

Copyright

by

Michael Paul Scherrer

2013

The Dissertation Committee for Michael Paul Scherrer
certifies that this is the approved version of the following dissertation:

**From the Inside Out: Determining Sequence
Conservation within the context of Relative Solvent
Accessibility**

Committee:

Claus Wilke, Supervisor

Robin Gutell

Lauren Meyers

Sara Sawyer

Jeffrey Barrick

**From the Inside Out: Determining Sequence
Conservation within the context of Relative Solvent
Accessibility**

by

Michael Paul Scherrer, B.A., A.S.

DISSERTATION

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2013

For my grandfather Stanley, who remains my enduring patron. Your jack of
all trades has finally become a master of one.

Acknowledgments

In truth, I did very little of the heavy lifting required to complete this document. The real heroes are those that surround me, who with Sisyphean shoulders, shielded me from the shadows of boulders while I skipped my pebble across the shimmering surface of Science. Their names are too numerous to remember, yet too seldom I forget...

Thank you to my advisor, Claus Wilke - your mentorship has been significant to my development throughout the years and I know that the probability of finding an advisor like you occurs less than 5%.

Thank you to my lab mates, Art Covert and Tong Zhou - you were always there with open ears and free hands whenever I needed them.

Thank you to my committee members for all of your feedback: Jennifer Morgan, Robin Gutell, Jeffrey Barrick, Sara Sawyer, and Lauren Meyers.

And finally, thank you to my colleagues: Marguerite Hunt, Bartram Smith, and Chintan Modi. I will see you again on the outside.

From the Inside Out: Determining Sequence Conservation within the context of Relative Solvent Accessibility

Publication No. _____

Michael Paul Scherrer, Ph.D.
The University of Texas at Austin, 2013

Supervisor: Claus Wilke

Evolutionary rates vary vastly across intraspecific genes and the determinants of these rates is of central concern to the field of comparative genomics. Tradition has held that preservation of protein function conserved the sequence, however mounting evidence implicates the biophysical properties of proteins themselves as the elements that constrain sequence evolution. Of these properties, the exposure of a residue to solvent is the most prevalent determinant of its evolutionary rate due to pressures to maintain proper synthesis and folding of the structure. In this work, we have developed a model that considers the microenvironment of a residue in the estimation of its evolutionary rate. By working within the structural context of a protein's residues, we show that our model is better able to capture the overall evolutionary trends affecting conservation of both the coding sequences and the protein structures from a genomic level down to individual genes.

Table of Contents

Acknowledgments	v
Abstract	vi
List of Tables	x
List of Figures	xi
Chapter 1. Introduction	1
1.1 The Evolution of Coding Sequences	1
1.1.1 Rate Variation Among Genes	2
1.1.2 Rate Variation Among Sites within Genes	4
1.2 Modeling Coding Sequence Evolution	9
1.2.1 Evolutionary Rate	9
1.2.2 Codon Models	12
The GY94 and MG94 Codon Models	13
The General Time Reversible Model	15
1.2.3 Estimation of Parameters	15
1.2.4 Hypothesis Testing	16
1.3 Significance	17
Chapter 2. Modeling coding-sequence evolution within the context of residue solvent accessibility	19
2.1 Background	19
2.2 An RSA-dependent Markov model of coding sequence evolution	21

2.3	A linear RSA dependency for all estimated parameters provides the best model fit	24
2.4	GY94 model provides a better model-fit than MG94 model	27
2.5	Effect of relative solvent accessibility on synonymous and non-synonymous substitution rates	30
2.6	The effect of core size and expression level on evolutionary rate	32
2.7	Discussion	35
2.8	Methods	42
2.8.1	Homology mapping and categorization of genes	42
2.8.2	Calculation of evolutionary rates	43
2.8.3	Statistical analysis	44
Chapter 3.	Modeling the mutation rates of coding-sequences under the constraints of solvent accessibility	45
3.1	Background	45
3.2	An RSA-dependent General Time Reversible model of coding-sequence evolution	48
3.3	The transition/transversion parameter, κ , is best modeled in the transition position	51
3.4	Individually estimated codon frequencies increase model performance over genomic codon frequencies	56
3.5	Mutation parameters have varying relationships with solvent accessibility across species	57
3.6	Discussion	61
3.7	Methods	68
Chapter 4.	Assessing the correlates of coding-sequence evolution in individual genes	69
4.1	Background	69
4.2	Determinates of average evolutionary rate across individual genes	71
4.3	Expression level and core size both impose an additional selective constraint beyond solvent accessibility	73
4.4	Expression level and GC content are both predictors of structural constraint	75
4.5	Discussion	77
4.6	Methods	81

Chapter 5. Conclusion	83
5.1 Contribution	83
5.1.1 Consideration of protein structure provides for better evolutionary rate estimates	84
5.1.2 Nucleotide-level processes are determinants of protein evolution	84
5.1.3 Individual determinants of evolutionary rate are revealed with a structural paradigm	85
5.2 Future	85
Vita	107

List of Tables

2.1	Fitted models, in order of ascending AIC	27
2.2	Effect of the number of bins on parameter estimates .	28
3.1	Overall nucleotide and amino acid sequence divergence	48
3.2	Fitted mutational model combinations, in order of ascending AIC	55
3.3	Fitted mutational model combinations for per-bin codon frequencies, in order of ascending AIC	58
4.1	Gene attributes and their definitions	72
4.2	Multivariate regression of gene attributes predicting average evolutionary rate, ω_A	73
4.3	Logistic regression of gene attributes predicting the presence of an ω slope	75
4.4	Multivariate regression of gene attributes predicting structural constraint, ω_1	77
4.5	Summary of gene attributes and their significance in predicting ω rates	79

List of Figures

1.1	The relative solvent accessibility of a residue. The relative solvent accessibility (RSA) of a residue is a continuous metric defined as the degree of surface area that is accessible to the solvent, normalized by the maximum area observed across all residues of that type. Here, the RSA of buried residues (red) and surface residues (blue) have been mapped onto the three-dimensional structure of cystathionine gamma-lyase from the yeast <i>S. cerevisiae</i>	6
1.2	Effects of degree of burial on fixation. Here, a cartoon diagram describes the results of two mutations types of mutations. Two similar residues, shown as a white triangle, with one in the core of the gray protein and one on the surface. When an amino acid substitution is made (shown as a red square) in the core of the protein, the change disrupts the stability of the protein, causing misfolding and elimination (right). The substitution for a surface residue causes no such disruption, leading to fixation of the residue (left).	8
1.3	The divergence of two orthologous sequences over evolutionary time. Evolutionary rate is defined as the average number of site substitutions between sequences since divergence over time, t . Two types of coding sequence mutations can be acquired during divergence, namely synonymous mutations (a change in the coding sequence only) and nonsynonymous mutations (resulting in an amino acid substitution downstream). The rates of nonsynonymous and synonymous change are derived from the alignments of orthologous sequences. Mutations can also be classified as either transitions ($A \rightleftharpoons G$ or $C \rightleftharpoons T$) or transversions (purine \rightleftharpoons pyrimidine). Here, the types of substitutions are labeled on a pairwise sequence alignment of two orthologs.	11

1.4	The ratio of nonsynonymous and synonymous substitution rates. The ratio of the nonysnonymous (dN) and synonymous (dS) rates of change is defined by $\omega = dN/dS$ and characterizes the selection acting on the protein sequence. Synonymous mutations have been conventionally thought to be “silent”, thus having no effect on the protein structure, and is used here as a normilization factor. The ratio $\omega = dN/dS$ can be interpreted as an indicator of purifying selection on the structure if it is less than one and positive selection if it is greater than one. Here, synonymous (blue) and nonsynonymous (red) mutations have been mapped onto an alignment of cystathione gamma-lyase from the yeasts <i>S. cerevisiae</i> and <i>S. paradoxus</i> .	12
2.1	Examples of RSA-dependent sequence-evolution models considered. All models have three parameters, evolutionary-rate ratio ω , branch length t , and transition-transversion ratio κ . All three parameters can be estimated as an individual value within each RSA bin (per-bin), as a linear function of RSA (linear), or as a constant across all RSA values (constant). The examples here are illustrated for $n = 10$ RSA bins. (A) All parameters are estimated per-bin. (B) ω is estimated as a linear function, t is estimated per-bin, and κ is estimated as a constant. (C) All parameters are estimated as a linear function.	24
2.2	Evolutionary-rate ratio increases linearly with RSA. The solid line shows $\omega = dN/dS$ versus RSA as estimated by the best mdoel (linear ω , linear t , and per-bin κ). The dots show the same for the best model with per-bin ω (which has linear t and per-bin κ). Both models are consistent with each other and strongly support a linear relationship between ω and RSA.	29
2.3	Comparison of the GY94 and MG94 models. The solid line shows $\omega = dN/dS$ versus RSA, as estimated by the GY94 model. The dashed line shows the same for the MG94 model. Under the MG94 model, ω shows moderate curvature. The GY94 model provides a better fit to the data ($\Delta\text{AIC} = 14$).	31
2.4	Evolutionary rates dN and dS. (A) The nonsynonymous rate, dN , correlates strongly with RSA under both the mutational-opportunity definition and the physical-sites definition. (B) The synonymous rate, dS , shows a moderate negative correlation with RSA under the mutational-opportunity definition and no slope under the physical-sites definition. The fitted model had linear ω , linear t , and per-bin κ .	33

2.5	Dependency of $\omega = dN/dS$ on protein core size and expression level. (A) Core size affects evolutionary rate on the surface of the protein but not in the core. (B) Expression level affectes evolutionary rate both on the surface and in the core. However, it has a bigger effect on the surface of the protein. In both figures, the solid lines were estimated jointly from the data using a linear dependency of ω on RSA. Points for individual bins are shown for illustration purposes only. They were estimated using a per-bin model for ω . In the figures above, the dashed black line represents the genome-wide trend, as shown in Figure 2.2, and is provided as a reference.	36
2.6	Joint analysis of the effects of both core size (small or large) and expression level (high or low) on the relationship between $\omega = dN/dS$ and RSA. Only the fitted lines are shown. Surprisingly, for low-expression genes, small-core proteins evolve faster than large-core proteins. This relationship is reversed in a larger dataset obtained with less-stringent criteria (see text).	37
3.1	Combinations of base pairs for transitions and transversions. Mutations can be classified as either transitions ($A \rightleftharpoons G$, $C \rightleftharpoons T$) or transversions (purine \rightleftharpoons pyrimidine). The transition to transversion rate ratio is described by the parameter κ . Transitions occur between similar bases and there are two possible transition mutations. Transversions are exchanges of dissimilar nucleotides and the four possible transitions mutations can be further categorized into the traditional Watson-Crick paring bases or bases that do not pair in a DNA double helix. . . .	51
3.2	Examples of RSA-dependent nucleotide-level models considered. All models have five nucleotide-level parameters (transition-transversion ratio, κ ; transition type ratio, T_{iN} ; transversion type ratio, T_{vN} ; paired transversion ratio, T_{vP} and the unpaired transversion ratio, T_{vU}), in addition to the linear parameters ω and t . All five parameters can be estimated as an individual value within each RSA bin (per-bin), as a linear function of RSA, or as a constant across all RSA values (constant). For T_{iN} , we introduce an additional quadratic relationship with RSA. The examples here are illustrated for $n = 10$ RSA bins. (A) The top scoring model in yeast. κ and T_{vP} are per-bin, T_{vN} and T_{vN} are linear, while T_{vN} is quadratic. (B) Here, κ is the only linear parameter. T_{vN} and T_{vP} are held constant, while T_{iN} and T_{vU} vary per-bin. (C) Here, there is no RSA-dependency for T_{vN} . κ and T_{vU} are constant, while T_{iN} and T_{vP} vary linearly with RSA.	52

3.3	Fitted models across species, in order of AIC. All parameters have three possible relationships with RSA (per-bin, linear, and constant), with the exception of T_{iN} which has a fourth (quadratic). In total, there were 324 models using these parameter combinations that were estimated for yeast, mouse, fly, and worm. Here, the top 15 models are shown for each of the species and the ΔAIC from the best model is indicated. A model which uses per-bin κ and quadratic T_{iN} parameter fits best in yeast and fly, while there is not much difference between the top models for mouse and worm.	61
3.4	Dependency of varying κ parameters on RSA across species. In yeast and fly, the transition-transversion rate ratio κ decreases with solvent accessibility and should be estimated per-bin. For mouse and worm, κ shows almost no relationship with RSA and all models fit equally as well. The numbers in parentheses are the differences between the top scoring model of a parameter type with the overall top scoring model.	62
3.5	The effects of RSA on several transition parameter types, T_{iN}, across species. While the transition rate ratio, T_{iN} shows almost no relationship with solvent accessibility in mouse, both yeast and fly display a quadratic relationship with RSA. A quadratic relationship with solvent accessibility is also apparent in worm, although a per-bin model fits equally as well. The numbers in parentheses are the differences between the top scoring model of a parameter type with the overall top scoring model.	63
3.6	The relationship of transversion parameter types, T_{vN}, with RSA across species. For yeast and fly, the transversion rate ratio, T_{vN} , shows a negative linear relationship with solvent accessibility. RSA affects T_{vN} individually per-bin in worm, while there appears to be no varying relationship between transversion type and solvent accessibility in mouse. The numbers in parentheses are the differences between the top scoring model of a parameter type with the overall top scoring model.	64
4.1	The average evolutionary rate, ω_A, vs. the structural constraint slope, ω_1. Here, ω_A is plotted against ω_1 for the 160 genes that displayed a significant ω_1 . The constraint imposed across solvent accessibilities is relaxed with increasing average evolutionary rate of a gene.	76

- 4.2 **The effects of expression level and %GC content on the structural constraint parameter, ω_1 .** Both expression level and %GC content are significant predictors of ω_1 in our multi-variate regression model. **(A)** Lowly expressed genes have the most variation in ω_1 , while highly expressed genes have very little variation in ω_1 . This shows that as expression level increases, so does the constraint across the entire protein structure. **(B)** The 4 categories of %GC content are shown with their ω_1 . As %GC content increases, the variation in ω_1 decreases, indicating that high GC content imposes constraint on the surface residues of a protein.

Chapter 1

Introduction

1.1 The Evolution of Coding Sequences

The field of molecular evolution examines how both genes and genomes change over evolutionary time. Since the advent of fully sequenced genomes and solved protein structures, researchers have found through comparative analysis that some genetic sequences vary vastly from one another while others have almost no determinate differences. Even in cases where compared protein-coding sequences (genotypes) are nearly unrecognizable as related, the resulting protein structures that they encode (phenotypes) are still essentially the same. The challenge, then, has been to elucidate the mechanisms that underly this sequence variation.

Related genes have diverged from some common ancestor over evolutionary time and the average number of mutations that have occurred across the gene during this time is the definition of evolutionary rate. Per-gene evolutionary rates vary dramatically even within a single species and this rate

variation is a trait that is universal to all species. Much work has already been done to elucidate the factors that underly such evolutionary rate variations across whole genes. In terms of protein evolution, the metric that is used is to assess the rate of change in sequence conservation with highly conserved sequences being preserved through natural selection. Selection acting among coding sequences has been found to occur at the nucleotide and amino-acid levels, as well as during the synthesis of the nascent polypeptide during translation and has been used to explain the rate variation among genes.

However, proteins themselves are not static entities and it is unrealistic to assume that each site in a protein is equally mutable. Sites in the protein that are strongly conserved experience less amino-acid changes, while sites that are less conserved will accumulate amino-acid substitutions more rapidly. While most protein sites evolve under purifying selection and are thus strongly conserved, a small fraction of these sites are able to acquire substitutions without detriment. Regions and sites in a protein are under different selective pressures and the biophysical properties of a protein's residues themselves have been implicated in the selective pressures constraining coding sequences. In this work, we investigate how the rate of coding sequence evolution among sites is driven by both structural and mutational level processes.

1.1.1 Rate Variation Among Genes

While the coding sequences of proteins determine their three-dimensional structures, the pressure to maintain proper folding and function causes the pro-

tein structures themselves to be better conserved than their sequences [6, 17]. Even with distantly related homologous proteins, the structures remain very similar and function is preserved even though the diverged sequences may appear quite different. The classic view of protein evolution suggests that a protein’s function should predict its evolutionary rate; however, protein function is a poor predictor overall and should only be considered on a case-by-case basis [90].

Although the protein structure itself heavily constrains the coding sequence, selection also occurs during the synthesis of the nascent protein structure. In nearly every organism where data is available, expression level has repeatedly been found to be a dominant predictor of evolutionary rate [32, 64, 67]. Highly expressed genes, measured by mRNA abundance, evolve at a rate that is much slower than their lowly expressed counterparts. A high expression level increases the fitness costs related to protein synthesis, misfolding, aggregation, and non-specific interactions [129, 117] due to the fitness requirements of large volume synthesis of highly expressed proteins. With all other factors equivalent, a mutation in a highly expressed gene has a large fitness cost simply due the number of copies of the protein, as mutations that lead to misfolding, aggregation, and toxicity will be far more damaging to a cell than with low level protein synthesis.

While it is tempting to implicate protein structure and function in constraining the evolutionary rates of highly expressed genes, the rate differences have been found to occur between highly and lowly expressed paralogous genes,

which share both of these features [30]. Instead, the reduced evolutionary rate of highly expressed genes is hypothesized to occur due to selection against protein misfolding at the translational level. The mistranslation-induced protein-misfolding hypothesis puts forth the notion that the majority of selection pressure acting on coding sequences arises during translation to avoid the toxic effects of mistranslated and misfolded proteins [32].

Selection against misfolding during translation requires translational accuracy, robustness, and reliability, therefore slowing the rate of substitution. Coding sequences must be translated with a lower error rate (accuracy) as well as code for proteins that are more tolerant of errors (robustness) though increased thermodynamic stability [32, 30, 115]. Furthermore, selection seems to act toward sequences with reliable folding kinetics that reduce the risk of error-free sequences becoming misfolded [32, 33]. Therefore, mutations that affect the creation of the newly-forming protein during translation will have an effect on the rate at which the coding sequence evolves.

1.1.2 Rate Variation Among Sites within Genes

The classic view of protein evolution was that amino acid sequence was primarily constrained by the maintenance of protein function; more recently, however, the biophysical properties of the residues themselves have been shown to be a greater determinant of selective constraint. Factors such as interactions between residues within the protein architecture and the accessibility of these residues to their microenvironment play a much greater role in determining

sequence conservation.

One determinant of amino acid substitution rate is the accessibility of a residue within the protein to the surrounding solvent. Buried residues are less mutable than their solvent exposed counterparts [81, 41, 74], with residues in the core of the protein evolving at a rate almost half that of those on the surface [39]. The relative solvent accessibility (RSA) of a residue is defined as the degree of surface area of a residue that can be traced by a solvent molecule and is continuous metric ranging from 0 (completely buried) to 1 (completely exposed) (Figure 1.1). A recent analysis by Franzosa and Xia [39] demonstrated that the evolutionary rate of protein residues scale in a near perfect linear fashion with their relative solvent accessibility (RSA). The effects of residue burial on the rate of evolution is independent of both residue hydrophobicity and secondary structure of a site [41, 39]. Thus, a mutation in the core of the protein is much less likely to go to fixation as one on its surface (Figure 1.2).

As surface residues are more mutable than those that are buried in the core of the protein, it is reasonable to expect proteins with a larger proportion of solvent exposed residues to evolve more rapidly in general. However, the proportion of buried sites in a protein has a positive correlation with its evolutionary rate and core size explains up to a tenth of this rate variation [39, 10, 131]. The explanation for this paradox is that proteins with a large core size benefit from increased overall stability of the structure, thereby alleviating mutational constraints on their surface residues [39, 10]. The increased site

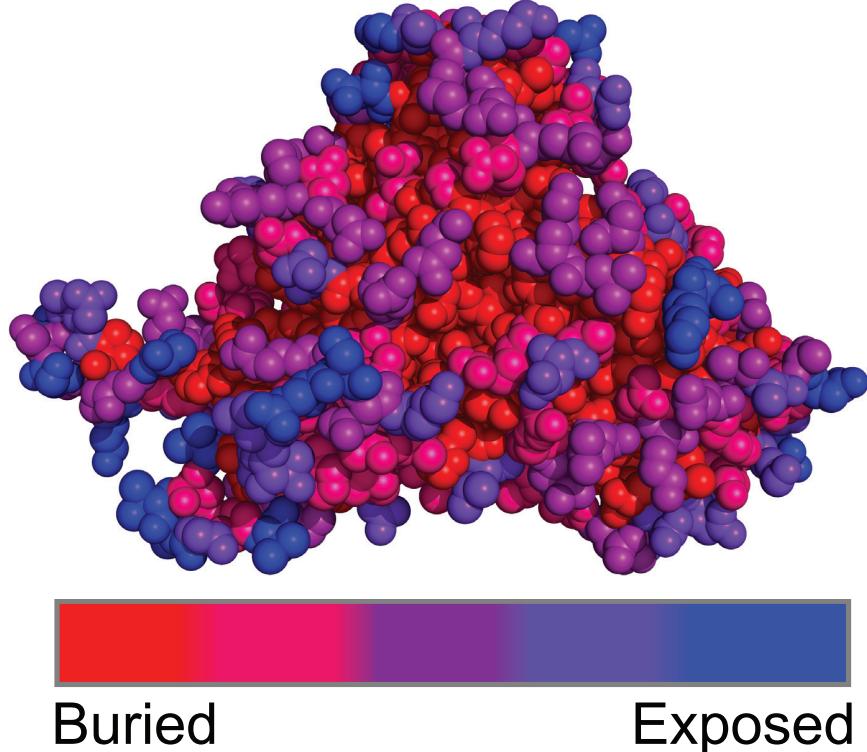


Figure 1.1: The relative solvent accessibility of a residue. The relative solvent accessibility (RSA) of a residue is a continuous metric defined as the degree of surface area that is accessible to the solvent, normalized by the maximum area observed across all residues of that type. Here, the RSA of buried residues (red) and surface residues (blue) have been mapped onto the three-dimensional structure of cystathionine gamma-lyase from the yeast *S. cerevisiae*.

variation on the surface of large core proteins is sufficient to account for their increased evolutionary rate over the more thermodynamically fragile small core proteins. However, regardless of the benefits of thermodynamic stability due to a large core, the core residues of both large and small core proteins remain equally conserved [39].

