EVOLUTIONARY DYNAMICS UNDER A STABILITY-CONSTRAINED MODEL

by

Noor Youssef

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2021

To my mother, whose strength and generosity know no bounds.

To my father, who sees every problem as a playful challenge to be overcome by craftiness and a smile.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The space of possible proteins is vast. For all but the smallest proteins, the number of sequences exceeds the number of atoms in the observable universe. Evolution—through the forces of natural selection, drift, and mutation—samples this space, leading to proteins with diverse structures and functions. Evolutionary biologists interested in understanding the history of a protein must identify signals from patterns of substitutions and decipher their likely causes. However, the true evolutionary process is often unknown. Simulations of protein evolution allow us to investigate various emergent phenomena with complete knowledge of the generating parameters in hand. Additionally, using plausible simulating models, we can assess the accuracy of inference procedures which, by necessity, make simplifying assumptions about the process of sequence evolution. In this dissertation, I focus on stability constraints of proteins using a modelling framework grounded in the formalisms of thermodynamics and population genetics theory. In Chapter 2, I show that stability-constrained evolution recapitulates various patterns present in natural alignments. I demonstrate that epistasis due to stability leads to elevated substitution rates compared to site-independent evolution and discuss the underlying mechanisms causing this increase. Additionally, I investigate the accuracy of rate inference from commonly used inference models. While the amount of among-site rate variability is often underestimated, the inferred rates correlated with the most common rates across sites. In Chapter 3, I explore the dynamics of resident amino acid propensities and show that decreases in propensities can occur due to epistasis, challenging claims that such a trend must have adaptive origins. In Chapter 4, I conduct a literature review on nonadaptive phenomena that lead amino acid preferences to change over time. Finally, in Chapter 5, I investigate the evolutionary response to destabilizing substitutions across and within protein structures. I find that destabilizing substitutions at buried residues often require a longer time for the effects to be mitigated than destabilizations at exposed sites. I end the dissertation by discussing the implications of epistasis on protein evolution and future research directions.

LIST OF ABBREVIATIONS AND SYMBOLS USED

ASA	Accessible surface area
C-SI	C-series site-independence model
CLM3	M3 models with covarion-like component
DMS	Deep mutational scanning
HKY85	Hasegawa *et al.* (1985) model
JC69	Jukes and Cantor (1969) model
M3(k)	M-series models from Yang *et al.* (2000)
MLE	Maximum likelihood estimate
MSA	Multiple sequence alignmnets
MutSel	Mutation-selection model
RSA	Relative solvent accessibility
S2S	Sequence-to-sequence space
S-SD	Stability-infromed site-dependence model
S-SI	Stability-infromed site-independence model
WCN	Weighted contact number
avg[$\pi^h_{a\|res}$]	Average propensity of a while it is resident
CM_k	Contact matrix in stucture k
dN	Nonsynonymous substitution rate
dS	Synonymous substitution rate
ΔG	Thermodynamic stability
$\Delta\Delta G$	Mutational effect on stability
ΔE^2	Variance in free energy in unfolded microstates
E_F	Free energy in the folded macrostate
E_U	Free energy in the infolded macrostate
$\bar{E}$	Mean free energy in unfolded microstates
ϵ_{MJ}	contact potentials from Miyazawa and Jernigan (1985)
f_i	Fitness of mutant i

$f_a^h(S)$	Fitness of sequence S given amino acid a at site h
$F^h(S)$	Vector of site-specific fitness landscape at site h
H^h	Shannon entropy of site-specific landscape
κ	Transition-transversion rate ratio
k_F	Native folded three-dimensional structure
$\{k_U\}$	Set of alternative structures
L	Protein length
μ_{ij}	Mutation Rate from i to j
N	Number of taxa
N_U	Number of unfolded microstates
N_e	Effective population size
N_c	Census population size
$\mathcal{N}_x$	Set of nonsynonymous single-step codons
π_n	Stationary Frequency of nucleotide n
π_x	Stationary Frequency of codon x
π_a	Stationary Frequency of amino acid a
$\pi_a^{(0)}$	Stationary Frequency in the absence of selection pressure
$\pi_{a\|new}^h$	Propensity for a when it was first accepted
P_{fix}	Fixation probability
P_{ij}	Transition probability from state i to j
Q	Transition rate matrix
q_{ij}	Transition rates between states i and j
$\rho(E)$	Distribution of free energies
s_{ij}	Selection coefficient
T_{res}	Amino acid residence time
ω	Inferred nonsynonymous-to-synonymous substitution rate ratio
Z_U	Partition function

ACKNOWLEDGEMENTS

Choice in research is a luxury. I am, therefore, greatly indebted to my supervisors Joseph Bielawski and Edward Susko for giving me fair latitude in choosing my research projects. Joe, your constant enthusiasm and curiosity about science are contagious. Ed, your skill for communicating complex topics in ways that are accessible is unmatched. I am very thankful for Andrew Roger; you have set an excellent example of how science should be: fun, thorough, and collaborative. I would also like to thank Mark Johnston for always providing insightful and timely responses. I am grateful to my supervisory committee, the members of the Biology Department, and the members of the Centre for Comparative Genomics and Evolutionary Bioinformatics (CGEB) for showing me the joy felt by those whose privilege it is to uncover a little about how the universe works.

To my family, the Youssef and Aly clans, nothing would be the same without you. You bring colour to my life and keep me (in)sane. Mama and baba, I owe everything to you. Thank you for all that you do for us. Special thanks to Omar Youssef for the many hours you have devoted to listening to me talk about science and for reading every book I recommend. You, my brother, are a scientist by design. Lamis Youssef, ringing in the New Year together makes the rest of the year a tough act to follow. Yet we always manage to live up to it. Ameto Sanaa, you are the most generous hostess and the glue that holds us together. Most of all, thank you for bringing into the world the very best of humans: Maged, Khlood, Khaled, and Sara. Our cousins' trips and study dates sprinkled the last few years with unforgettable memories.

Scott McCain, you are a man of eager curiosity. Thank you for keeping me company when I was "in the cloud" (c.f. Uri Alon's 2013 TED talk) and for reviewing a draft of every chapter. You are an endless resource of support and ingenuity. To the once friends who are now family, Sarah Martakoush and Taylor Hersh: this journey would not have been the same without you. Sarah, you have been a constant pillar of support during every step of my secondary education. I truly cannot look back at the past few years without laughing at a memory we have shared. Taylor, the second half of Taynoor, I am grateful to have shared a home filled with great food and even better company. I am also thankful for the many friendships that flourished through the Biology Organization of Graduate

Students (BOGS) and our shared love for science: Lisette Delgado, Cat Bannon, Megan Roberts, Loay Jabre and other BOGS members. An academic enjoys a license to always be learning; we are the lucky ones, allowed to spend our days in a continuing course of education.

CHAPTER 1

INTRODUCTION

> The whole organic world is the result of innumerable different combinations and permutations of relatively few factors . . . These factors are the unit which the science of heredity has to investigate. Just as physics and chemistry go back to molecules and atoms, the biological sciences have to penetrate these units in order to explain . . . the phenomena in the living world.
>
> —Hugo de Vries

The diversity in nature is awe-inspiring, from tiny microbes to enormous whales. Despite differences in their size, shape, life span, and almost every aspect of their existence, all living things utilize the same biochemical system as genetic material. How, then, does this diversity arise? The objective of evolutionary biologists is to propose and investigate a mechanistic explanation for the intricate characteristics of different living forms. This chapter, with brevity in mind, covers the historical background of evolution by natural selection—from Darwin's *Origin* through the *Synthesis* and ending with modern computational evolutionary biology. In this chapter, I also introduce, in detail, two modelling frameworks used throughout the remainder of the dissertation: the Mutation-Selection (MutSel) framework and the stability-informed framework. This introduction sets the stage for the chapters that follow.

1.1 The origin of the synthesis

The essence of evolution by natural selection is commonly attributed to Charles Darwin (1809-1882)[1]. In his famous book *On the Origin of Species by Means of Natural Selection* (1859), or *Origin* in short, Darwin laid down the foundations for evolution by natural selection. With our modern-day vantage point, the theory appears intuitive: when animals reproduce, they sometimes produce variants that differ from the parents. Because of competition over scarce resources, individuals that are better adapted for an environment are "naturally selected". In other words, they are more likely to survive and leave viable offspring. This process is gradual, occurring slowly over time. Nevertheless, it is this process that produces the adaptive complexity of living forms that surround us today.

The basic elements of evolution by natural selection embody three principles, as summarized by Lewontin (1970):

1. Phenotypic variation: Individuals in a population differ in morphology, physiology, and behaviour.

2. Differential fitness: Phenotypically different individuals have different rates of survival and reproduction.

3. Heritability: Similarity in phenotype between parents and offspring.

At the time, Darwin's theory was largely conceptual since the molecular underpinnings were not yet discovered—How do phenotypic differences arise? How are organismal attributes transmitted to offspring? In 1865, six years after Darwin's *Origin*, Gregor Mendel (1822-1884) presented work at the Natural Science Society in Brno, demonstrating the existence of heritable traits (Mukherjee, 2016). Unfortunately, for decades, his work went largely unnoticed[2]. Mendelian genetics was not linked to Darwinian evolution until its rediscovery in the early 1900s by de Vries (1848-1935), Bateson (1861-1926), and Mogran (1866-1945)[3]. Their works laid the foundations for the discipline we now call

[1]The basic concept of biological evolution predates Darwin by almost two millennia. The pre-Socratic Greek philosopher Anaximander (610 - 546 BC) believed in a progression of animal forms, hypothesizing that life originated in the sea.

[2]In the same meeting, and soon after Mendel's presentation, a botanist discussed the *Origin* and Darwin's theory of evolution. However, the link between Mendel's and Darwin's work was not then appreciated.

[3]Despite being informed of Mendel's work, de Vries pointedly neglected to cite it in his publication on plant hybrids. On the other hand, Bateson was an ardent champion of Mendel. For this, he was nicknamed "Mendel's bulldog".

Genetics—a term coined by Bateson in 1905 (Richmond, 2001). (De Vries also coined a now recognizable term of his own, *mutation*.) Rather than champion Darwin's theory of evolution, the Mendelian mechanisms of inheritance drove it to the brink of extinction. For decades to follow, evolutionary biologists partook in debates regarding the relative roles of mutations (*mutationist school*[4]) and natural selection (*selectionist school*) in explaining the variation in organismal forms. The mutationists asserted that most organismal differences are attributable to mutations alone and rejected Darwin's doctrine of *natura non facit saltum* (Stoltzfus and Cable, 2014). The selectionists maintained that the gradual accumulation of mutations leads to phenotypic differences on which natural selection acts. The former quickly gained ground, and the latter dwindled in numbers.

It was not until the development of rigorous mathematical models by Ronald Fisher (1890-1962), John B. S. Haldane (1892-1964), and Sewall G. Wright (1889-1988), among others, that mutational theory, the laws of inheritance, and natural selection were integrated into a unified framework. Their works during the mid-twentieth century led to the birth of the *modern synthesis*—coined by Julian Huxley (1887-1975) in his 1942 book *Evolution: The Modern Synthesis* (see Mayr and Provine (1981) for a relatively more recent review). Several ideas developed in this period are essential for the work completed in this dissertation—such as the notions of *random genetic drift* and the *effective population size*.

The size of a population has significant consequences for its evolutionary dynamics. Its importance was adequately emphasized by Wright (1931): "There remains one factor of the greatest importance in understanding the evolution of a Mendelian system. This is the size of the population". Due to finite population sizes, gene frequencies will fluctuate over time merely by chance. Wright referred to this phenomenon as *random genetic drift* (Wright, 1931). Random drift is more pronounced in smaller populations, and its influence dwindles as population size increases. This leads to the question: how should the size of a population be defined? The rate of genetic drift is proportional to the *effective population size*, N_e, rather than the census population size, N_c. The effective population size is a theoretical quantity that can be conceptualized as the size of an *idealized* population that would exhibit the same intensity of genetic drift as the natural population. In an idealized population, N_e will equal N_c. The definition of an idealized population assumes:

1. an equal number of breeding males and females

[4]Also called, *Mutationstheorie*, or mutation theory, by de Vries (De Vries, 1919).

2. that the population is panmictic (exhibiting random mating)

3. an equal expectation of offspring for each indivudual

4. that the number of breeding individuals is constant

Violations in these assumptions, which are common in natural populations, leads to $N_e < N_c$. The effective population size N_e is hence the parameter of interest when modelling the evolutionary process.

