evol. 27:1546-60.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet. 11(7):487-98.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. Protein Engineering 8:641-5.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. Mol Biol Evol. 24:1464-79.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13(10):2229-35.

Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. 2006. ProTherm and ProNIT: thermodynamic databases for proteins and

protein-nucleic acid interactions. Nucleic Acids Res. 34(Database issue):D204-6.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307-20.

Le SQ, Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst Biol. 59:277-87.

Lemos B, Meiklejohn CD, Cáceres M, Hartl DL. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution. 59(1):126-37.

Liberles DA, Tisdell MD, Grahnen JA. 2011. Binding constraints on the evolution of enzymes and signalling proteins: the important role of negative pleiotropy. Proc Biol Sci. 278(1714):1930-5.

Liò P, Goldman N. 2002. Modeling mitochondrial protein evolution using structural information. J Mol Evol. 54(4):519-29

Lüthy R, McLachlan AD, Eisenberg D. 1991.Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins. 10(3):229-39.

Meyerovich M, Mamou G, Ben-Yehuda S. 2010. Visualizing high error levels during gene expression in living bacterial cells.
Proc Natl Acad Sci U S A. 107(25):11543-8.

Miyata T, Hayashida H, Yasunaga T, Hasegawa M. 1979. The preferential codon usages in variable and constant regions of immunoglobulin genes are quite distinct from each other. Nucleic Acids Res. 7:2431-8.

Momen-Roknabadi A, Sadeghi M, Pezeshk H, Marashi SA. 2008. Impact of residue accessible surface area on the prediction of protein secondary structures. BMC Bioinformatics. 9:357.

Muse SV, Gaut BS. 1994. A likelihood method for comparing synonymous and non-synonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11:715-724.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. Annu Rev Genomics Hum Genet. 11:265-89.

Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. 1(2):216-26.

Overington J, Johnson MS, Sali A, Blundell TL. 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc Biol Sci. 241(1301):132-45.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927-31.

Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol. 18:750-6.

Parisi G, Echave J. 2004. The structurally constrained protein evolution model accounts for sequence patterns of the LbetaH superfamily. BMC Evol Biol. 4:41.

Parisi G, Echave J. 2005. Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. Gene. 345:45-53.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007.Splicing and the evolution of proteins in mammals. PLoS Biol. 5(2):e14.

Pitera JW, Kollman PA. 2000. Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. Proteins. 41:385-97.

Plaxco KW, Larson S, Ruczinski I, Riddle DS, Thayer EC, Buchwitz B, Davidson AR, Baker D. 2000. Evolutionary conservation in protein folding kinetics. J Mol Biol. 298(2):303-12.

Prevost M, Wodak SJ, Tidor B, Karplus M. 1991. Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96----Ala mutation in barnase. Proc Natl Acad Sci USA. 88:10880-4.

Rakoczy EP, Kiel C, McKeone R, Stricher F, Serrano L. 2011. Analysis of disease-linked rhodopsin mutations based on structure, function, and protein stability calculations. J Mol Biol. 405(2):584-606.

Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. 2009. Protein sequences encode safeguards against aggregation. Hum Mutat. 30(3):431-7.

Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csürös M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H, Liu Y, Messina DN, Shen Y, Song HX, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AF, Ullmer B, Wang H, Xing J, Burhans R, Cheng Z, Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han SG, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu LL, Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'brien WE, Prüfer K, Stenson PD, Wallace JC, Ke H, Liu XM, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS. 2007.Evolutionary and biomedical insights from the rhesus macaque genome. Science. 316(5822):222-34.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol. 20:1692-704.

Rodrigue N, Kleinman CL, Philippe H, Lartillot N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Mol Biol Evol. 26:1663-76.

Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 347:207-17.

Rodrigue N, Philippe H, Lartillot N. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol. 23:1762-75.

Rykunov D, Fiser A. 2010. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics. 11:128.

Saitou N and Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution.* 4:406-425.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 9:56-68.

Schneider A, Cannarozzi GM, Gonnet GH. 2005. Empirical codon substitution matrix. BMC Bioinformatics. 6:134.

Schneider R, Sander C. 1996. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res. 24(1):201-5.

Self SG, Liang K-L. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc. 82:605–10.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33:1141-53.

Sharp PM, Stenico M, Peden JF, Lloyd AT. 1993. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans. 21:835-41.

Sippl MJ. 1993. Recognition of errors in three-dimensional structures of proteins. Proteins. 17:355-62.

Sippl MJ. 1995. Knowledge-based potentials for proteins. Curr Opin Struct Biol. 5:229-35.

Sokal, RR and Michener CD. 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 38: 1409-1438.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinf. 22:2688–90.

Stansfield I, Jones KM, Herbert P, Lewendon A, Shaw WV, Tuite MF. 1998. Missense translation errors in *Saccharomyces cerevisiae.* J Mol Biol. 282:13-24.

Valdar WS, Thornton JM. 2001. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins. 42(1):108-24.

Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. 2005. GROMACS: fast, flexible, and free. J Comput Chem. 26:1701-18.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691-9.

Willie E, Majewski J. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. 20:534-8.

Wolf YI, Gopich IV, Lipman DJ, Koonin EV. 2010. Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes. Genome Biol Evol. 2:190-9.

Worth CL, Blundell TL. 2009. Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. Proteins. 75(2):413-29.

Worth CL, Blundell TL. 2010. On the evolutionary conservation of hydrogen bonds made by buried polar amino acids: the hidden joists, braces and trusses of protein architecture. BMC Evol Biol. 10:161.

Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A. 109(14):E831-40.

Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol. 6:421.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Molecular biology and Evolution. 15: 1600-11.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39:306-14.

Yin S, Ding F, Dokholyan NV. 2007. Eris: an automated estimator of protein stability. Nat Methods. 4:466-7.

Zhang J, Maslov S, Shakhnovich EI. 2008. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. Mol Syst Biol. 4:210.

Zoller S, Schneider A. 2011. A new semiempirical codon substitution model based on principal component analysis of mammalian sequences. Mol Biol Evol. 29(1):271-7.

APPENDIX A- SUPPLEMENTARY FIGURES AND TABLES FOR CHAPTER 2

Supplementary Figure Legends:

Figure A1-A2: $\Delta\Delta G$ distributions obtained across all sites in 1XG2 to the 20 amino acids. Approximations to $\Delta\Delta G$ values obtained across sites when introducing point mutations into the wildtype pectin methyl-transferase (PDBID 1XG2) sequence using P_PROSA (A1) or FoldX (A2). The distributions of $\Delta\Delta G$ energy values are graphed for the 6 structural environments studied in this paper: Exposed Extended (B), Exposed Helix (H), Exposed Other (O), Buried Extended (b), Buried Helix (h) and Buried Other (o).

Figure A3: Subsample selected from the 300 test sets used in Le and Gascuel, 2009. A representative subsample of 48 sequence alignments (large black points) was selected from amongst the 300 test alignments used in Le and Gascuel (2009) (small black points).

Figure A4: Dominance of partition models over standard single partition models : (a) ω Determines the weight given to a partition-specific model in the evaluation of the rate matrix and stationary frequencies. Estimates for ω tend to be highest for 2 partition models, lowest for 3 partition models and intermediate for 6 partition models. While incorporating energies into the model tended to reduce the magnitude of ω, ω tended to be large indicating that a partition-specific model was often preferred to a model without partitioning.

Figure A5: mean and standard deviation of $\boldsymbol{\alpha_M}$: $\boldsymbol{\alpha_{LG}}$: each model M studied in the paper resulted in a maximum likelihood estimate for an 8Γ rates across sites parameter $\boldsymbol{\alpha_M}$. We graph the ratio of $\boldsymbol{\alpha_M}$ to that obtained using the standard LG model for models utilizing P_PROSA, FoldX, C_PROSA and FoldX+C_PROSA energies. No significant effect on alpha was determined.

Figure A6: mean and standard deviation of the $\mathbf{normalized\_branch\_length_M}$: $\mathbf{normalized\_branch\_length_{LG}}$: each model M studied in the paper resulted in a maximum likelihood estimate for the branch lengths. We graph the ratio of the average branch length ($\boldsymbol{nbl_M}$) under a model M to that obtained using the standard LG model ($\boldsymbol{nbl_M}$: $\boldsymbol{nbl_{LG}}$).

Figure A7-A10: Median stationary frequencies obtained for the 2P_LG, 3P_LG and 6P_LG partition models utilizing P_PROSA (A7), C_PROSA (A8), FoldX (A9) or FoldX+C_PROSA (A10) energies:

(a) Median stationary frequencies derived for the 2P_LG model in exposed and buried structural classes obtained from a subsample of 1092 sites from each structural category. **(b)** Median stationary frequencies derived for the 3P_LG model derived similarly for subsamples of 1092 sites from each structural category. **(c)** Median stationary frequencies derived for the 6P_LG model derived similarly for subsamples of 1092 sites from each structural category. Median stationary frequencies are again presented for the LG+ΔG model (Black dashed lines). In each case, we compare to the median stationary frequency from 1092 sites sampled randomly from our dataset under the LG+ΔG model (Black dashed lines).

Figure A11-A14: Boxplots for the distributions of stationary frequencies: Boxplots of the stationary frequency distributions observed for various amino acids for the 4 models analyzed. The stationary frequencies have been presented separately for each structural category evaluated by the model. Each figure was generated by subsampling 1092 sites from the appropriate model/structural category combination. Boxplots are presented for P_PROSA (A11), C_PROSA (A12), FoldX (A13) or FoldX+C_PROSA (A14) energies.

Figure A15-A18: Median Q ratio matrices for models utilizing a combination of FOLDX and PROSA energies. After running our structurally constrained partition models in all 48 datasets we randomly sampled 1092 sites being analyzed under a particular structural-category/model combination. We calculate the ratio of the median Q_{ij} entries found using the various models analyzed in this paper to the median entries found when using the standard LG model. For these figures we removed the top 5% of outliers to ensure the median as a measure of central tendency. Clearly the major contributor to model improvement is solvent accessibility. Plots are presented for P_PROSA (A15), C_PROSA (A16), FoldX (A17) or FoldX+C_PROSA (A18) energies.

Figure A19-A22: mean average deviation away from the median Q_{ij} matrices. After running our structurally constrained partition models in all 48 datasets we randomly sampled 1092 Q_{ij} matrices from sites being analyzed under a particular structural-category/model combination. We calculate the mean average deviation from the median Q_{ij} as a measure of variance for the model. For these figures we removed the top 5% of outliers to ensure the median as a measure of central tendency. Clearly the major contributor to model improvement is solvent accessibility. Plots are presented for P_PROSA (A19), C_PROSA (A20), FoldX (A21) or FoldX+C_PROSA (A22) energies.

Figure A23: Comparing back to the general structurally constrained model can more boldly reveal regions of the protein structure selected for function. Looking for patches on the protein surface that are poorly modeled under our model may lead to functionally relevant insights. Using FoldX + C_Prosa energies, we plot log(Site likelihood) differences (ld) between the same site modeled under two models and map the result onto the structure of pectin methyl-transferase (1XG2). In **(a)** we examine the effect of incorporating energies into the model by comparing LG to LG+ΔG while in **(b)** we examine the effect of partitioning by comparing LG+ΔG to 6P_LG + ΔG. Blue regions represent regions where ld>1 indicating that the more complex model gave a greater likelihood for that site. Red regions indicate ld<0 indicating that the simpler model fit that site best. Colors in between have 0<ld<1. Partitioning reveals a large poorly modeled patch on the surface that corresponds well to substrate and product binding sites **[(c) ,(d)]** inferred by superposition from related protein structures (PDBID: 2NTB and 2NSP) as well as an inhibitor binding site present in the original crystal structure **(e)**. **(f)** Residues potentially engaged in hydrogen bonds (cut-off: 3.4Å), hydrophobic interactions (cut-off: 5Å), or simply making space for these interaction partners (cut-off: small residues "GASPVTCILN" within 5Å) seem to be poorly modeled with respect to the rest of the protein. Many more sites here seem poorly modeled when compared to those in figure 6.

Table A1: Alignment statistics for datasets used in to assess structurally constrained models.