Most proteins function through interactions with other proteins, either

through obligate interactions (where the protein is part of a larger complex of proteins) or through transient interactions (where the protein briefly docks to another to perform a task, i.e. signal transduction). For these protein interactions to be successful, the residues that participate in the interface need to be selected to work in concert. It follows, then, that residues that participate in protein-protein interactions are conserved more strongly than those that do not participate in an interface [107, 73, 57, 35]. Both the fraction of residues involved in an interface [57] as well as selection against nonspecific protein-protein interactions affect the evolutionary rate interface residues[26, 129], with obligate interfaces constraining amino acid sequence more than transient interfaces [73]. While participation in an interface reduces solvent accessibility when two proteins are docked, Franzosa and Xia [39] showed that participation in an interface introduces an additional selective pressure beyond a reduction in RSA. Protein-protein interactions, therefore, impose an associated structural constraint that is similar to yet independent of RSA.

While protein evolution has traditionally been thought to occur at the amino acid level, the same selective pressures play out at the nucleotide level with synonymous mutations. Frequently considered to be “silent”, selection on synonymous sites occurs through the ribosomal machinery and mRNA working in concert to produce a polypeptide. Selection at the nucleotide level affects codon bias, which in turn regulates both translational speed and accuracy.

Translational speed is tightly linked to tRNA abundance, as highly abundant tRNA molecules are readily available to the ribosome during trans-

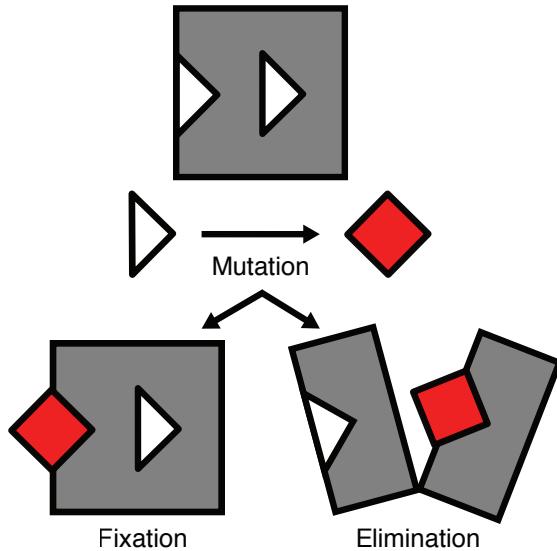


Figure 1.2: Effects of degree of burial on fixation. Here, a cartoon diagram describes the results of two mutation types of mutations. Two similar residues, shown as a white triangle, with one in the core of the gray protein and one on the surface. When an amino acid substitution is made (shown as a red square) in the core of the protein, the change disrupts the stability of the protein, causing misfolding and elimination (right). The substitution for a surface residue causes no such disruption, leading to fixation of the residue (left).

lation and therefore translated faster. Synonymous mutations that change the codon to a rare tRNA molecule could cause the ribosomal machinery to sputter and stall [60, 128], which then interferes with the folding kinetics of the newly forming protein. Codon choice is therefore necessary to regulate ribosomal translation speed though a supply and demand balance of tRNAs to allow for proper folding of the protein [83, 60, 128, 20, 58].

The choice of codon can also affect translational accuracy as different codons are known to have varying rates of error [63]. More accurate codons

have a selective advantage over those with a higher error rate as they provide greater assurance of translational accuracy; however, a trade-off between accuracy and adaptation is highly dependent on the residue in question. For sensitive residues, such as those located in the core of a protein, translational accuracy is necessary to prevent errors in protein folding and elimination. Less essential residues, such as those in a loop or disordered region of the protein that have no specific function would not require such assurance for accurate translation. Thus, translational accuracy and codon bias is directly linked to structural selection and an association between optimal codons has been shown for the rate, function, and solvent accessibility of the residues they encode [32, 2, 101, 132, 111].

1.2 Modeling Coding Sequence Evolution

1.2.1 Evolutionary Rate

Evolutionary rate is defined as the average number of mutations accumulated by sites in diverging sequences over evolutionary time. Thus, if the rate of evolution is constant throughout time, the distance between the two sequences is a simple linear function of the time since divergence. However, such a simplistic function is applicable only to closely related sequences and will underestimate the amount of incorporated changes between the sequences as the distance between them increases. A variable site could have occurred through a single substitution or multiple substitutions. Sites that

appear to be similar could have arisen though independent yet parallel events, converged after multiple substitutions, or changed to the original state through back substitutions. In consequence, many of these changes may be obscured as site substitutions occur throughout evolutionary time.

To accurately depict the underlying substitution process between sequences, it becomes necessary to model the changes from one evolutionary unit to another using a probabilistic paradigm. What defines an evolutionary unit? Such analyses can take place at three different levels: at the DNA level using nucleotides, from a protein perspective using amino acids, or a combination of the two by using the triplets of coding sequences. While nucleotide based methods are the most common, they ignore the heterogeneous patterns inherent to the genetic code and are therefore unsuited to the analysis of coding sequences [95, 12]. Conversely, amino acid models omit the finer details that are revealed by the coding triplet. Although they contain no physico-chemical information about the amino acids they encode, codon models have been shown to provide dramatically improved estimates over their amino acid counterparts for coding sequences [94].

The inherent power of codon models lies in their ability to differentiate mutations occurring at the nucleotide level (synonymous changes) versus those that affect the amino acid sequence (nonsynonymous changes). Due to the degeneracy of the genetic code, most mutations will be synonymous, or “silent”, having no effect on the three dimensional structure of the protein. Fewer mutations are nonsynonymous and express the expected change in phe-

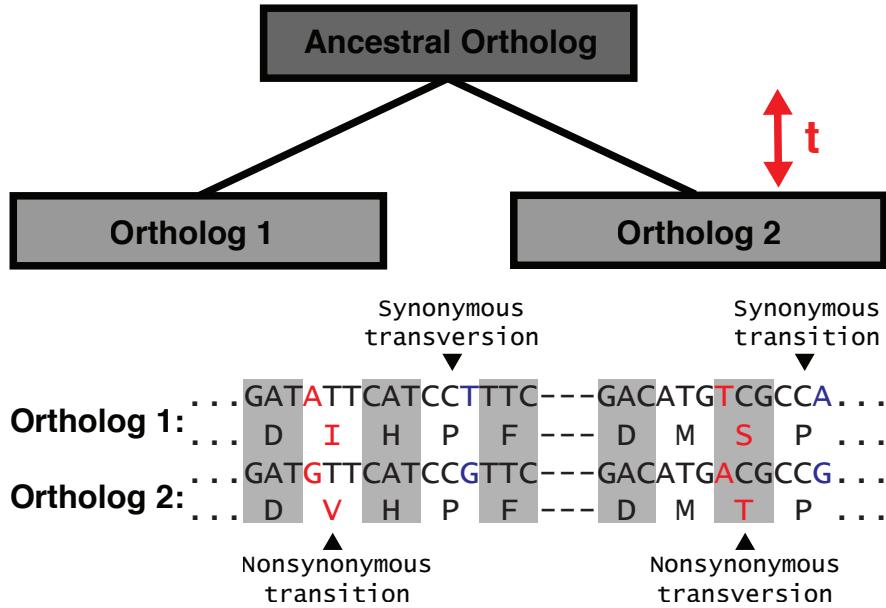


Figure 1.3: The divergence of two orthologous sequences over evolutionary time. Evolutionary rate is defined as the average number of site substitutions between sequences since divergence over time, t . Two types of coding sequence mutations can be acquired during divergence, namely synonymous mutations (a change in the coding sequence only) and nonsynonymous mutations (resulting in an amino acid substitution downstream). The rates of nonsynonymous and synonymous change are derived from the alignments of orthologous sequences. Mutations can also be classified as either transitions ($A \Rightarrow G$ or $C \Rightarrow T$) or transversions (purine \Rightarrow pyrimidine). Here, the types of substitutions are labeled on a pairwise sequence alignment of two orthologs.

notype from divergence. Figure 1.3 demonstrates these two types of changes that can occur between pairwise sets of codons. Such a distinction allows for the comparison of phenotypic changes to the background genotypic ticking of the evolutionary machinery through the use of the ratio between nonsynonymous (dN) and synonymous (dS) substitution rates, which indicates the selective pressures acting on the amino acid sequence [59, 53].

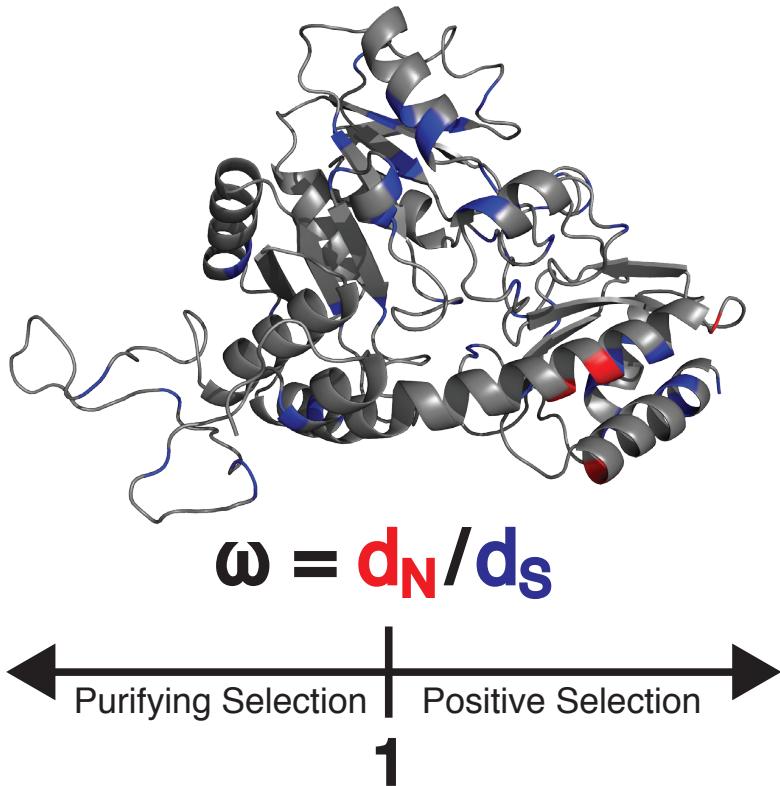


Figure 1.4: The ratio of nonsynonymous and synonymous substitution rates. The ratio of the nonsynonymous (dN) and synonymous (dS) rates of change is defined by $\omega = dN/dS$ and characterizes the selection acting on the protein sequence. Synonymous mutations have been conventionally thought to be “silent”, thus having no effect on the protein structure, and is used here as a normalization factor. The ratio $\omega = dN/dS$ can be interpreted as an indicator of purifying selection on the structure if it is less than one and positive selection if it is greater than one. Here, synonymous (blue) and nonsynonymous (red) mutations have been mapped onto an alignment of cystathione gamma-lyase from the yeasts *S. cerevisiae* and *S. paradoxus*.

1.2.2 Codon Models

In a codon model, the unit of evolution is the codon triplet rather than a single nucleotide or amino acid. The substitution of one codon for another

in a protein coding sequence is described by a random process which has no memory, known as a Markov model. In such a process, as the coding sequence changes through time, t , the probability of change from one codon to another depends only on the current state of the sequence rather than any past states. Models of codon substitutions are represented by a 61×61 matrix, $Q = \{Q_{ij}\}$, which defines the instantaneous rate of change from codon i to codon j for each of the 61 sense codons. Stop codons are omitted from the matrix as functional proteins are typically intolerant of such nonsense mutations.

Over some time $t > 0$, the probability of change from codon i to codon j is given by the transition probability matrix $P(t) = \{p_{ij}(t)\} = e^{Qt}$, which is solved using the differential equation $dP(t)/dt = P(t)Q$ [21]. The calculation of the transition probability matrix relates the model to the observed data in that the substitution process runs through continuous time, t , over a tree describing the phylogenetic relationships between sequence alignments. At time $t = 0$, since no changes to the sequence have occurred, the matrix Q is equal to the identity matrix I , a case where the diagonals Q_{ii} are all equal to 1 and all other elements Q_{ij} are 0 —representing no evolution. Conversely, when $t \rightarrow \infty$, the steady state frequencies for π_j , or equilibrium, has been reached.

The GY94 and MG94 Codon Models

The GY94 [42] and the MG94 [77] models were the first two codons models introduced and both use the distinction between nonsynonymous and

synonymous changes in their estimates. A simplified version of the GY94 model is defined as

$$Q_{ij} = \begin{cases} 0, & \text{if more than one change} \\ \pi_j, & \text{if synonymous transversion} \\ \kappa\pi_j, & \text{if synonymous transition} \\ \omega\pi_j, & \text{if nonsynonymous transversion} \\ \omega\kappa\pi_j, & \text{if nonsynonymous transition} \end{cases} \quad (1.1)$$

where κ is the transition/transversion rate ratio, ω is the nonsynonymous/synonymous rate ratio, and π_j is the equilibrium frequency of codon j . The parameters κ and π_j describe the mutational process occurring at the DNA level, while the parameter ω describes nonsynonymous selection occurring at the protein level.

By contrast, the MG94 model is defined as

$$Q_{ij} = \begin{cases} 0, & \text{if more than one change} \\ \alpha\pi_{jn}, & \text{if synonymous substitution} \\ \beta\pi_{jn}, & \text{if nonsynonymous substitution} \end{cases} \quad (1.2)$$

where α and β are defined as separate synonymous and nonsynonymous substitution rates, respectively. The equilibrium frequency π_{jn} refers to the nucleotide n of codon j . The two main differences between the models are (1) the GY94 model corrects for transition/transversion bias through the use of the parameter κ and (2) the GY94 model considers rates as proportional to the frequency of codon j rather than proportional to the frequency of the nucleotide n in codon j . Thus, the MG94 model has fewer parameters and is less realistic, but is more computationally tractable and more reliable in smaller samples where estimation of codon frequencies may be inaccurate [5].

The General Time Reversible Model

The General Time Reversible Model (GTR) is the most comprehensive nucleotide rate matrix that satisfies the assumption of time reversibility [103]. A nucleotide model describes substitutions acting at the DNA-level rather than at the intermediary codon triplet and is useful for describing the mutation process. In contrast to the 61×61 matrix that defines the possible sense codon changes, a nucleotide model is a 4×4 matrix that defines changes between the four different nucleotides. The GTR model is defined as

$$Q = \begin{bmatrix} * & \alpha_{AC}\pi_C & \alpha_{AG}\pi_G & \alpha_{AT}\pi_T \\ \alpha_{AC}\pi_A & * & \alpha_{CG}\pi_G & \alpha_{CT}\pi_T \\ \alpha_{AG}\pi_A & \alpha_{CG}\pi_C & * & \alpha_{GT}\pi_T \\ \alpha_{AT}\pi_A & \alpha_{CT}\pi_C & \alpha_{GT}\pi_G & * \end{bmatrix} \quad (1.3)$$

where the diagonals are placed so that the sum of the row is equal to 0. The GTR model has four nucleotide frequency parameters, π_i , that runs over the 4 nucleotides (A, G, C, and T). There are six substitution rate parameters, α_{ij} , that define the rate of change from one nucleotide to another. As the GTR model has more free parameters than all of the other nucleotide models, it is able to describe the evolution of sequences more realistically than simpler models.

1.2.3 Estimation of Parameters

As the parameters for any given model will vary based on the sequence data, we estimate the values of the parameters for each individual data set using maximum likelihood (ML). Under ML, parameter estimates are given

by their likelihood function, $L = P(D|P, M)$, which is the probability of the observed data D given the values of the parameters P and the codon model M . Using a set of aligned sequences and a phylogenetic tree for the species in the alignment, ML will optimize the branch lengths of the tree as well as the model parameters. The end result is a set of model parameters for which the likelihood of the aligned sequences evolving at the estimated rates is at its maximum. The values estimated by ML typically are close to the average frequencies for the rate parameters in the model. The likelihood score produced by ML can then be used to evaluate the fit of a particular model to the actual data.

1.2.4 Hypothesis Testing

Although no model is a perfect reflection of reality, it becomes crucial to use further inferences to determine which parameter estimates most accurately describe the underlying evolutionary processes. We use two methods of determining the better fitting model in this work: the likelihood ratio test (LRT) and the Akaike information criterion (AIC) [1, 14].

The LRT compares nested hypotheses defined by the number of parameters in each model. The model with the greater number of parameters (H_A) will always fit at least as well as the “nested” model with fewer parameters (H_0). The differences in likelihoods between the models is given by the log-likelihood ratio statistic, $G = -2 \ln(L(\theta_0|D)/L(\theta_A|D))$, where $L(\theta_0|D)$ and $L(\theta_A|D)$ are the likelihood scores of the null and alternative models given the

data D , respectively. The test statistic G is approximated by a χ^2 probability distribution with degrees of freedom defined by the differences between the number of free parameters in the models.

When multiple or non-nested models are compared, the AIC of each model is calculated by $AIC = 2k - 2\ln(L)$, where k is the number of free parameters in the model and L is the log-likelihood score. Models are ranked by their AIC with lower scores indicating a better goodness-of-fit of the model and extra parameters inflicting a score penalty.

1.3 Significance

The evolution of protein coding sequences occurs at both the amino acid and nucleotide levels within the context of proper protein structure and folding. While rapid developments have been made in recent years to advance our understanding of the evolutionary process from a molecular perspective, there is still no unifying model to explain the importance of different effects in shaping coding sequences. Describing the effects of the various mechanisms affecting evolutionary rate is a major challenge in molecular evolution and is one that requires increasingly realistic models that incorporate protein biochemistry as well as nucleotide level mutational processes. The overarching hypothesis of this work, then, is that protein evolution is driven by both structural and mutational level processes.

In the second chapter, we developed a simple model that combines

protein structure with sequence evolution and apply it across the yeast genome. We ask whether introducing a structural dependency into an evolutionary model will provide substantially better rate estimates.

In the third chapter, we define nucleotide level processes in terms of solvent accessibility and ask whether mutation biases are affected by protein structure in increasingly diverged genomes.

In the final chapter, we turn our attention toward the evolutionary mechanisms affecting individual genes. We ask what the contributions and interactions of these mechanisms are in terms of protein structure.

Our methods present a unified statistical framework for comparing evolutionary rates and effects for different proteins within the context of solvent accessibility. We demonstrate that protein structure is an important tool for comparative sequence analysis and can improve our understanding of the molecular processes that drive sequence evolution.

Chapter 2

Modeling coding-sequence evolution within the context of residue solvent accessibility

2.1 Background

Substitution patterns in protein-coding genes are shaped by the 3-dimensional structure of the expressed proteins. To account for this influence of structure on sequence evolution, evolutionary biologists increasingly aim to combine sequence analysis with structural information or to develop models of sequence evolution that incorporate structural features of the expressed protein. Some authors calculate amino-acid substitution matrices as a function of protein structure [81, 61] or correlate sequence variability in alignments with structural features [74, 29]. Others subdivide proteins into broad categories by solvent exposure (buried/exposed) or secondary structure (α -helix, β -sheet, etc.) and then use standard maximum likelihood models of

sequence evolution to infer evolutionary rates as a function of structural features [105, 41, 10, 131, 39]. Some authors employ more complex methods that allow for non-independence among sites, and use energy functions to model how substitutions at one site influence substitutions at others [89, 92, 93, 91]. Finally, a few groups have attempted a variety of other approaches to link sequence variability with protein structure [15, 25, 70, 18].

These various analyses differ in their specific results as well as in the approaches taken. However, one pattern consistently emerges: Residues in the core of the proteins are more conserved than on the surface. This finding agrees with our understanding of protein biochemistry. Substitutions in the core of a protein are more likely to disrupt fold stability than substitutions on the surface and the loss of the structural integrity of a protein is frequently the underlying cause of loss of function [127, 11]. Further, the observed relationship between residue buriedness and evolutionary conservation seems surprisingly simple. When evolutionary rate is plotted as a function of relative solvent accessibility, (RSA, a number between 0 and 1 measuring how exposed a residue is to the solvent surrounding the protein), one finds a near-perfect linear relationship [39, 88]. Inspired by the observed linear relationship between evolutionary conservation and RSA, we here take the standard Goldman-Yang model of coding-sequence evolution (GY94, [42]) and introduce to it a dependency of the model parameters on RSA. We find that the RSA-dependent GY94 model provides a substantially better fit to yeast sequence data than the standard, RSA-independent model. We further find that for several model

parameters, a simple, linear dependency on RSA provides the best fit. In particular, the ratio of non-synonymous to synonymous evolutionary rates ω is a linear, increasing function of RSA. Thus, we can characterize protein evolutionary rates by the slope and intercept of the ω -RSA relationship rather than by just a single ω value. We show that the slope and intercept of the ω -RSA relationship vary among proteins with different structures or different expression levels.