Of underlying importance during this period was the calculation of the probability of ultimate success (i.e., fixation) or elimination of a mutant. In 1927, Haldane, using a method developed by Fisher (1922), resolved that the probability of fixation of a beneficial mutation with selective advantage s will be approximately $2s$. Alternatively, Wright (1931) and Fisher (1958) estimated that if a mutation has little effect on fitness, then the probability of its eventual fixation in a diploid population will be $1/2N_e$. In 1962, Motoo Kimura unified the different probability estimates, accounting for arbitrary dominance regimes and fluctuating selection coefficients, such that the probability with which a mutant j becomes fixed in a diploid population with wildtype variant i is given by

$$P_{fix} = \frac{1 - \exp(-2\ s_{ij})}{1 - \exp(-4N_e s_{ij})} \tag{1.1}$$

where N_e is the effective population size, $s_{ij} = f_j - f_i$ is the relative fitness effect of the mutant, and f_j is the fitness of mutant j. In equation (1.1) N_e acts as a tuning parameter for the relative intensities of selection and random genetic drift. If a mutation is selectively neutral ($s_{ij} \approx 0$), then the probability of it going to fixation just by random drift is equal to $1/2N_e$, as expected from Wright (1931) and Fisher (1958). For a positive s_{ij} and a very large N_e, the probability of fixation of the beneficial mutant will be equal to $2s_{ij}$, Haldane's result. Lastly, deleterious mutations, where $s_{ij} < 0$, are selected against such that the probability of fixation is less than that of a neutral mutation.

Darwin was a self-proclaimed mathematical novice. "I attempted mathematics... but I got on very slowly. The work was repugnant to me", he states in his autobiography (Darwin, 1958). Yet he was fully aware of its merits: "I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics; for men thus endowed seem to have an extra sense" (Darwin, 1958). It was through the works of scientists "thus endowed" that evolutionary theory, by the

mid-twentieth century (over a century after the first publication of the *Origin*), had come to be grounded in mathematical rigour.

1.2 Sequence space and epistasis

A key component of evolution by natural selection is, as stated earlier, differential fitness, where phenotypic differences among individuals lead some to survive and reproduce more frequently than others. However, in practice, the effects of mutations on organismal fitness are complex and challenging to measure (see Chapter 4 for further discussion). Alternatively, the fitness of a protein molecule is more directly amenable to precise measurements (e.g., enzymatic activity, binding affinity, folding stability). All else being equal, the fitness of an individual carrying a particular mutant protein will correlate with the fitness of the protein.

In 1970, John Maynard Smith presented an intuitive analogy to protein evolution, a word game where the objective is to move from one meaningful word to another by changing a single letter at a time. Imagine, for example, a path between WORD and GENE. One such path is WORD $\rightarrow$ WORE $\rightarrow$ GORE $\rightarrow$ GONE $\rightarrow$ GENE; this was the example trajectory provided by Maynard Smith (1970). The parallels with protein evolution are: each meaningful word represents a functional protein; each letter, an amino acid; and every letter change, a substitution.

This simple analogy reveals various salient aspects of protein evolution. First, what constitutes a meaningful word? Must they be English words defined, for example, in the *Oxford English Dictionary*? Whatever this requirement may be, it would correspond to the sequence-level selective pressures acting on proteins. Such a sequence-level fitness function maps sequences to fitness values (or words to meaning). The space of all possible sequences of a given length L is immense and multi-dimensional. The space contains 20^L possible amino acid sequences and every sequence has $19 \times L$ single-step neighbours. Multi-dimensional spaces are complex and produce many counter-intuitive phenomena (Gavrilets, 2003, 2004). While various advancements have been made on this front (e.g., Steinberg and Ostermeier (2016)), they remain challenging to work with and interpret.

A more tractable approach is to define the fitness landscape at an individual position in a word or a site in the protein (Bazykin, 2015). These site-specific fitness landscapes are fully defined by a vector of length 20 representing the fitness value of each amino

acid or a vector of 26 for every letter in the word game analogy. During the path from WORD to GENE, the fitness landscape at the first position changes as the background sequence varies. Letters that produce meaningful words in one background fail to do so in another (e.g., WORE and WONE). Similar site-specific dynamics occur throughout protein evolution. The dependency of the fitness effect of a mutation on the genetic background is referred to as *epistasis*[5].

Sequence space is rich beyond measure: for an average-sized protein of length 300, the total number of sequences (20^{300}) is vaster than all the stars in the observable universe[6]. Evolution threads paths through sequence space leading to a diversity of life forms. Can we decipher from these sequences the evolutionary past? What do the observed patterns of substitutions indicate? To address these questions, we must rely on plausible models of sequence evolution.

1.3 Models of sequence evolution

Probabilistic models that characterize evolutionary dynamics have been valuable in evolutionary biology. The process of sequence evolution is Markovian in nature; the probability of a mutation, and ultimately fixation (or elimination), depends only on the current state and not on past states. Therefore, continuous-time Markov chains are commonly used to model the evolutionary process (e.g., Muse and Gaut (1994); Goldman and Yang (1994); Yang and Nielsen (2002); Kosakovsky Pond and Frost (2005)). Markov chains can model the evolution of particular codon positions in a protein (e.g., Goldman and Yang (1994)), or nucleotide positions within a codon position (e.g., Jukes and Cantor (1969)), or at a larger scale, the evolution of the entire protein sequence (e.g, Youssef *et al.* (2020)). Markov models are stochastic, memoryless processes that describe transitions between different states.

Jukes and Cantor (1969) presented a simple Markov model, referred to as JC69, which assumes that all nucleotides are equally likely, and that every nucleotide has the same rate of changing into any other nucleotide. A slightly more complex nucleotide model was presented by Hasegawa *et al.* (1985), referred to as HKY85, which accounts for differences

[5]The term epistasis was also coined by William Bateson in 1909.

[6]One of the earliest estimates of the number of stars in the universe comes from Archimedes (287-212 BC), the author of the *The Sand Reckoner*. He estimated that 10^{63} grains of sand are required to fill the universe, that is 10^{83} atoms; a number eerily close to our estimates today ($10^{78} - 10^{82}$).

in nucleotide frequencies $\{\pi_A, \pi_C, \pi_T, \pi_G\}$ and for differences in the rates of transition within purines (A $\leftrightarrow$ G) and within pyrimidines (T $\leftrightarrow$ C) versus the rates between the purines and pyrimidines (A,G $\leftrightarrow$ T, C).

The transition rates between states in a Markov chain are given by an instantaneous rate matrix Q. This matrix is populated by elements q_{ij} describing the transition rates from state i to state j in an infinitesimally small amount of time. For example, the transition matrix specifying the HKY85 model is given by:

$$Q = \begin{pmatrix} . & \pi_C & \pi_T & \kappa\pi_G \\ \pi_A & . & \kappa\pi_T & \pi_G \\ \pi_A & \kappa\pi_C & . & \pi_G \\ \kappa\pi_A & \pi_C & \pi_T & . \end{pmatrix} \tag{1.2}$$

where κ is the transition-transversion rate ratio, π_j is the stationary frequency of nucleotide j, and the diagonal elements $q_{ii} = -\sum_j q_{ij}$ are specified such that the row sum equals zero. The Q-matrix fully defines the dynamics of the Markov chain.

The JC69 (Jukes and Cantor, 1969) and HKY85 (Hasegawa *et al.*, 1985) models are two examples of DNA-level (i.e., having nucleotide states) Markov models. Such models are fully characterized by a 4×4 Q-matrix, see for example equation (1.2). Codon or amino acid models can analogously be specified by a 61×61 (where each state represents a sense codon) or 20×20 transition rate Q-matrices, respectively. The transition rates between states in a Markov model can be informed by population genetics parameters. The following section describes one such application, referred to as the mutation-selection (MutSel) framework (Halpern and Bruno, 1998).

1.3.1 The mutation-selection (MutSel) framework

The MutSel framework assumes an idealized Wright-Fisher population with fixed effective population size (N_e) and a weak mutation-strong selection regime such that a mutation is either fixed (or eliminated) before the introduction of a second mutant into the population (Fisher, 1922; Wright, 1931). The transition rates, q_{ij}, between states are equal to the product of the rate of a novel mutation j occurring in the population, $2N_e\mu_{ij}$, and its subsequent rate of fixation, P_{fix}:

$$q_{ij} \propto 2N_e\mu_{ij}P_{fix} \tag{1.3}$$

where μ_{ij} represents the mutation rate from variant $i \rightarrow j$, and P_{fix} is defined from population genetics theory using equation (1.1).

Throughout this dissertation, I modelled the evolutionary process as occurring between protein sequences. The state-space of a sequence model is made up of 20^L possible states where L is the length of the protein. It is impossible to fully define the $20^L \times 20^L$ Q-matrix for all but the smallest proteins. Instead, given that the process is currently at sequence i, it is feasible to calculate the transition rates to all single-nucleotide step neighbouring sequences[7]. To model the process of sequence evolution, the probability of a transition into another state j is calculated as $P_{ij} = q_{ij} / \sum_{j \neq i} q_{ij}$. At each time step, the substitution to the next state is determined by a random draw from a multinomial distribution with probabilities P_{ij}.

Grounded in population genetics theory, the MutSel framework is a powerful tool for exploring evolutionary dynamics. Nevertheless, its plausibility as a model for generating sequences is contingent upon appropriate definitions of the underlying parameters (e.g., selection coefficients, s_{ij}). To assign fitness values to amino acid sequences, I used a stability-informed biophysical model.

1.3.2 The stability-informed model

Proteins are biomolecules that must obey the physical laws of our universe. Over the last decade, biophysical models of proteins have been useful for understanding the evolutionary dynamics of proteins (e.g., Williams *et al.* (2006a); Goldstein (2011); Pollock *et al.* (2012); Shah *et al.* (2015); Youssef *et al.* (2020)). Most proteins are marginally stable, teetering on the verge of unfolding. The marginal stability of proteins was first interpreted as evidence of selection favouring lower stability values, allowing the protein greater flexibility to change configurations (DePristo *et al.*, 2005). However, using a stability-informed evolutionary model, it became evident that marginal stability emerges due to the interplay between mutation, selection, and drift (Goldstein, 2013). Hence, marginal stability is an expected emergent property even if evolution does not actively favour it. Biophysical models of protein evolution have also lead to a deeper understanding of intricate evolutionary dynamics, leading to characterizations of phenomena such as contingency (Shah *et al.*, 2015), entrenchment (Shah *et al.*, 2015), evolutionary Stokes shift (Pollock *et al.*, 2012;

[7]Since each amino acid is encoded by three nucleotides, an amino acid sequence of length L is encoded by a nucleotide sequence of length $3L$. Then, the number of single-nucleotide step neighbours is only $3 \times 3L$.

Goldstein and Pollock, 2017), and more recently, the evolutionary anti-Stokes shift. These phenomena will be discussed in detail in Chapter 3.

To illustrate how stability is calculated, consider a protein of length L with sequence $S = \{a^1, \ldots, a^h, \ldots a^L\}$, where a^h specifies the amino acid at position h. The amino acid sequence of a protein influences its physicochemical properties (such as the protein's mechanism of action, its structure, and its stability). The linear amino acid polypeptide chain defined by S folds into a unique three-dimensional structure k_F within the aqueous solution of a cell. The calculation of stability is easiest to illustrate in the case of a simple two-state folding system, where the protein molecule is either in the folded (F) or unfolded (U) configurations, or *macrostates*. A macrostate defines a state with macroscopically measurable parameters (e.g., temperature, volume, pressure). Each macrostate is composed of numerous *microstates*, defining the positions of individual particles. As an analogy, consider the diagram presented in figure 1.1A. Suppose that the measurable parameter is the number of black (or white) boxes. There are five possible macrostates, having 0, 1, 2, 3, or 4 black boxes. The 2-black box macrostate can be obtained from six different microstates (i.e., the different arrangement of black boxes). For molecules, the boxes in the analogy described above correspond to individual atoms or particles. Different orientations of atoms result in numerous microstates for any given molecule.

The stability, ΔG, of a sequence is defined as the difference in free energy between the folded E_F and unfolded E_U macrostates:

$$\Delta G = E_F - E_U \tag{1.4}$$

It is widely accepted that the native folded state corresponds to the free energy minimum (Kaffe-Abramovich and Unger, 1998). Therefore, the energy required to maintain the sequence in the correctly folded native state of a protein (E_F) is lower than the energy in an unfolded configuration (E_U). As such, protein stability, ΔG, is usually negative and the process of folding occurs spontaneously (figure 1.1B).