PDB	#sites	#Taxa	ntbl	Ex	Bu	B	α	O	Exβ	Exα	ExO	Buβ	Buα	BuO
1ONL	103	53	0.14	60	43	42	15	45	17	11	31	25	4	14
1QUP	121	18	0.14	85	36	47	19	55	23	14	48	24	5	7
1QWD	121	61	0.13	86	35	62	4	46	37	3	37	25	1	9
1S0L	122	33	0.09	82	40	56	0	66	22	0	60	34	0	6
1BYR	126	25	0.16	76	50	38	41	46	14	24	37	24	17	9
2F1F	131	74	0.08	83	48	55	38	37	24	27	31	31	11	6
1QQ0	146	35	0.14	110	36	37	0	70	16	0	55	21	0	15
1VDD	151	93	0.09	97	54	17	56	78	4	32	61	13	24	17
1Y1O	153	41	0.12	103	50	50	28	42	20	15	35	30	13	7
2DKF	163	88	0.16	75	88	46	31	86	13	15	47	33	16	39
1C8U	165	84	0.12	110	55	55	36	74	28	20	62	27	16	12
1H4U	172	12	0.16	111	61	75	14	75	45	2	56	30	12	19
2A8E	194	14	0.12	125	69	37	78	78	10	54	60	27	24	18
1HG3	209	22	0.22	102	107	35	79	94	1	42	58	34	37	36
1HYN	213	35	0.08	154	59	38	47	73	13	27	59	25	20	14
1XNF	230	34	0.16	150	80	0	166	58	0	96	48	0	70	10
1L9X	243	14	0.19	132	111	64	72	80	17	39	49	47	33	31
2DP5	256	21	0.04	248	8	10	21	225	7	20	221	3	1	4
1E6E	265	44	0.15	150	115	37	101	124	15	50	82	22	51	42
2GA9	267	18	0.13	178	89	32	126	85	15	70	69	17	56	16
1XG2	282	89	0.09	144	138	103	7	172	26	4	114	77	3	58
1Q2B	293	68	0.09	165	128	110	7	175	45	5	114	65	2	61
2H85	303	28	0.07	209	94	72	65	166	41	37	131	31	28	35
1WUF	304	33	0.15	163	141	68	101	117	29	51	65	39	50	52
1CRZ	306	47	0.10	198	108	118	33	155	55	16	127	63	17	28
2DGK	310	57	0.12	140	170	40	97	159	11	45	70	29	52	89
1HYO	316	46	0.14	172	144	79	63	172	22	35	113	57	28	59
1T4D	327	91	0.09	186	141	59	129	139	12	73	101	47	56	38
1QZ9	332	42	0.13	199	133	37	122	172	8	72	118	29	50	54
3GLY	337	37	0.12	169	168	12	164	161	4	58	107	8	106	54
1O98	339	93	0.15	182	157	52	119	167	10	65	106	42	54	61
2GLF	352	34	0.15	180	172	86	81	185	24	44	112	62	37	73
1Q78	354	40	0.09	236	118	34	133	167	17	73	126	17	60	41
2F82	357	43	0.08	184	173	72	138	147	17	66	101	55	72	46
1R9Z	364	57	0.10	209	155	70	128	159	18	78	106	52	50	53
1CIY	383	38	0.10	234	149	70	179	130	38	93	99	32	86	31
1ZVL	388	47	0.07	231	157	50	130	208	21	68	142	29	62	66
1DKM	392	16	0.08	233	159	63	147	182	24	70	139	39	77	43
2BH9	392	48	0.10	245	147	90	108	194	29	59	157	61	49	37
2F7F	432	56	0.14	258	174	95	148	186	37	82	136	58	66	50
2A3L	439	49	0.10	323	116	29	127	198	4	73	161	25	54	37
1BFD	453	38	0.16	251	202	75	159	218	17	78	155	58	81	63
2OLB	472	44	0.14	263	209	89	140	243	45	72	146	44	68	97
2H4M	477	20	0.13	317	160	0	294	145	0	156	123	0	138	22
1QBA	520	40	0.09	289	231	123	113	284	58	54	177	65	59	107
1LNS	617	15	0.17	338	279	99	182	336	32	86	220	67	96	116
1QLN	739	11	0.20	473	266	46	355	317	18	200	234	28	155	83
2GHO	897	44	0.06	534	363	94	76	727	41	28	465	53	48	262

* Abbreviations: exposed (Ex), buried (Bu), Extended (β), Helix (α), Other (O), buried extended (Buβ), buried helix (Buα), buried other (BuO), exposed extended (Exβ), exposed helix (Exα) and exposed (ExO).

Table A2:

Model	Number of parameters $\|\mathbf{T}\| = \mathbf{1}$	Number of parameters $\|\mathbf{T}\| = \mathbf{2}$
1P	1	1
1P+ΔG	2	3
2P	40	40
2P+ΔG	42	44
3P	62	62
3P+ΔG	62	65
6P	116	116
6P+ΔG	122	128

Table A3: Comparing 3P+ΔG models to LG: Partition-specific significance factors and the % of likelihood differences greater than 0.

		All sites		Extended		Helix		Other	
Energy type	**Residue Hydrophobicity Class[a]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**
P_PROSA	All Sites	1	65%	1.19	67%	1.18	66%	0.78	63%
	Hydrophillic	1.05	67%	0.31	55%	1.03	65%	1.28	71%
	Hydrophobic	0.85	63%	1.71	75%	1.01	65%	0.12	54%
	Other	1.42	65%	1.1	62%	2.4	71%	0.73	60%
C_PROSA	All Sites	1	66%	1.14	68%	1.2	68%	0.79	64%
	Hydrophillic	0.99	67%	0.21	53%	1.02	65%	1.2	71%
	Hydrophobic	0.92	66%	1.69	77%	1.08	69%	0.24	56%
	Other	1.39	65%	1.04	64%	2.3	72%	0.77	61%
FoldX	All Sites	1	76%	1.4	79%	1.05	78%	0.78	72%
	Hydrophillic	0.57	67%	0.49	65%	0.51	65%	0.63	69%
	Hydrophobic	1.41	85%	1.95	90%	1.45	89%	1.01	79%
	Other	1.02	70%	1.21	69%	1.27	77%	0.74	65%
FoldX + C_PROSA	All Sites	1	77%	1.26	80%	1.04	79%	0.85	75%
	Hydrophillic	0.75	72%	0.52	66%	0.67	70%	0.86	75%
	Hydrophobic	1.20	83%	1.68	88%	1.25	85%	0.84	76%
	Other	1.17	73%	1.33	71%	1.50	80%	0.85	68%
		All sites		Extended		Helix		Other	
Counts	All Sites	15028		3087		5019		6922	
	Hydrophillic	6605		1024		2016		3565	
	Hydrophobic	6826		1793		2408		2625	
	Other	1597		270		595		732	

a: Sites containing > 80% hydrophobic/hydrophilic residues labeled as such. All other sites pooled into an Other category.

b: partition-specific significance factor = = $\overline{ld}_C/\overline{ld}_D$

c: % of sites in the category with likelihood differences greater than 0.

Table A4: Comparing 6P+ΔG models to LG: Partition-specific significance factors and the % of likelihood differences greater than 0.

		All sites		Exposed		Buried	
Energy type	**Residue Hydrophobicity Class[a]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**	**PSSF[b]**	**%>0[c]**
P_PROSA	All Sites	1	66%	1.06	64%	0.93	68%
	Hydrophillic	1.11	66%	1.45	70%	0.23	56%
	Hydrophobic	0.9	67%	0.14	52%	1.26	73%
	Other	0.97	60%	1.18	60%	0.42	60%
C_PROSA	All Sites	1	70%	1.05	69%	0.94	71%
	Hydrophillic	1.05	71%	1.38	76%	0.22	57%
	Hydrophobic	0.88	69%	0.14	54%	1.22	77%
	Other	1.31	66%	1.45	66%	0.94	64%
FoldX	All Sites	1	79%	0.9	74%	1.12	85%
	Hydrophillic	0.79	75%	0.87	75%	0.6	73%
	Hydrophobic	1.17	85%	0.86	75%	1.32	90%
	Other	1.11	71%	1.08	69%	1.21	76%
FoldX+C_PROSA	All Sites	1	80%	0.93	76%	1.08	84%
	Hydrophillic	0.85	76%	0.96	78%	0.56	72%
	Hydrophobic	1.10	84%	0.72	73%	1.28	90%
	Other	1.21	74%	1.19	73%	1.25	77%

		All sites	Exposed	Buried
Counts	All Sites	15028	8106	6922
	Hydrophillic	6605	4757	1848
	Hydrophobic	6826	2181	4645
	Other	1597	1168	429

a: Sites containing > 80% hydrophobic/hydrophilic residues labeled as such. All other sites pooled into an Other category.

b: partition-specific significance factor = $= \overline{ld}_C / \overline{ld}_D$

c: % of sites in the category with likelihood differences greater than 0.