2.2 An RSA-dependent Markov model of coding sequence evolution

Previous works assessing the relationship between evolutionary rate and RSA subdivided sites into groups with comparable RSA and then calculated evolutionary rates separately for each group [39, 88]. This approach yields a set of independent evolutionary-rate estimates that can be plotted against representative RSA values for each group. While this approach has provided valuable new insight, it is not satisfactory from a methodological perspective. First, some model parameters (such as parameters describing the nucleotide-level mutation process, e.g. the transition-transversion bias) could be conserved among groups, yet they are estimated individually for each group. Second, a consistent framework for hypothesis testing is lacking. For example, in order to test whether evolutionary rates vary linearly with RSA, one would have to do a regression analysis on the previously estimated rates. In this regression analysis, sample size corresponds to the number of RSA groups

rather than to the number of sites in the original data set. Consequently, the P value resulting from the regression would likely be incorrect.

To resolve these shortcomings, we developed a variant of the GY94 model [42] in which model parameters are functions of RSA. We write the infinitesimal generator $Q = (Q_{ij})$ of the Markov process describing the substitution process as (for $i \neq j$)

$$Q_{ij} = \begin{cases} 0, & \text{if more than one change} \\ \pi_j, & \text{if synonymous transversion} \\ \kappa(r)\pi_j, & \text{if synonymous transition} \\ \omega(r)\pi_j, & \text{if nonsynonymous transversion} \\ \kappa(r)\omega(r)\pi_j, & \text{if nonsynonymous transition} \end{cases} \quad (2.1)$$

where κ is the ratio of transitions to transversions, ω is the ratio of the nonsynonymous to synonymous substitution rates, and r stands for the RSA of a site. The indices i and j run over all 61 sense codons, and j is the frequency of codon j . (We do not estimate site-specific codon frequencies). The finite-time transition matrix is given by

$$P = \exp[t(r)Q], \quad (2.2)$$

where t corresponds to evolutionary time, in arbitrary units. The parameter t measures the branch length in the phylogenetic tree; it is broadly related to the rate of synonymous substitutions. On first glance, it might be surprising that we allow t to vary with RSA. However, as we will see below, models with site-dependent t fit the data better than models with a single t across all sites. The reason for the improved fit is that RSA influences both amino acid level processes and nucleotide-level processes.

We implemented this model in the phylogenetic modeling language HyPhy [84]. One problem we faced is that HyPhy does not allow a continuous co-variable (such as r) in the model matrix. To overcome this technical problem, we binned RSA values into n bins and represented all RSA values within bin k by the bin mid-point, which we denote by r_k . In this way, we approximate a single matrix $Q(r)$ that changes continuously with r by a set of n discrete matrices $Q_k = Q(r_k)$, with $k = 1, \dots, n$. HyPhy allows us to simultaneously fit multiple discrete matrices, and it also allows us to share parameters among these matrices. In the limit of large n , our discretized model converges to the model that is continuous in r .

Our model contains three fitted parameters: $\omega(r)$, $\kappa(r)$, and $t(r)$. For each parameter, we considered three types of RSA dependency. First, a parameter can be constant, i.e., not actually depend on RSA. In this case, we have $\omega(r) = \omega_0$, $\kappa(r) = \kappa_0$, or $t(r) = t_0$. Second, a parameter can be a linear function of RSA. In this case, we have $\omega(r) = \omega_0 + \omega_1 r$, $\kappa(r) = \kappa_0 + \kappa_1 r$, or $t(r) = t_0 + t_1 r$. (But note that we actually only use n discrete RSA values r_k , because of the binning procedure). Finally, we can allow for separate ω , κ , and t values in each bin. (We refer to this case as *per-bin* parameter estimation). In this case, we fit n distinct ω values, one for each bin (which we refer to as ω_{r_k}), and likewise for κ and t . Figure 2.1 illustrates the various modeling choices for ω , κ , and t , in various combinations.

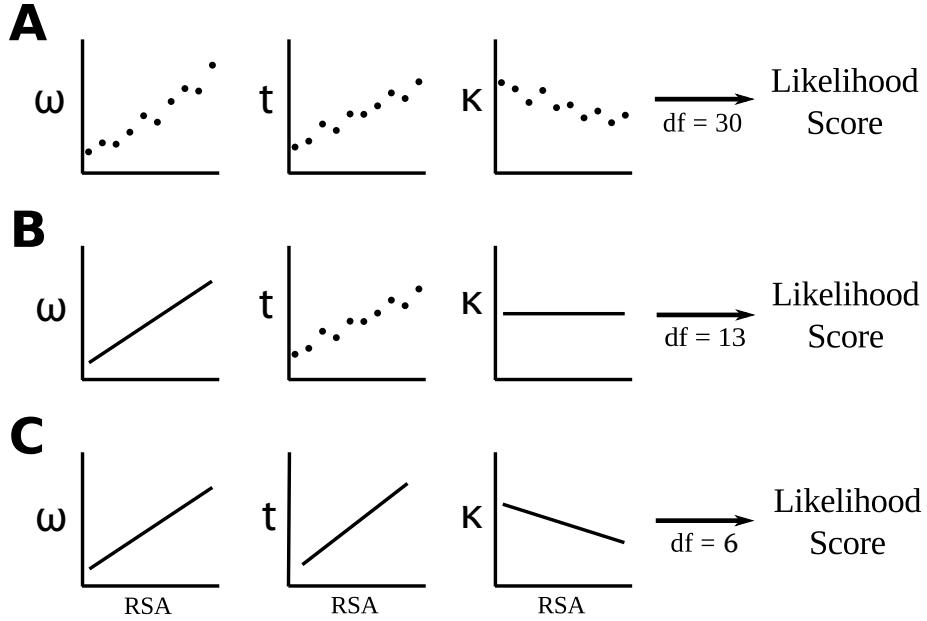


Figure 2.1: Examples of RSA-dependent sequence-evolution models considered. All models have three parameters, evolutionary-rate ratio ω , branch length t , and transition-transversion ratio κ . All three parameters can be estimated as an individual value within each RSA bin (per-bin), as a linear function of RSA (linear), or as a constant across all RSA values (constant). The examples here are illustrated for $n = 10$ RSA bins. **(A)** All parameters are estimated per-bin. **(B)** ω is estimated as a linear function, t is estimated per-bin, and κ is estimated as a constant. **(C)** All parameters are estimated as a linear function.

2.3 A linear RSA dependency for all estimated parameters provides the best model fit

We fitted our model to a data set of yeast sequences with available structural information. We identified 587 *Saccharomyces cerevisiae* genes with a known ortholog in *Saccharomyces paradoxus* and with a representative structure in the Protein Data Bank (PDB). We calculated RSA for each site as described in [10]. Unless noted otherwise, we used $n = 20$ evenly spaced RSA

bins.

Since we considered three different functional forms of RSA dependence (constant, linear, and per-bin) for each of the three parameters ω , κ , and t , we had 27 possible models. We fit all these models to our data set and ranked them by their Akaike Information Criterion (AIC [1, 14]). Results for all models are shown in Table 2.1. The top-scoring model was one in which ω and t depended linearly on RSA while κ was estimated per-bin. The differences in AIC were quite substantial among models, and the top-scoring model was clearly better than the next best model (in which all parameters were estimated as linear functions).

In general, we found that all parameters varied significantly with RSA. The top eight models did not contain a single model in which even one parameter was constant over RSA. This result shows that it is not sufficient to just make ω a function of RSA, but that the transition-transversion bias κ and branch length t also depend on RSA. Among the models with constant parameters, models with constant t ranked the highest. Models with constant ω consistently ranked the lowest. This result highlights the strong dependency of amino-acid substitution patterns on RSA.

Whenever the transition-transversion bias κ was allowed to vary with RSA, either linearly or per-bin, we found that it generally had a negative slope (decreased with increasing RSA). The branch length t tended to have a positive slope (increased with increasing RSA), unless κ was made constant, in which case t assumed a negative slope (Table 2.1).

Figure 2.2 shows ω as a function of RSA as estimated for the overall best model (with linear ω and t and per-bin κ). We see that the estimates from both models are highly consistent with each other, and that the per-bin estimates strongly support a linear relationship between ω and RSA. To assess the effect of the binning procedure on model estimation, we re-fitted the fully linear model (with linear ω , κ , and t) using different numbers of bins, from $n = 4$ to $n = 20$. Parameter estimates were nearly independent of n and varied smoothly in n (Table 2.2). We obtained similar results when we used a model with linear ω and t and per-bin κ .

Surprisingly, the log-likelihood did not vary smoothly in n (Table 2.2). For example, we observed the overall best likelihood score for $n = 11$, while $n = 10$ had a comparatively poor likelihood score. We believe that aliasing issues caused the discontinuity in likelihood scores. A sites RSA can be high or low relative to the range of RSA values within a bin. After a small change in the number of bins (for example, from $n = 10$ to $n = 11$), some sites that previously had a relatively low RSA for their bin will now have a relatively high RSA or vice versa. If those sites are particularly variable or particularly conserved, the change in their location relative to the bin center can substantially affect the quality of the model fit. For this reason, we do not think that it is reasonable to select the number of bins based on the likelihood score of the model. Instead, we opted for using a relatively large bin number ($n = 20$), which more accurately approximates a smooth dependency of model parameters on RSA.

Table 2.1: Fitted models, in order of ascending AIC

ω	t	κ	lnL	df	AIC	t slope	κ slope
linear	linear	per-bin	-839713.86	24	1679476	+	-
linear	linear	linear	-839736.74	6	1679485	+	-
per-bin	linear	per-bin	-839701.37	42	1679487	+	-
per-bin	linear	linear	-839722.37	24	1679493	+	-
linear	per-bin	linear	-839723.27	24	1679495	+	-
linear	per-bin	per-bin	-839707.75	42	1679499	+	-
per-bin	per-bin	linear	-839710.08	42	1679504	+	-
per-bin	per-bin	per-bin	-839694.42	60	1679509	+	-
linear	constant	linear	-839757.23	5	1679524	0	-
per-bin	constant	linear	-839740.64	23	1679527	0	-
linear	constant	per-bin	-839742.62	23	1679531	0	-
per-bin	constant	per-bin	-839727.25	41	1679537	0	-
linear	linear	constant	-839825.99	5	1679662	-	0
per-bin	linear	constant	-839809.70	23	1679665	-	0
linear	per-bin	constant	-839817.06	23	1679680	-	0
per-bin	per-bin	constant	-839800.41	41	1679683	-	0
linear	constant	constant	-839867.98	4	1679744	0	0
per-bin	constant	constant	-839856.43	22	1679757	0	0
constant	linear	per-bin	-840468.84	23	1680984	+	-
constant	per-bin	per-bin	-840459.99	41	1681002	+	-
constant	per-bin	linear	-840479.14	23	1681004	+	-
constant	linear	linear	-840524.57	5	1681059	+	-
constant	linear	constant	-840697.41	4	1681403	+	0
constant	per-bin	constant	-840688.35	22	1681421	+	0
constant	constant	linear	-840738.77	4	1681486	0	-
constant	constant	constant	-840740.37	3	1681487	0	0
constant	constant	per-bin	-840726.86	22	1681498	0	0

2.4 GY94 model provides a better model-fit than MG94 model

The GY94 model describes evolutionary rates using the two parameters t and ω . An alternative model, the Muse-Gaut model (MG94 [77]), uses instead

Table 2.2: Effect of the number of bins on parameter estimates

<i>n</i>	ω_0	ω_1	t_0	t_1	κ_0	κ_1	$\ln L$
4	0.1205	0.0106	0.7110	2.4706	-2.5487	5.3465	-839824.56
5	0.1208	0.0116	0.6967	2.4734	-2.5948	5.3547	-817178.82
6	0.1162	0.0135	0.7012	2.4828	-2.5361	5.3136	-839781.41
7	0.1149	0.0143	0.7034	2.4849	-2.5102	5.2976	-839764.54
8	0.1138	0.0148	0.7269	2.4805	-2.5336	5.2996	-839760.69
9	0.1123	0.0154	0.7062	2.4900	-2.4831	5.2759	-835407.29
10	0.1129	0.0156	0.7020	2.4898	-2.5003	5.2811	-839745.29
11	0.1132	0.0159	0.6742	2.4879	-2.4497	5.2669	-797981.33
12	0.1119	0.0161	0.6706	2.5007	-2.4451	5.2571	-837291.42
13	0.1110	0.0162	0.7114	2.4902	-2.4846	5.2703	-836692.33
14	0.1108	0.0164	0.6956	2.5005	-2.4632	5.2532	-837806.63
15	0.1115	0.0164	0.6959	2.4941	-2.4759	5.2653	-839684.07
16	0.1102	0.0167	0.7174	2.4897	-2.4858	5.2666	-839740.91
17	0.1098	0.0169	0.7146	2.4886	-2.4609	5.2562	-835852.76
18	0.1097	0.0170	0.7074	2.4942	-2.4652	5.2548	-839148.15
19	0.1100	0.0169	0.7038	2.4937	-2.4785	5.2627	-839318.45
20	0.1097	0.0171	0.7038	2.4943	-2.4732	5.2592	-839736.74

the parameters α and β . The parameter α in MG94 corresponds to t in GY94 and the parameter β in MG94 corresponds to $t\omega$ in GY94. If we fit a model without site variability (all parameters are constant across sites), the MG94 model and the GY94 model are identical. However, when we allow for site variability, the two models become different. The GY94 model is usually set up with a constant t and a variable ω [80, 125]. This set-up implicitly assumes that the synonymous rate is constant across sites whereas the nonsynonymous rate is variable. The MG94 model, on the other hand, has been used to explicitly model both nonsynonymous and synonymous site variability [85].

Here, we have allowed both ω and t to vary with RSA, so we have

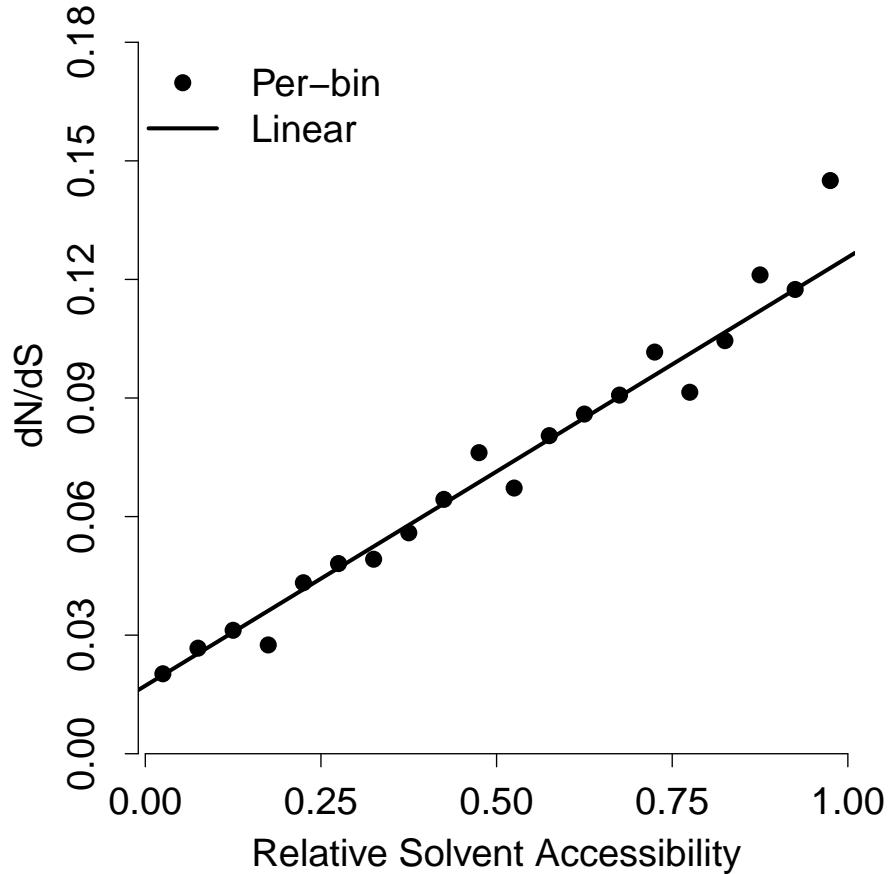


Figure 2.2: Evolutionary-rate ratio increases linearly with RSA. The solid line shows $\omega = dN/dS$ versus RSA as estimated by the best model (linear ω , linear t , and per-bin κ). The dots show the same for the best model with per-bin ω (which has linear t and per-bin κ). Both models are consistent with each other and strongly support a linear relationship between ω and RSA.

considered both nonsynonymous and synonymous rate variation. However, in using the GY94 model, we have assumed that the two quantities that vary linearly with RSA are the synonymous rate and the ratio of the nonsynonymous to synonymous rates. *A priori*, it is just as reasonable to assume that

the synonymous rate α and the nonsynonymous rate β are linear functions of RSA. In this case, the ratio $\omega = \beta/\alpha$ would of course not be linear in RSA.

To assess whether the nonsynonymous rate β or the ratio $\omega = \beta/\alpha$ is linear in RSA, we fitted a model in which α and β were linear functions of RSA. (κ was estimated per-bin). The resulting relationship of ω vs. RSA was similar but not identical to the one observed for linear ω (Figure 2.3). The log-likelihood score for this model fit was -839720.75, compared to a log-likelihood score of -839713.86 for the model with linear ω . The two models are not nested, so we cannot compare them using a likelihood ratio test. However, they are comparable via AIC, and the model with linear ω was clearly better ($\Delta\text{AIC} = 14$).

2.5 Effect of relative solvent accessibility on synonymous and nonsynonymous substitution rates

The previous subsections have shown that substitution rates at both synonymous and nonsynonymous sites are affected by RSA, and that the ratio $\omega = dN/dS$ changes linearly with RSA. If ω is linear in RSA and both dN and dS vary with RSA, then we expect dN and dS individually to not be linear in RSA.

The quantities dN and dS are not parameters that are estimated in the model fit. Instead, they are derived quantities that we can calculate once

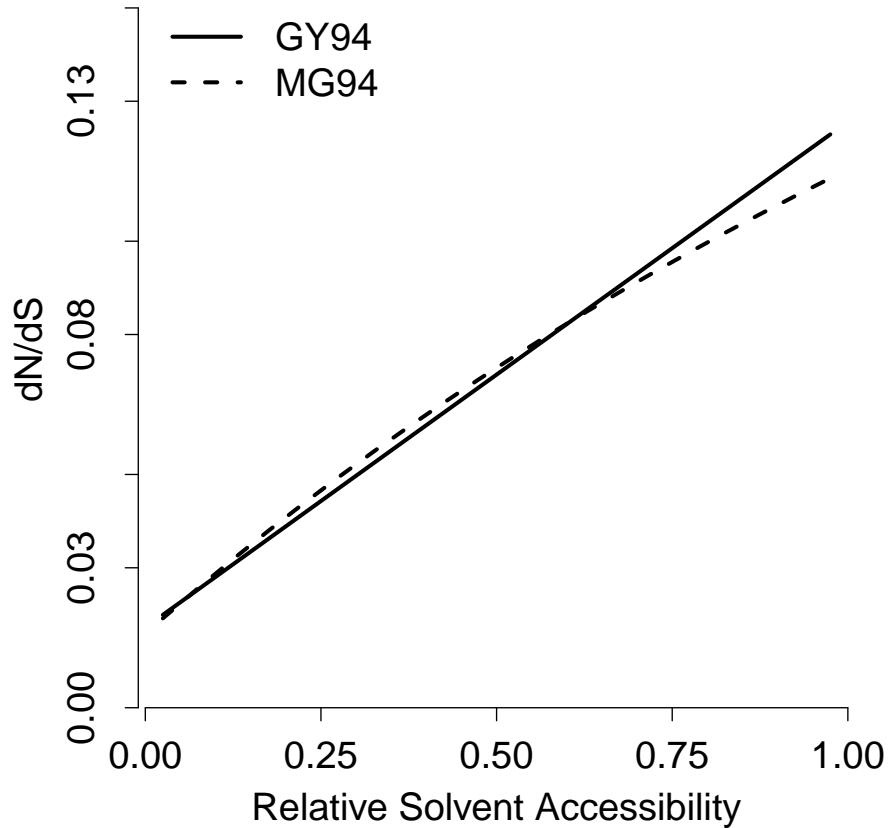


Figure 2.3: Comparison of the GY94 and MG94 models. The solid line shows $\omega = dN/dS$ versus RSA, as estimated by the GY94 model. The dashed line shows the same for the MG94 model. Under the MG94 model, ω shows moderate curvature. The GY94 model provides a better fit to the data ($\Delta\text{AIC} = 14$).

the model has been fit to the data. One complication in calculating dN and dS arises, however: There are multiple definitions of these parameters. For example, dS is defined as the number of synonymous differences divided by the number of synonymous sites in the sequence. We obtain the number of synonymous differences by summing over the elements in the matrix Q [123].

The number of synonymous sites can be obtained in two different ways. First, we can simply count the number of sites at which mutations can cause either a synonymous or a nonsynonymous change. This method of counting gives us the physical-sites definition of dS [9]. Second, we can weigh each site with the probability that a synonymous mutation will occur at this site under the fitted model. This method of counting sites gives us the mutational-opportunity definition of dS [9]. The same two definitions exist for dN .

The mutational-opportunity and the physical-sites definitions gave nearly identical results for dN (Figure 2.4A). In both cases, dN showed a strong increasing trend with RSA, with a slight deviation from linearity for higher RSA values. By contrast, the two definitions gave somewhat different results for dS . Under the mutational-opportunity definition, dS was decreasing with RSA, whereas under the physical-site definition it showed no obvious trend (Figure 2.4B).