The free energy E_k associated with sequence S in a given structure k can be approximated as the sum of pairwise energy potentials for amino acids in contact in the tertiary structure,

$$E_k = \sum_{x<y} \varepsilon_{MJ}(a^x, a^y)\ \mathrm{CM}_k^{x,y} \tag{1.5}$$

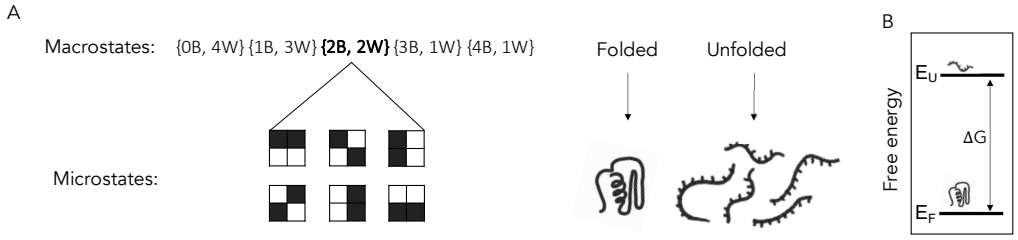


Figure 1.1: Illustration of thermodynamic principles of protein stability calculations. (A) An analogy for the description of macrostates and microstates. Macrostates represent any system with measurable macroscopic properties (e.g., temperature, volume). Alternatively, microstates define the position of individual particles. For each macrostate, there is a large number of corresponding microstates. In this analogy, the macrostate can be defined as the number of black (or white boxes). There are five different macrostates: zero black boxes and 4 white $\{$0B, 4W$\}$, or one black box and 3 white $\{$1B, 3W$\}$ and so on. Consider the $\{$2B, 2W$\}$ macrostate. There are a total of six different microstate configurations that yield two black and two white boxes. With regards to proteins, the two macrostates are the folded and unfolded configurations. The folded state corresponds to the known native three-dimensional structure of the protein. Alternatively, there are a large number of unfolded configurations. (B) Protein stability is calculated as the difference in free energy between the folded E_F and unfolded E_U states.

where ε_{MJ} are the contact potentials determined by Miyazawa and Jernigan (1985), and CM_k is the contact matrix specifying interactions between sites in structure k such that $\text{CM}_k^{x,y}$ = 1 if site x and y are in contact and 0 otherwise[8]. The contact potentials ε_{MJ} are based on a statistical analysis of known protein structures (Miyazawa and Jernigan, 1985). Due to the nature of the statistical analysis, these potentials represent potentials of mean force and implicitly include hydrophobic interactions and other effects of the solvent.

The free energy associated with the folded state E_F can be calculated using equation (1.5), with $k = k_F$ as the known native structure. Usually, however, there is not a single unfolded configuration. It is intractable to characterise the entire set of possible unfolded structures, making it challenging to estimate E_U directly. Instead, a subset of structures $\{k_U\}$ is selected to characterise the distribution of thermodynamic properties of the ensemble of unfolded microstates. Then, the free energy in the unfolded macrostate is

[8]Sites are considered to be in contact if the C_β atoms are within 7Å. If the amino acid present is glycine, distance is calculated with reference to the C_α atom.

given by the Helmholtz free energy equation:

$$E_U = -\beta^{-1} \ln Z_U \tag{1.6}$$

where $\beta = 1/kT$, k is the Boltzmann constant, T is absolute temperature, and Z_U is the partition function over the ensemble of unfolded microstates. Assuming that the free energy of the ensemble of unfolded structures approximately follows a Gaussian distribution, $\{k_U\}$ is used to estimate the mean $\bar{E}$ and standard deviation ΔE^2. These parameters define the distribution, $\rho(E)$, of free energies, E, over unfolded states:

$$\rho(E) = \frac{1}{\sqrt{2\pi\Delta E^2}} \exp\left[\frac{-(E-\bar{E})^2}{2\Delta E^2}\right] \tag{1.7}$$

The partition function sums over all unfolded energies which is equivalently a sum of all possible energies, weighted by how frequently they arise:

$$Z_U = \sum_{i=1}^{N_u} \exp(-\beta E_i) \tag{1.8}$$

$$\approx N_U \int \rho(E) \exp(-\beta E) dE \tag{1.9}$$

$$= N_U \exp(\frac{1}{2}\beta^2 \Delta E^2 - \beta \bar{E}) \tag{1.10}$$

where N_U is the number of unfolded microstates. The stability of a sequence S can then be rewritten as

$$\Delta G = E_F + \beta^{-1} \ln Z_U \tag{1.11}$$

$$= E_F - \bar{E} + \frac{1}{2}\beta \Delta E^2 + \beta^{-1} \ln N_U \tag{1.12}$$

A common assumption in biophysical models of protein evolution is that fitness is equal to the proportion of correctly folded proteins at thermodynamic equilibrium. From thermodynamic theory, the probability of a system (or molecule) occupying a macrostate m is described by the Boltzmann distribution:

$$P_m = \frac{e^{-\beta E_m}}{\sum_n e^{-\beta E_n}} \tag{1.13}$$

where E_m is the free energy associated with being in macrostate m, and the sum over n represents all macrostates. In the case of a two-state folding system, the probability of observing a sequence in the native structure at thermodynamic equilibrium will be

$$P_F = \frac{e^{-\beta E_F}}{e^{-\beta E_F} + e^{-\beta E_U}} \tag{1.14}$$

$$= \frac{e^{-\beta E_F}}{e^{-\beta E_F} + e^{-\beta E_U}} \left(\frac{e^{\beta E_U}}{e^{\beta E_U}} \right) \tag{1.15}$$

$$= \frac{e^{-\beta \Delta G}}{e^{-\beta \Delta G} + 1} \tag{1.16}$$

This framework allows us to estimate the stability of a sequence in a given structure. Therefore, the stability-informed model acts as a mapping between amino acid sequences and fitness values (analogous to the role of the *Oxford English Dictionary* in Maynard Smith's word game analogy). We can, therefore, use this framework in conjunction with the MutSel model for simulations of the evolutionary process that are grounded in the formalisms of thermodynamics and population genetics.

1.4 Thesis outline

In this dissertation, I combine the stability-informed and MutSel frameworks to simulate hypothetical evolutionary trajectories and study the emerging dynamics. I begin by validating the stability-informed model by comparing predicted properties with patterns in natural data (Chapter 2). Then, I investigate the impact of stability-mediated epistasis on substitution rates. I find that epistasis tends to inflate the rate at which substitutions accumulate compared to the rates under site-independent evolution. Nevertheless, inferred rates from commonly used models (e.g., ω-based models) are not systematically biased due to epistasis: the accuracy of substitution rate inference is comparable when alignments are generated with and without epistasis.

In Chapter 3, I investigate how the propensity for a resident amino acid changes over time. Through this analysis, I defined a new phenomenon, the *evolutionary anti-Stokes shift* where the preference for the resident amino acid decreases as the protein evolves. This observation challenges previous claims that epistasis cannot explain such reductions in preferences (Popova *et al.*, 2019; Stolyarova *et al.*, 2020). I show that in the absence of an adaptive change (i.e., given a constant environment) the number of sites for which

the propensity of the resident amino acid increases is balanced by the number of sites for which the propensity of the resident amino acid decreased. I also observe that epistasis reduces the magnitude of propensity shifts through a significant negative auto-correlation in propensity changes. Increases in propensities tend to be followed by decreases (and vice versa).

In Chapter 4, I perform a comprehensive review of theoretical and experimental work related to nonadaptive changes in site-specific fitness landscapes. Analysis of natural sequence alignments often show evidence of declining levels of homoplasy (convergence, reversions, and rates of parallel evolution) with divergence time. The levels and trends in empirical data are inline with expectations from epistatic models. Furthermore, I report on results from site-directed mutagenesis, and available deep mutational scanning datasets. Experimental studies reveal that changes in site-specific fitness landscapes are often minor in magnitude, even over long evolutionary time scales. Importantly, this review identifies evidence in the literature in support of a balance in the proportion of sites for which the propensity of the resident amino acid increases or decreases.

Finally, in Chapter 5, I investigate how proteins, and sites within proteins, adjust to destabilizing substitutions. I find differences in the number of compensatory substitutions that are required to adjust for the destabilization among proteins and across sites within a protein. I report on the structural features that explain the disparities in response to destabilizations. Together, the chapters in this thesis advance our understanding of the implications of stability and epistasis on protein evolution.

CHAPTER 2

CONSEQUENCES OF STABILITY-INDUCED EPISTASIS FOR SUBSTITUTION RATES

This work was published previously in the journal *Molecular Biology and Evolution* (Youssef *et al.*, 2020).

2.1 Abstract

Do interactions between residues in a protein ($i.e.$ epistasis) significantly alter evolutionary dynamics? If so, what consequences might they have on inference from traditional codon substitution models which assume site-independence for the sake of computational tractability? To investigate the effects of epistasis on substitution rates, I employed a mechanistic mutation-selection model in conjunction with a fitness framework derived from protein stability. I refer to this as the stability-informed site-dependent (S-SD) model, and developed a new stability-informed site-independent (S-SI) model that captures the average effect of stability constraints on individual sites of a protein. Comparison of S-SI and S-SD offers a novel and direct method for investigating the consequences of stability-induced epistasis on protein evolution. I developed S-SI and S-SD models for three natural proteins and showed that they generate sequences consistent with real alignments. The analyses revealed that epistasis tends to increase substitution rates compared to the rates under site-independent evolution. I then assessed the epistatic sensitivity of individual sites and discovered a counterintuitive effect: highly connected sites were less influenced by epistasis relative to exposed sites. Lastly, I show that, despite unrealistic assumptions,

traditional models perform comparably well in the presence and absence of epistasis, and provide reasonable summaries of average selection intensities. While epistatic models are critical to understanding protein evolutionary dynamics, epistasis might not be required for reasonable inference of selection pressure when averaging over time and sites.

2.2 Introduction

Most proteins must fold into a native structure in which they are moderately stable before they are able to perform their biological function. Protein stability depends on the sequence of amino acids and their interactions in the folded three-dimensional structures. Because of these interactions, evolutionary selective constraints to maintain adequate stability result in epistatic dependencies between residues. Specifically, epistasis manifests as a dependency in the fitness effect of a mutation on the background protein sequence in which it arose. For example, let $f_a^h(S)$ be the fitness of the protein provided amino acid a is occupying site h in the context of background sequence S. Then, $F^h(S) = \langle f_1^h(S), ..., f_{20}^h(S) \rangle$ is the site-specific vector of amino acid fitness values specifying the fitness landscape at site h. Following a substitution at another position in the protein, so that the background sequence changes from S to X, the fitness of the same amino acid will subsequently change, $f_a^h(S) \neq f_a^h(X)$. Therefore, in the presence of epistatic dependencies the fitness landscape at a site is subject to fluctuations as substitutions occur at other sites (figure 2.1A). Stability constraints typically result in global epistasis, meaning that a change in the incumbent amino acid at one site induces shifts in the fitness landscapes at many, often all, other sites in the protein (Starr and Thornton, 2016). While such interdependencies inevitably occur, the magnitude and frequency of these shifts, and their impact on protein evolution, remain controversial.

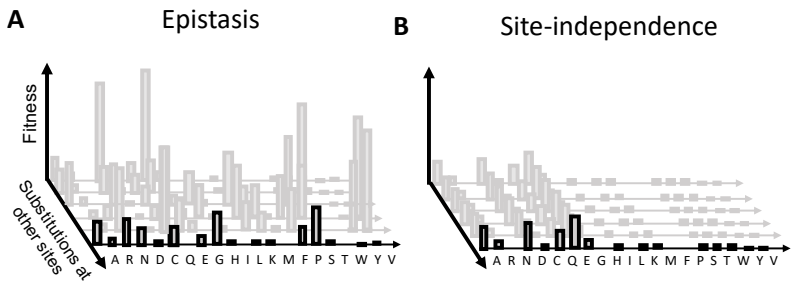


Figure 2.1: Site-specific fitness landscape dynamics under epistatic (A) and site-independent (B) models. (A) Epistasis results in a changing site-specific fitness landscape as substitutions occur at other positions in the protein. (B) Site-independent evolution implies a static (constant) fitness landscape.

Using extensive computational experiments, Pollock *et al.* (2012) found that stability-induced epistasis results in frequent and substantial shifts in amino acid fitness landscapes. To the contrary, Ashenberg *et al.* (2013) used computational and experimental approaches and reported that while stability-induced fluctuations in site-specific amino acid fitness landscapes do occur, they are relatively minor in magnitude and are therefore inconsequential with regards to long term evolutionary dynamics. This controversy has spurred multiple follow-up experiments, finding support for both claims and little consensus (Shah *et al.*, 2015; Risso *et al.*, 2015; Ferrada, 2019; Starr *et al.*, 2018). It remains unclear if and how stability-induced epistasis influences protein evolution.

Models used to infer evolutionary parameters from natural protein alignments commonly assume site-independence and other simplifying assumptions ($e.g.$ time-stationary substitution rates, and low levels of among-site rate heterogeneity) for the sake of computational tractability. In this study, I focus on the widely used codon substitution models which infer selection pressure as ω, the normalized ratio of nonsynonymous substitutions to the ratio of synonymous substitutions (Goldman and Yang, 1994; Muse and Gaut, 1994); I refer to these as ω-based models. Natural proteins evolve under complex evolutionary dynamics that are not entirely captured by traditional ω-based models ($e.g.$ epistatic interactions between sites). If epistasis between residues in a protein does have a dramatic effect on protein evolution, then the validity of inference from site-independence models might be negatively impacted.