Figure A1: C_PROSA energies

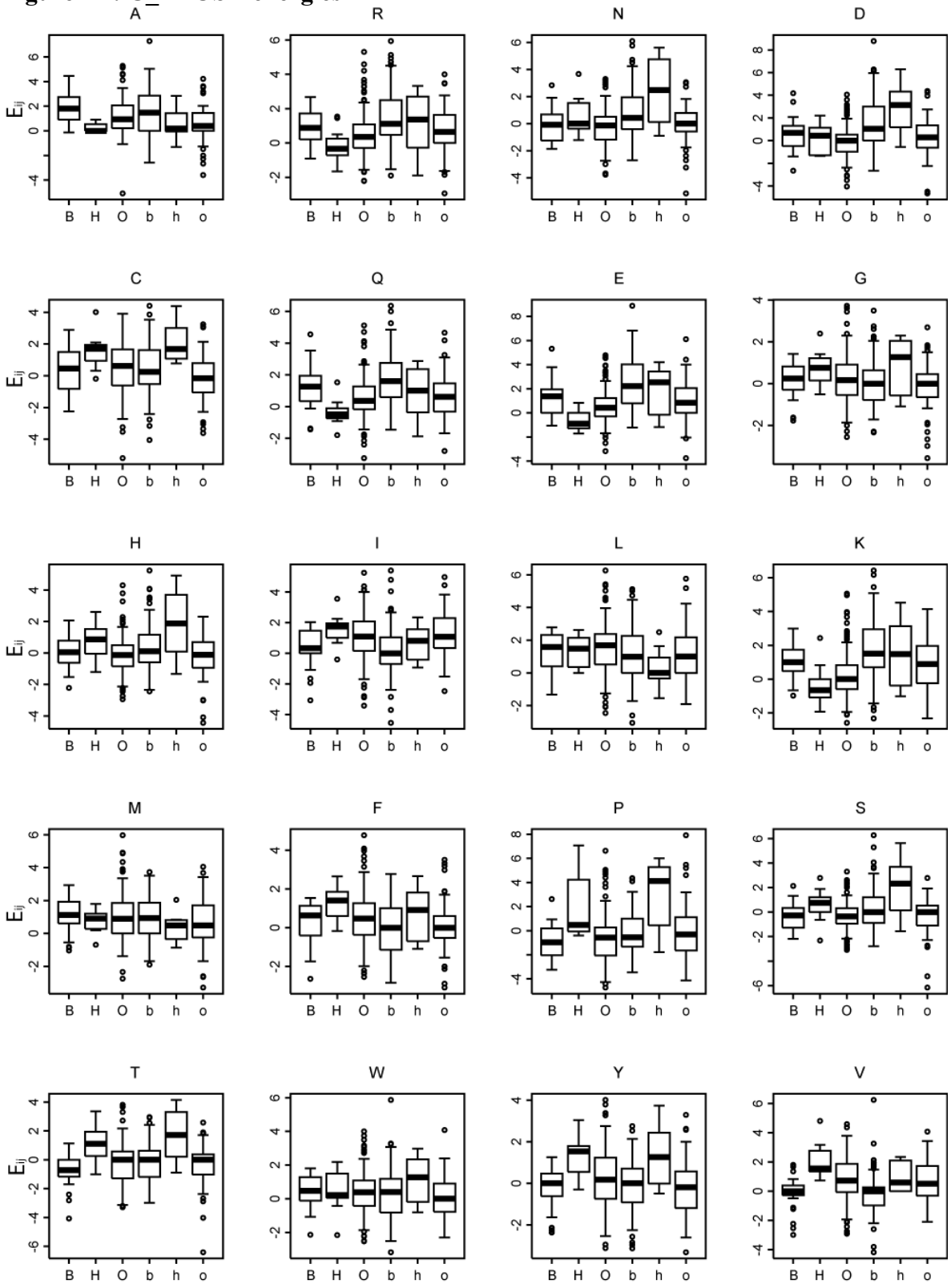


Figure A2: FoldX energies

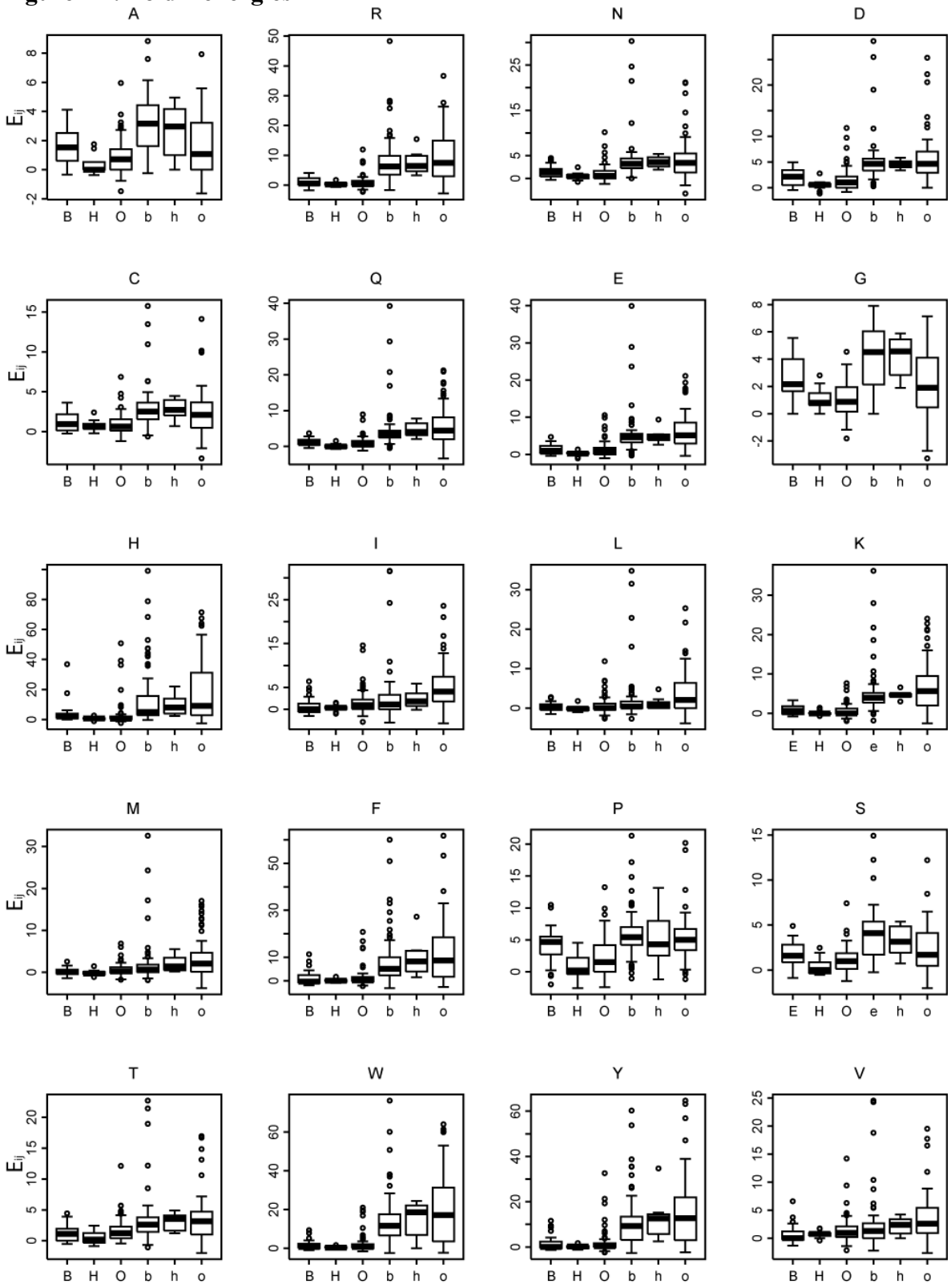


Figure A3

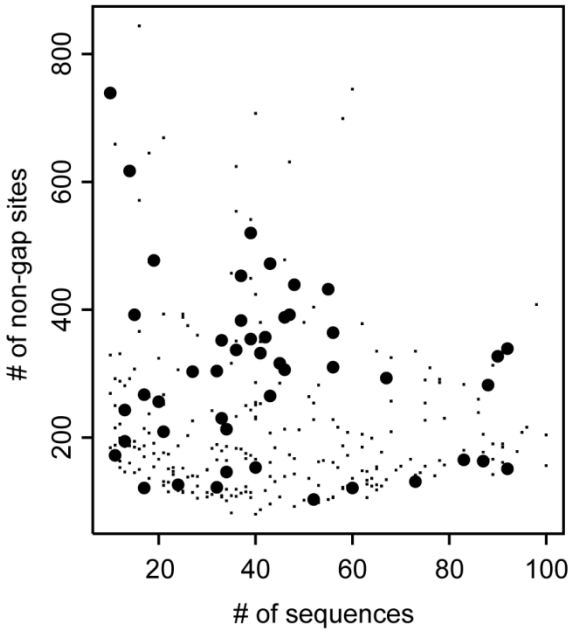


Figure A4

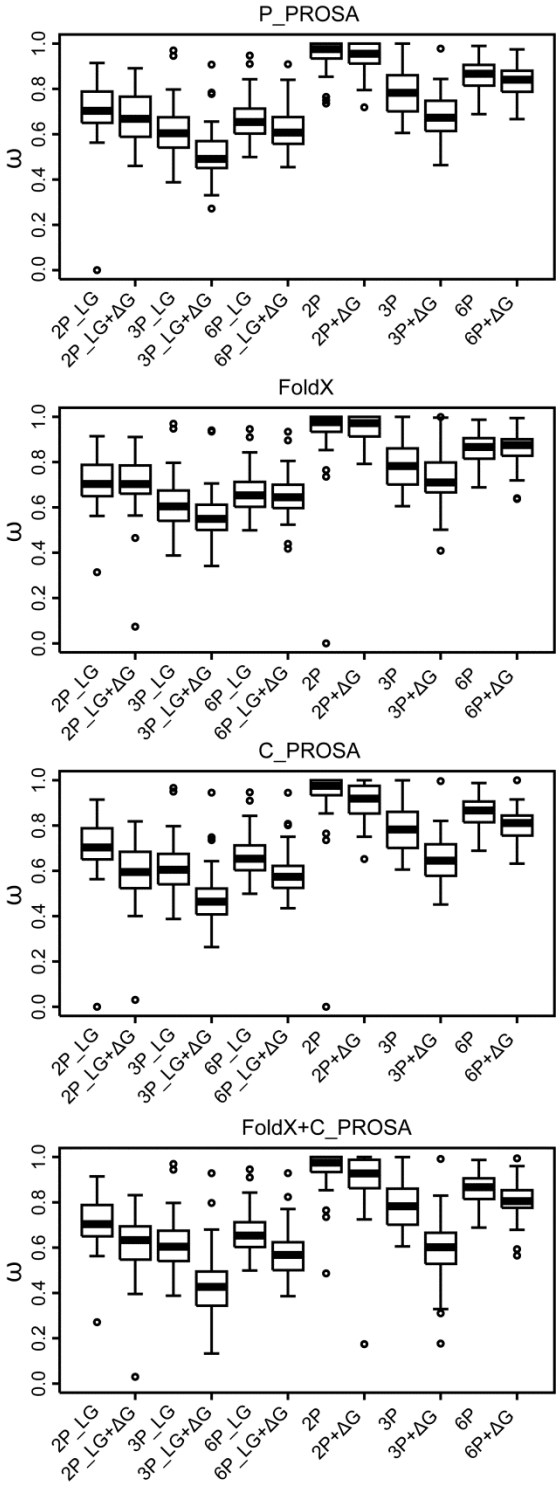

P_PROSA
FoldX
C_PROSA
FoldX+C_PROSA
ω
0.0
0.2
0.4
0.6
0.8
1.0
2P_LG
2P_LG+ΔG
3P_LG
3P_LG+ΔG
6P_LG
6P_LG+ΔG
2P
2P+ΔG
3P
3P+ΔG
6P
6P+ΔG

Figure A5

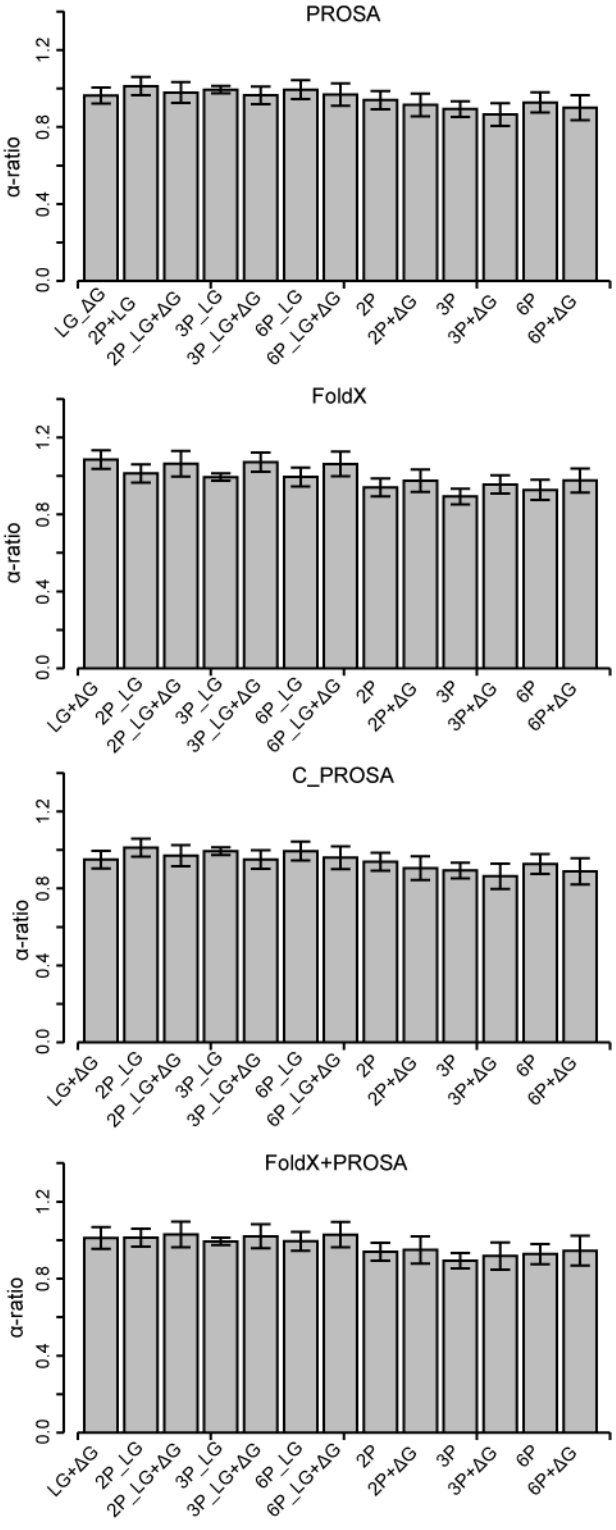

PROSA
α-ratio
0.0
0.4
0.8
1.2
LG_ΔG
2P+LG
2P_LG+ΔG
3P_LG
3P_LG+ΔG
6P_LG
6P_LG+ΔG
2P
2P+ΔG
3P
3P+ΔG
6P
6P+ΔG
FoldX
α-ratio
0.0
0.4
0.8
1.2
LG+ΔG
2P_LG
2P_LG+ΔG
3P_LG
3P_LG+ΔG
6P_LG
6P_LG+ΔG
2P
2P+ΔG
3P
3P+ΔG
6P
6P+ΔG
C_PROSA
α-ratio
0.0
0.4
0.8
1.2
LG+ΔG
2P_LG
2P_LG+ΔG
3P_LG
3P_LG+ΔG
6P_LG
6P_LG+ΔG
2P
2P+ΔG
3P
3P+ΔG
6P
6P+ΔG
FoldX+PROSA
α-ratio
0.0
0.4
0.8
1.2
LG+ΔG
2P_LG
2P_LG+ΔG
3P_LG
3P_LG+ΔG
6P_LG
6P_LG+ΔG
2P
2P+ΔG
3P
3P+ΔG
6P
6P+ΔG

Figure A6

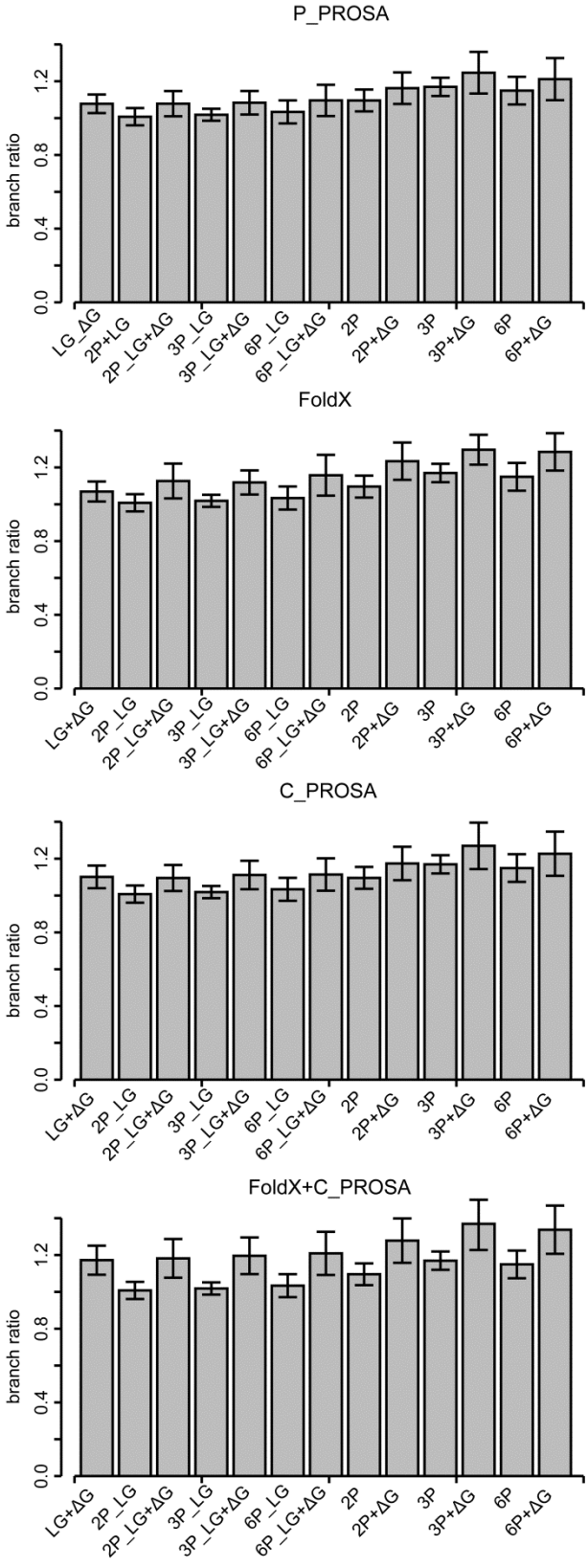

P_PROSA
FoldX
C_PROSA
FoldX+C_PROSA
branch ratio
0.0
0.4
0.8
1.2
LG_ΔG
2P+LG
2P_LG+ΔG
3P_LG
3P_LG+ΔG
6P_LG
6P_LG+ΔG
2P
2P+ΔG
3P
3P+ΔG
6P
6P+ΔG
LG+ΔG
2P_LG
2P_LG+ΔG
3P_LG
3P_LG+ΔG
6P_LG
6P_LG+ΔG
2P
2P+ΔG
3P
3P+ΔG
6P
6P+ΔG