2.6 The effect of core size and expression level on evolutionary rate

In yeast, the primary determinant of evolutionary rate is gene expression level [30, 31]. A second determinant is protein structure, measured either by contact density [10] or by core size [39]. Thus, we investigated how the slope and intercept of the linear function $\omega = \omega_0 + \omega_1 r$ changed with protein core size (measured by average RSA) and with gene expression level (measured by mRNA abundance).

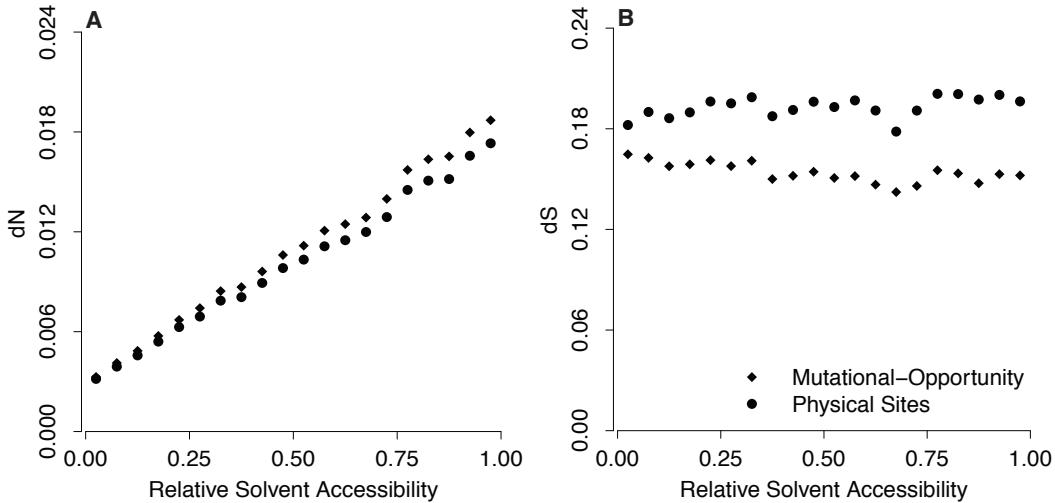


Figure 2.4: Evolutionary rates dN and dS . (A) The nonsynonymous rate, dN , correlates strongly with RSA under both the mutational-opportunity definition and the physical-sites definition. (B) The synonymous rate, dS , shows a moderate negative correlation with RSA under the mutational-opportunity definition and no slope under the physical-sites definition. The fitted model had linear ω , linear t , and per-bin κ .

Franzosa and Xia showed that the slope of ω changed with core size while the intercept remained nearly unchanged. We repeated their analysis by identifying the 33% largest and smallest cores and fitting a joint evolutionary model to these proteins. We fitted one line for each κ and t but fitted two separate lines for ω , one for the large-core proteins ($\omega^{lc} = \omega_0^{lc} + \omega_1^{lc}r$) and one for the small-core proteins ($\omega^{sc} = \omega_0^{sc} + \omega_1^{sc}r$), as shown in Figure 2.5. We found that small-core proteins displayed a smaller slope than large-core proteins ($\omega_1^{sc} = 0.082$ vs. $\omega_1^{lc} = 0.127$). This difference in slopes was significant (likelihood ratio test, $P = 6.41 \times 10^{-9}$). By contrast, the intercepts were not significantly different (likelihood ratio test, $P = 0.136$), and we found

$$(\omega_0^{sc} = \omega_0^{lc} = 0.018).$$

The two slopes we found were more similar to each other than the ones found by Franzosa and Xia [39]. The main difference between our data set and theirs was that we used more stringent criteria to match sequences to structures. To verify that we could reproduce the results of Ref. [39], we relaxed our criteria for alignment length to 70%, thereby increasing our dataset to 870 sequence-structure pairs. For this larger data set, we found a similar slope for large-core proteins as found before ($\omega_0^{lc} = 0.124$), but the slope for small-core proteins was reduced ($\omega_0^{lc} = 0.058$). These slopes were consistent with the findings of Ref. [39].

We carried out a similar analysis on high-expression and low-expression genes, fitting a separate line to each group of proteins ($\omega^{he} = \omega_0^{he} + \omega_1^{he}r$ for high-expression genes, $\omega^{le} = \omega_0^{le} + \omega_1^{le}r$ for low-expression genes). We found a substantial difference in slope between these two groups of genes ($\omega_0^{he} = 0.047$ vs. $\omega_0^{le} = 0.164$). The difference was significant (likelihood ratio test, $P = 1.75 \times 10^{-62}$). We also found a difference in intercept ($\omega_1^{he} = 0.011$ vs. $\omega_1^{le} = 0.023$) and this difference was significant as well (likelihood ratio test, $P = 6.05 \times 10^{-12}$). Similar results were found when we used codon adaptation index as a proxy for gene expression level (data not shown).

Finally, we carried out a joint analysis of core size and expression level by extracting four groups of proteins from our data set: proteins with (1) high expression level and large core, (2) high expression level and small core, (3) low expression level and large core, and (4) low expression level and small core.

Figure 2.6 shows the resulting model fit. Clearly, expression level plays a larger role in determining evolutionary rate than core size. However, the model with core-size-dependent slope showed a better fit than a model in which the slope depended only on expression level (likelihood ratio test, $P = 5.33 \times 10^{-4}$). Surprisingly, the effect of core size on slope was reversed for high- and low-expression genes. For high-expression genes, proteins with smaller core size showed a larger slope in ω than did proteins with smaller core size, consistent with prior results. by contrast, low-expression proteins with larger core size showed a smaller slope than did proteins with smaller core size. However, this unexpected pattern disappeared when we repeated the above analysis on our expanded data set with 870 sequence-structure pairs. There, the large-core-size proteins had the larger slope in all cases, consistent with prior results (data not shown).

2.7 Discussion

We have developed a method that models the evolutionary rate of a coding sequence within the context of the protein’s 3-dimensional structure. Our method is a simple extension of the standard GY94 model, modified such that all parameters are functions of relative solvent accessibility (RSA). We have found that the evolutionary-rate ratio $\omega = dN/dS$, the branch length t , and the transition-transversion bias κ all depend on RSA. The overall best fitting model had a linear relationship of ω and t with RSA, while κ showed small deviations from strict linearity. In the second-best model, all parameters

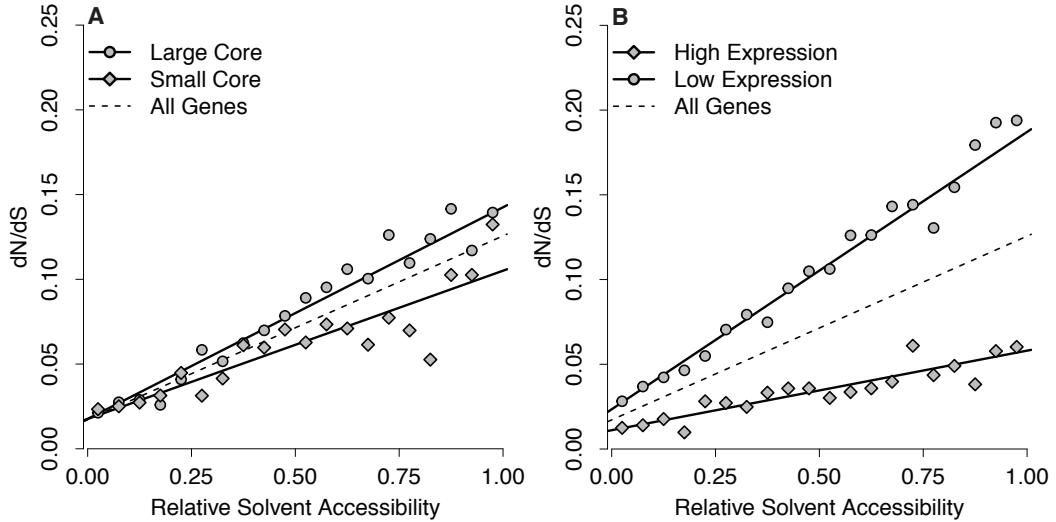


Figure 2.5: Dependency of $\omega = dN/dS$ on protein core size and expression level. (A) Core size affects evolutionary rate on the surface of the protein but not in the core. (B) Expression level affects evolutionary rate both on the surface and in the core. However, it has a bigger effect on the surface of the protein. In both figures, the solid lines were estimated jointly from the data using a linear dependency of ω on RSA. Points for individual bins are shown for illustration purposes only. They were estimated using a per-bin model for ω . In the figures above, the dashed black line represents the genome-wide trend, as shown in Figure 2.2, and is provided as a reference.

had a linear relationship with RSA.

Our method presents a unified statistical framework for comparing RSA-dependent model parameters among different groups of proteins. Using this framework, we have shown that protein core size affects only the slope of ω as a function of RSA, but not the intercept. The most buried residues have —on average —the same ω value regardless of protein core size. By contrast, expression level affects ω for even the most buried residues.

We have found that the variation in ω with RSA is substantial; for the

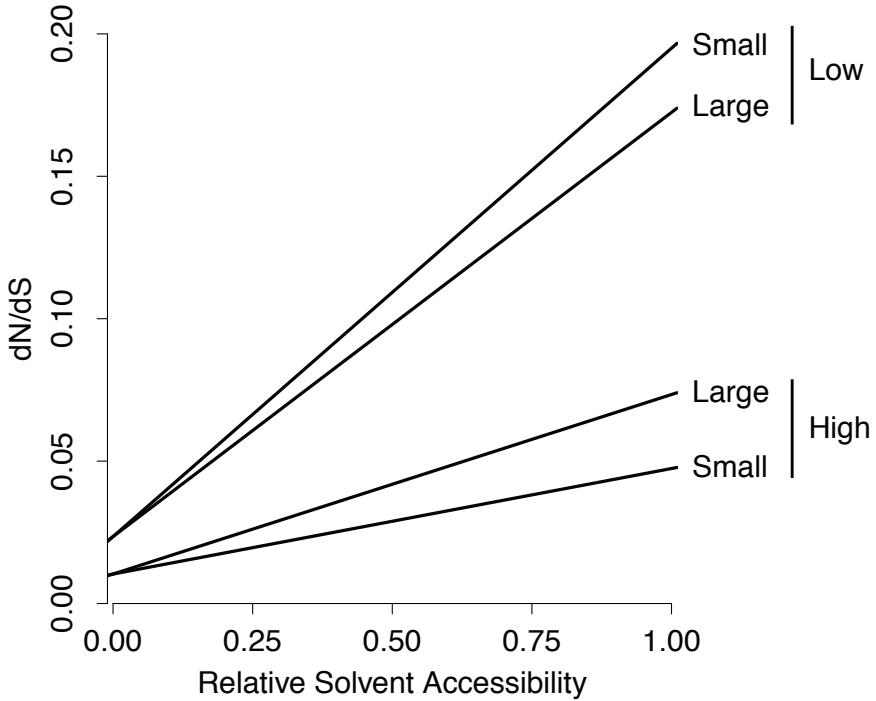


Figure 2.6: Joint analysis of the effects of both core size (small or large) and expression level (high or low) on the relationship between $\omega = dN/dS$ and RSA. Only the fitted lines are shown. Surprisingly, for low-expression genes, small-core proteins evolve faster than large-core proteins. This relationship is reversed in a larger dataset obtained with less-stringent criteria (see text).

most exposed residues, ω was on average 5-10 times larger than it was for the most buried residues. This observation highlights the importance of incorporating protein structure into models of coding-sequence evolution. Traditional models of rate variation [125, 123, 86] cannot distinguish between rate variation caused by other factors (e.g., positive or negative selection on sites of functional importance). As an obvious extension to the work presented here, we can combine the present model with more traditional models of rate vari-

ation among sites with similar RSA. This work has been presented in Ref. [72].

Our findings here are broadly consistent with the findings of Franzosa and Xia [39]. We have confirmed the linear relationship between dN/dS and RSA in an independently derived data set; we have also confirmed that proteins with larger core size show a faster increase of dN/dS with increasing RSA than proteins with smaller core size. Our work goes beyond Franzosa and Xia's findings by demonstrating that the evolutionary rate of fully buried residues is independent of protein core size, that expression level affects evolutionary rate at all RSA values, and that the GY94 model provides a better fit than the MG94 model when RSA-dependent evolutionary rates are considered. Our work also suggests that nucleotide-level processes vary systematically with protein structure.

In the joint analysis of core size and expression level, we made the unexpected observation that the effect of core size on the slope of ω is reversed for genes with low expression level. However, this observation disappeared in a larger data set obtained under slightly less stringent criteria for matching sequences to PDB structures. We can offer no good explanation for this observation. It could be a statistical fluke. The number of genes in each of the four groups (low expression and small core size, low expression and large core size, high expression and small core size, high expression and large core size) is relatively small in this analysis, so a few unusual proteins could skew the analysis. What exactly is the cause of this unexpected observation may have

to be clarified in future analyses, either using expanded data sets —as more structures become available —or using data from different organisms.

Our approach is conceptually related to other recent works attempting to combine protein structure with sequence evolution [89, 92, 93, 91]. These works imposed structural constraints on sequence evolution via sophisticated energy functions describing how protein fold stability changes as amino acids are replaced. In comparison, our approach is much more simplistic. However, we believe that this simplicity has substantial benefits. First, our approach is simple and fast. All the models we have used here can be fit within 1015 minutes on an off-the-shelf laptop. Second, our approach yields results that can be interpreted easily. Instead of a single ω value per gene, we obtain two values, an intercept and a slope. The intercept tells us to what extent selection constrains the most buried residues; the slope tells us by how much selection relaxes as we move towards more exposed residues. Third, our approach can be implemented with relative ease in existing modeling frameworks such as HyPhy [84].

Following Franzosa and Xia [39], we used a model that fit a single rate ratio ω , regardless of which amino acids were substituted into which other ones. A recent study has shown that such models can always be improved upon with amino-acid dependent transition rates, even if amino acids are grouped into exchangeability categories at random [28]. This finding is not entirely surprising, considering that amino-acid substitution matrices have consistently been found to depend substantially on the amino-acid identity (e.g. Refs.

[24, 52, 114]). Therefore, it would be desirable to develop codon-level substitution models that accurately capture this rate variation, without adding too many additional parameters. Approaches that have been suggested include automatically grouping amino acids into exchangeability categories [65, 27] and decomposing amino-acid substitution rates into components corresponding to biophysical properties of amino acids (LCAP model, Ref. [19]). Yet substitution rates also depend on protein structure [81, 61, 41, 62], and thus one would want to incorporate structure into these models as well. One study developed a variant of the LCAP model where parameters were fit separately to buried and exposed sites and found to be significantly different [18]. Since we have seen here that substitution rates seem to depend continuously (and linearly) on RSA, it might be worth it to investigate a variant of the LCAP model in which rate parameters are linear functions of RSA. Such a model would have the same number of parameters as the model in Ref. [18], but would quite possibly provide a better fit to the data. Alternatively, one could attempt to incorporate an RSA-dependence into models that automatically group amino acids [65, 27].

We found that in our model, both t and ϑ varied with RSA. We believe that this finding reflects the effect of selection on nucleotide-level processes. First, equilibrium amino-acid frequencies vary with RSA [88, 87], and this variation will have some effect on equilibrium codon frequencies. Second, protein structure also seems to exert a direct selection pressure on synonymous codon choice [104, 60, 20, 58, 128, 132, 66], most likely through an interaction be-

tween the translation process and protein folding. A more realistic model could represent this relationship between protein structure and the nucleotide-level substitution process more accurately, for example via a structure-dependent variant of the FMutSel model [124] or by extending models such as the LCAP model [18, 19] to contain structure-dependent terms for nucleotide-level processes.

The challenge in developing any such models will be to make them realistic yet sufficiently simple so they can be fit to moderately sized data sets. An alternative, simpler strategy could be to calculate equilibrium codon frequencies in an RSA-dependent manner. We considered calculating codon frequencies per bin and found that doing so generally improved AIC scores but did not eliminate the need for RSA-dependent t or τ , nor did it alter any of our other results in a substantive way (not shown).

Our method requires a solved crystal structure to calculate RSA values. Although the Protein Data Bank (PDB) has been growing rapidly over the past decade, the number of available structures is still small compared to the number of available sequences. For example, many of the yeast sequences we used in our analysis did not have a corresponding structure. For those sequences, we relied on homologous protein structures solved in related organism. Homology mapping performs relatively well in predicting relative solvent accessibility [132] but clearly it is not perfect. Further, certain proteins or regions of proteins, such as membrane proteins or intrinsically disordered regions, can usually not be crystallized. Thus, our method cannot be applied to

such proteins or regions of proteins.

Our method assumes that RSA remains constant throughout evolution. Yet every amino-acid replacement will cause some distortion in the protein structure [17], and RSA values at homologous sites will slowly diverge with increasing sequence divergence [132]. In the future, if either the number of available PDB structures increases drastically or if atom-level computational modeling of protein structures becomes sufficiently reliable, we will be able to study how changes in structure correlate with evolutionary rate.

2.8 Methods

2.8.1 Homology mapping and categorization of genes

In order to construct a large data set of sequences with corresponding structures, we obtained open reading frames (ORFs) of the yeast *Saccharomyces cerevisiae* from the Saccharomyces Genome Database [16] and aligned them with orthologous *Saccharomyces paradoxus* sequences using MUSCLE [36]. Each ORF was translated and searched against the Protein Data Bank (PDB) [7] using the PSI-BLAST algorithm [4] and then paired with the structural chain with the lowest alignment E-value. To ensure that enough of the yeast protein was represented in the chain and that the PDB structure was a reasonable homology model, we only considered pairs with > 80% alignment length and > 40% sequence identity for analysis. Our final data set had 587 sequence-structure pairs. A data set with relaxed criteria used > 70% alignment

length and $> 40\%$ sequence identity. This data set had 870 sequencestructure pairs.

The percent solvent-accessible surface area (ASA) for each aligned residue was calculated using DSSP [54]. We obtained relative solvent accessibility (RSA) by normalizing ASA values with the surface areas of an extended Gly-X-Gly peptide [22].

2.8.2 Calculation of evolutionary rates

The codons from the yeast alignments were binned by the RSA value of their respective residues, as described [39]. Protein core size was estimated by the average RSA value over all residues in a protein. We considered a structure to have a large core if its average RSA ranked within the bottom third of all average RSA values and to have a small core if ranked within the top third of all average RSA values [39]. Yeast expression data measured in mRNA abundance per cell was obtained from [48]. Codon adaptation index (CAI), a measure of the strength of codon usage bias, was used as an alternative for expression level, since the latter may be biased by laboratory growth conditions of the yeast cells [96]. Both expression level and CAI were ranked and divided into thirds with the top third representing high-expression genes and the bottom third low-expression genes.

We implemented the model described by Equation 2.1 in the HyPhy batch language [84]. We estimated codon frequencies (π_j) using $F3 \times 4$ model.

We calculated synonymous (dS) and nonsynonymous (dN) substitution rates according to the mutational-opportunity and the physical-sites definitions, as described [123, 9].

2.8.3 Statistical analysis

We used the Akaike information criterion (AIC) [1, 14] to rank models by their quality of fit. For pairwise comparison of nested models, we also carried out likelihood-ratio tests. All statistical analyses were performed using the statistics software R [50].

Chapter 3

Modeling the mutation rates of coding-sequences under the constraints of solvent accessibility

3.1 Background

Substitution mutations occur when one nucleotide is replaced by another and such changes are classified as either transitions ($A \rightleftharpoons G$, $C \rightleftharpoons T$) or transversions ($A \rightleftharpoons C$, $A \rightleftharpoons T$, $G \rightleftharpoons C$, $G \rightleftharpoons T$). Although mutations are seen to arise randomly through the DNA replication process, this randomness tends not apply pragmatically at the molecular level as not all nucleotides are comparatively convertible. Since sequenced genomes have become available, transition mutations have been observed to occur far more frequently than transversion mutations, despite the fact that there are more possible transversions than transitions [40, 109, 110]. In models of nucleotide substitution, tran-

sition/transversion bias is defined as the ratio between the overall rate of transition changes (T_i) to the overall rate of transversion mutations (T_v). The transition/transversion bias, although more pronounced in animal mitochondrial DNA (mtDNA) than in nuclear coding sequences, is a universal property of DNA evolution [110]. Therefore, in order to understand the divergence of species and the marks of natural selection, it becomes necessary to understand and interpret the mutational biases that have arisen between species.

Mutation bias arises primarily through the biochemical structures of the nucleotide bases and the properties of their chemical bonds in complementary pairing. While purines must always pair with purines and pyrimidines with pyrimidines, mutations occur when a disfavored tautomeric form of one of the four bases is incorporated during DNA replication. Transitions arise through purine-pyrimidine mispairs while transversions from purine-purine mispairs; pyrimidine-pyrimidine mispairing does not occur [113, 110, 40]. Since mispairings differ in their free energies of pairing and their abilities to incorporate into a helical structure, in addition to transversion mutations necessitating a base rotation for mispairing, transitions are favored over transversions starting with the DNA structure [106, 99, 40].

If mutation bias is inherent to the process of DNA replication, why does the transition/transversion rate ratio scale with the relative solvent accessibility (RSA) of a protein’s residues as seen in Chapter 2? When a substitution mutation occurs in a protein coding region, there is a possibility of this mutation resulting in an amino acid replacement downstream. Among point

mutations that occur in the universal genetic code at the third position, only 3% of transition mutations result in a nonsynonymous substitution in comparison to the 41% of transversions that cause amino acid replacements. At the first and second positions, both transitions and transversions cause nonsynonymous changes in nearly all cases; however, transition mutations tend to cause changes to amino acids with similar chemical properties [108]. Therefore, if selection for protein stability acts to conserve the chemical composition of a protein, transitions will be favored over transversions [43, 79]. Since destabilizing mutations in the core of the protein will be more disruptive than those on the surface, we would expect more transitions to occur in triplets that code for buried residues.