Does epistasis substantially influence the rate at which proteins evolve? And if so,

how reliable are inferences from traditional ω-based models which assume that sites evolve independently? Addressing these questions are the main objectives of this chapter. To do this, I model the evolutionary process from first principles of population genetics theory using the mutation-selection (MutSel) framework (Halpern and Bruno, 1998; Yang and Nielsen, 2008). Unlike ω-based models, MutSel models account for differences in amino acids fitness values and allow for more realistic levels of among-site rate heterogeneity by assigning each site a unique fitness landscape (F^h). MutSel frameworks are commonly used as a method for simulating plausible evolutionary dynamics (Rodrigue *et al.*, 2010; Spielman and Wilke, 2015; Jones *et al.*, 2017, 2018, 2020; Ashenberg *et al.*, 2013). Nevertheless, these are site-independent models and therefore do not directly model the dynamics of epistasis. With appropriate fitness values, they can in theory be used to model the marginal effects of stability and/or other selective pressures on a site. The challenge then lies in determining plausible site-specific fitness landscapes.

Several ways of calculating amino acid fitness values have been proposed. For example, Spielman and Wilke (2015) derived amino acid fitness values based on empirical site-specific frequencies from large alignments of homologous proteins. Alternatively, Jones *et al.* (2018) assigned amino acid fitness values such that the estimated probability density function of the scaled fitness effects ($s_{ij} = 2N_e(f_j - f_i)$ for amino acids i and j and effective population size N_e) matches the distribution inferred from empirical data. Hereafter, these approaches are referred to as site-wise MutSel. Under the site-wise MutSel formulations, site-specific fitness landscapes average the selective pressure acting on a site, assuming site-independent evolution, and therefore time-stationary fitness landscapes (figure 2.1B). Changes in site-specific fitness landscape are interpreted as a change in selection pressure (either due to a change in environment or a change in protein function).

Determining fitness landscapes has also been addressed mechanistically by combining the MutSel approach with biophysical models of protein folding where fitness values depend on protein stability or the proportion of correctly folded proteins at thermodynamic equilibrium (Goldstein and Pollock, 2016; Pollock *et al.*, 2012; Goldstein and Pollock, 2017; Ashenberg *et al.*, 2013). While comparable at first glance, the biophysical approaches differ extensively from the site-wise MutSel applications. Importantly, the biophysical models account for temporal variation in site-specific fitness landscapes that emerges as a consequence of global stability-induced epistasis (Figure 2.1A). Accounting

for these temporal dynamics is essential for understanding how epistasis influences protein evolution. While the evolution of natural proteins is certainly shaped by additional structural and functional constraints, for most proteins, proper folding into a native structure is prerequisite to being able to carry out their biological function.

To investigate the influence of epistasis on protein substitution rates, I use the MutSel evolutionary model in conjunction with a biophysical model of protein folding. I refer to this as the stability-informed site-dependent (S-SD) model since stability calculations inherently take into account epistatic interactions between sites. I develop an analogous stability-informed site-independent (S-SI) model where proteins evolve under equivalent stability mediated selection pressures but having independent and constant fitness landscapes (figure 2.1B). Specifically, from each S-SD evolutionary simulation, I calculated the average fitness landscapes at each site over different background sequences. I then use these site-specific average landscapes as the unique and constant landscapes for each site in the S-SI simulations (figure 2.2). Therefore, for each S-SD alignment I generated an analogous S-SI alignment under the same average selection constraints but without the temporal dynamics characteristic of epistasis. Therefore, the S-SI versus S-SD model comparison allows for a novel and direct way of investigating the influence of stability-induced epistasis on evolutionary dynamics. To permit comparison with models that do not account for stability, I include a third independent and identically distributed across sites framework where site-specific fitness landscapes are derived from the C-series frequency profiles (Le *et al.*, 2008); I refer to this as the C-series site-independent (C-SI) model.

The conditions of the simulations are derived from multiple sequence alignments for three natural protein-coding genes with PDB structures 1qhw, 2ppn, and 1pek. The three protein structures differ in important ways. The 2ppn protein folds following a two-state folding process and therefore conforms to one of core thermodynamic model assumptions. The 1qhw structure was used to maintain consistency with previous studies which used the same structure (Goldstein and Pollock, 2016; Pollock *et al.*, 2012; Goldstein and Pollock, 2017). Lastly, the 1pek protein is comparable in length to the 1qhw protein, however, the 1pek protein is more densely packed. I begin by validating the stability-informed models and show that simulated alignments are phenomenologically comparable to the real protein alignments based on various metrics. I then use the S-SI and S-SD models to investigate the difference in dynamics when sites evolve with epistatic interactions or

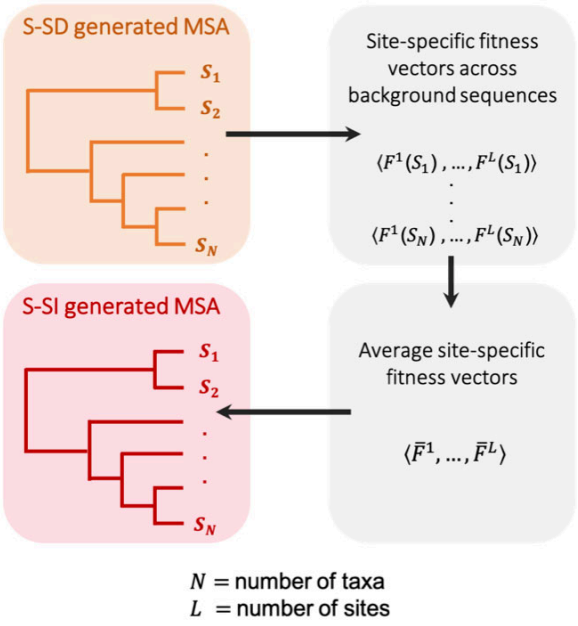


Figure 2.2: Derivation of the stability-informed site-independent (S-SI) model. First, I generated multiple sequence alignment (MSA) under stability-informed site-dependent model (S-SD; see Methods section for details). Then, at each site I calculated $F^h(S) = \langle f_1^h(S), ..., f_{20}^h(S)\rangle$, the site-specific fitness vector where $f_a^h(S)$ is the fitness of amino acid a at site h given background sequence S. This was repeated across all extant sequences $S_1, ..., S_N$. Next, I calculated $\bar{F}^h$ the average fitness landscape at site h across background sequences. I generate under S-SI with $\bar{F}^h$ as the independent and constant fitness landscapes (see Methods section for details). N is the number of taxa in the protein-specific alignment (14, 14, and 12 for proteins 1qhw, 2ppn, and 1pek), and L is the number of sites in the protein-specific alignment (300, 107, and 279 for proteins 1qhw, 2ppn, and 1pek).

independently. I find that epistasis results in minor elevations in substitution rates over the whole protein. However, site-wise analysis reveals that the impact of epistatic interactions on substitution rates can be substantial at individual sites. I describe a mechanism whereby epistasis increases substitution rates compared to the rates under site-independent evolution. Lastly, I report that while models that treat site-wise variation in ω as a random variable underestimate the degree of among-site rate heterogeneity, the estimated ω rates tend to accurately identify the most common substitution rates across sites. Therefore, despite their simplicity, ω-based inference models preformed comparably well in the presence and absence of epistasis.

2.3 Results

The objective of this chapter is to (1) investigate how epistasis influences the rates at which proteins evolve (measured by the number of substitutions), and (2) the impact of epistasis on inference procedures (in particular with regards to ω rate estimation). To this end, I generated sequence alignments using three simulation models: C-series site-independent (C-SI), stability-informed site-independent (S-SI) and stability-informed site-dependent (S-SD). The simulation models differ in how fitness values are calculated (stability-informed, S-, or estimated from C-series profiles, C-) and whether they model sites as evolving independently or with epistatic interaction (-SI vs -SD, respectively).

For three protein-coding genes with known protein structures (PDB code: 1qhw, 1pek, and 2ppn), I obtained multiple sequence alignments (MSA) and estimated a corresponding phylogenetic tree (figure 2.3). I fit the MSA and phylogenetic tree to the M3 (k=3) codon substitution model to estimate protein-specific mutation parameters (table 2.1). Then, I generated fifty protein-specific alignments under C-SI, S-SI, and S-SD, using the corresponding mutation parameters and phylogenetic tree. I fit the real and simulated alignments to ω-based models and inferred the substitution rates. The ω-based codon models use the maximum likelihood framework to estimate rate parameter ω conditioned on a known phylogeny and multiple sequence alignment. Briefly, the M-series ω-based models partition sites into c categories and estimate substitution rates $\omega_1 < .. < \omega_c$, and proportions $p_1, ..., p_c$ (Yang *et al.*, 2000) (the models are described in more detail in the Methods section). The validity of the simulation model can then be assessed by comparing the inferred ω rates from the simulated and real alignments. I also calculated the true rates

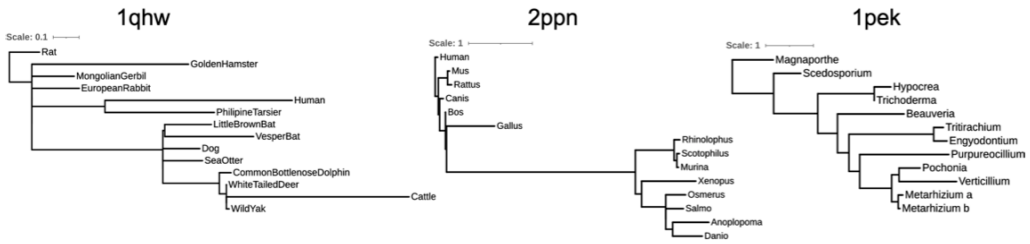


Figure 2.3: Phylogenetic trees for the 1qhw, 2ppn, and 1pek natural protein alignments. The topologies were inferred using IQ-TREE (Nguyen *et al.*, 2014) with ModelFinder (Kalyaanamoorthy *et al.*, 2017) and ultrafast bootstrapping (Minh *et al.*, 2013). Branch lengths, measured as the expected number of single nucleotide substitution per codon site, were inferred from codon model M3(k=3) (Yang *et al.*, 2000). This figure was generated using iTOL (Letunic and Bork, 2021).

(dN/dS) directly from the generating models. Comparison between dN/dS and ω allows us to assess the performance of the commonly used ω models. An outline of the methods is provided in figure 2.4.

Table 2.1: Protein-specific mutation parameters estimated from the natural alignments for proteins 1qhw, 2ppn, and 1pek under ω-based model M3 ($k = 3$).

	1qhw	**2ppn**	**1pek**
κ	4.372	2.503	0.904
π_A	0.205	0.268	0.188
π_C	0.318	0.245	0.346
π_G	0.280	0.294	0.258
π_T	0.197	0.192	0.208
number of taxa	14	14	12
number of sites	300	107	279
Tree length	4.93	8.04	13.88

$\kappa =$ transition-to-transversion ratio

$\pi_n =$ stationary frequency of nucleotide n

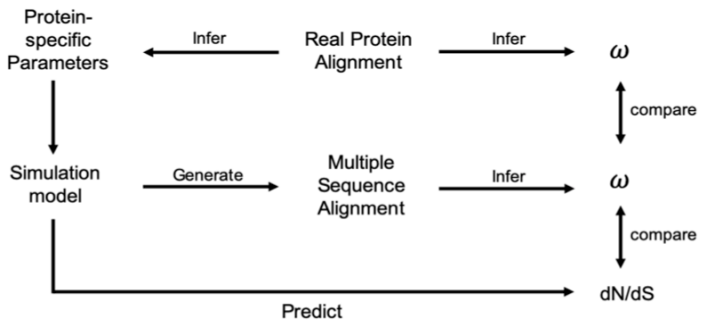


Figure 2.4: Flowchart of method design. Real protein alignments were fitted to M-series models to obtain maximum likelihood estimates of substitution rates (ω) and estimates of protein-specific parameters (phylogeny, κ, π_A, π_C, π_G, and π_T). The protein-specific parameters were then used to generate fifty alignments under each of the simulation models: C-series site-independent (C-SI), stability-informed site-independent (S-SI), and stability-informed site-dependent (S-SD). The validity of the simulation model was assessed by comparing the inferred ω rates from the simulated alignments to the ω estimates from the corresponding real protein alignment. To assess the performance of inference models, expected substitution rates, dN/dS, were calculated directly from the simulation models and compared to the inferred ω values. Diagram modified from Spielman and Wilke (2015).