Figure A7

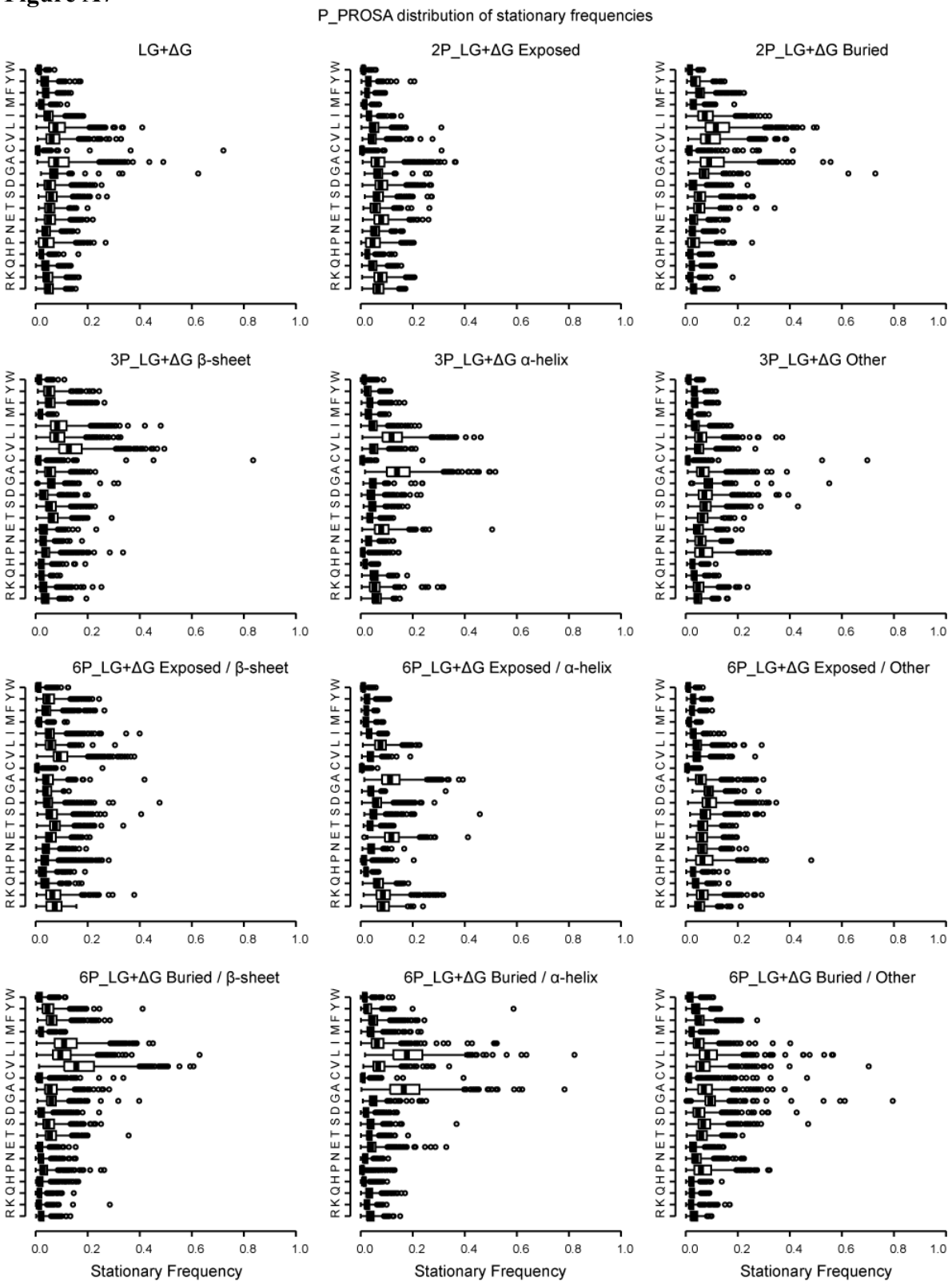

P_PROSA distribution of stationary frequencies
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
R K Q H P N E T S D G A C V L I M F Y W
0.0 0.2 0.4 0.6 0.8 1.0
Stationary Frequency

Figure A8

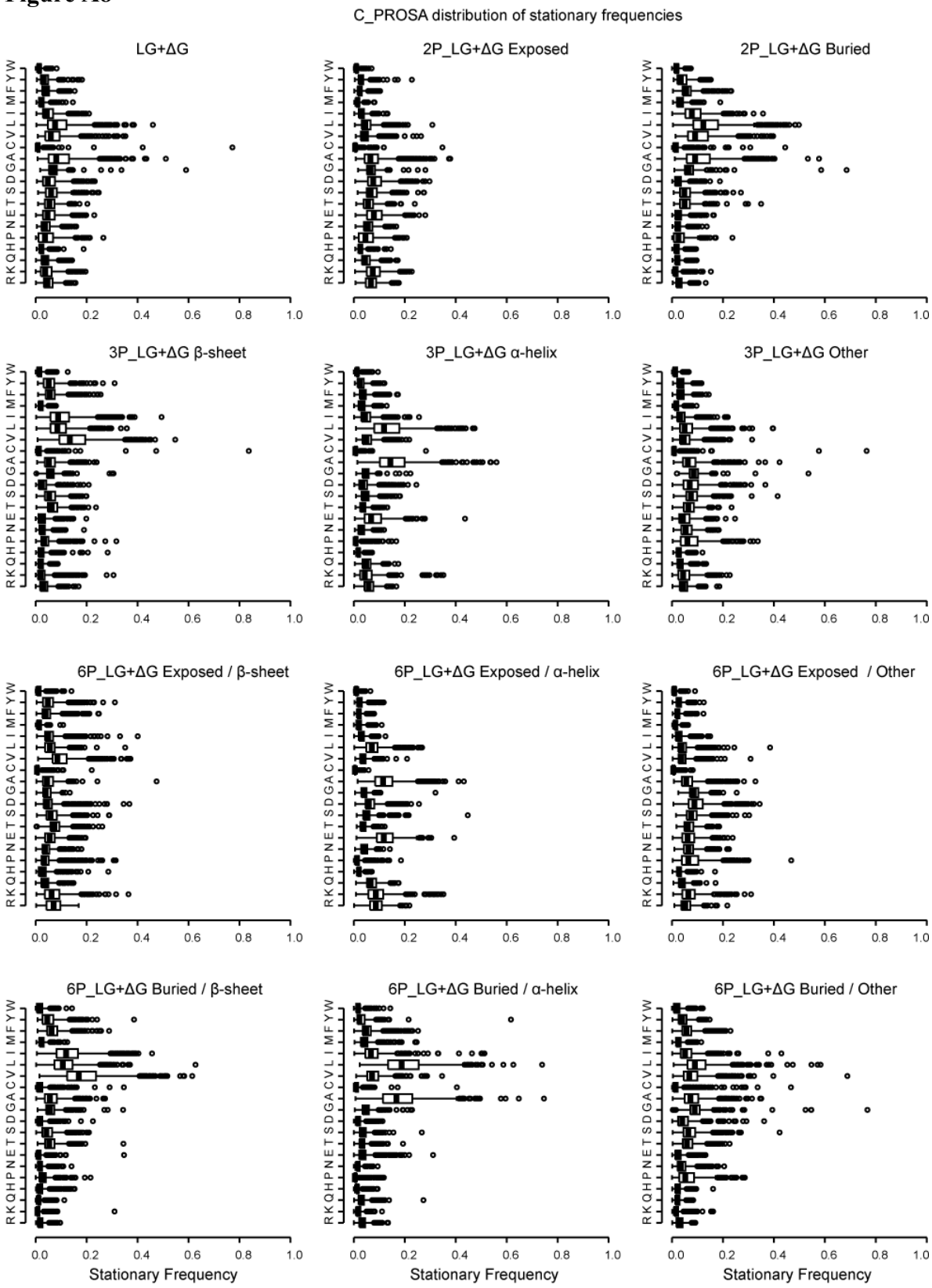

C_PROSA distribution of stationary frequencies
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
R K Q H P N E T S D G A C V L I M F Y W
0.0 0.2 0.4 0.6 0.8 1.0
Stationary Frequency

Figure A9

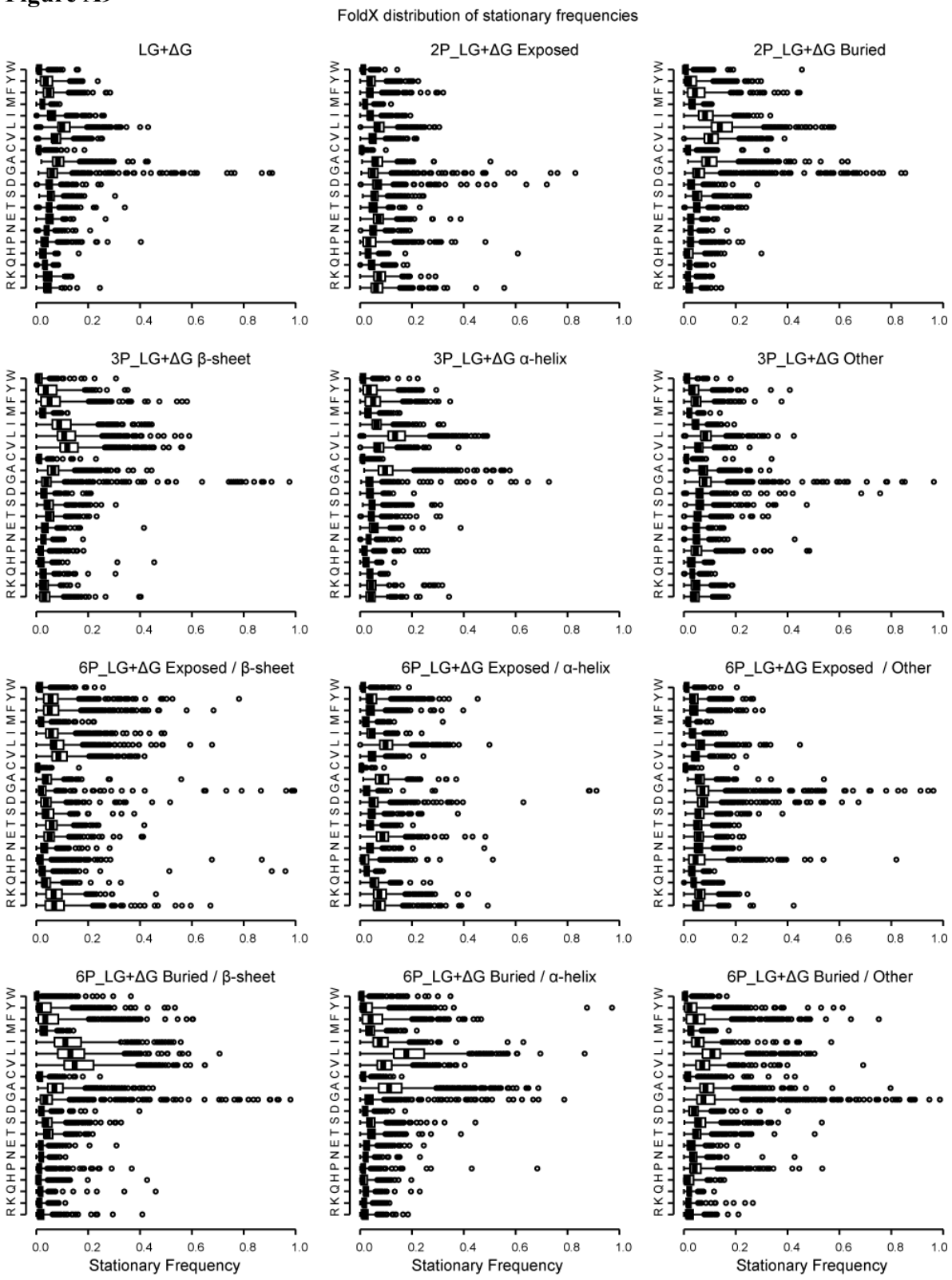

FoldX distribution of stationary frequencies
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
R K Q H P N E T S D G A C V L I M F Y W
0.0 0.2 0.4 0.6 0.8 1.0
Stationary Frequency