At high levels of divergence, saturation of transition mutations causes the appearance of an apparent equal transition to transversion ratio [13, 75, 23]. Table 3.1 shows the overall nucleotide and amino acid sequence divergence as well as divergence times obtained from the literature for four pairwise alignments between model organisms (yeast: ~5 MYA [56]; fly: ~10-12 MYA [69]; mouse: ~20 MYA [120]; worm: ~80-110 MYA [47]). We believe that the relationship between κ and solvent accessibility is related to the time since divergence between the orthologous species in question.

To determine the effects of solvent accessibility on the nucleotide-level substitution process, we chose to extend our analysis in Chapter 2 by using a model that introduces structure-dependent terms for nucleotide-level processes. Here, we find that the transition/transversion parameter is best mod-

eled in the transition position in our RSA-dependent model in yeast. Furthermore, we find that estimating codon frequencies individually across the RSA gradient increases the model performance substantially. Finally, we show that the nucleotide-level parameters all vary with solvent accessibility in a manner that is highly dependent on species.

Table 3.1: Overall nucleotide and amino acid sequence divergence

Species	Divergence Time	Nucleotide Divergence	AA Divergence
Yeast	~ 5 MYA	7.95%	3.03%
Fly	~ 10-12 MYA	6.54%	3.78%
Mouse	~ 20 MYA	7.24%	7.29%
Worm	~ 80-110 MYA	21.16%	11.65%

3.2 An RSA-dependent General Time Reversible model of coding-sequence evolution

Our work in Chapter 2 resulted in a statistical framework for assessing the compatibility of RSA-dependent model parameters across different groups of proteins. Such an approach has shown the linear relationship between RSA and the two evolutionary rate parameters, protein selection (ω) and branch length (t), to be directly accessible within the estimates of a codon model. What is less clear from our prior analysis is how the nucleotide-level processes, specifically described by the transition-transversion rate ratio κ , vary with solvent accessibility. While our analysis indicates a dependency between RSA and κ , it is unclear whether the relationship is one that can be described by a linear function or if κ has a discrete value for each RSA division (modeled

per RSA bin). Furthermore, κ is a ratio of the number of transitions to transversions if modeled for transitions and the reciprocal if modeled at the transversion position. However, the position of κ at either the transition or transversion position has been inconsistent between previous methodology and may affect the outcomes of future analyses [121, 84, 123, 122].

To determine the best way to model the mutation parameter κ , we developed a variant of the General Time Reversible (GTR) [103] that has been adapted to work with coding sequences through the addition of GY94 parameters. Under this model, there are 6 free parameters that describe the instantaneous rate of change between the four types of nucleotides. We chose this model as it provides the most general description of nucleotide changes occurring during the mutational process. We use the standard continuous time Markov chain description of coding sequence evolution. The transition probability matrix for $t > 0$ is given by the matrix exponential e^{Qt} , where the rate matrix $Q = Q_{ij}$ describes the instantaneous rate of change from codon i to codon j , for all $i \neq j$

$$Q_{ij} = \begin{cases} 0, & \text{if more than one change} \\ \theta_{ij}(r)\pi_j, & \text{if synonymous change} \\ \omega(r)\theta_{ij}(r)\pi_j, & \text{if nonsynonymous change} \end{cases} \quad (3.1)$$

where ω is the ratio of the nonsynonymous to synonymous substitution rates and r stands for the RSA of a site. The nucleotide-level mutation parameters (such as the transition-transversion ratio, κ) are described by the matrix θ , which is a component of our 61×61 matrix Q and represents the possible nucleotide changes in each codon. The matrix θ defines four additional param-

eters: T_{iN} , which is the ratio of the rate of mutation between purine \rightleftharpoons purine verses pyrimidine \rightleftharpoons pyrimidine; T_{vN} , which is the rate ratio between Watson-Crick paired nucleotides and unpaired nucleotides; T_{vP} , the rate ratio between A \rightleftharpoons T and G \rightleftharpoons C; and T_{vU} , the rate ratio between A \rightleftharpoons C or G \rightleftharpoons T mutations (refer to Figure 3.1). Equation 3.2 provides an example 4 x 4 matrix, θ , while Equation 3.3 shows the each parameter in the inverse positions of those in Equation 3.2.

$$\theta = \begin{bmatrix} * & T_{vU} & \kappa T_{iN} & T_{vN} T_{vP} \\ T_{vU} & * & T_{vN} & \kappa \\ \kappa T_{iN} & T_{vN} & * & 1 \\ T_{vN} T_{vP} & \kappa & 1 & * \end{bmatrix} \quad (3.2)$$

$$\theta = \begin{bmatrix} * & \kappa T_{vN} & 1 & \kappa \\ \kappa T_{vN} & * & \kappa T_{vP} & T_{iN} \\ 1 & \kappa T_{vP} & * & \kappa T_{vN} T_{vU} \\ \kappa & T_{iN} & \kappa T_{vN} T_{vU} & * \end{bmatrix} \quad (3.3)$$

For each parameter, we consider four types of RSA dependency, r . In each of these examples, ρ is used as a example parameter for those found in the model. First, a parameter can be estimated as a constant that does not actually depend on RSA ($\rho(r) = \rho_0$). Second, a parameter can be estimated individually with each RSA division, or per-bin, resulting in n discrete RSA values of r_k , where k is the mid-point of the RSA bin. Third, a parameter can be described by a linear function of RSA ($\rho(r) = \rho_0 + \rho_1 r$). Finally, we introduce a quadratic relationship with RSA, such that $\rho(r) = \rho_a r^2 - \rho_b r + \rho_c$. Figure 3.2 provides examples of models with each of these parameter types.

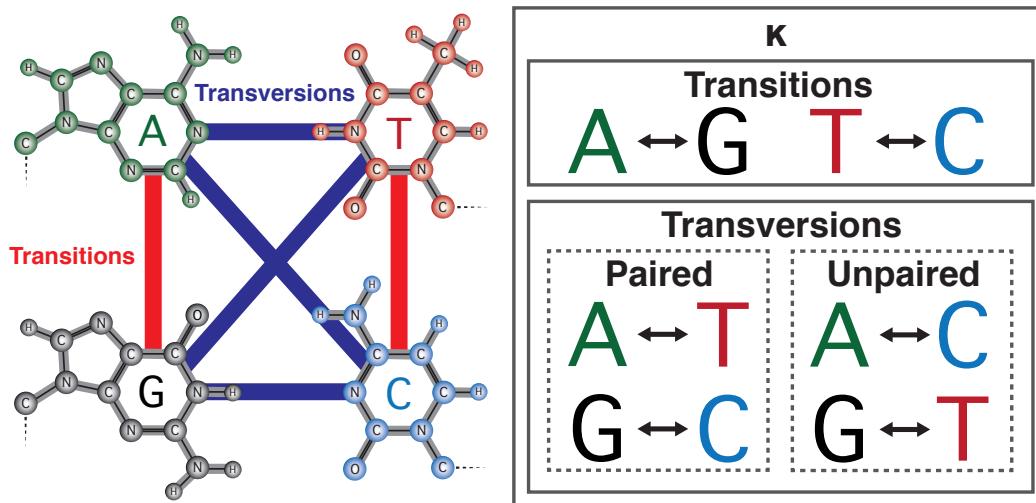


Figure 3.1: Combinations of base pairs for transitions and transversions. Mutations can be classified as either transitions ($A \rightleftharpoons G$, $C \rightleftharpoons T$) or transversions (purine \rightleftarrows pyrimidine). The transition to transversion rate ratio is described by the parameter κ . Transitions occur between similar bases and there are two possible transition mutations. Transversions are exchanges of dissimilar nucleotides and the four possible transitions mutations can be further categorized into the traditional Watson-Crick paring bases or bases that do not pair in a DNA double helix.

3.3 The transition/transversion parameter, κ , is best modeled in the transition position

The transition/transversion parameter κ describes the mutational process and is modeled at either the transition or transversion position in the substitution matrix. Typically a κ estimate at one position is the reciprocal of the other and both positions provide equivalent model fits. However, the introduction of RSA-dependent parameters removes this reciprocity between the transition and transversion positions. To determine the best position in which to model the κ parameter, we fitted our more general model to our data

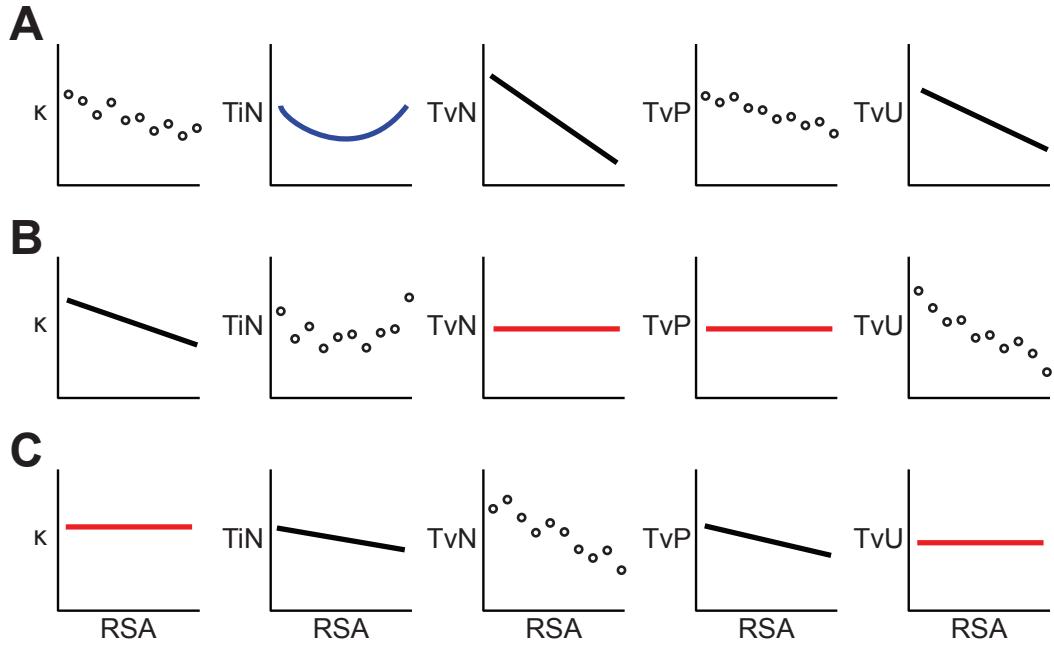


Figure 3.2: Examples of RSA-dependent nucleotide-level models considered. All models have five nucleotide-level parameters (transition-transversion ratio, κ ; transition type ratio, T_{iN} ; transversion type ratio, T_{vN} ; paired transversion ratio, T_{vP} and the unpaired transversion ratio, T_{vU}), in addition to the linear parameters ω and t . All five parameters can be estimated as an individual value within each RSA bin (per-bin), as a linear function of RSA, or as a constant across all RSA values (constant). For T_{iN} , we introduce an additional quadratic relationship with RSA. The examples here are illustrated for $n = 10$ RSA bins. **(A)** The top scoring model in yeast. κ and T_{vP} are per-bin, T_{vN} and T_{vN} are linear, while T_{vN} is quadratic. **(B)** Here, κ is the only linear parameter. T_{vN} and T_{vP} are held constant, while T_{iN} and T_{vU} vary per-bin. **(C)** Here, there is no RSA-dependency for T_{vN} . κ and T_{vU} are constant, while T_{iN} and T_{vP} vary linearly with RSA.

set of 587 yeast sequence-structure pairs from the orthologs of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*. The RSA for each site in the structure was calculated as previously described [39] and our $n = 20$ RSA bins were spaced evenly as before.

The parameters in our model corresponding to the RSA-dependent GY94 model, ω , t , and κ , were all fitted as a linear function of RSA while the nucleotide GTR parameters (T_{iN} , T_{vN} , T_{vP} , and T_{vU}) were all estimated per-bin. The κ parameter was placed either in the transition or transversion position in our model matrix. Each of the four nucleotide parameters was placed in one of two possible positions: (1) T_{iN} could refer either to a purine or pyrimidine transition pair. (2) T_{vN} was placed either at the positions between Watson-Crick paired nucleotides for transversions (referred to as paired) or between nucleotides that do not pair in the traditional fashion (unpaired). (3) T_{vP} refers to transversion changes between either A=T or G=C for traditionally paired bases, while (4) T_{vU} references unpaired transversion mutations from either A=C or G=T. As a result, there were 32 possible model combinations.

In general, the types of parameter combinations were distributed evenly over the top 50% scoring models (Table 3.2). However, the best model as shown by AIC is one that placed the mutation parameter κ in the transition position of the matrix. Furthermore, κ transition models comprise the majority of the top 25% scoring models, a result that heavily suggest modeling κ as the ratio of transitions to transversions rather than vice versa. Amongst the nucleotide parameters, models that place the transition parameter T_{iN} in the purine position universally score higher than their pyrimidine position counterparts. The transversion parameter T_{vN} provides a better score when placed between non-Watson-Crick bases in the matrix while the paired transversion

parameter T_{vP} is more appropriate between G↔C changes. The position of the unpaired transversion parameter T_{vU} had no effect on the model score as the log-likelihood was equal between models when all other parameters were the same. In such cases, only a minuscule change in the estimated t and κ slopes could be detected.

As previously shown in Chapter 2, the transition-transversion bias κ tended to have a negative slope with RSA. In addition to this, the branch length t had a positive slope with RSA unless κ was made constant, in which case it became negative. Here, the slopes of t and κ are dependent on the position of the parameter κ in the model, with κ having a negative slope for transitions and a positive slope for transversions. Inversely, t displays a positive slope for transitions and a negative slope for transversions.

Table 3.2: Fitted mutational model combinations, in order of ascending AIC

κ	TiN	TvN	TvP	TvU	lnL	df	AIC	t slope	κ slope
transition	purine	unpaired	GC	AC	-839468.00	86	1679108	+	-
transition	purine	unpaired	GC	GT	-839468.00	86	1679108	+	-
transversion	purine	unpaired	GC	AC	-839470.41	86	1679113	-	+
transversion	purine	unpaired	GC	GT	-839470.41	86	1679113	-	+
transition	purine	paired	GC	GT	-839473.03	86	1679118	+	-
transition	purine	paired	AT	GT	-839473.03	86	1679118	+	-
transition	purine	unpaired	AT	GT	-839474.60	86	1679121	+	-
transition	purine	unpaired	AT	AC	-839474.60	86	1679121	+	-
transversion	purine	paired	GC	GT	-839475.76	86	1679124	-	+
transversion	purine	paired	AT	GT	-839475.76	86	1679124	-	+
transversion	purine	unpaired	AT	AC	-839475.97	86	1679124	-	+
transversion	purine	unpaired	AT	GT	-839475.97	86	1679124	-	+
transition	purine	paired	GC	AC	-839477.20	86	1679126	+	-
transition	purine	paired	AT	AC	-839477.20	86	1679126	+	-
transversion	purine	paired	GC	AC	-839478.46	86	1679129	-	+
transversion	purine	paired	AT	AC	-839478.46	86	1679129	-	+
transversion	pyrimidine	unpaired	GC	GT	-839497.52	86	1679167	-	+
transversion	pyrimidine	unpaired	GC	AC	-839497.52	86	1679167	-	+
transversion	pyrimidine	unpaired	AT	GT	-839502.47	86	1679177	-	+
transversion	pyrimidine	unpaired	AT	AC	-839502.47	86	1679177	-	+
transversion	pyrimidine	paired	AT	GT	-839503.18	86	1679178	-	+
transversion	pyrimidine	paired	GC	GT	-839503.18	86	1679178	-	+
transition	pyrimidine	unpaired	AT	AC	-839504.90	86	1679182	-	-
transition	pyrimidine	unpaired	AT	GT	-839504.90	86	1679182	-	-
transition	pyrimidine	unpaired	GC	GT	-839505.00	86	1679182	+	-
transition	pyrimidine	unpaired	GC	AC	-839505.00	86	1679182	+	-
transition	pyrimidine	paired	GC	AC	-839505.30	86	1679183	-	-
transition	pyrimidine	paired	AT	AC	-839505.30	86	1679183	-	-
transversion	pyrimidine	paired	AT	AC	-839505.77	86	1679184	-	+
transversion	pyrimidine	paired	GC	AC	-839505.77	86	1679184	-	+
transition	pyrimidine	paired	AT	GT	-839512.07	86	1679196	+	-
transition	pyrimidine	paired	GC	GT	-839512.07	86	1679196	+	-

3.4 Individually estimated codon frequencies increase model performance over genomic codon frequencies

In models of codon evolution, the indices i and j run over all of the 61 sense codons and the parameter π_j is the frequency of codon j in the coding alignments. The codon frequencies π_j are estimated empirically from the sequence data from three sets of frequencies of the four types of nucleotides at each of the three codon positions (the F3x4 model), which provides a reasonable description of the data without sacrificing computational tractability. In our previous analyses, codon frequencies were estimated from the concatenated "genomic" sequence alignments for all the 587 yeast sequence-structure pairs in our data set. These genomic codon frequencies were shared globally across all RSA bins in our model.

Here, we repeat our above permutation analysis for the 32 combinations of parameter locations using codon frequencies estimated individually for each RSA bin. In comparison with the genomic codon frequency models, estimating codon frequencies per RSA bin reduces the AIC score by nearly 36,000 in all model permutations (Table 3.3). This provides strong evidence that codon frequencies should be calculated individually for each RSA bin. Furthermore, models that place the κ parameter in the transition position now dominate the top scoring models in our analysis, which reinforces the conclusion that κ should be modeled for transitions.

Other changes of note between the genomic and individual codon frequency model types are the disappearance of a clear division between the purine and pyrimidine positions of T_{iN} in the top scoring models. Although models with a purine T_{iN} parameter still outperform their pyrimidine counterparts in general, superior performance of the purine models is no longer a universal trait. The top scoring models now place the T_{vN} parameter in the paired position, while all other parameters between the two top scoring models remain the same. Furthermore, the t and κ slopes for each model are consistent whether genomic or individual codon frequencies are used.

3.5 Mutation parameters have varying relationships with solvent accessibility across species

As the mutational mechanisms could vary across different species, we chose to analyze the fit of various parameter types to pairwise alignments of multiple species. To this end, we fitted our models to the previous data set of the 587 yeast sequence-structure pairs in addition to the following orthologous species sets: 329 sequence-structure pairs between *Mus musculus* and *Rattus norvegicus*; 763 orthologous coding sequences between *Drosophila melanogaster* and *Drosophila erecta*; and 425 *Caenorhabditis elegans* and *Caenorhabditis remanei* orthologs. All models were estimated using individual codon frequencies and parameter positions corresponding to the top scoring model from Table 3.3 (κ : transition, T_{iN} : purine, T_{vN} : paired, T_{vP} : GC, T_{vU} : AC).

Table 3.3: Fitted mutational model combinations for per-bin codon frequencies, in order of ascending AIC

κ	TiN	TvN	TvP	TvU	lnL	df	AIC	t slope	κ slope
transition	purine	paired	GC	AC	-821527.48	86	1643227	+	-
transition	purine	paired	AT	AC	-821527.48	86	1643227	+	-
transition	purine	unpaired	GC	GT	-821553.35	86	1643279	+	-
transition	purine	unpaired	GC	AC	-821553.35	86	1643279	+	-
transition	purine	unpaired	AT	AC	-821565.87	86	1643304	+	-
transition	purine	unpaired	AT	GT	-821565.87	86	1643304	+	-
transition	purine	paired	AT	GT	-821582.94	86	1643338	+	-
transition	purine	paired	GC	GT	-821582.94	86	1643338	+	-
transition	pyrimidine	paired	AT	GT	-821586.91	86	1643346	-	-
transition	pyrimidine	paired	GC	GT	-821586.91	86	1643346	-	-
transversion	pyrimidine	unpaired	GC	GT	-821592.88	86	1643358	-	+
transversion	pyrimidine	unpaired	GC	AC	-821592.88	86	1643358	-	+
transversion	pyrimidine	paired	GC	GT	-821593.23	86	1643358	-	+
transversion	pyrimidine	paired	AT	GT	-821593.223	86	1643358	-	+
transversion	pyrimidine	unpaired	AT	AC	-821596.93	86	1643366	-	+
transversion	pyrimidine	unpaired	AT	GT	-821596.93	86	1643366	-	+
transversion	pyrimidine	paired	GC	AC	-821597.93	86	1643368	-	+
transversion	pyrimidine	paired	AT	AC	-821597.93	86	1643368	-	+
transition	pyrimidine	unpaired	AT	GT	-821606.56	86	1643385	-	-
transition	pyrimidine	unpaired	AT	AC	-821606.56	86	1643385	-	-
transversion	purine	paired	AT	GT	-821608.94	86	1643390	+	-
transversion	purine	paired	GC	GT	-821608.94	86	1643390	+	-
transversion	purine	unpaired	GC	AC	-821609.23	86	1643390	+	+
transversion	purine	unpaired	GC	GT	-821609.23	86	1643390	+	+
transition	pyrimidine	unpaired	GC	AC	-821613.84	86	1643400	-	-
transition	pyrimidine	unpaired	GC	GT	-821613.84	86	1643400	-	-
transversion	purine	paired	AT	AC	-821620.93	86	1643414	+	+
transversion	purine	paired	GC	AC	-821620.93	86	1643414	+	+
transversion	purine	unpaired	AT	GT	-821623.18	86	1643418	+	-
transversion	purine	unpaired	AT	AC	-821623.18	86	1643418	+	-
transition	pyrimidine	paired	AT	AC	-821630.04	86	1643432	-	-
transition	pyrimidine	paired	GC	AC	-821630.04	86	1643432	-	-

The selection parameters ω and t were both modeled as a linear function of RSA.