2.3.1 Stability-informed models generate sequence alignments consistent with real data

2.3.1.1 Evaluating the relationship between substitution rates and structural features

Buried residues, towards the core of the protein, are more densely packed having higher weighted contact number (WCN) and lower relative solvent exposure (RSA) compared to surface residues. Analyses of natural protein alignments often reveal significant correlations between site-specific substitution rates and structural properties such as RSA and WCN: buried sites tend to be more conserved with lower substitution rates compared to exposed sites (Yeh *et al.*, 2014; Shahmoradi *et al.*, 2014; Marcos and Echave, 2015; Echave *et al.*, 2015). I was interested in assessing if any of the generative models recapitulate this phenomenon. I measured the expected site-specific substitution rate (dN^h/dS^h) directly from the fitness landscapes using equation (2.7) for C-SI and S-SI, and equation (2.8) for the S-SD. I refer to dN^h/dS^h as the expected substitution rate throughout the study since it represents the theoretically predicted substitution rate at evolutionary equilibrium (Spielman and Wilke, 2015).

Under both stability-informed frameworks (S-SI and S-SD) a significant positive correlation was found between RSA and dN^h/dS^h, and a significant negative correlation was found between WCN and dN^h/dS^h (figure 2.5). The correlations between RSA and dN^h/dS^h were slightly higher for rates predicted under the S-SD framework compared to the correlations based on the S-SI simulations. Similarly, correlations between WCN and rates predicted under the S-SD were more negative compared to rates predicted under S-SI. In contrast, the site-specific rates expected under the C-SI framework did not correlate significantly with RSA or WCN.

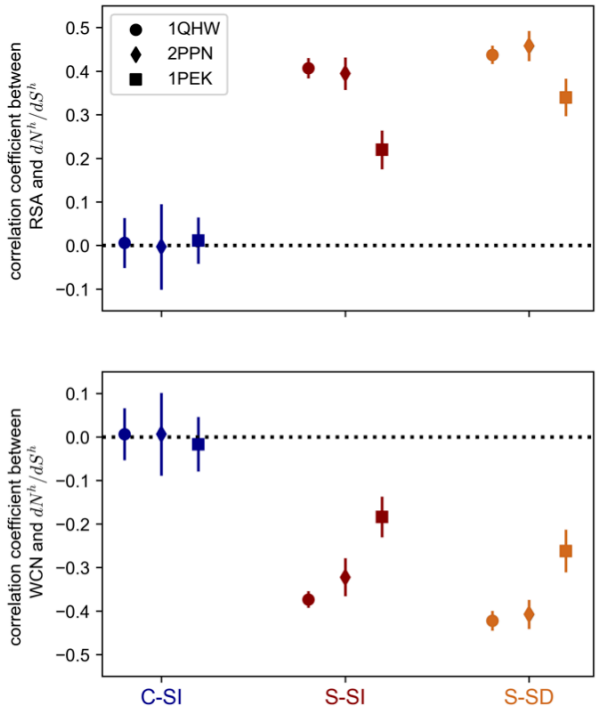


Figure 2.5: Stability-informed models (S-SI and S-SD) reproduce empirically observed correlations between substitution rates and structural features. Fifty alignments were generated with three protein-specific parameters (1qhw, 2ppn, 1pek) under models C-series site-independent (C-SI), stability-informed site-independent (S-SI), and stability-informed site-dependent (S-SD). For each alignment I calculated the Pearson correlation between the expected site-specific substitution rates dN^h/dS^h and relative solvent accessibility (RSA, top panel), and weighted contact number (WCN, bottom panel). Plotted are the mean correlation coefficients (and standard deviation) across trials.

Since the true substitution rates are unknown for the natural proteins, I used traditional

codon models to infer substitution rates ω, measured as the normalized ratio of nonsynonymous to synonymous substitutions. In order to assess the correlation between RSA or WCN and substitution rates, I used the posterior mean ω^h from the best fitting M-series model as the site-specific rate estimate. The posterior mean ω^h at a site is calculated as $(\omega_1 \times P_1^h) + (\omega_2 \times P_2^h) + ... + (\omega_k \times P_c^h)$, where P_c^h is the posterior probability of the site corresponding to rate class ω_c. I found a significant positive correlation between posterior mean ω^h and RSA in the 1qhw and 1pek real protein alignments (correlation coefficient was 0.39 and 0.53 respectively; both p-values < 0.001); and a significant negative correlation between rates and WCN (correlation coefficient was -0.35 and -0.43 for the 1qhw and 1pek alignments respectively; both p-values < 0.001). I found no significant correlation between rates and structural properties (RSA or WCN) for the 2ppn alignment. The small size of the 2ppn gene, and the unusual mixture of long and short edges in its phylogeny (figure 2.3), is likely problematic for posterior estimation of ω, which could explain the insignificant correlations.

Various alternative methods have been developed to infer site-specific substitution rates from multiple sequence alignments ($e.g.$ Kosakovsky Pond and Frost (2005); Meyes and vonHaeseler (2003); Massingham and Goldman (2005); Murrell *et al.* (2012)). However, the estimated rates are subject to large variability when the number of taxa is relatively small. These methods are therefore not suitable to infer site-specific rates for the alignments used here (number of taxa = 14, 14, and 12 for 1qhw, 2ppn and 1pek). Using large alignments (number of taxa = 300) of more than 200 proteins, Marcos and Echave (2015) estimated the correlations between rates and RSA, and between rates and WCN. The range of correlations coefficients between RSA and site-specific rates was between 0.26 and 0.75; the range of correlation coefficients between WCN and site-specific rates was -0.19 and -0.73. The correlation coefficients I report for both rate measures (dN^h/dS^h and posterior mean ω^h) are within the range reported in Marcos and Echave (2015). Overall, I found that the stability-informed models are able to recapitulate the empirically observed correlations between structural properties and rates, which suggests that accounting for folding stability captures important structural features that are absent in the stability-naïve C-SI framework derived from the widely used C-series profiles.

2.3.1.2 Comparing inferred substitution rates and sequence variability between real and simulated data

In order to use the simulations as a means of investigating the influence of epistasis on rates, I needed to first verify that the generative models produce plausible substitution rates. In other words, I needed to compare substitution rates from the generative models to the rates experienced by real proteins. I fitted simulated and real alignments to codon model M3($k = 2$) to obtain estimates of substitution rates. A value of $\omega \approx 1$ is indicative of neutral or nearly neutral evolution where nonsynonymous mutations are fixed at an equal rate to synonymous mutations. An ω value < 1 is representative of purifying selection, and $\omega > 1$ is indicative of positive selection.

Analyses of the natural 1qhw, 2ppn and 1pek alignments revealed evidence for purifying selection with $\omega_1 < \omega_2 < 1$ for all three natural alignments (figure 2.6). The 2ppn protein alignment had the lowest rate estimates with $\omega_1 = 0.00$ and $\omega_2 = 0.09$, and respective proportions $p_1 = 0.67$ and $p_2 = 0.33$. The 1qhw and 1pek alignments had comparable rate estimates with $\omega_1 =$ 0.01 and 0.02, and $\omega_2 =$ 0.30 and 0.24, respectively; however, the proportion of sites belonging to the more stringent selection regime (ω_1) was approximately 10% higher for the 1qhw alignment ($p_1 = 0.71$) compared to the 1pek alignment ($p_1 = 0.64$).

Alignments generated under the stability-informed models (S-SI and S-SD) were also consistent with purifying selection, with $\omega_1 < \omega_2 < 1$ for all simulated protein-specific alignments (figure 2.6, first row). The ω values inferred from the S-SI generated alignments were on average significantly lower than rates estimated from the analogous protein-specific S-SD simulations, and more consistent with the ω values estimated from the natural protein alignments (figure 2.6; Bonferroni corrected p-values < 0.001 for all comparisons). With the exception of the 1pek protein, the natural alignments were consistently inferred to be under more stringent selection regimes with slightly lower substitution rates. For the 1pek simulations, the ω_2 estimate from the real alignment (ω_2 = 0.24) alignment was higher than the distribution of estimates from the S-SI alignments (figure 2.6, first row). Nonetheless, the proportion of quickly evolving sites (p_2) was lower in the real alignment (figure 2.6, second row). This suggests that in the real 1pek protein a small proportion of sites were evolving faster than expected under stability constraints. However, when considering all sites in the alignment, by comparing the single rate estimated under M0, I found that the rates were largely comparable: ω was 0.06 for the real 1pek alignment and the mean

ω estimate over the fifty S-SI trials was 0.07. In contrast, rates inferred from the C-SI simulations were significantly higher than estimates from the other simulations and from the real proteins (Bonferroni corrected p-values < 0.001 for all comparisons). For the C-SI generated alignments, the ω estimates were suggestive of neutral or weak selection regimes (figure 2.6, first row).

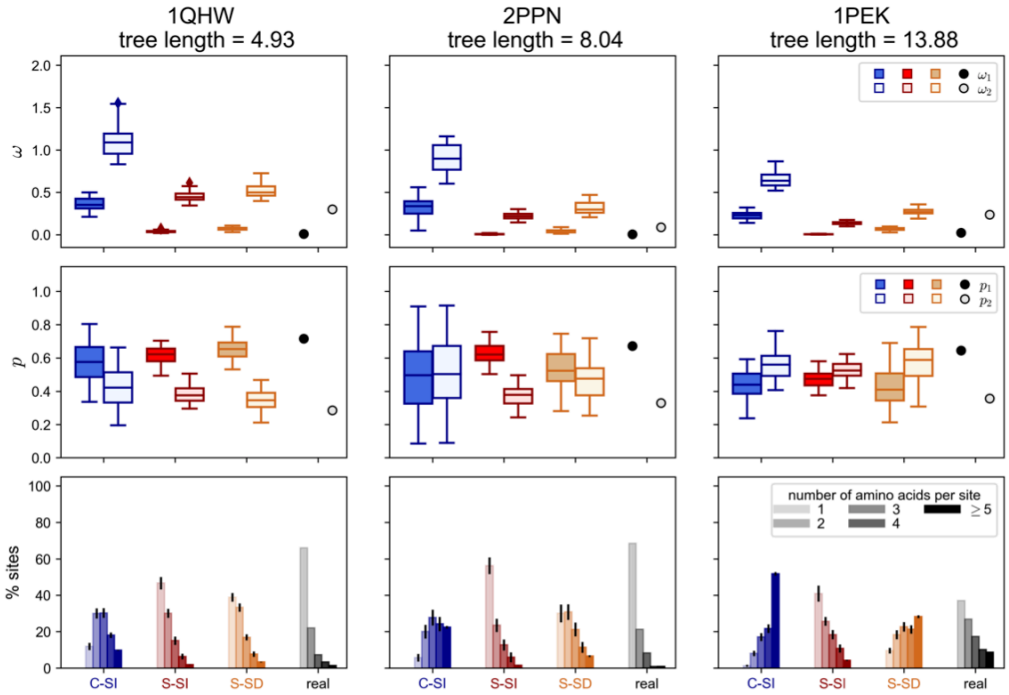


Figure 2.6: Stability informed models (S-SI and S-SD) generate alignments consistent with real data with respect to substitution rates and amino acid variability. For each of three natural protein (1qhw, 2ppn, 1pek corresponding to the three columns), I generated fifty protein-specific alignments under simulation models C-series site-independent (C-SI), stability-informed site-independent (S-SI), and stability-informed site-dependent (S-SD). The first row reports the estimated substitution rates ω_1 (dark) and ω_2 (light) inferred from M3($k = 2$). The ω distributions are of the fifty model- and protein-specific aignments; the dots are the estimates from the real protein alignments. The second row reports the proportion of sites in each rate category, p_1 (dark) and p_2 (light). The third row plots the distributions of the number of amino acids observed per alignment site.

Consistent with having the highest ω rate estimates, the C-SI generated alignments were the most variable with regards to the number of amino acids observed per site (figure 2.6, third row). Across the three protein-specific simulations, the proportion

of fully conserved sites (one amino acid per site) were significantly lower than those observed from the stability-informed simulations (Bonferroni corrected p-values < 0.001 for all comparisons). Furthermore, the average fraction of sites with ≥ 5 amino acids were significantly higher. While the S-SD generated alignments were more conserved than the analogous C-SI simulation, the alignments were more variable compared to the corresponding S-SI simulations and real alignment. For the 1qhw and 2ppn alignments generated under S-SD, the distributions of the number of amino acids per site were largely consistent with the corresponding real protein alignment; however, the 1pek-specific S-SD simulations were strikingly more variable (figure 2.6, third row). This is consistent with results from Goldstein *et al.* (2015) which showed that under the S-SD model, the number of amino acids per site is expected to increase with tree length (branch lengths are measured as the expected number of single nucleotide substitutions per codon site). In general, I found that the S-SI simulations were the most consistent with the real alignments. In both the S-SI simulated alignments and the natural alignments (1) the most common site pattern included only one amino acid for all protein alignments, and (2) the 2ppn proteins were the most conserved compared to the 1qhw and 1pek proteins. The number of amino acids per site was on average slightly more conserved for the real alignments than the S-SI simulations which is consistent with the natural proteins being subject to additional selective constraints beyond folding stability.