Figure A10

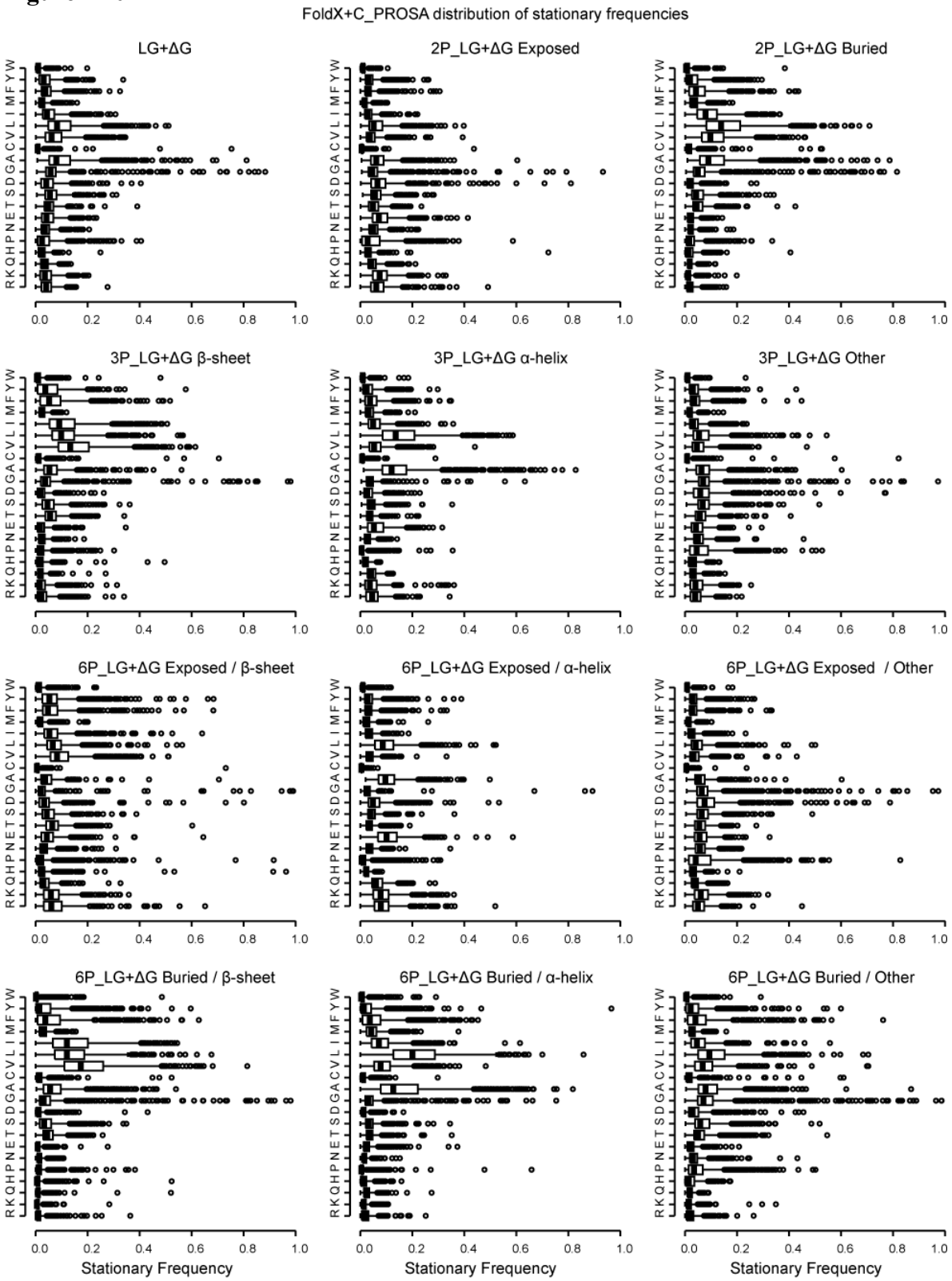

FoldX+C_PROSA distribution of stationary frequencies
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
R K Q H P N E T S D G A C V L I M F Y W
0.0 0.2 0.4 0.6 0.8 1.0
Stationary Frequency

Figure A11

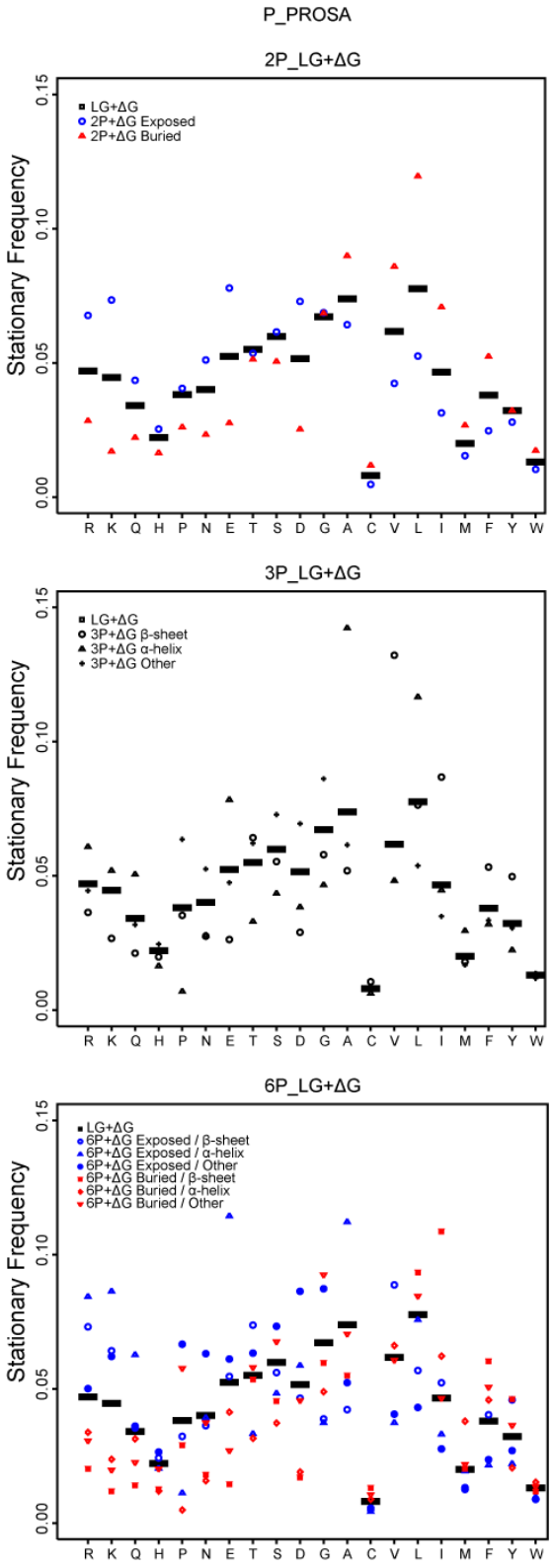

P_PROSA
2P_LG+ΔG
LG+ΔG
2P+ΔG Exposed
2P+ΔG Buried
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
3P_LG+ΔG
LG+ΔG
3P+ΔG β-sheet
3P+ΔG α-helix
3P+ΔG Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
6P_LG+ΔG
LG+ΔG
6P+ΔG Exposed / β-sheet
6P+ΔG Exposed / α-helix
6P+ΔG Exposed / Other
6P+ΔG Buried / β-sheet
6P+ΔG Buried / α-helix
6P+ΔG Buried / Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W

Figure A12

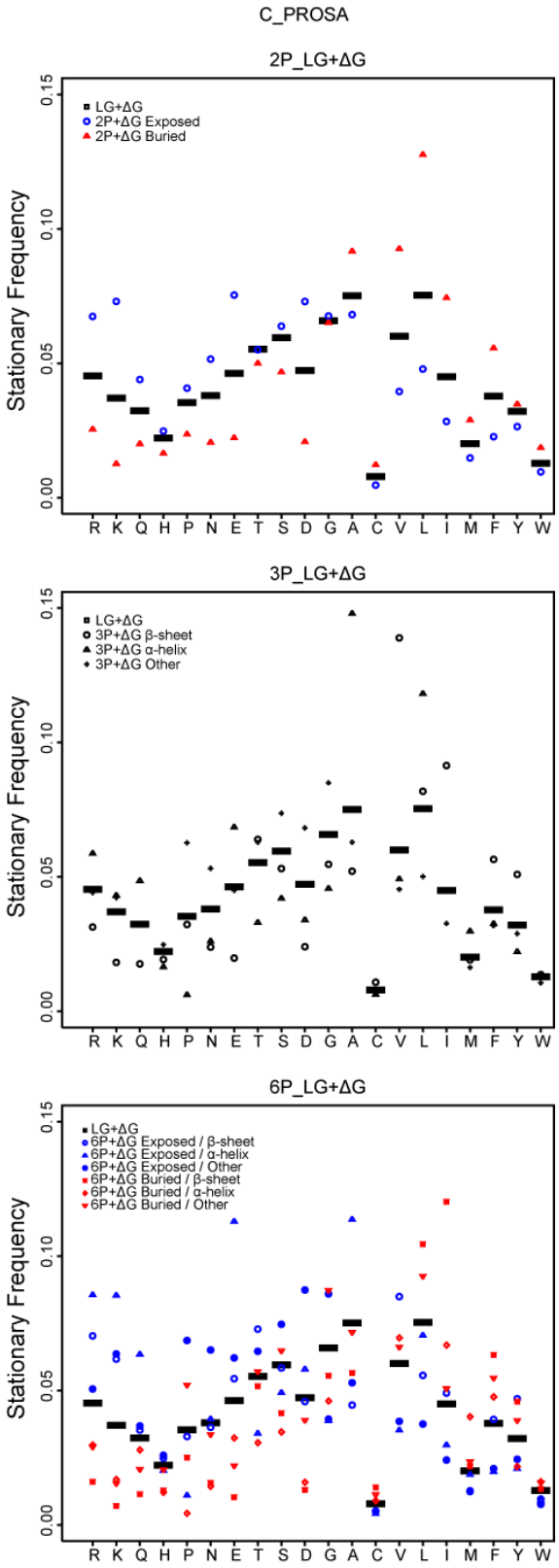

C_PROSA
2P_LG+ΔG
LG+ΔG
2P+ΔG Exposed
2P+ΔG Buried
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
3P_LG+ΔG
LG+ΔG
3P+ΔG β-sheet
3P+ΔG α-helix
3P+ΔG Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
6P_LG+ΔG
LG+ΔG
6P+ΔG Exposed / β-sheet
6P+ΔG Exposed / α-helix
6P+ΔG Exposed / Other
6P+ΔG Buried / β-sheet
6P+ΔG Buried / α-helix
6P+ΔG Buried / Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W

Figure A13

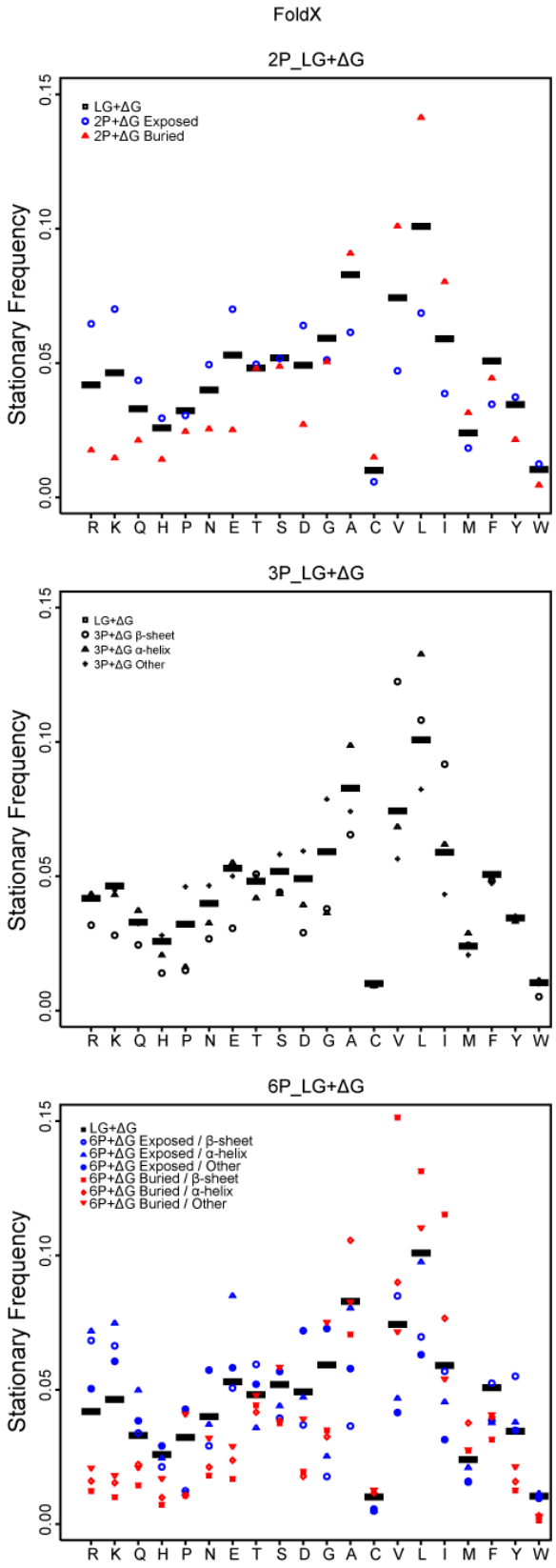

FoldX
2P_LG+ΔG
LG+ΔG
2P+ΔG Exposed
2P+ΔG Buried
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
3P_LG+ΔG
LG+ΔG
3P+ΔG β-sheet
3P+ΔG α-helix
3P+ΔG Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
6P_LG+ΔG
LG+ΔG
6P+ΔG Exposed / β-sheet
6P+ΔG Exposed / α-helix
6P+ΔG Exposed / Other
6P+ΔG Buried / β-sheet
6P+ΔG Buried / α-helix
6P+ΔG Buried / Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W