For the five mutation parameters in the model, we considered three dif-

ferent functional forms of RSA dependence (per-bin, constant, and linear) with the exception of the parameter T_{iN} , for which we also considered a quadratic relationship. For all species (yeast, mouse, fly, and worm), we fit all of the 324 possible model combinations and ranked them by their AIC (Figure 3.3). The top scoring models varied considerably between species, with yeast showing the most sensitivity to model selection. In yeast, the top model differed by an AIC of 21,091 from the lowest scoring model, while all other species showed less of a discrepancy between the top- and bottom-scoring models (mouse: $\Delta AIC = 71$, fly: $\Delta AIC = 122$, worm: $\Delta AIC = 137$). In general, there was not much difference between the top-scoring models in mouse and worm as evaluated by AIC in spite of variation in parameters types (a change of $\Delta AIC = 2$ is considered a significant difference). In particular, it seems that only two parameters are necessary for describing the mouse nucleotide-level processes as most of the parameter types were constant. For yeast and fly, however, the nucleotide process seems very sensitive to model selection as AIC score increases very rapidly and there are very few models that require parameter constants. Model selection, therefore, is dependent on the species in question.

The per-bin permutation of the transition/transversion ratio κ consistently ranks in the top-scoring models across species. For the two model sensitive species, yeast and fly, the per-bin κ is a prerequisite to model fit and displays a decreasing trend with RSA ((Figure 3.4). This result is consistent with our previous analysis, where both t and κ varied with RSA with inverse

trends to one another. In mouse and worm, both the per-bin and constant models have equivalent AIC scores, implying that the κ is not dependent on solvent accessibility in these species.

Interestingly, the transition rate ratio, T_{iN} , displays a quadratic relationship with RSA in yeast, fly, and worm (Figure 3.5). The quadratic T_{iN} model is a universal trait of the top-scoring fly models and fit nearly perfectly in all species but mouse. Such a result indicates that purine mutations are much more prevalent than mutations between pyrimidines in the core of the protein, a trend that reverses for the middle solvent exposed residues to increases again for the outer most surface residues. In mouse, purine mutations occur only slightly more than pyrimidine mutations with no influence from solvent exposure.

The transversion rate ratio, T_{vN} , decreases linearly with increasing solvent exposure in yeast and fly (Figure 3.6). In these two species, exchanges between $A \rightleftharpoons T$ and $G \rightleftharpoons C$ occur much more frequently in the core of the protein compared to the surface residues. Although the results are less clear in mouse and worm, at no point in any of the species do any of the fall below 1, indicating that $A \rightleftharpoons T$ and $G \rightleftharpoons C$ transversions are more prevalent over $A \rightleftharpoons C$ and $G \rightleftharpoons T$ transversions in our data.

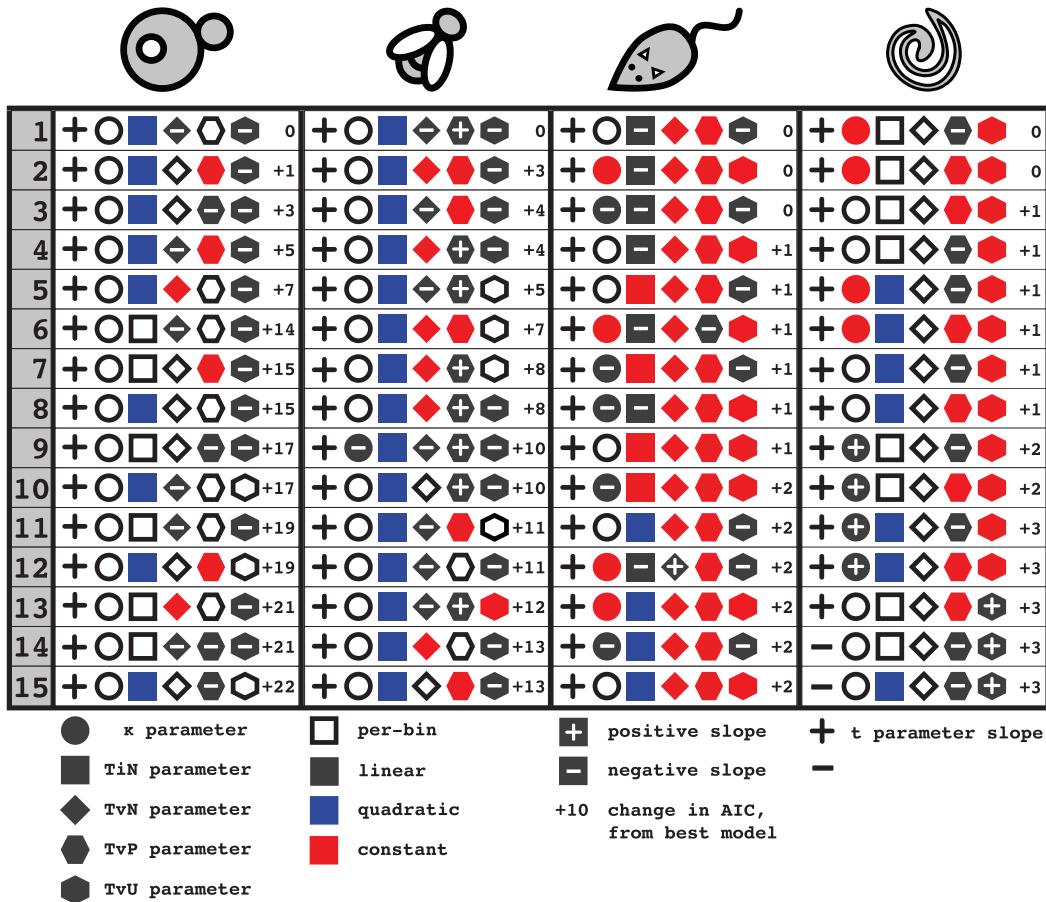


Figure 3.3: Fitted models across species, in order of AIC. All parameters have three possible relationships with RSA (per-bin, linear, and constant), with the exception of T_{iN} which has a fourth (quadratic). In total, there were 324 models using these parameter combinations that were estimated for yeast, mouse, fly, and worm. Here, the top 15 models are shown for each of the species and the ΔAIC from the best model is indicated. A model which uses per-bin κ and quadratic T_{iN} parameter fits best in yeast and fly, while there is not much difference between the top models for mouse and worm.

3.6 Discussion

We have extended the method developed in Chapter 2 to model DNA-level mutational processes within the context of a protein's 3-dimensional

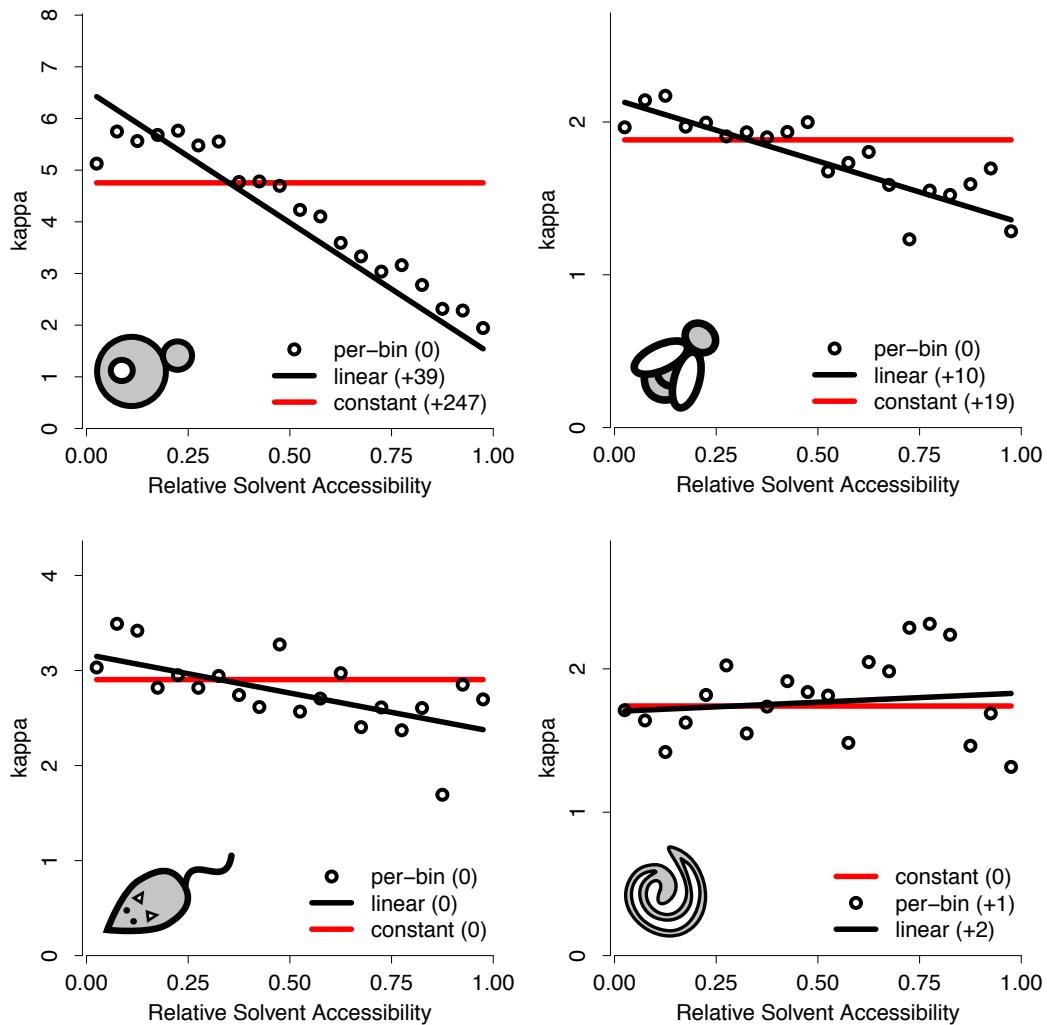


Figure 3.4: Dependency of varying κ parameters on RSA across species. In yeast and fly, the transition-transversion rate ratio κ decreases with solvent accessibility and should be estimated per-bin. For mouse and worm, κ shows almost no relationship with RSA and all models fit equally as well. The numbers in parentheses are the differences between the top scoring model of a parameter type with the overall top scoring model.

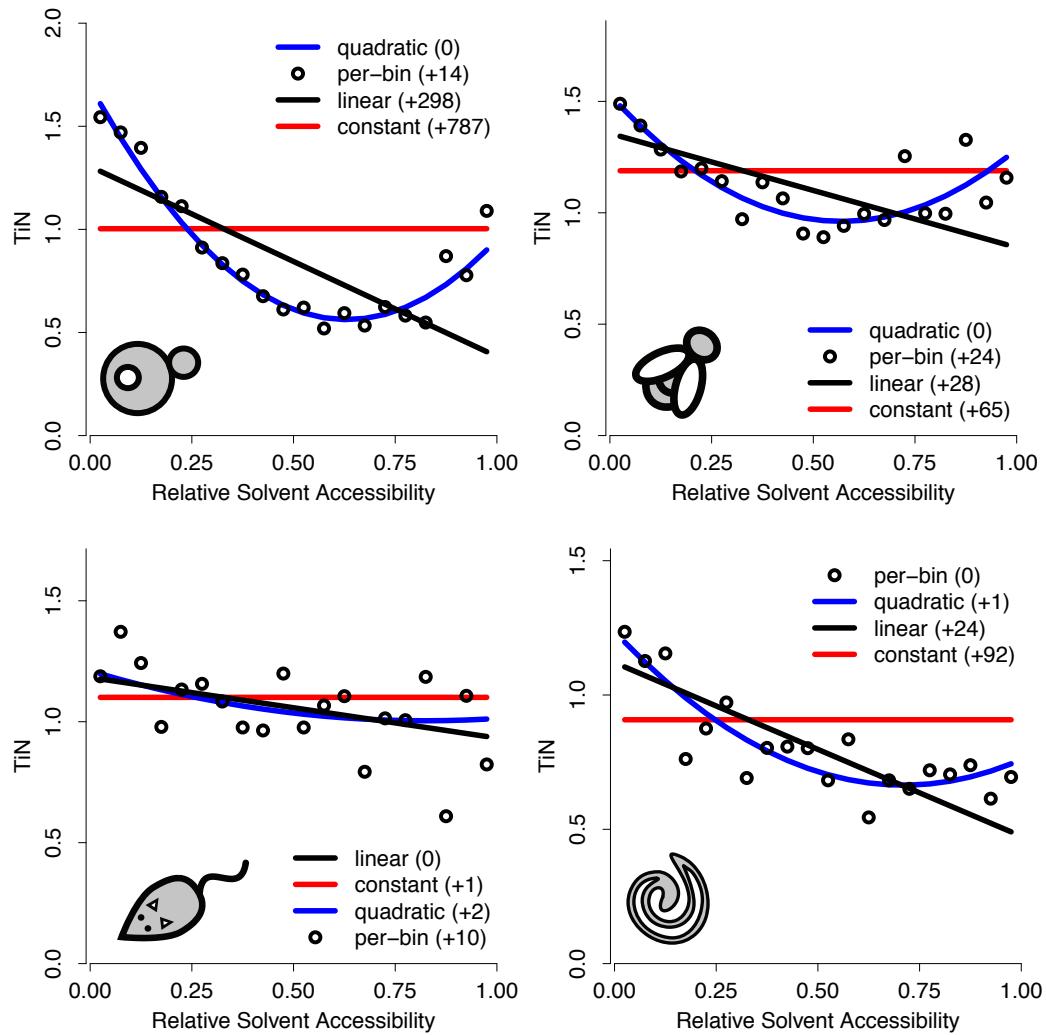


Figure 3.5: The effects of RSA on several transition parameter types, T_{iN} , across species. While the transition rate ratio, T_{iN} shows almost no relationship with solvent accessibility in mouse, both yeast and fly display a quadratic relationship with RSA. A quadratic relationship with solvent accessibility is also apparent in worm, although a per-bin model fits equally as well. The numbers in parentheses are the differences between the top scoring model of a parameter type with the overall top scoring model.

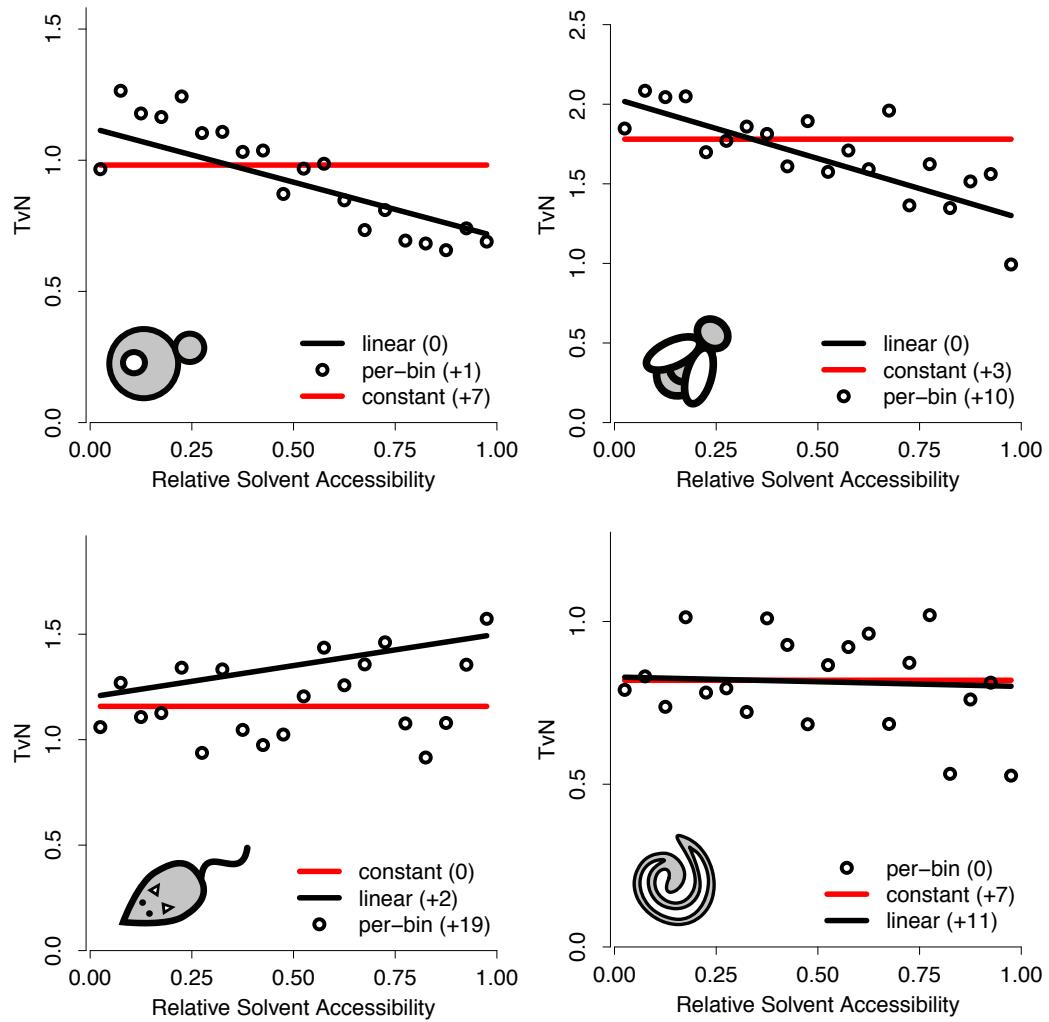


Figure 3.6: The relationship of transversion parameter types, T_{vN} , with RSA across species. For yeast and fly, the transversion rate ratio, T_{vN} , shows a negative linear relationship with solvent accessibility. RSA affects T_{vN} individually per-bin in worm, while there appears to be no varying relationship between transversion type and solvent accessibility in mouse. The numbers in parentheses are the differences between the top scoring model of a parameter type with the overall top scoring model.

structure. Our method modifies the General Time Reversible (GTR) model by introducing solvent dependent mutation parameters and provides for a more accurate representation of the relationship between protein structure and mutation bias. We have found that nucleotide-level mutational processes are dependent on relative solvent accessibility (RSA), although this dependency varies with the species in question. Thus, the selective pressure exerted by the need for protein stability extends down to DNA-level processes.

We found that the transition/transversion bias parameter κ scales negatively with RSA, although in general this relationship is best modeled individually per RSA bin. Models which place κ in the transition position provide a better fit when an RSA-dependency is introduced.

In addition, we found that the κ parameter has an RSA-dependent slope that is opposite of that found with the branch length, t . For models in which κ is located at the transition position, the slope of κ is negative and t is positive, while the inverse is found for transversions. The transition/transversion bias κ has been previously found to correlate negatively with sequence distance in primates [126], and such a result may be due to saturation of transitions with increasing divergence time.

Our findings show that the equilibrium codon frequencies, π_j , also vary with RSA as models with codon frequencies estimated per-bin fit significantly better than using pooled genomic codon frequencies. Such a finding is reflective of selection on codon choice by the protein structure. Previously, amino-acid equilibrium frequencies have been found to vary with RSA [87, 88], and such

variation should extend to the nucleotide level in equilibrium codon frequencies.

Our results are consistent with the idea that selection for codon usage is directly linked to the process of protein synthesis. Selection by the protein structure on synonymous sites has been repeatedly demonstrated and most likely occurs through and interaction between protein translation and folding [104, 60, 20, 58, 128, 132, 66]. Due to the degeneracy of the genetic code, all of the amino acids except for methionine and tryptophan are encoded by multiple codons, which can be divided into optimal and non-optimal categories. Selection for translational accuracy and efficiency has been observed across taxa and appears through bias in codon usage [51, 98, 100, 31, 101, 132]. Optimal codons are presumed to be translated much faster and more accurately than non-optimal codons [2, 32]. Conversely, non-optimal codons can be used to slow down protein synthesis.

The mistranslation-induced protein misfolding hypothesis postulates that selection occurs for optimal codons at sites where errors in translation are disruptive to the protein structure, leading to misfolding and aggregation that is detrimental to the cell [32]. Optimal codons are therefore typically associated with buried sites, where substitutions and translational errors are more likely to be destabilizing to the protein structure [132]. Furthermore, non-optimal codons have recently been found to be enriched in linker elements such as α -helices and hydrogen-bonded turns, which connect the more rigidly folding elements of a protein; such a result is seen to represent a balance be-

tween time needed for folding versus prevention of aggregation [83]. If enrichment of transitions in the core of the protein are seen to prevent destabilizing nonsynonymous change, it remains to be seen how mutation bias parameters associate with different structural elements within the protein itself.

Interestingly, we find that nucleotide level parameters have an RSA-dependence even when using a per-bin codon model that does not explicitly define such a dependence. Such a result indicates that selection at the nucleotide level is not solely due to codon choice and that equilibrium codon frequencies might not be fully descriptive in certain contexts.

In our model analysis between species, we determined that model choice is highly dependent on the species in question. In yeast and fly, the transition bias κ is much higher in the core residues compared to surface residues; however, this effect seems to disappear for mouse and worm, where model choices provide nearly equivalent results. We believe this result to be an effect of our choice in pairwise species comparisons. At high levels of divergence, saturation of transition mutations causes the appearance of an apparent equal transition to transversion ratio [13, 75, 23]. The relationship of κ with RSA in our analysis is consistent with divergence times between our pairwise species, as κ becomes almost constant with increasing divergence times.