2.3.2 Epistasis increases substitution rates compared to site independent evolution

2.3.2.1 Comparing expected substitution rates in the presence and absence of epistatic interactions

Values of ω estimated from the S-SD alignments were on average higher than estimates from the S-SI simulations (figure 2.6, first row). This suggests that epistasis, as modelled in the S-SD framework, might lead to an increase in substitution rates compared to site-independent evolution. However, it remains unclear if the observed increase in rates is a genuine outcome of epistasis or a consequence of inference model misspecification. To address this, I compared the expected site-specific substitution rates calculated directly from the S-SI and S-SD generating frameworks. Consistent with the finding that epistasis increased the inferred substitution rates, the distributions of expected dN^h/dS^h were more positively skewed (higher) when epistasis was included (S-SD) for all three protein-specific

simulations compared to the rates expected had sites evolved independently (S-SI; figure 2.7).

Rate distributions predicted from the S-SI model often displayed three peaks at dN^h/dS^h values representative of highly stringent selection regimes ($dN^h/dS^h \approx 0.00$), moderate selection pressures ($dN^h/dS^h \approx 0.25$), and more relaxed selection ($dN^h/dS^h \approx 0.4$). The position of the peaks differed only slightly depending on the protein-specific simulation (figure 2.7, second row). Rate distributions estimated from S-SD were bimodal with considerably fewer sites under highly stringent selection ($dN^h/dS^h \approx 0$) compared to the analogous S-SI protein-specific distribution (figure 2.7). Furthermore, more sites were under weak selection pressures under S-SD compared to S-SI; the percentage of sites with $dN^h/dS^h > 0.5$ under (S-SI, S-SD) were (8.5%, 17.2%), (2.9%, 4.2%) and (3.9%, 10.8%) for the 1qhw, 2ppn, and 1pek simulations respectively.

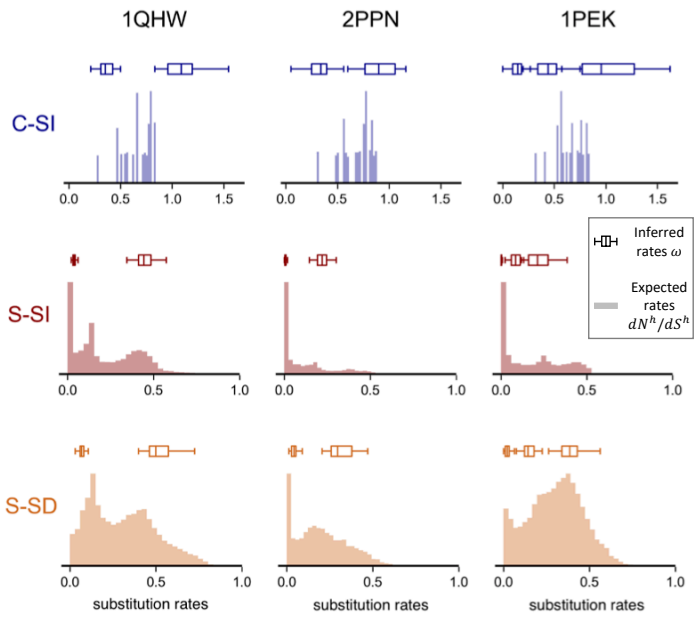


Figure 2.7: M-series inference models capture the most common substitution rates across sites. Histograms represent the distributions of expected site-specific substitution rates, dN^h/dS^h, calculated from simulation models C-SI, S-SI, and S-SD (row) with protein-specific parameters (column). The boxplots represent the distribution of maximum likelihood rate estimates, $\omega_1 < \omega_2$, under M3 ($k = 2$) for proteins 1qhw and 2ppn, and M3 ($k = 3$) for protein 1pek ($\omega_1 < \omega_2 < \omega_3$). Note the difference in x-axis range in the top row (0.0 to 1.5) and the bottom rows (0.0 to 1.0).

An advantage of the S-SI and S-SD frameworks is that for each site evolving with epistatic dependencies (under the temporally-dynamic S-SD), we are able to model an analogous site evolving independently and under the same average stability restrictions (under the time-homogenous S-SI). To assess the magnitude of the effect of epistasis on evolutionary rates, I calculated the difference in substitution rates under epistasis (S-SD) and site-independence (S-SI). Averaged over all sites in the alignment, the mean differences in rates were 0.07, 0.08 and 0.11 for the 1qhw, 2ppn, and 1pek simulations respectively, implying that across the whole protein epistasis had a modest effect on substitution rates. However, site-wise analyses of rate differences revealed that epistasis increased the expected substitution rate at 88.8%, 89.5%, and 84.3% of sites in the 1qhw,

2ppn, and 1pek simulations. The largest differences in dN^h/dS^h rates were observed at sites subject to stringent selection regimes under site-independence ($dN^h/dS^h < 0.2$, figure 2.8). The less frequent and more minor reductions in rates due to epistasis occurred at sites evolving close to neutrality with $dN^h/dS^h \approx 1$ under site-independence.

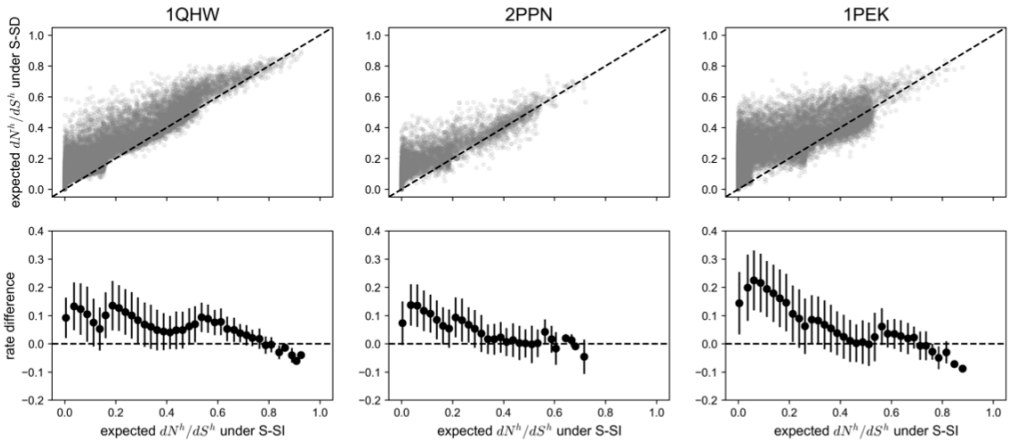


Figure 2.8: Epistasis results in an increase in the expected substitution rate at a site, dN^h/dS^h, compared to the expectation under site-independent evolution. Analysis was completed for three protein structures: 1qhw, 2ppn, and 1pek (columns). Top panels show the relationship between dN^h/dS^h under a stability-informed site-independent (S-SI) model (rates calculated using equation (2.7)) and a stability-informed site-dependent (S-SD) model (rates calculated using equation (2.8)). Epistasis increased substitution rates at 88.8%, 89.5%, and 84.3% of sites in the 1qhw, 2ppn, and 1pek proteins. Bottom panels show the difference in dN^h/dS^h under S-SD compared to the rate under S-SI. Positive values indicate that rates are expected to be higher when epistatic interactions are included. The mean differences in rates were 0.07, 0.08 and 0.11 for the 1qhw, 2ppn, and 1pek simulations respectively.

2.3.2.2 Evaluating the relationship between epistatic sensitivity and structural features

The previous result suggests that epistasis has a variable impact across sites. I was therefore interested in assessing the properties which made a site more or less sensitive to epistasis. To do this, I calculated a site's "epistatic sensitivity" by measuring the variability in the expected substitution rate given different background sequences. Since the vast majority of randomly generated sequences have zero probability of folding correctly, I used the sequences from the S-SD protein-specific alignments as the set of possible background sequences. Therefore, the number of background sequences was $50 \times N$, where $N = \{14, 14, 12\}$ is the number of taxa for the 1qhw, 2ppn, and 1pek simulations respectively.

If the substitution rate at a site was minimally influenced by the background sequence, then I expect little variation in dN^h/dS^h values. Alternatively, if the rate at a site was heavily influenced by the residues present at other positions, I expect higher variance in the dN^h/dS^h values depending on the background protein sequence. I found that the degree of epistatic sensitivity correlated significantly with structural properties, specifically relative solvent accessibility (RSA) and weighted contact number (WCN). The correlation coefficient between RSA and epistatic sensitivity was 0.34, 0.39 , and 0.32 (all p-values $<$ 0.001) for the 1qhw, 2ppn, and 1pek protein structures. Similarly, a significant correlation was observed between WCN and epistatic sensitivity with a correlation coefficient of -0.38, -0.42, and -0.22 for the 1qhw, 2ppn, and 1pek protein structures, respectively (all p-values < 0.001). The relationship between epistatic sensitivity and the number of contacts is shown in figure 2.9. Therefore, the results suggest that sites near the core of the protein structure, with low solvent exposure (RSA) and high packing density (WCN), were more robust to changes in the background protein sequence compared to solvent-exposed residues (high RSA and low WCN).

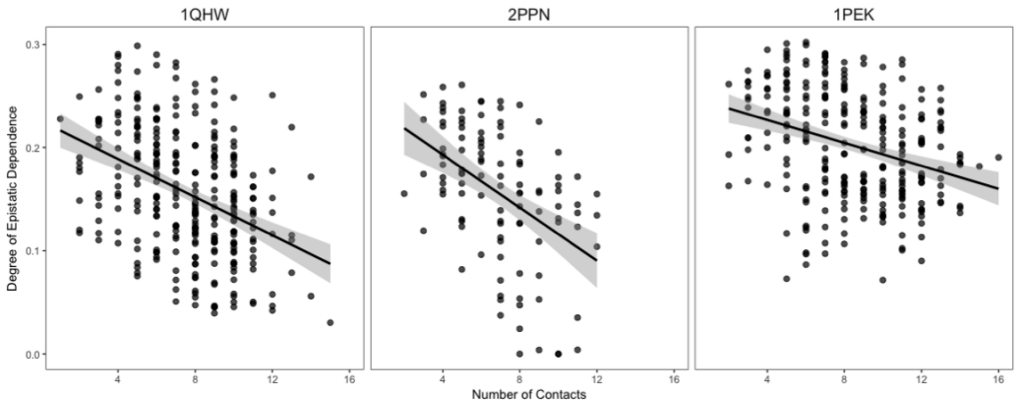


Figure 2.9: Relationship between epistatic sensitivity and number of contacts. Epistasis sensitivity is measured as the standard deviation in expected rates (dN^h/dS^h) across all $50 \times N$ background sequences, where N is the number of taxa = 14, 14, and 12 for the 1qhw, 2ppn, and 1pek proteins respectively. The lines represent a linear regression and the shaded area the 95% confidence interval.

The observation that highly connected sites are less influenced by epistasis may initially appear counterintuitive. However, consider a highly connected site at which the fitness landscape needs to be compatible with the amino acid residues present at

several interacting positions. A change at a few of the many neighboring amino acids has negligible affect on a fitness landscape that is otherwise highly constrained by its many contacts; hence there are minimal impacts on dN^h/dS^h values. I illustrate this using a buried site and an exposed site in the 1qhw protein (figure 2.10A). For buried site 41 (RSA = 0.01 and WCN = 1.27), the standard deviation in dN^h/dS^h was 0.04 across all 50×14 background sequences. The fitness landscape at site 41 given four background sequences with increasing divergence levels are plotted in figure 2.10B (top panels). Amino acid isoleucine (I) was consistently the fittest at site 41, followed by amino acids valine (V) and leucine (L) across the different background sequences. At equilibrium, the site will primarily be occupied by the optimal amino acid (I) and most nonsynonymous mutations will be deleterious resulting in a low dN^h/dS^h as expected given the correlations between RSA (or WCN) and dN^h/dS^h (figure 2.5). By contrast, consider a surface site which tends to have fewer contacts. A substitution at one of the few interacting positions is more likely to induce a larger shift in amino acid preferences and consequently alter the expected substitution rate. This is illustrated in the bottom panels of figure 2.10B, which show the fitness landscapes at surface site 73 of the 1qhw protein (RSA = 0.82, WCN = 0.79, standard deviation in dN^h/dS^h = 0.11).