Figure A14

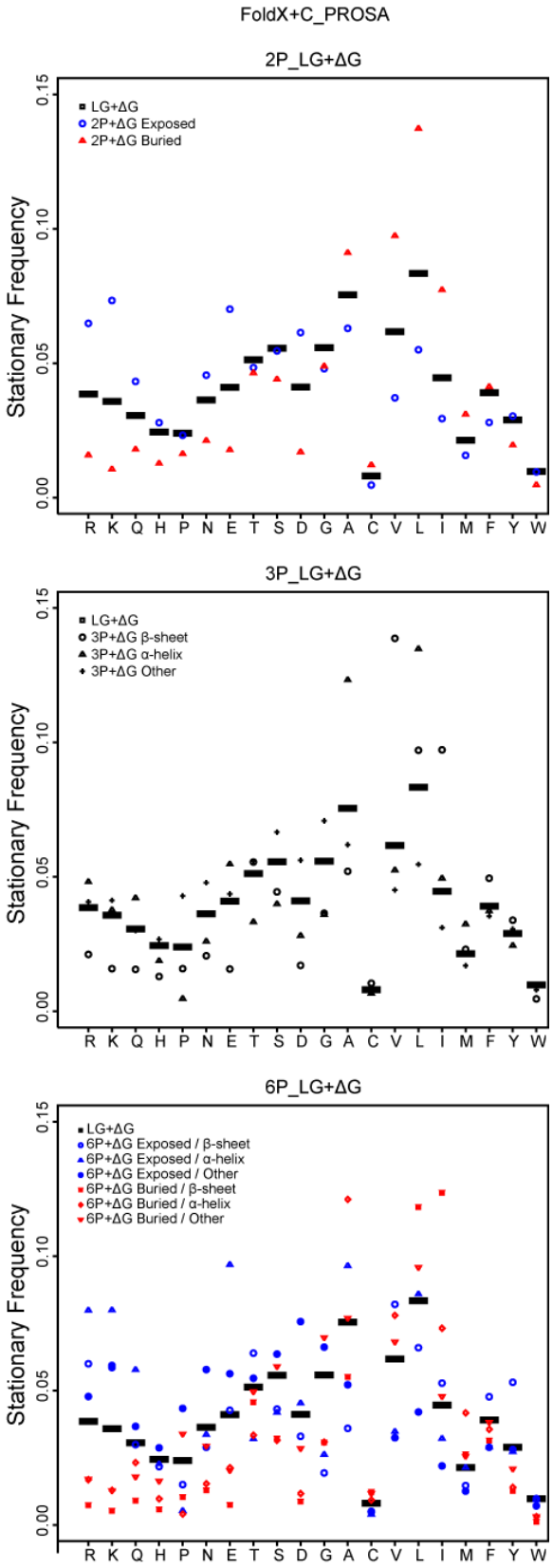

FoldX+C_PROSA
2P_LG+ΔG
LG+ΔG
2P+ΔG Exposed
2P+ΔG Buried
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
3P_LG+ΔG
LG+ΔG
3P+ΔG β-sheet
3P+ΔG α-helix
3P+ΔG Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W
6P_LG+ΔG
LG+ΔG
6P+ΔG Exposed / β-sheet
6P+ΔG Exposed / α-helix
6P+ΔG Exposed / Other
6P+ΔG Buried / β-sheet
6P+ΔG Buried / α-helix
6P+ΔG Buried / Other
Stationary Frequency
0.00
0.05
0.10
0.15
R K Q H P N E T S D G A C V L I M F Y W

Figure A15

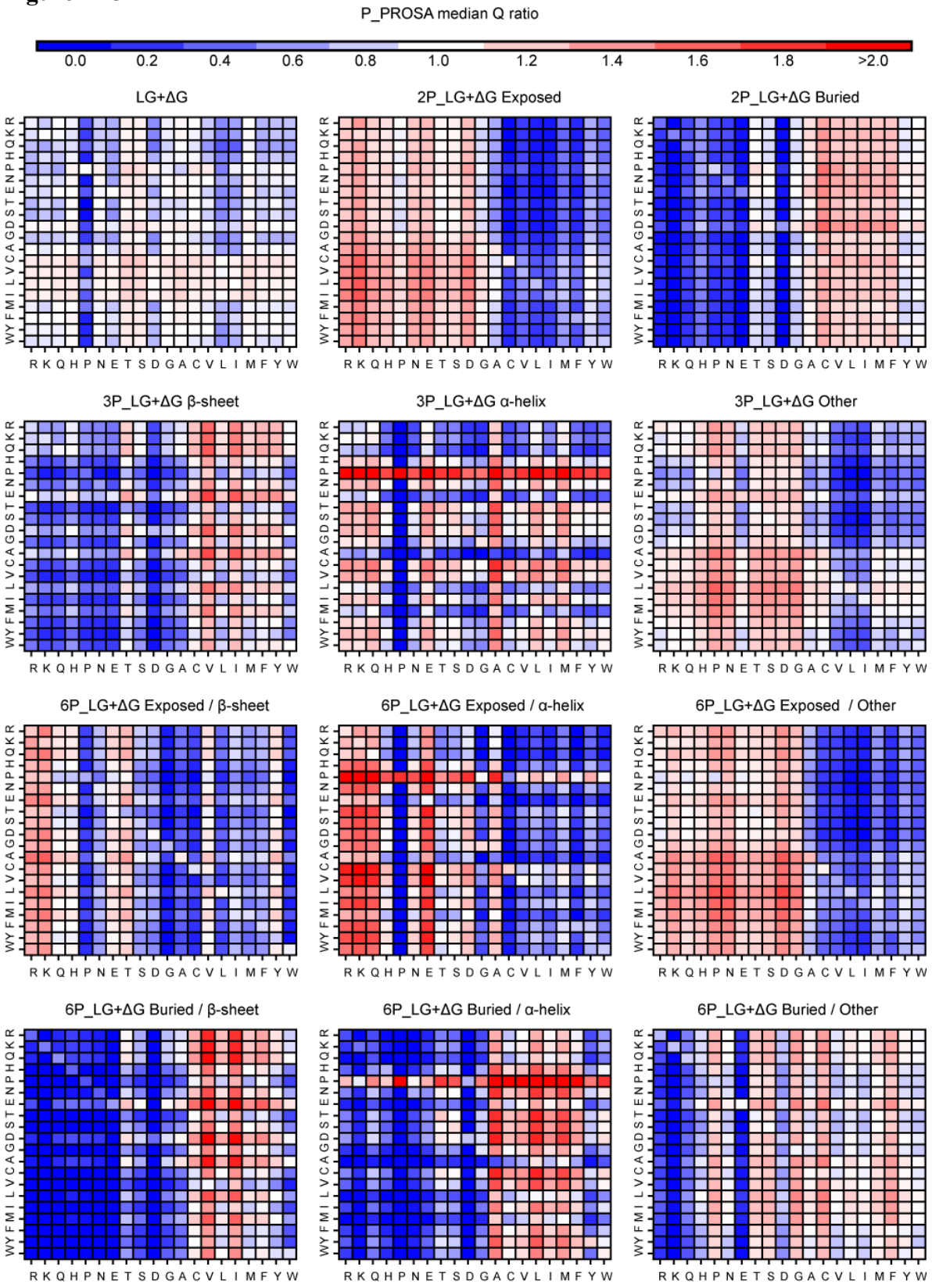

P_PROSA median Q ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
>2.0
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure A16

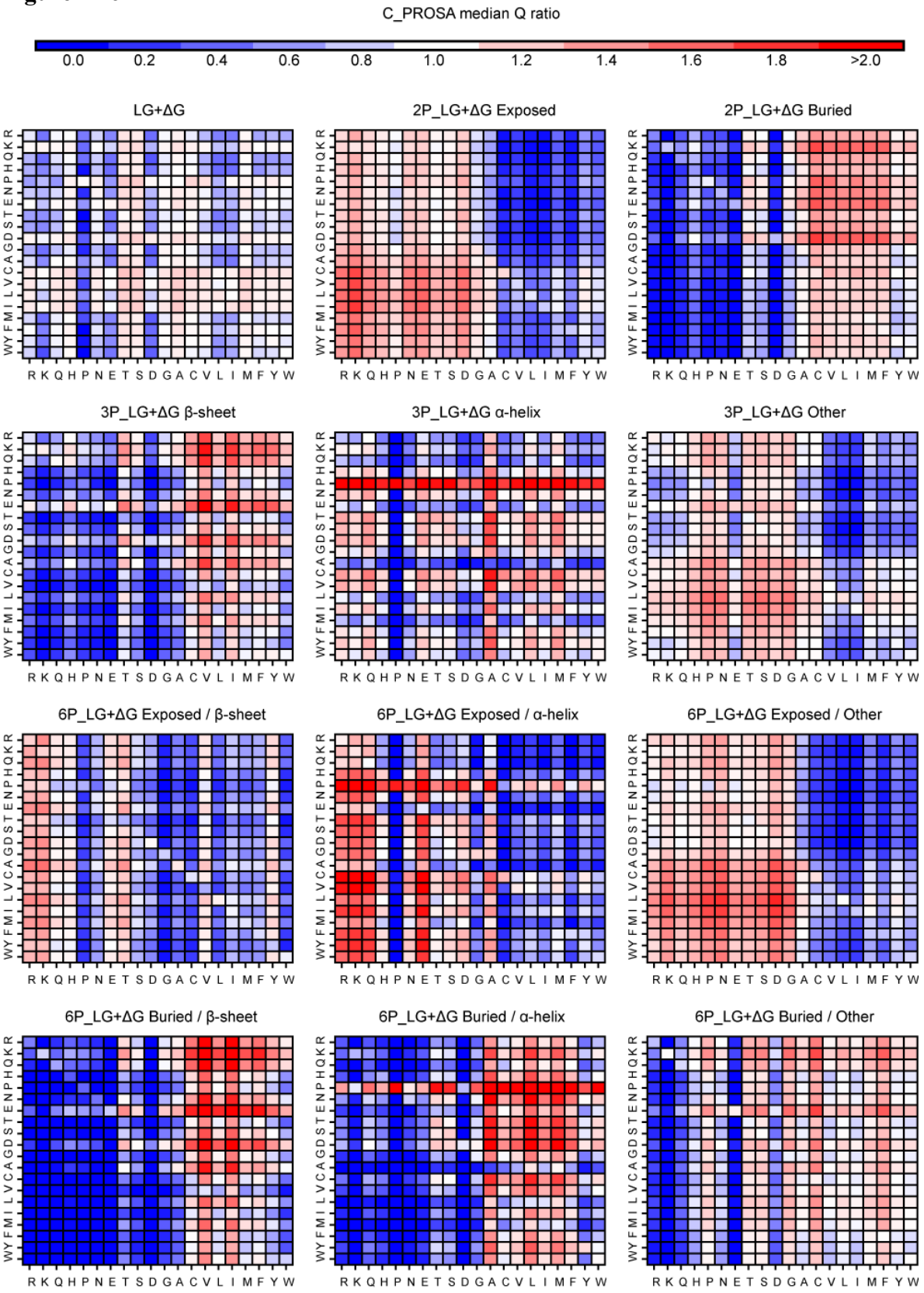

C_PROSA median Q ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
>2.0
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure A17

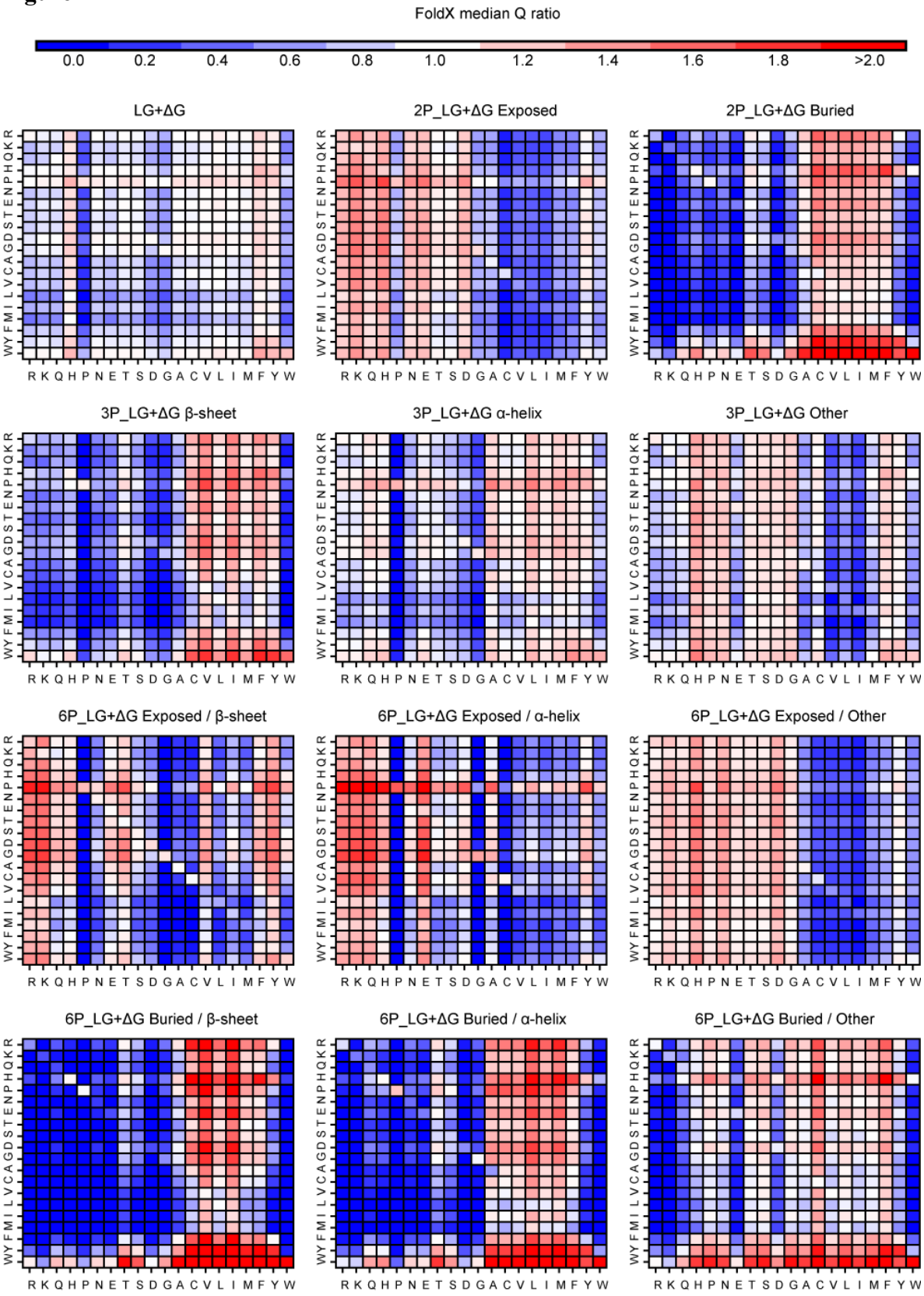

FoldX median Q ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
>2.0
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure A18