Divergence time also seems to have an effect on the other mutation parameters in our model, with the relationships of both T_{iN} and T_{vN} with RSA becoming less pronounced with increasing divergence time. Surprisingly, T_{iN} has a quadratic relationship with RSA, indicating that purine to purine tran-

sitions occur much more frequently in the core of the protein, increasing again somewhat for surface residues. While there is no solid mechanistic explanation for such a trend, the quadratic relationship all but disappears in mouse and it is unclear whether this trend will apply universally to mammals.

3.7 Methods

All analyses were conducted in the same manner as those presented in Chapter 2. The additional data sets were obtained from the following locations: the 329 sequence-structure pairs between *Mus musculus* and *Rattus norvegicus* were downloaded from Ensembl [38]; the 763 orthologous coding sequences between *Drosophila melanogaster* and *Drosophila erecta* were obtained from the Eisen Lab (<http://rana.lbl.gov/drosophila/documents.html>); and the 425 *Caenorhabditis elegans* and *Caenorhabditis remanei* orthologs were acquired from WormBase [45].

Chapter 4

Assessing the correlates of coding-sequence evolution in individual genes

4.1 Background

Proteins evolve at vastly different rates even within the same species [44, 71]. While some proteins are conserved even across great periods of divergence (e.g., eukaryotic ribosomal and histone proteins), others are hardly recognizable as orthologs even in closely related species [55, 78]. Intraspecific rate variation of proteins has been found regardless of the species under study and the question then becomes, what are the driving forces behind such rate variation?

With the availability of reliable genomic and structural measurements in many different species, molecular biologists are now able to assess the mechanisms that shape coding sequences over evolutionary time. Previous

studies focused on assessing the correlation of a singular genetic feature with the evolutionary rate of a protein. For example, the evolutionary rate of a protein has been found to correlate with expression level in yeast and vertebrates through translational selection for optimal codon usage [82, 3, 31, 102]. The number of tissues in which a gene is expressed is also a correlate, as broadly expressed housekeeping genes tend to evolve more slowly than tissue specific genes [46, 34, 130, 116]. The composition of the protein structure itself plays a role in sequence conservation, as structures with a greater fraction of buried sites have the benefit of greater stability and therefore can evolve more rapidly [10, 11, 68, 131, 39]. Finally, synonymous substitution rates tie together mutation and selection, with genes evolving slowly through substitution bias at the third codon position, which in turn is correlated with GC content [97, 37, 49, 118].

However, disentangling the relative contribution of these effects has been difficult as many of these correlates tend to also correlate with one another. Here, we focus on the properties of individual genes and the mechanisms behind their evolutionary rates within the context of the structural constraint imposed by the solvent accessibility of their protein. We identified features which constrain sequence evolution at the structural, translational, and nucleotide levels and assess their effects on evolutionary rate using a multivariate statistical framework. We find that the expression level of a gene and the core size of its protein impose additional constraints beyond that imposed by solvent accessibility.

4.2 Determinates of average evolutionary rate across individual genes

To determine which properties predict the average evolutionary rate of a gene, we constructed a data set of 284 coding sequence alignments of mammalian orthologs curated from the Ensembl database. Each alignment ranges from 10 to 36 representative mammalian species to provide sufficient coverage to detect differences between the shorter individual gene sequences. We calculated each of the rate parameters (ω , t , and κ) as an average across the entire alignment using the standard GY94 model without an RSA dependency. We refer to these parameter types as ω_A , t_A , and κ_A .

We collected several attributes that have been known to correlate significantly with evolutionary rate in the literature and performed a multivariate regression analysis to determine the dependency of the protein selection parameter ω on each of the attributes (Table 4.1). Here, we chose expression level (measured in mRNA abundance) and expression breadth (the number of human tissues in which the transcript is present) as measures of protein abundance. The length of the transcript as well as its percent GC content were chosen as attributes that represent the coding sequence, while the average relative solvent accessibility (RSA) of the protein was chosen to represent the core size of the protein. The number of orthologs contained in each alignment is also present in our linear model. Thus, our model contains attributes that affect evolutionary rate from a structural, translational, and nucleotide-level

perspective.

Table 4.1: Gene attributes and their definitions

Attribute	Definition
Expression level	The expression level measured in mRNA abundance.
Expression breadth	The number of tissues in which the transcript is expressed.
Transcript length	The length of the aligned transcripts in base pairs.
Average RSA	The average RSA for all residues in the protein structure.
% GC Content	The percent GC content in the aligned transcripts.
# of Orthologs	The number of orthologs in the alignment.

We find that expression level, transcript length, percent GC content, and the number of orthologs are all significant predictors of the selection acting on the protein. Furthermore, our model accounts for 29% of the variation in evolutionary rate between the mammalian genes. As the expression level of a gene increases, ω_A decreases ($\beta_1^{\text{Exp}} = -0.031$); thus, we find that highly expressed genes evolve more slowly. We also find that the percent GC content of a coding sequence scales negatively with its evolutionary rate ($\beta_1^{\text{GC}} = -0.006$), and thus genes with higher GC content will evolve slowly. Our model also shows that longer genes will also evolve more slowly than shorter genes ($\beta_1^{\text{TLen}} = -4.693 \times 10^{-7}$), while expression breadth and average RSA were not significant predictors of evolutionary rate (Table 4.2).

As the number of orthologs was a significant predictor of evolutionary rate in our model, we sought to determine if our results were an artifact of the number of representative species in our alignments. We therefore repeated our analysis for only those genes which had more than 30 mammalian species represented in the alignment. In this case, the number of orthologs is no longer

Table 4.2: Multivariate regression of gene attributes predicting average evolutionary rate, ω_A

Attribute	Estimate	Standard Error	P-value
Expression level	-0.031	0.007	5.410×10^{-6}
Expression breadth	-5.930×10^{-4}	5.965×10^{-4}	0.320
Transcript length	-4.693×10^{-7}	1.540×10^{-7}	0.003
Average RSA	-0.142	0.179	0.429
% GC Content	-0.006	0.001	3.260×10^{-6}
# of Orthologs	-0.008	0.002	1.410×10^{-6}

significant in the regression model while expression level, transcript length, and percent GC content all remain significant predictors of evolutionary rate (data not shown). Thus, the above attributes affect evolutionary rate independently of the number of orthologs used in the alignments.

4.3 Expression level and core size both impose an additional selective constraint beyond solvent accessibility

In the previous chapters, we estimated the ω parameter as a linear function of relative solvent accessibility (RSA), which yields both a slope and an intercept instead of an individual value. The intercept explains to what extent completely buried residues are conserved, while the slope shows us by how much selection is relaxed as we move toward the surface residues of the protein. Thus, the slope parameter, ω_1 , is a measure of structural constraint acting on the protein. Here, we ask whether there are different attributes that affect genes that have a significant ω_1 compared to those that do not.

We fitted our RSA-dependent rate models to a data set consisting of our 284 mammalian orthologs that were paired with a known structure in the Protein Data Bank (PDB). We calculated the RSA for each site in the proteins as described in Chapter 2. Our binning procedure used $n = 10$ evenly-spaced RSA bins in contrast to our $n = 20$ bins in the previous chapters to accommodate for the shorter individual sequences.

The two types of models we fitted for each alignment both used a parameter shared across RSA bins for κ and t , while ω was fitted as a linear function of RSA. However, in the first model we allowed ω_1 to vary with RSA while in the second one, the slope was held constant ($\omega_1 = 0$) across the RSA bins. We then performed a likelihood ratio test of the model fits to each gene and corrected the resulting P-values using the False Discovery Rate (FDR) control. We found that of the 284 orthologous coding sequences in our analysis, 160 had a significant ω_1 compared to the other 124 genes which did not.

Using these two slope and non-slope categories, we conducted a logistic regression to determine which of our gene attributes could differentiate between the two classes of genes. We find that the expression level and the average RSA of a gene both are significant predictors of whether or not a gene demonstrates a relationship of ω_1 with RSA, while all other gene attributes are not significant (Table 4.3). As the expression level of a gene increases, it is less likely to have a significant ω_1 ($\beta_1^{\text{Exp}} = -0.286$); thus, for highly expressed genes, expression level introduces an additional constraint over that imposed

by solvent accessibility.

Table 4.3: Logistic regression of gene attributes predicting the presence of an ω slope

Attribute	Estimate	Standard Error	P-value
Expression level	-0.287	0.108	0.008
Expression breadth	-0.006	0.009	0.483
Transcript length	-3.013×10^{-6}	2.410×10^{-6}	0.211
Average RSA	-8.774	2.984	0.003
% GC Content	-0.018	0.021	0.394
# of Orthologs	0.032	0.025	0.202

The average evolutionary rate, ω_A , correlates strongly and positively with the measure of structural constraint, ω_1 ($r = 0.554$, $P = 2.2 \times 10^{-16}$). Thus, as the average evolutionary rate of a gene increases, the constraint across the residues from the core to the surface becomes more relaxed (Figure 4.1).

A similar trend is noted for the average RSA of a gene, with the probability of having a significant ω_1 decreasing with increasing average RSA ($\beta_1^{\text{avgRSA}} = -8.843$). Proteins that have a low average RSA are described as having a large core while small core proteins have a high average RSA; therefore, large core proteins will have a greater evolutionary rate differential between core and surface residues than proteins with a small core.

4.4 Expression level and GC content are both predictors of structural constraint

To determine how other evolutionary effects might affect the measure of overall structural constraint, ω_1 , we selected the 160 sequence-structure pairs

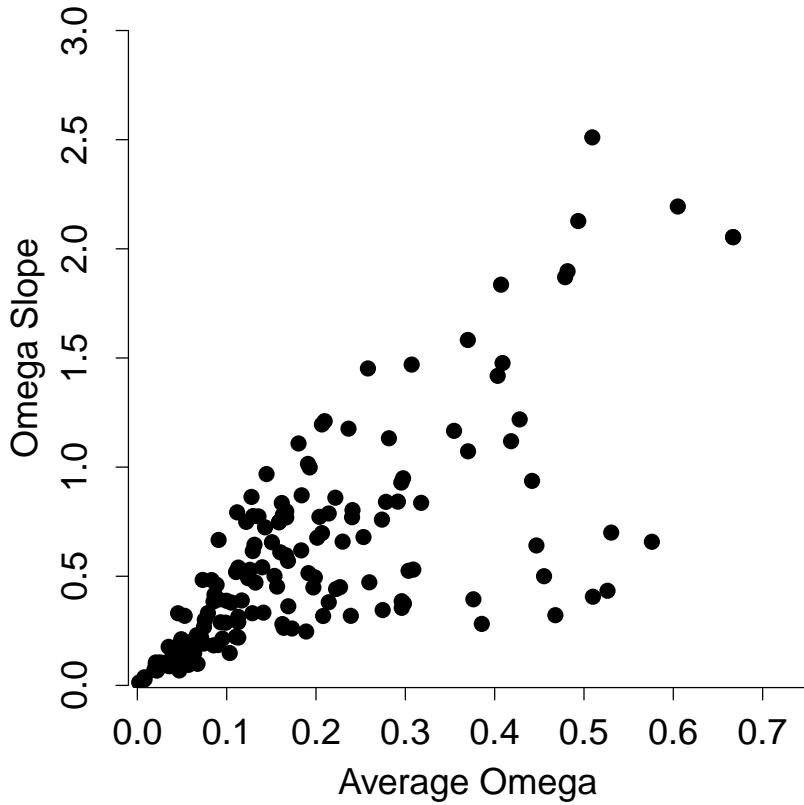


Figure 4.1: The average evolutionary rate, ω_A , vs. the structural constraint slope, ω_1 . Here, ω_A is plotted against ω_1 for the 160 genes that displayed a significant ω_1 . The constraint imposed across solvent accessibilities is relaxed with increasing average evolutionary rate of a gene.

with a significant ω_1 for further analysis. Here, we conducted a multivariate regression analysis using our selected gene attributes as explanatory variables for ω_1 . We find that expression level and GC content are both significant predictors of structural constraint (Table 4.4). Our linear model accounts for 13% of the variation in ω_1 .

Table 4.4: Multivariate regression of gene attributes predicting structural constraint, ω_1

Attribute	Estimate	Standard Error	P-value
Expression level	-0.094	0.042	0.028
Expression breadth	-0.001	0.003	0.723
Transcript length	-1.636×10^{-6}	9.930×10^{-7}	0.102
Average RSA	-0.440	1.015	0.665
% GC Content	-0.014	0.007	0.032
# of Orthologs	-0.017	0.008	0.039

For expression level, the surface residues in the protein become more conserved as the protein becomes more highly expressed ($\beta_1^{\text{Exp}} = -0.094$). For genes that are highly expressed, variation in ω_1 is far reduced compared to genes that are less abundant. Thus, we find that the surface residues of highly expressed proteins are less variable than their lowly expressed counterparts (Figure 4.2A).

We find a similar trend for percent GC content, with structural constraint across the protein increasing as GC content increases ($\beta_1^{\text{GC}} = -0.014$). For genes with a very high %GC content in their transcripts, ω_1 becomes drastically reduced and the surface residues of the protein become more constrained (Figure 4.2B).

4.5 Discussion

We have applied a statistical framework that models the evolutionary rates of individual genes in terms of the structural constraint imposed by the relative solvent accessibility (RSA) of their protein structures. By control-

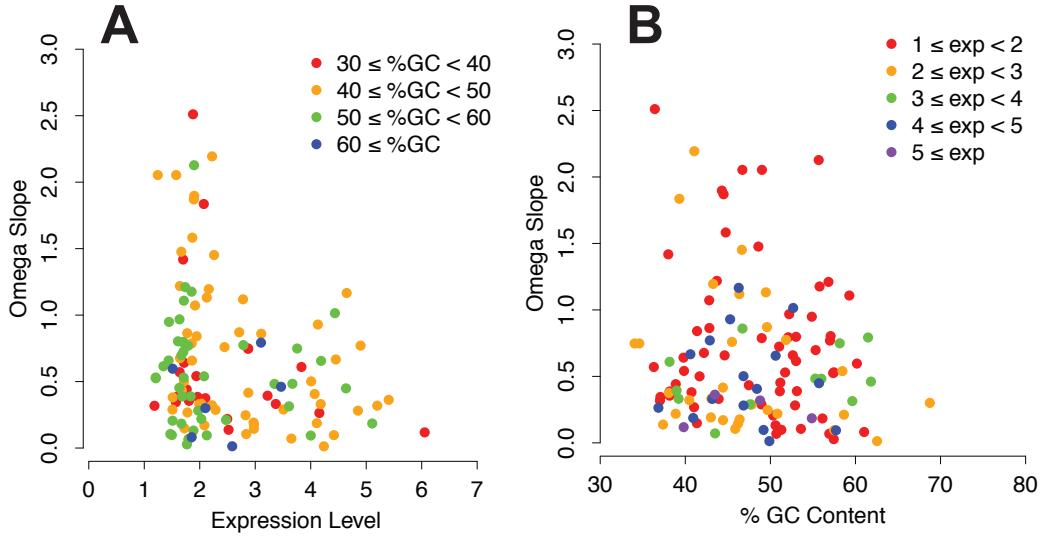


Figure 4.2: The effects of expression level and %GC content on the structural constraint parameter, ω_1 . Both expression level and %GC content are significant predictors of ω_1 in our multivariate regression model. **(A)** Lowly expressed genes have the most variation in ω_1 , while highly expressed genes have very little variation in ω_1 . This shows that as expression level increases, so does the constraint across the entire protein structure. **(B)** The 4 categories of %GC content are shown with their ω_1 . As %GC content increases, the variation in ω_1 decreases, indicating that high GC content imposes constraint on the surface residues of a protein.

ling for structural effects and using multivariate statistical methods, we are able to refine the contribution of various genetic properties in determining the evolutionary rate of mammalian genes.

The results of the multivariate statistical methods are summarized in Table ???. We have conducted a multivariate regression of gene attributes in predicting the average evolutionary rate (vs. ω_A), a logistic regression in predicting the presence of an ω slope (ω_1 or $\omega_1 = 0$), and a multivariate regression in predicting structural constraint (ω_1). Here, significant gene attributes

in predicting the types of ω rates are indicated by asterisks (*), with an increasing number of asterisks indicating greater significance. Non-significant gene attributes are indicated with a dash (-).

Table 4.5: Summary of gene attributes and their significance in predicting ω rates

Attribute	vs. ω_A	ω_1 or $\omega_1 = 0$	vs. ω_1
Expression level	***	**	*
Expression breadth	-	-	-
Transcript length	**	-	-
Average RSA	-	**	-
% GC Content	***	-	*
# of Orthologs	***	-	*

We found that the average evolutionary rate of a gene, ω_A , is significantly predicted by expression level, transcript length, and GC content. These results are broadly consistent with the findings in previous literature. Highly expressed genes have been repeatedly found to evolve slowly, even when paralogous genes with similar structure and function are examined [82, 3, 30, 31, 102]. Expression level is one of the most dominant correlates of evolutionary rate and may explain up to 30% of the total rate variation in yeast [30].

We also find that the evolutionary rate of a coding sequence decreases with increasing percent GC content. GC content relates to evolutionary rate through synonymous codon choice, specifically at the third codon position [76, 8] and while both codon bias and GC content are negatively correlated with evolutionary rate, they are both positively correlated with one another [119].

Both expression level and GC bias are associated with the maintenance of translational speed and accuracy to prevent misfolding and aggregation of the protein during synthesis [32].

Surprisingly, we found that genes with longer transcripts evolve more slowly than shorter genes. Transcript length corresponds to the final size of the folded protein and selection for shorter gene lengths should be expected from previous studies [10, 112]. However, length has also been shown to inversely relate with expression level [112] and might be an interacting term in our model. Such a result is a caveat for controlling individual factors that might relate and presents a challenge to even advanced statistical techniques.

When comparing genes with a significant omega slope, ω_1 , we find that the two factors that differentiate between these groups are the expression level and average RSA of a gene. The term ω_1 is used as a measure of the change in evolutionary rate between residues in the core of the protein and those on the surface, and therefore is a measure of structural constraint imposed by solvent accessibility. We have shown in Chapter 2 that there is a difference in ω_1 between both highly expressed genes and lowly expressed genes, as well as between large core proteins (low average RSA) and small core proteins (high average RSA). This result is consistent with our findings in Chapter 2, where highly expressed genes and small core proteins evolve much faster than their lowly expressed and large core counterparts. Core size and expression level impose an additional constraint beyond just solvent accessibility, and in both highly expressed genes and small core proteins, surface residues are evolving

under nearly as much constraint as those imposed in the core of the protein.

We next asked how the slope of the relationship of evolutionary rate across the solvent accessibility of the protein is affected by the predictor variables in our analysis. We find that both expression level and percent GC content are significant predictors of ω_1 , with the overall structural constraint across solvent accessibilities becoming reduced with increasing expression level and %GC content. As both of these properties are directly linked to the translational process, it appears that selection for translational speed and accuracy might be extending its effects to the protein structure itself. Thus, while surface residues tend to be more variable than residues that are more buried in the protein structure, translational selection is limiting their variability beyond what would be expected from RSA alone.

While our method here presents a powerful technique for disentangling the relative effects of evolutionary rate correlates, our models only explain about a tenth of the variation in evolutionary rate while controlling for relative solvent accessibility. Such techniques as presented here might be sensitive to noise between the genetic features under study as well as incomplete data coverage.

4.6 Methods

In order to construct a data set of individual genes with a large number of representative species, we downloaded the open reading frames (ORFs) of

all 36 mammalian species in the Ensembl database [38] and created sequence-structure pairs using *Homo sapiens* as the master ortholog with the methods described in Chapter 2. This resulted in a data set of 284 individual mammalian genes, with orthologs of *Homo sapiens* ranging from 15 to 35 representative species.

Expression level in mRNA abundance and expression breadth in number of tissues were acquired for *Homo sapiens* from Ensembl’s BioMart. The average relative solvent accessibility (RSA) was calculated using DSSP as in Chapter 2 and averaged over the entire protein structure. Transcript length and percent GC content were calculated directly from the *Homo sapiens* ORF in each gene alignment. The number of orthologs is the number of species represented in each alignment.

Multivariate and logistic regressions were conducted using the statistical software package R [50].

Chapter 5

Conclusion

5.1 Contribution

While the sequences of proteins within a species change at dramatically different rates, elucidating the mechanisms that drive this rate variation have been a paramount challenge to the field of molecular biology. Of these mechanisms, the biophysical properties of the proteins themselves have been found to play a large role in determining sequence conservation. Specifically, the degree of burial of a residue in a protein structure measured by its relative solvent accessibility (RSA) is a dominant factor of sequence constraint. In this work, we investigated how the rate of coding sequence evolution is driven by both structural and mutational level processes using a framework that considers RSA of a protein’s residues. Our contributions to the field in this regard are as follows.

5.1.1 Consideration of protein structure provides for better evolutionary rate estimates

In the second chapter, we asked whether consideration the structure of a protein could more accurately describe its evolutionary rate. To this end, we developed a model of sequence evolution that introduces an explicit structural term representing the exposure of each residue in a protein to the solvent. Our model is a variant of the GY94 codon model in which each parameter can be defined as a function of the RSA of a residue. Using a data set that models the genomic trends in yeast, we find that the evolutionary rate parameter ω fits better as a linear function of RSA than if it were estimated independently of solvent accessibility. We also show that the branch length t and transition-transversion ratio κ also vary with RSA. Thus, we show that the evolutionary rate of a protein is better characterized within a structural paradigm rather than just as an average over a sequence.