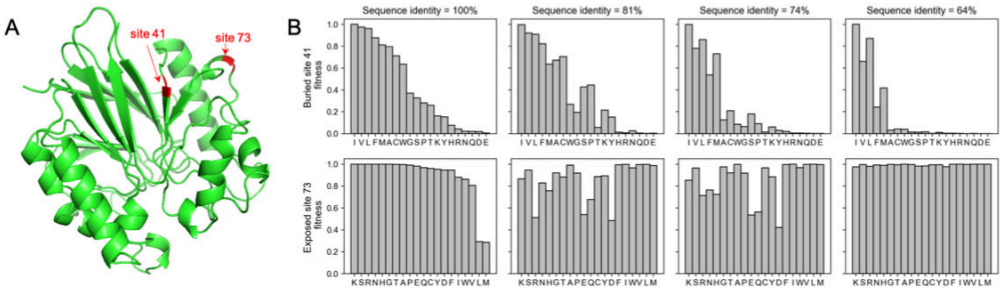


Figure 2.10: Buried sites are more robust to changes in the background protein sequence compared to exposed sites. (A) The structure of the 1qhw protein. Arrows indicate the location of buried site 41 (RSA = 0.01 and WCN = 1.27) and exposed site 73 (RSA = 0.82 and WCN = 0.79). (B) The fitness landscapes at buried site 41 (top panels) and exposed site 73 (bottom panels) given different background sequences (columns). The reported sequence identities are in reference to the background sequence used to determine the landscapes in the left-most column.

2.3.3 Traditional ω-based codon substitution models perform well despite their site-independence assumption

2.3.3.1 Assessing the accuracy of substitution rate inference under M-series codon models

I have thus far shown that epistasis impacts substitution rates; however, traditional codon models used to infer selection pressures assume that sites evolve independently. Does neglecting to account for epistasis bias inference from traditional ω-based models? Furthermore, ω-based models assume that a small number of rate categories is sufficient to account for the among-site rate heterogeneity. It is therefore important to compare errors in estimation due to epistasis to the baseline estimation errors arising from unmodelled variability in rates across sites. Comparing the inferred substitution rates (ω) from the S-SI simulations to the theoretical rate expectations dN^h/dS^h, allows us to assess the inference of rates in the presence of among-site rate heterogeneity but without temporal changes in rate due to epistasis. The S-SD simulations allow us to assess the performance of ω-based models in the presence of among-site rate heterogeneity and epistasis.

First, I used the M3(k) versus M3($k+1$) likelihood ratio test to determine the number of significant rate categories from each alignment (table 2.2). Three factors influence the number of significant rate categories: simulation model, protein length, and tree length. Within each protein-specific simulation, I found that the C-SI alignments had the lowest number of significant tests for three rate categories compared to S-SI and S-SD simulations. This is perhaps expected since the C-SI simulations had less heterogeneity in rates across sites compared to the stability informed models. Each C-SI alignment had at most 20 unique rate categories, whereas under S-SI and S-SD each site had a unique fitness landscape(s) (see Methods section for details). Second, within each generating framework, the 2ppn-specific simulations had the lowest number of significant results for three rate categories. The 2ppn alignments were much smaller with only 107 codon sites compared to the 1qhw (300 codon sites) and 1pek (279 codon sites) alignments. This suggests that there is less power to detect additional rate components with fewer sites. Lastly, despite similar numbers of codon sites, a larger number of the 1pek-specific simulations displayed significant evidence for three rate categories compared to the 1qhw-specific simulations. There are two potential reasons for this observation: (1) the number of rate categories is influenced by the protein structure such that the 1pek contact map induces more variation in rates across sites compared to the 1qhw structure; or (2) there is more power to identify

rate heterogeneity with deeper trees (1pek tree length = 13.88, 1qhw tree length = 4.93). To distinguish between these two possibilities, I conducted an additional experiment: I generated 1qhw-specific alignments under the three generative frameworks (C-SI, S-SI, and S-SD) along the 1qhw phylogeny with double the branch length (blx2, table 2.2) and 1qhw-specific mutation parameters (table 2.1). From these additional simulations, I found an increase in detection of three rate categories across all generative models. More importantly, the number of significant tests for three rate categories were now comparable to those from the 1pek-specific simulations (table 2.2). These results support the notion that deeper trees provide more informative site patterns for the detection of among-site rate heterogeneity.

Table 2.2: Model contrasts for real and simulated alignments from three proteins (1qhw, 2ppn, and 1pek). The 1qhw blx2 results are from simulations on the 1qhw tree with double the branch length. Reported are the number of alignments out of fifty for which the specified likelihood ratio test was significant. Alignments were generated under simulation models C-series site-independent (C-SI), stability site-independent (S-SI), and stability site-dependent (S-SD). The mean total tree lengths from M3($k = 3$) are also reported.

Model Contrast	**1qhw**	**1qhw blx2**	**2ppn**	**1pek**
Real				
M0 vs M3($k = 2$)	yes	–	yes	yes
M3($k = 2$) vs M3($k = 3$)	yes	–	no	yes
M3($k = 3$) vs M3($k = 4$)	no	–	no	no
M3($k = 2$) vs CLM3	yes	–	no	yes
BUSTED($\omega_3 < 1$) vs BUSTED	no	–	no	yes
Tree length	4.93	–	8.04	13.88
C-SI				
M0 vs M3($k = 2$)	50	50	50	50
M3($k = 2$) vs M3($k = 3$)	6	19	1	28
M3($k = 3$) vs M3($k = 4$)	0	0	0	3
M3($k = 2$) vs CLM3	7	30	17	33
BUSTED($\omega_3 < 1$) vs BUSTED	0	0	0	0
Mean tree length	5.27	10.48	7.55	13.32
S-SI				
M0 vs M3($k = 2$)	50	50	50	50
M3($k = 2$) vs M3($k = 3$)	21	42	7	39
M3($k = 3$) vs M3($k = 4$)	0	15	0	3
M3($k = 2$) vs CLM3	10	23	14	22
BUSTED($\omega_3 < 1$) vs BUSTED	0	0	0	0
Mean tree length	4.99	9.35	7.15	12.45
S-SD				
M0 vs M3($k = 2$)	50	50	50	50
M3($k = 2$) vs M3($k = 3$)	15	42	16	43
M3($k = 3$) vs M3($k = 4$)	2	0	0	4
M3($k = 2$) vs CLM3	25	47	35	50
BUSTED($\omega_3 < 1$) vs BUSTED	0	0	1	0
Mean tree length	5.04	9.65	7.57	14.18

Overall, I found that the number of rate categories inferred using the M3(k)–M3($k+1$) likelihood ratio test were consistent with the number of peaks observed in the corresponding dN^h/dS^h distribution. I next asked whether the inferred substitution rates (ω) corresponded to the expected rates (dN^h/dS^h). For the 1qhw and 2ppn specific simulation, two rate categories were most commonly detected in the S-SI simulations. The first rate

category was reflective of the sites subject to highly stringent selection regimes with low substitution rates ($\omega_1 \approx 0$). The second rate category often took on values representative of the average of the tail of the dN^h/dS^h distribution (figure 2.7, second row). For the S-SD simulations, the inferred ω values were consistent with the most common rates with ω_1 values comparable to the first peak in the dN^h/dS^h distribution and ω_2 approximating the second peak (figure 2.7, third row). More than half of the 1pek-simulated alignments showed significant evidence for three rate categories; 28/50, 39/50, and 43/50 under C-SI, S-SI and S-SD respectively (table 2.2). Consequently, for the 1pek simulations, I compared the distributions of expected dN^h/dS^h rates to the ω_1, ω_2, and ω_3 distributions estimated under M3($k = 3$), and found that the rates inferred using traditional codon models tended to capture the most common rate categories ($i.e.$, the distribution of ω values corresponded to peaks in the dN^h/dS^h distributions, figure 2.7). Therefore, in the presence and absence of epistasis, the ω estimates were consistent with the most common rate expectations.

The distributions of dN^h/dS^h under S-SD and S-SI are rich distributions showing variation like that of a continuous distribution (figure 2.7). Due to computational limitations (related to use of the pruning algorithm), ω-based models can only approximate these distributions discretely. Some care is thus required in defining the target of ω-based model estimation. I assessed the performance of ω-based models in two additional ways. First, I looked at the correlations between expected site-specific rates (dN^h/dS^h) and the posterior mean ω^h inferred based on the best fitting M-series model. For rates calculated based on the stability-informed models (S-SI and S-SD) the correlations were significant in all fifty model- and protein-specific trials (table 2.3).

Table 2.3: Pearson correlations between expected site-specific substitution rates (dN^h/dS^h) and inferred site-specific rates (posterior mean ω^h). Reported are the number of simulated alignments (50 total) for which the correlation was significant.

	1qhw		**2ppn**		**1pek**	
	#sig	mean r	#sig	mean r	#sig	mean r
C-SI	49	0.24	28	0.30	49	0.39
S-SI	50	0.68	50	0.75	50	0.79
S-SD	50	0.67	50	0.70	50	0.74

Second, under M3($k = 2$), ω_c is interpretable as the substitution rate averaged over

time and across sites belonging to the rate class $c = 1$ or 2. Therefore, a potential way of addressing the performance of M3($k = 2$) is by resolving sites according to the posterior probability of belonging to rate class (P_c) and calculating the average expected rate as

$$dN_c/dS_c = 1/n \sum_h P_c^h dN^h/dS^h \tag{2.1}$$

I compared the expected dN_c/dS_c to the inferred ω_c values for respective rate class c; the relative error in rate estimates are plotted in figure 2.11. As expected, the errors were lowest for alignments generated under C-SI, since the generating model was the most consistent with inference model assumptions (rates under C-SI are independent and identically distributed). Nonetheless, the ω_1 values were often underestimated. Based on the results of Spielman and Wilke (2015), I suspect that the underestimation is due to the asymmetry in the mutation models ($\mu_{ij} \neq \mu_{ji}$) present in all protein-specific simulations (table 2.1). Importantly, and consistent with results from figure 2.7, the relative error in ω estimates were comparable across S-SI and S-SD simulations. This supports the previous conclusion that the performance of ω-based models is somewhat robust to epistatic effects.

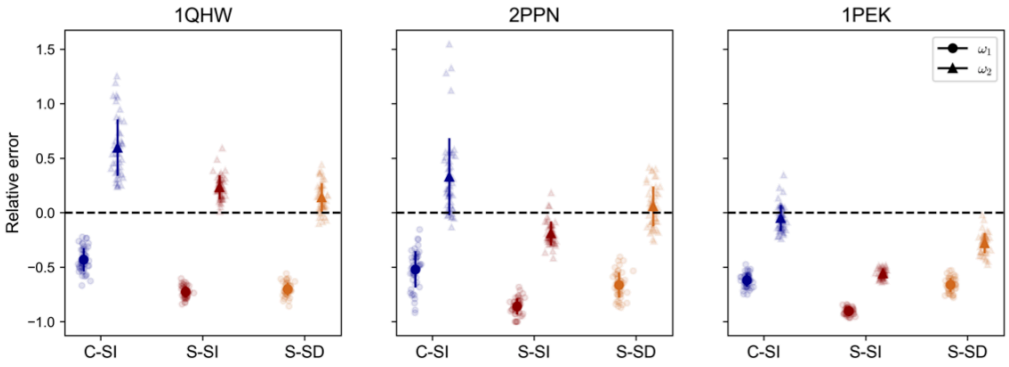


Figure 2.11: The accuracy of rate estimation under M-series model is comparable when alignments are generated with and without epistasis. Plotted is the relative error ($\frac{\omega_c}{dN_c/dS_c} - 1$) in rate estimation under M3 (k=2) for alignments generated under C-series site-independent model (C-SI), stability-informed site-independent (S-SI), and stability-informed site-dependent (S-SD) for each of the three proteins (1qhw, 2ppn, and 1pek). The lighter points represent the relative error from each of the fifty model- and protein-specific trials. The darker points are the average values across trials and the bars are the standard deviation.

2.3.3.2 Detecting temporal fluctuations in substitution rates and positive selection

By framing the S-SI and S-SD models within the MutSel framework, differences in site-wise evolutionary dynamics between the site-independence assumption and epistatic evolution become apparent. Under the traditional site-wise MutSel framework, the substitution process is modelled independently at each position and hence the fitness effect of a mutation is not influenced by the background protein sequence with fixed site-specific fitness landscapes (figure 2.1B). Shifts in fitness landscapes (non-stationary dynamics) are interpreted as evidence of adaptive events where external changes in environment or gene function result in changes in the amino acid preferences at the site (dos Reis, 2015; Jones *et al.*, 2017). However, if a site is subject to epistatic interactions, the site-specific fitness landscape, and hence the expected substitution rate at the site, are influenced by the residues present at other positions. Epistasis, therefore, implies a nonstationary substitution process over time such that the fitness landscape at a site constantly changes because of substitutions at other positions (figure 2.1A), even when there are no adaptive events.

I was therefore interested in assessing whether traditional ω-based inference models are able to detect temporal rate fluctuations due to epistasis. However, it is important to note that using the MutSel framework, Jones *et al.* (2017) previously observed that site-independent evolution can result in a detectable signal for temporal variation in substitution rates (at evolutionary equilibrium) by a process reminiscent of Wright's nonadaptive phase of shifting. This occurs when a site accepts a mutation due to drift to a suboptimal amino acid which is then followed by a transient period of higher rates of nonsynonymous fixations as the site evolves towards the peak of the landscape. Additionally, they found that these dynamics can result in site patterns consistent with positive selection when tested using the BUSTED($\omega_3 < 1$) versus BUSTED likelihood ratio test. It is therefore important to compare the results due to epistasis to the baseline detection rates expected due to nonadaptive shifting balance.