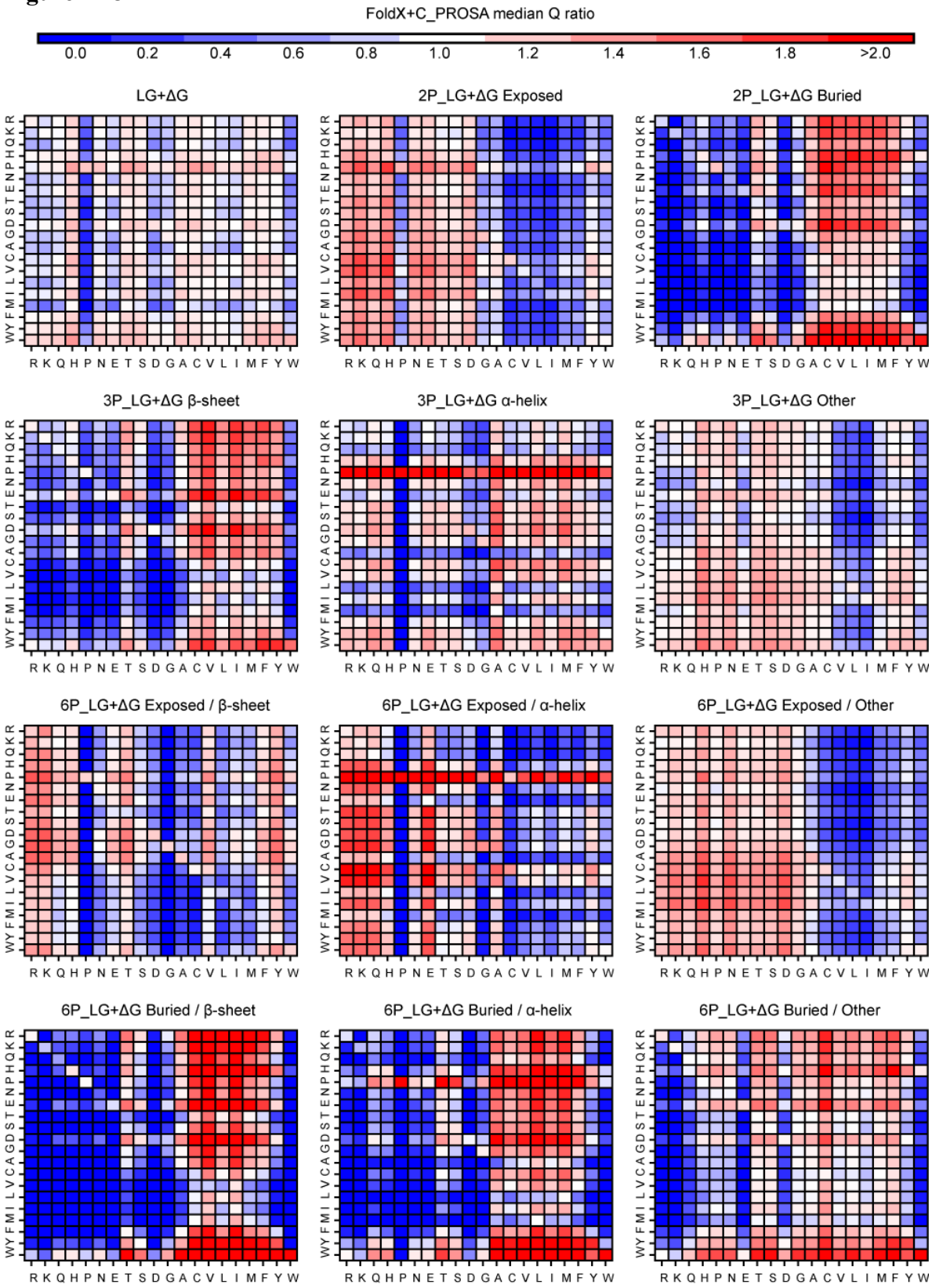

FoldX+C_PROSA median Q ratio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
>2.0
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure A19:

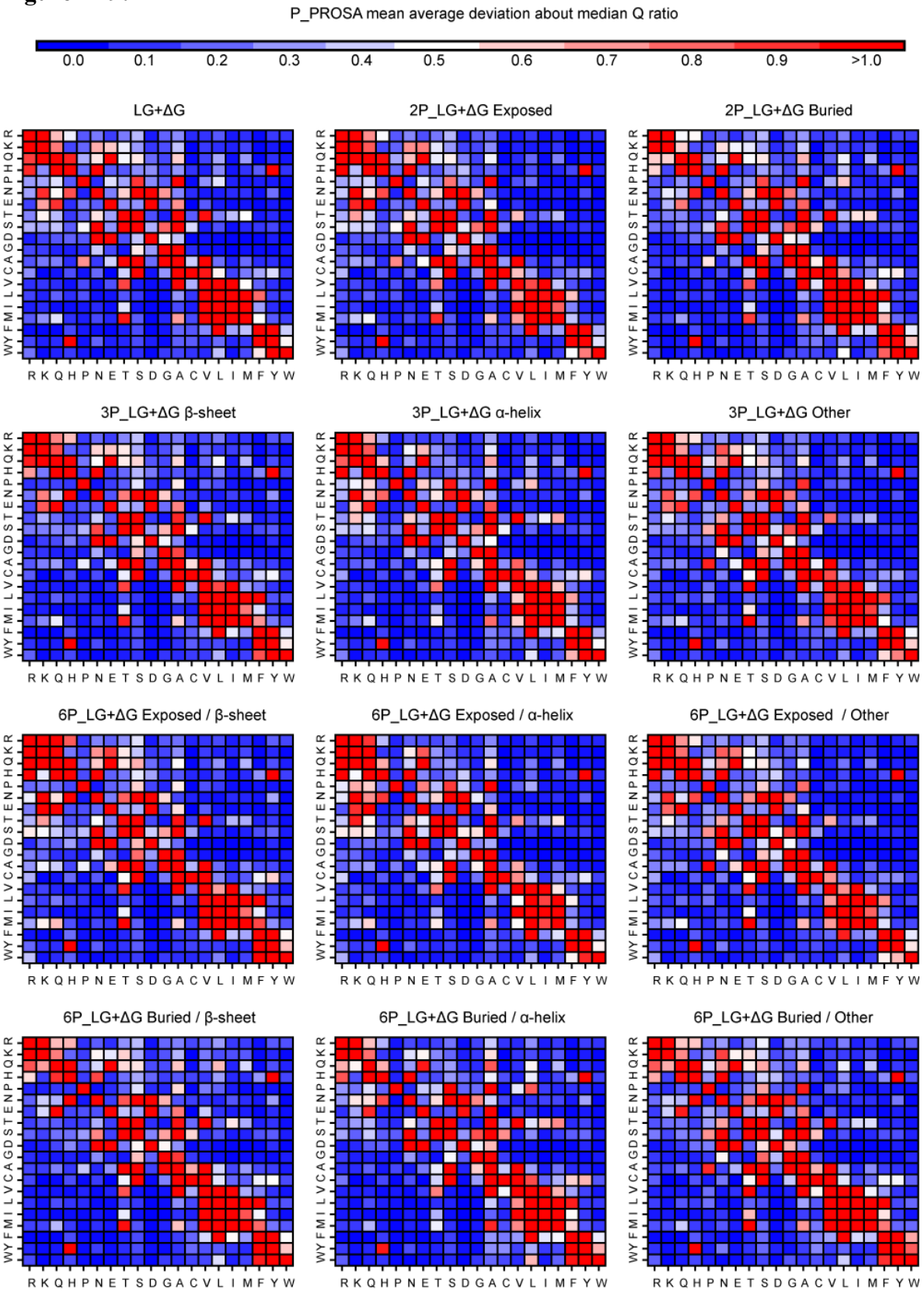

P_PROSA mean average deviation about median Q ratio
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
>1.0
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure A20:

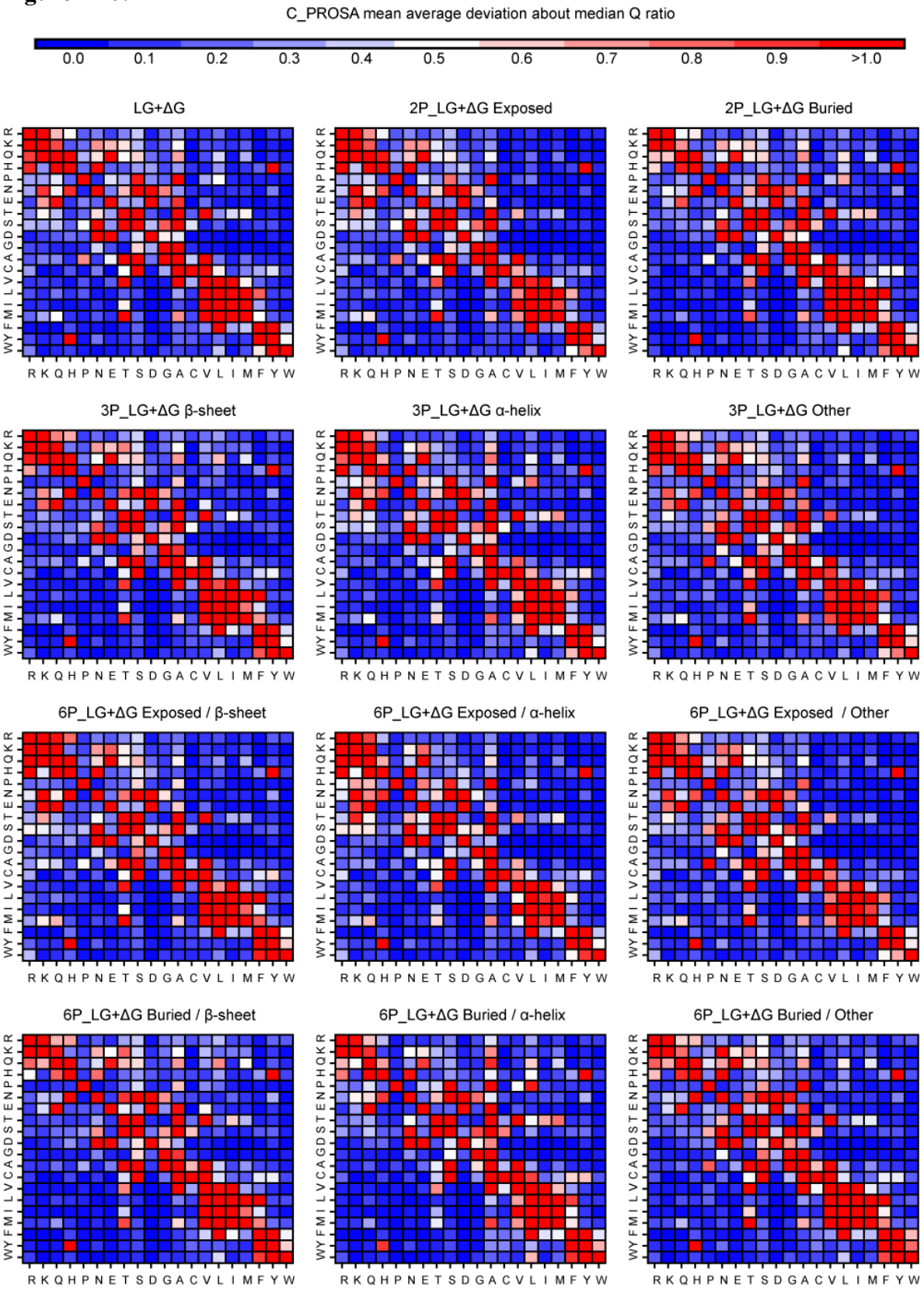

C_PROSA mean average deviation about median Q ratio
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
>1.0
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure A21:

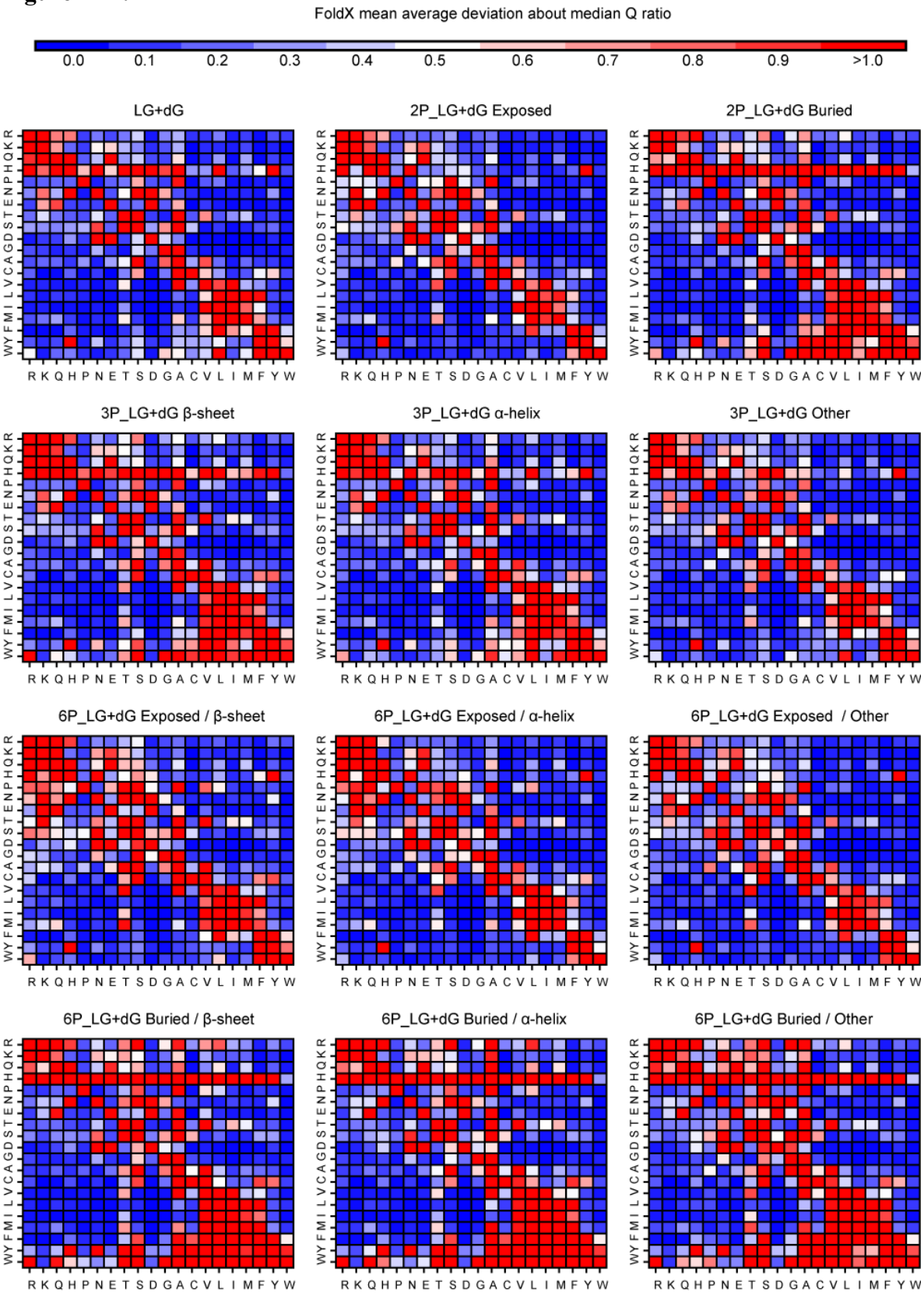

FoldX mean average deviation about median Q ratio
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
>1.0
LG+dG
2P_LG+dG Exposed
2P_LG+dG Buried
3P_LG+dG β-sheet
3P_LG+dG α-helix
3P_LG+dG Other
6P_LG+dG Exposed / β-sheet
6P_LG+dG Exposed / α-helix
6P_LG+dG Exposed / Other
6P_LG+dG Buried / β-sheet
6P_LG+dG Buried / α-helix
6P_LG+dG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure A22:

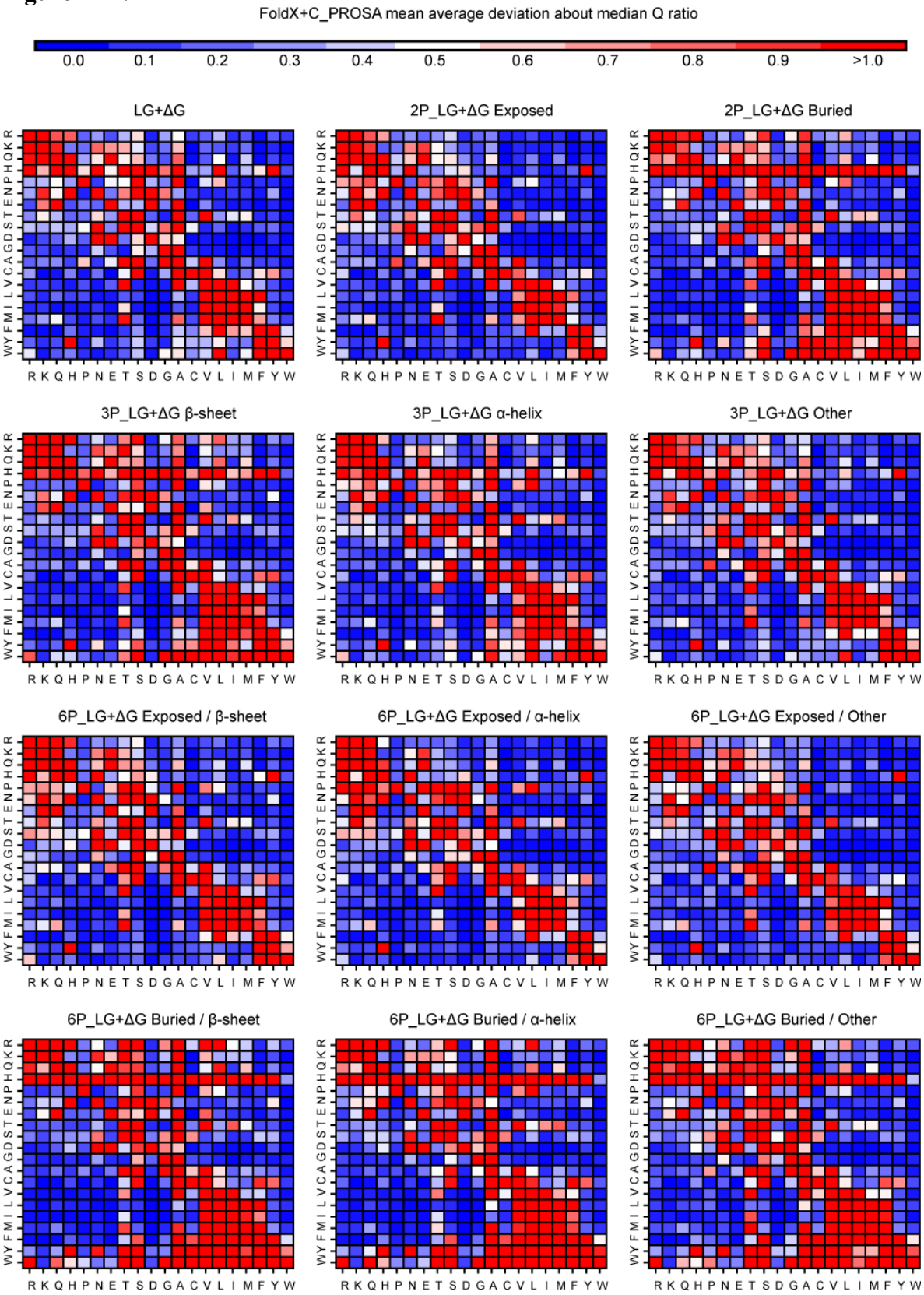
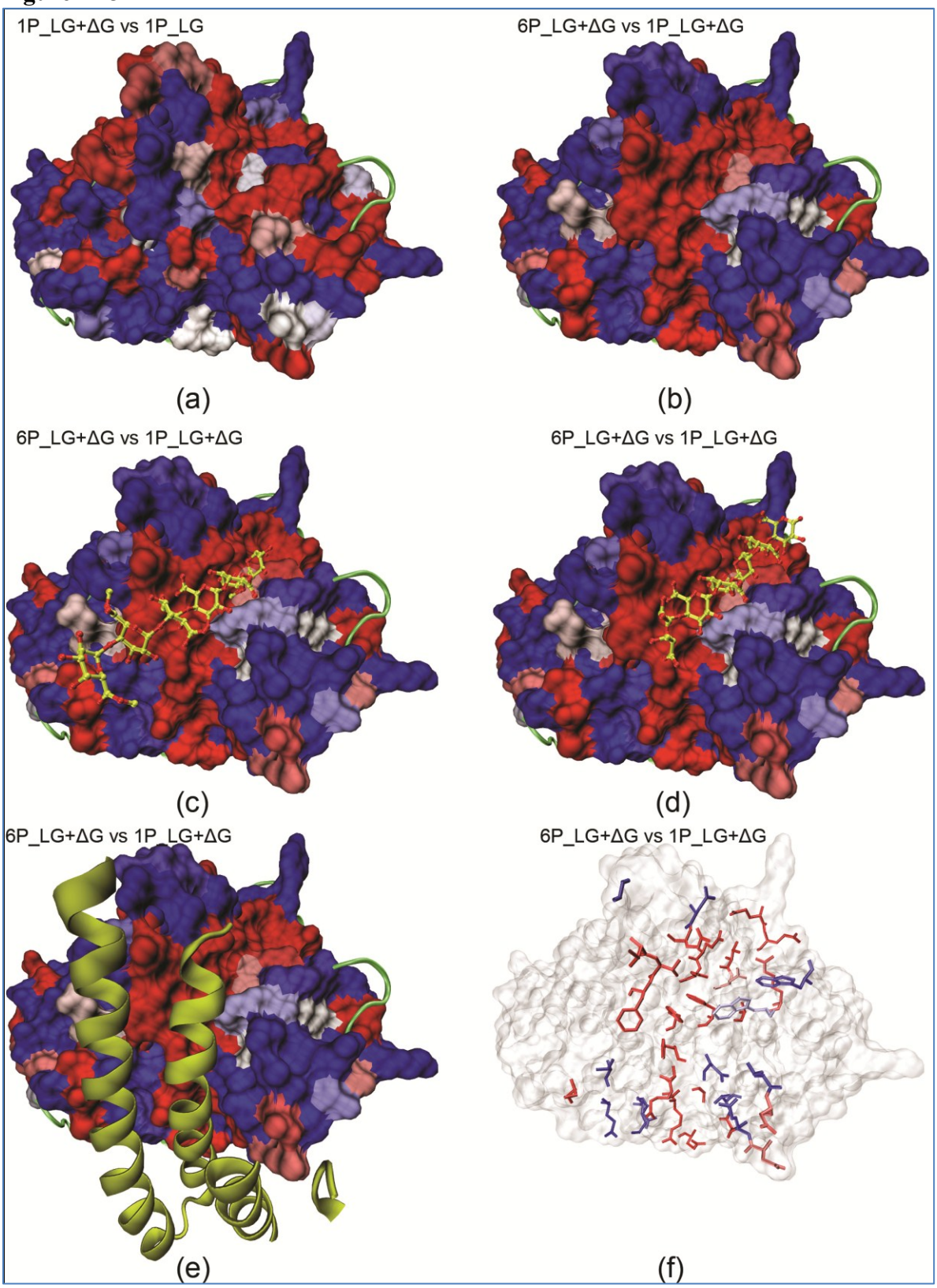

FoldX+C_PROSA mean average deviation about median Q ratio
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
>1.0
LG+ΔG
2P_LG+ΔG Exposed
2P_LG+ΔG Buried
3P_LG+ΔG β-sheet
3P_LG+ΔG α-helix
3P_LG+ΔG Other
6P_LG+ΔG Exposed / β-sheet
6P_LG+ΔG Exposed / α-helix
6P_LG+ΔG Exposed / Other
6P_LG+ΔG Buried / β-sheet
6P_LG+ΔG Buried / α-helix
6P_LG+ΔG Buried / Other
W Y F M I L V C A G D S T E N P H Q K R
R K Q H P N E T S D G A C V L I M F Y W

Figure A23


1P_LG+ΔG vs 1P_LG
(a)
6P_LG+ΔG vs 1P_LG+ΔG
(b)
6P_LG+ΔG vs 1P_LG+ΔG
(c)
6P_LG+ΔG vs 1P_LG+ΔG
(d)
6P_LG+ΔG vs 1P_LG+ΔG
(e)
6P_LG+ΔG vs 1P_LG+ΔG
(f)