5.1.2 Nucleotide-level processes are determinants of protein evolution

In the third chapter, we asked whether nucleotide-level mutational biases are affected by the structure of a protein. To answer this question, we introduced model parameters that accurately model changes occurring at the nucleotide-level as a function of a residue's RSA. Using data sets that reflect the genomic coding sequences of yeast, fly, mouse, and worm, we find that the transition-transversion ratio κ changes with solvent accessibility in a manner

that is highly dependent on the divergence time of the species under study. We also show that mutational biases are important in their effects on protein structure, but the precise mechanisms by which they act require further study. Thus, we show that nucleotide-level processes are an essential determinant of the amino acid composition of a protein in terms of its structure.

5.1.3 Individual determinants of evolutionary rate are revealed with a structural paradigm

Finally, in the fourth chapter, we ask whether the linear relationship between ω and RSA can be used to determine the effects of other genetic characteristics on the evolutionary rate of a protein. To this aim, we applied our structural paradigm to individual mammalian genes with numerous orthologous sequences. Using a curated set of evolutionary rate correlates, we show that the core size and expression level of a protein impose an additional constraint beyond that of residue burial. Furthermore, we show that expression level and nucleotide composition in terms of GC content will reduce the evolutionary rate of a protein overall. Thus, we show that a structural paradigm is an important tool for elucidating additional factors that constrain sequence evolution.

5.2 Future

While the role of protein structure in determining its evolutionary rate has been demonstrated throughout this work, much remains to be discovered

in the future.

Our work is limited by the number of solved protein structures available that can be incorporated with the primary sequence data. Techniques that marry sequence and structure data in future evolutionary rate analyses will flourish as advances in structural determination reveal more about the protein landscape.

Protein structures themselves are not static entities and participate in various functions within a living organism. Eventual approaches that integrate structural information with biological function will paint a more vivid portrait of the mechanisms that animate and construct the cellular infrastructure.

Finally, the results we present here are theoretical extrapolations taken from the anecdotal narratives of a relatively small set of model organisms. In the expedition for the primal universals of coding sequence evolution, empirical experimentation —both *in vitro* and *in silico*—must plant the flag finally upon the protein precipice.

Bibliography

- [1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:6:716–723, 1974.
- [2] H. Akashi. Synonymous codon usage in *Drosophila melanogaster*: natural selection and accuracy. *Genetics*, 136:927–935, 1994.
- [3] H. Akashi. Translational selection and yeast proteome evolution. *Genetics*, 164:1291–1303, 2003.
- [4] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:17:3389–3402, 1997.
- [5] M. Anisimova and C. Kosiol. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution*, 26:255–271, 2009.
- [6] M. Bajaj and T. Blundell. Evolution and the tertiary structure of pro-

teins. *Annual Review of Biophysics and Bioengineering*, 13:453–492, 1984.

- [7] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Wessig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:1:235–242, 2000.
- [8] G. Bernardi. Isochores and the evolutionary genomics of vertebrates. *Sociological Methods and Research*, 241:3–17, 2000.
- [9] N. Bierne and A. Eyre-Walker. The Problem of Counting sites in the Estimation of the Synonymous and Nonsynonymous Substitution Rates: Implications for the Correlation Between the Synonymous Substitution Rate and Codon Usage Bias. *Genetics*, 165:1587–1597, 2003.
- [10] J.D. Bloom, D.A. Drummond, F.H. Arnold, and C.O. Wilke. Structural determinants of the rate of protein evolution in yeast. *Molecular Biology and Evolution*, 23:1751–1761, 2006.
- [11] J.D. Bloom, S.T. Labthavikul, C.R. Otey, and F.H. Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences USA*, 103:5869–5874, 2006.
- [12] L. Bofkin and N. Goldman. Variation in evolutionary processes at different codon positions. *Molecular Biology and Evolution*, 24:513–521, 2007.

- [13] W.M. Brown, E.M. Prager, A. Wang, and A.C. Wilson. Mitochondrial DNA sequences of primates, tempo and mode of evolution. *Journal of Molecular Evolution*, 18:225–239, 1982.
- [14] K.P. Burnham and D.R. Anderson. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33:261–304, 2004.
- [15] C.D. Bustamante, J.P. Townsend, and D.L. Hartl. Solvent Accessibility and Purifying Selection Within Proteins of *Escherichia coli* and *Salmonella enterica*. *Molecular Biology and Evolution*, 17:2:301–308, 2000.
- [16] J.M. Cherry, C. Alder, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein. SGD: Saccharomyces Genome Database. *Nucleic Acids Research*, 26:1:73–79, 1998.
- [17] C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *European Molecular Biology Organization*, 5:823–826, 1986.
- [18] G.C. Conant and P.F. Stadler. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular Biology and Evolution*, 26:1155–1161, 2009.

- [19] G.C. Conant, G.P. Wagner, and P.F. Stadler. Modeling amino acid substitution patterns in orthologous and paralogous genes. *Molecular Phylogenetics and Evolution*, 42:298–307, 2007.
- [20] P. Cortazzo, C. Cervenansky, M. Marin, C. Reiss, R. Ehrlich, and A. Deana. Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochemical and Biophysical Research Communications*, 293:537–541, 2002.
- [21] D.R. Cox and H.D. Miller. *The theory of stochastic processes*. Chapman and Hall, 1977.
- [22] T.E. Creighton. *Proteins: structures and molecular properties*. Freeman, 1992.
- [23] R.H. Crozier and Y.C. Crozier. The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics*, 133:97–117, 1993.
- [24] M.O. Dayhoff, E.V. Eck, and C.M. Park. A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5:89–99, 1972.
- [25] A.M. Dean, C. Neuhauser, E. Grenier, and G.B. Golding. The pattern of amino acid replacements in α/β -barrels. *Molecular Biology and Evolution*, 19:1846–1864, 2002.

- [26] E.J. Deeds, O. Ashenberg, J. Gerardin, and E.I. Shakhnovich. Robust protein-protein interactions in crowded cellular environments. *Proceedings of the National Academy of Sciences USA*, 104:14952–14957, 2007.
- [27] W. Delport, K. Scheffler, G. Botha, M.B. Gravenor, S.V. Muse, and S. Kosakovsky Pond. CodonTest: modeling amino acid substitution preferences in coding sequences. *PLoS Computational Biology*, 6:e1000885, 2010.
- [28] W. Delport, K. Scheffler, M.B. Gravenor, S.V. Muse, and S. Kosakovsky Pond. Benchmarking multi-rate codon models. *PLoS One*, 5:e11157, 2010.
- [29] N.V. Dokholyan and E.I. Shakhnovich. Understanding hierarchical protein evolution from first principles. *Journal of Molecular Biology*, 23:327–337, 2001.
- [30] D.A. Drummond, J.D. Bloom, C. Adami, and C.O. Wilke. Why highly expressed genes evolve slowly. *Proceedings of the National Academy of Sciences USA*, 102:14338–14343, 2005.
- [31] D.A. Drummond, A. Raval, and C.O. Wilke. A single determinant dominates the rate of protein evolution. *Molecular Biology and Evolution*, 312:289–307, 2006.
- [32] D.A. Drummond and C.O. Wilke. Mistranslation-induced protein mis-

folding as a dominant constraint on coding-sequence evolution. *Cell*, 134:341–352, 2008.

- [33] D.A. Drummond and C.O. Wilke. The evolutionary consequences of erroneous protein synthesis. *Nature Review Genetics*, 10:715–724, 2009.
- [34] L. Duret and D. Mouchiroud. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution*, 17:68–74, 2000.
- [35] M. Eames and T. Kortemme. Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Structure*, 14:1442–1451, 2007.
- [36] R.C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:5:1792–1797, 2004.
- [37] A. Erye-Walker and M. Bulmer. Synonymous substitution rates in enterobacteria. *Genetics*, 140:1407–1412, 1995.
- [38] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Khri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. Singh Riat, G.R.S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y.A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova,

S. White, M. Wilson, S.P. Wilder, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernndez-Surez, J. Harrow, J. Herrero, T.J.P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S.M.J. Searle. Ensembl 2012. *Nucleic Acids Research*, 40:D84–D90, 2012.

- [39] E.A. Franzosa and Y. Xia. Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular Biology and Evolution*, 26:10:2387–2395, 2009.
- [40] T. Gojobori, W-H. Li, and D. Graur. Patterns of nucleotide substitution in pseudogenes and function genes. *Journal of Molecular Evolution*, 18:360–376, 1982.
- [41] N. Goldman, J.L. Thorne, and D.T. Jones. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445–458, 1998.
- [42] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11:5:725–736, 1994.
- [43] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185:862–864, 1974.
- [44] N.V. Grishin, Y.I. Wolf, and E.V. Koonin. From complete genomes to

measures of substitution rate variability within and between proteins.

Genome Research, 10:991–1000, 2000.

- [45] T.W. Harris, I. Antoshechkin, T. Bieri, D. Blasiar, J. Chan, W.J. Chen, N. De La Cruz, P. Davis, M. Duesbury, R. Fang, J. Fernandes, M. Han, R. Kishore, R. Lee, H. Mller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E.M. Schwarz, M.A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, K. Yook, R. Durbin, L.D. Stein, J. Spieth, and P.W. Sternberg. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*, 38:D463–D467, 2010.
- [46] K.E.M. Hastings. Strong evolutionary conservation of broadly expressed protein isoforms in the Troponin I gene family and other vertebrate gene families. *Journal of Molecular Evolution*, 42:631–640, 1996.
- [47] L.W. Hillier, R.D. Miller, S.E. Baird, A. Chinwalla, L.A. Fulton, D.C. Koboldt, and R.H. Waterston. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biology*, 5, 2007.
- [48] F.C.P. Holstege, E.G. Jennings, J.J. Wyrick, T.I. Lee, C.J. Hengartner, M.R. Green, T.R. Golub, E.S. Lander, and R.A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95:717–728, 1998.
- [49] L.D. Hurst and E.J.B. Williams. Covariation of GC content and the

silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene*, 261:107–114, 2000.

- [50] R. Ihaka and R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.
- [51] T. Ikemura. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology*, 151:389–409, 1981.
- [52] D. Jones, W. Taylor, and J. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, 8:275–281, 1992.
- [53] T.H. Jukes and J.L. King. Evolutionary nucleotide replacements in DNA. *Nature*, 281:605–606, 1979.
- [54] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:12:2577–2637, 1983.
- [55] M. Kellis, B.W. Birren, and E.S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428:617–624, 2004.

- [56] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–254, 2003.
- [57] P.M. Kim and L.J. Lu and M.B. Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314:1882–1883, 2006.
- [58] C. Kimchy-Sarfaty, J.M. Oh, I.W. Kim, Z.E. Sauna, A.M Calcagno, S.V. Ambudkar, and M.M. Gottesman. A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science*, 315:525–528, 2007.
- [59] M. Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267:275–276, 1977.
- [60] A.A. Komar, T. Lesnik, and C. Reiss. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Letters*, 462:387–391, 1999.
- [61] J.M. Koshi and R.A. Goldstein. Context-dependent optimal substitution matrices. *Protein Engineering Design and Selection*, 8:641–645, 1995.
- [62] J.M. Koshi and R.A. Goldstein. Models of natural mutations including site heterogeneity. *Proteins*, 32:289–295, 1998.
- [63] E.B. Kramer and P.J. Farabaugh. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA*, 13:87–96, 2007.

- [64] D.M. Krylov, Y.I. Wolf, I.B. Rogozin, and E.V. Koonin. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are all correlated in eukaryotic evolution. *Genome Research*, 13:2229–2235, 2003.
- [65] N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21:1095–1109, 2004.
- [66] Y. Lee, T. Zhou, G.G. Tartaglia, M. Vendruscolo, and C.O. Wilke. Translationally optimal codons associate with aggregation-prone sites in proteins. *Proteomics*, 10:4163–4171, 2010.
- [67] B. Lemos, B.R. Bettencourt, C.D. Meiklejohn, and D.L. Hartl. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Molecular Biology and Evolution*, 22:1345–1354, 2005.
- [68] Y.S. Lin, W.L. Hus, J.K. Hwang, and W.H. Li. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Molecular Biology and Evolution*, 24:1005–1011, 2005.
- [69] M.Z. Ludwig, A. Palsson, E. Alekseeva, C.M. Bergman, J. Nathan, and M. Kreitman. Functional Evolution of a cis-Regulatory Module. *PLoS Biology*, 3, 2005.

- [70] L. Marsh and C.S. Griffiths. Protein structural influences in rhodopsin evolution. *Molecular Biology and Evolution*, 22:894–904, 2005.
- [71] J.O. McInerney. The causes of protein evolutionary rate variation. *Trends in Ecology and Evolution*, 21:230–232, 2006.
- [72] A.G. Meyer and C.O. Wilke. Integrating sequence variation and protein structure to identify sites under selection. *Molecular Biology and Evolution*, 2013.
- [73] J. Mintseris and Z. Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences USA*, 102:10930–10935, 2005.
- [74] L.A. Mirny and E.I. Shakhnovich. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of Molecular Biology*, 291:177–196, 1999.
- [75] C. Moritz, T.E. Dowling, and W.M. Brown. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology, Evolution, and Systematics*, 18:269–292, 1987.
- [76] S.V. Muse and B.S. Gaut. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences*, 84:166–169, 1987.
- [77] S.V. Muse and B.S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application

to the chloroplast genome. *Molecular Biology and Evolution*, 11:715–724, 1994.

- [78] J.H. Nadeau and D. Sankoff. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 147:1259–1266, 1997.
- [79] G.J.P. Naylor, T.M. Collins, and W.M. Brown. Hydrophobicity and phylogeny. *Nature*, 373:565–566, 1995.
- [80] R. Nielsen and Z. Yang. Likelihood models for detecting positive selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–936, 1998.
- [81] J. Overington, D. Donnelly, M.S. Johnson, A. Sali, and T.L. Blundell. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Science*, 1:216–226, 1992.
- [82] C. Pal, B. Papp, and L.D. Hurst. Highly expressed genes in yeast evolve slowly. *Genetics*, 158:927–931, 2001.
- [83] S. Pechmann and J. Frydman. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural and Molecular Biology*, 20:237–244, 2013.
- [84] S. Kosakovsky Pond, S.D.W. Frost, and S.V. Muse. HyPhy: hypothesis testing using phylogenetics. *Bioinformatics*, 21(5):676–679, 2005.

- [85] S. Kosakovsky Pond and S.V. Muse. Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution*, 22:2375–2385, 2005.
- [86] S. Kosakovsky Pond, K. Scheffler, M.B. Gravenor, A.F.Y. Poon, and S.D.W. Frost. Evolutionary fingerprinting of genes. *Molecular Biology and Evolution*, 27:520–536, 2010.
- [87] M. Porto, H.E. Roman, M. Vendruscolo, and U. Bastolla. Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. *Molecular Biology and Evolution*, 22:630–638, 2004.
- [88] D.C. Ramsey, M.P. Scherrer, T. Zhou, and C.O. Wilke. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, 188:479–488, 2011.
- [89] D.M. Robinson, D.T. Jones, H. Kishino, N. Goldman, and J.L. Thorne. Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, 20:1692–1704, 2003.
- [90] E.P. Rocha. The quest for the universals of protein evolution. *Trends in Genetics*, 22:412–416, 2006.
- [91] N. Rodrigue, C.L. Kleinman, H. Philippe, and N. Lartillot. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Molecular Biology and Evolution*, 26:1663–1676, 2009.

- [92] N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347:207–217, 2005.
- [93] N. Rodrigue, H. Philippe, and N. Lartillot. Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular Biology and Evolution*, 23:1762–1775, 2006.
- [94] T.K. Seo and H. Kishino. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. *Systematic Biology*, 57:367–377, 2008.
- [95] B. Shapiro, A. Rambaut, and A.J. Drummond. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution*, 23:7–9, 2006.
- [96] P.M. Sharp and W.H. Li. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15:1281–1295, 1987.
- [97] P.M. Sharp and W.H. Li. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Molecular Biology and Evolution*, 4:222–230, 1987.
- [98] P.M. Sharp, T. Tuohy, and K. Mosurski. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, 14:5125–5143, 1986.

- [99] N.K. Sinha and M.D. Haimes. Molecular mechanisms of substitution mutagenesis: an experimental test of the Watson-Crick and Topal-Fresco models of base mispairing. *Journal of Biological Chemistry*, 256:10671–10683, 1981.
- [100] M. Stenico, A.T. Lloyd, and P.M. Sharp. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Research*, 22:2437–2446, 1994.
- [101] N. Stoletzki and A. Erye-Walker. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Molecular Biology and Evolution*, 24:374–381, 2007.
- [102] S. Subramanian and S. Kumar. Gene expression intensity shapes evolutionary rates of protein encoded by the vertebrate genome. *Genetics*, 168:373–381, 2004.
- [103] S. Tavar. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.
- [104] T.A. Thanaraj and P. Argo. Ribosome-mediated translationional pause and protein domain organization. *Protein Science*, 5:1594–1612, 1996.
- [105] J.L. Thorne, N. Goldman, and D.T. Jones. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673, 1996.

- [106] M.D. Topal and J.R. Fresco. Complementary base pairing and the origin of substitution mutaitons. *Nature*, 263:285–289, 1976.
- [107] W.S.J. Valdar and J.M. Thornton. Protein-protein interfaces. Analysis of amino-acid conservation in homodimers. *Proteins: Structure, Function, and Genetics*, 42:108–124, 2001.
- [108] F. Vogel and M. Kopun. Higher frequencies of transitions among point mutations. *Journal of Molecular Evolution*, 9:159–180, 1977.
- [109] J. Wakeley. Substitution rate variation among sites and the estimation of transition bias. *Molecular Biology and Evolution*, 11:436–442, 1994.
- [110] J. Wakeley. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *Trends in Ecology and Evolution*, 11:158–163, 1996.
- [111] T. Warnecke and L.D. Hurst. GroEL dependency affects codon usage - support for a critical role of misfolding in gene evolution. *Molecular Systems Biology*, 6:340, 2010.
- [112] J. Warringer and A. Blomberg. Evolutionary constraints on yeast protein size. *BMC Evolutionary Biology*, 6:61–71, 2006.
- [113] J.D. Watson and F.H.C. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171:964–967, 1953.

- [114] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18:691–699, 2001.
- [115] C.O. Wilke and D.A. Drummond. Population genetics of translational robustness. *Genetics*, 173:473–481, 2006.
- [116] E.E. Winter, L. Goodstadt, and C.P. Pointing. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Research*, 14:54–61, 2004.
- [117] M.Y. Wolf, Y.I. Wolf, and E.V. Koonin. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biology Direct*, 3:40, 2008.
- [118] K.H. Wolfe, P.M. Sharp, and W.H. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337:283–285, 1989.
- [119] Y. Xia, E.A. Franzosa, and M.B. Gerstein. Integrated assessment of genomic correlates of evolutionary rate. *PLoS Computational Biology*, 5, 2009.
- [120] S. Yang, A.F. Smith, S. Schwartz, F. Chiaromonte, K.M. Roskin, D. Haussler, W. Miller, and R.C. Hardison. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Genome Research*, 14:517–527, 2004.

- [121] Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, 13(5):555–556, 1997.
- [122] Z. Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24:8:1586–1591, 1999.
- [123] Z. Yang. *Computational Molecular Evolution*. Oxford University Press, 2006.
- [124] Z. Yang and R. Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, 25:568–579, 2008.
- [125] Z. Yang, R. Nielsen, N. Goldman, and A.M.K. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449, 2000.
- [126] Z. Yang and A.D. Yoder. Estimation of the Transition/Transversion Rate Bias and Species Sampling. *Journal of Molecular Evolution*, 48:274–283, 1999.
- [127] P. Yue, Z. Li, and J. Moult. Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, 353:459–473, 2005.

- [128] G. Zhang, M. Hubalewska, and Z. Ignatova. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural and Molecular Biology*, 16:274–280, 2009.
- [129] J. Zhang, S. Maslov, and E.I. Shakhnovich. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Molecular Systems Biology*, 4:210, 2009.
- [130] L. Zhang and W.H. Li. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution*, 21:236–239, 2004.
- [131] T. Zhou, D.A. Drummond, and C.O. Wilke. Contact Density Affects Protein Evolutionary Rate from Bacteria to Animals. *Molecular Biology and Evolution*, 66:395–404, 2008.
- [132] T. Zhou, M. Weems, and C.O. Wilke. Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. *Molecular Biology and Evolution*, 26:7:1571–1580, 2009.

Vita

Michael Paul Scherrer was born in Yonkers, New York on 11 August 1982. He received an Associate of Science degree in Liberal Arts and a certificate in Plumbing, Heating, and Pipefitting from the State University of New York at Delhi in 2004. During his undergraduate career, he ran his own business as a general contractor and handyman. After curiosity had him enrolled in a botany course, he realized that all the little circles under the microscope were just solar panels and plumbing pipes. Deciding to switch fields to botany, he enrolled in the Molecular Biology program at the State University of New York at New Paltz and spent his time researching yeast kinetochore proteins and building hydroponic systems for the greenhouse. After receiving a Bachelor of Arts degree in Molecular Biology in 2007, he was accepted and enrolled in graduate studies in the Cell and Molecular Biology program at the University of Texas at Austin. He hopes to one day own a goat and a greenhouse.

No plants were researched in the making of this dissertation.

Permanent address: 18 Cherry Lane
Putnam Valley, New York 10579

This dissertation was typeset with L^AT_EX[†] by the author.

[†]L^AT_EX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's T_EX Program.