I used the M3($k = 2$) versus CLM3 model comparison to test for temporal variations in rates. M3($k = 2$) serves as the null model whereas the covarion-like CLM3 accounts for temporal switches between ω_1 and ω_2 by estimating a δ parameter interpretable as the expected number of rate switches per substitution. I found that the number of significant tests for temporal rate shifts was mainly influenced by two factors: the tree length and the generative model. Consistent with the results reported in Jones *et al.* (2017), I found that

the number of trials for which CLM3 was the better fitting model increased with tree length (table 2.2), this was true for all generative models and all protein-specific simulations. In regard to the generative model, within each set of protein-specific simulations, the number of trials with evidence for temporal switching was highest for the S-SD simulations compared to alignments generated from the site-independent frameworks (C-SI and S-SI). Furthermore, δ was estimated to be at least two times higher in the S-SD simulations compared to the S-SI simulations (table 2.4). For the 1qhw, 2ppn, and 1pek simulated alignments, δ was estimated to be (0.062, 0.148), (0.046, 0.182), and (0.031, 0.106) when simulated under (S-SI, S-SD). These results suggest that temporal variations in rates due to stability-induced epistasis produce a detectable signal in excess of the baseline signal expected due to nonadaptive shifting balance on static fitness landscapes.

Table 2.4: Mean maximum likelihood estimate (MLE) under CLM3 from fifty simulated alignments under models (C-SI, S-SI, or S-SD) with protein-specific parameters (1qhw, 2ppn, or 1pek)

Simulation model	**1qhw** Mean MLE	**2ppn** Mean MLE	**1pek** Mean MLE
C-SI	$\omega_1 = 0.268, \omega_2 = 0.983$	$\omega_1 = 0.321, \omega_2 = 5.458$	$\omega_1 = 0.232, \omega_2 = 2.743$
	$p_1 = 0.440, \delta = 0.385$	$p_1 = 0.579, \delta = 0.267$	$p_1 = 0.449, \delta = 0.159$
S-SI	$\omega_1 = 0.028, \omega_2 = 0.449$	$\omega_1 = 0.006, \omega_2 = 0.290$	$\omega_1 = 0.004, \omega_2 = 0.181$
	$p_1 = 0.589, \delta = 0.062$	$p_1 = 0.649, \delta = 0.046$	$p_1 = 0.489, \delta = 0.031$
S-SD	$\omega_1 = 0.052, \omega_2 = 0.520$	$\omega_1 = 0.024, \omega_2 = 0.424$	$\omega_1 = 0.033, \omega_2 = 0.314$
	$p_1 = 0.635, \delta = 0.148$	$p_1 = 0.587, \delta = 0.182$	$p_1 = 0.363, \delta = 0.106$

Surprisingly, none of the simulated alignments showed significant evidence of positive selection using the BUSTED($\omega_3 < 1$)–BUSTED likelihood ratio test, with the exception of only 1/50 S-SD generated alignments with 2ppn-specific parameters (table 2.2). This is in contrast with previous results where nonadaptive shifting balance produced evidence of positive selection in up to 40% of trials (Jones *et al.*, 2017). This suggests that shifting balance dynamics can be sufficiently different when fitness landscapes are informed by stability constraints rather than being randomly drawn from a normal distribution.

2.4 Discussion

I have examined the influence of stability-induced epistasis on expected and inferred substitution rates, and assessed the accuracy of rate estimation from traditional ω-based models. I found that epistasis resulted in minor elevations in substitution rates considering sites across the whole protein. However, the impact of epistasis on site-specific dynamics was prominent. A site evolving with epistatic effects on fitness had higher substitution rates compared to an analogous site evolving independently and under the same average stability constraints. Under site-independence, theory predicts that purifying selection will maintain the site on or near the fitness optima of the fixed fitness landscape ($i.e.$, the site will predominantly be occupied by the optimal amino acid). Most nonsynonymous mutations will be deleterious and are eliminated from the population resulting in low rates of nonsynonymous substitutions relative to the rates of synonymous substitutions (low dN^h/dS^h). In comparison, consider an epistatic site h and suppose that the site is occupied by the fittest residue, a, given the current background sequence S. Following a substitution at another position in the protein (so that the background sequence changes from S to X), the fitness landscape at site h will change (figure 2.1B). If the change maintains a as the fittest residue, then the substitution rate will remain low. On the other hand, if the change in landscape renders amino acid a suboptimal, then over some period of time the site will be occupied by a suboptimal amino acid. Therefore, the change in fitness landscape induces a change in the amino acid equilibrium frequencies. Since the expected substitution rate, dN^h/dS^h, is a function of the equilibrium frequencies (equation 2.7 and 2.8), and since epistatic sites are more likely to be occupied by suboptimal amino acid, the expected substitution rate will consequently be higher compared to site-independence. In other words, in the presence of epistasis, sites must constantly adapt to amino acid replacements occurring at other positions in the protein which results in higher substitution rates.

The observation that epistasis increased substitution rates contrasts with previous results discussed in Rodrigue and Lartillot (2017), which found that epistasis most often decreased substitution rates compared to site-independence. The discrepancy between the results presented here and theirs is likely because of differences in the way epistatic interactions are modelled and because of differences in expectations of what the rate would have been under site-independent evolution. Rodrigue and Lartillot (2017) model

epistasis as random deviations from multiplicative fitness, and consider the effect of an epistatic landscape by comparison with a randomly assigned fixed fitness landscape. Here, I implicitly model epistasis as a by-product of protein stability, and I compare the rates from a model that accounts for protein stability but no epistasis (stability-informed site-independent, S-SI) to a model that accounts for stability and includes temporal rate fluctuations due epistasis (stability-informed site-dependent, S-SD). As such, both an epistatic and an independently evolving stability-informed site are subject to the same average stability-constraints, however, the epistatic site experiences fluctuating fitness landscapes whereas the independent site is evolving on a fixed landscape (figure 2.1A & B). This approach allows for a direct way of investigating the influence of epistasis on protein evolution.

Since substitution rates are primarily determined from the fitness coefficients, I expect that dynamic fitness landscapes due to epistasis will induce fluctuations in substitution rates over time. The variation in rate may be transient, where preferences at the site shift from some subset of amino acids to another; for example, polar residues might be preferred in one background sequence whereas nonpolar residues might be preferred given another sequence. For a short period of evolutionary time, the substitution rate will be transiently high as the site adjusts to the new peak (dos Reis, 2015). Alternatively, a shift from a more-uniform to a more-rugged landscape (or vice versa) would result in a sustained difference in rate from high to low (or low to high). To test if such dynamics are detectable using traditional ω-based inference models, I conducted the M3-CLM3 likelihood ratio test on all simulated alignments. While there was evidence of temporal rate variations under epistasis, it is important to note that Jones *et al.* (2017) showed that evolution on fixed fitness landscapes can also result in detectable signal for temporal variations in rates. They described a process reminiscent of the nonadaptive phase of Wright's shifting balance where a deleterious substitution due to drift moves a site away from its fitness peak and is followed by a transient period of high rates of nonsynonymous substitutions as the site evolves back to the fitness optima. In this way, epistasis and shifting balance result in similar temporal rate dynamics; a site becomes occupied by a suboptimal amino acid and subsequent nonsynonymous mutations are fixed in order to readjust to the fitness peak. The difference, however, is that under site-independence the site is destabilized due to a chance deleterious substitution at the site. In contrast, under epistasis, the site is

destabilized because of a substitution at another position causing a shift in the underlying fitness landscape. I found that the intensity of temporal rate switching was on average at least two times higher because of epistasis compared to the switching rates due to shifting balance. The higher switching rates is perhaps expected since shifting balance dynamics are contingent on the rare fixation of deleterious mutations by drift, whereas epistasis subjects sites to continuous changes in fitness landscapes.

Moreover, nonadaptive shifting balance dynamics were previously shown to elevate ω rates to values greater than one (Jones *et al.*, 2017), resulting in the canonical signal for positive selection. Specifically, Jones *et al.* (2017) reported significant evidence for positive selection at $10 - 40\%$ of trials when branch lengths were sufficiently long (total tree length was at least 7 substitutions per codon site). Here, two of the three phylogenies used for simulations had a total tree length > 7 substitutions per codon site (the 2ppn and 1pek phylogenies). However, I found no evidence for positive selection when alignments were generated with stability-informed fitness landscapes (with and without epistasis). Importantly, these results suggest that realistic fitness landscapes based on stability constraints are not a source of conflation for the canonical signal for adaptive evolution ($\omega > 1$) when tested using traditional ω-based inference models.

Inference models operate on a set of assumptions that are certainly incorrect for real protein evolution. Two of the most pervasive assumptions are that sites evolve independently, and that the variability in rates among-site is accurately approximated by a small number of rate categories. I find that, despite not accounting for epistasis, ω-based inference models perform comparably well when alignments are generated with and without epistatic interactions. A potential explanation for the comparability in model performance is that the magnitude or frequency (or both) of changes in amino acid preferences as a by-product of stability-induced epistasis are minor throughout evolutionary history. This supports previous computational and experimental work showing that, with respect to their impact on protein stability, amino acid fitness effects tend to remain relatively well conserved over long evolutionary times (Risso *et al.*, 2015; Ashenberg *et al.*, 2013). While accounting for epistasis is essential for understanding how proteins evolve, the site-independence assumption does not appear to limit the utility or accuracy of traditional inference models at identifying average selective pressures acting on natural proteins.

To address the concern that among-site rate variation might not be well-approximated by a small number of rate categories, more sophisticated inference models based on the MutSel framework were developed that permit a unique substitution process at each alignment site (Tamuri *et al.*, 2012, 2014; Rodrigue *et al.*, 2010; Rodrigue and Lartillot, 2014). However, these frameworks are generally only applicable when large phylogenies (> 100 taxa) are available in order to reliably estimate site-specific parameters ($e.g.$, the amino acid frequencies at each site, 19 parameters per site). Therefore, inference from smaller datasets must rely on traditional ω-based inference models which group sites into a small number of categories and estimate a much smaller number of parameters. While I found that the full extent of site-wise rate heterogeneity was not detectable by traditional models, the number of significant rate categories was consistent with the number of peaks in the distributions of expected rates. This suggests that traditional inference models are capable of detecting among-site heterogeneity when a sufficient number of sites share similar rates. Additionally, and perhaps more importantly, the ω values estimated were comparable to the theoretical rate expectations at the two or three clusters of sites. Furthermore, I found that the posterior mean ω^h calculated from simple M-series models correlated significantly with the expected rates. Overall, the results from this chapter suggest that ω-based models sufficiently describe average selective pressures.

The mutation-selection (MutSel) framework and biophysical models are a step towards more mechanistically plausible generative frameworks. Nonetheless, our models are limited by any underlying assumptions about the evolutionary process that are inconsistent with real protein evolution. The population genetics theory underlying the MutSel framework assumes mutations enter a population at an extremely low rate followed by a near-instantaneous fixation or loss. As such, a system might not be well modeled by MutSel when the dynamics of standing polymorphism can impact substitution rates ($e.g.$, extended residence times for polymorphism, selective interference, stochastic tunneling in large population), or the mutation rate is high ($e.g.$ viral systems). As the goal was to model an evolutionarily conserved property (stability constraints) for lineages having low mutation rates and relatively small effective sizes, MutSel substitution dynamics are expected to be appropriate.

The principles of thermodynamics underlying the biophysical model assume a simple two-state folding process where proteins are either correctly folded or are unfolded.

Small monomeric proteins (< 100 amino acids) can fold in this way (Jackson, 1998); however larger proteins require stable intermediate structures to fold properly. Of the protein structures used here, and previously within this framework (Goldstein, 2011, 2013; Pollock *et al.*, 2012; Goldstein and Pollock, 2016, 2017), only the 2ppn protein has been experimentally shown to fold following the two-state process (Jackson, 1998). In fact, while it is the largest protein known to fold without the need of intermediate structures, it is the smallest protein to ever be used within this thermodynamic framework. More generally, the three structures used here differ in important ways ($e.g.$ biological function, protein length, packing density); nonetheless I observed similar consequences of epistasis on substitution rates which suggests that the results may be generalized across stable, globular proteins.

The current formulation of the biophysical model is limited to stable proteins with a known three-dimensional structure and therefore does not characterise the evolutionary dynamics of intrinsically disordered proteins or proteins with multiple conformations. The three-dimensional structure is used to approximate the free energy of a sequence in a given native state. Various methods have been developed to estimate stability values upon mutations ($e.g.$, FoldX (Guerois *et al.*, 2002), Rosetta (Rohl *et al.*, 2004)). In this study, I used the Miyazawa-Jernigan contact potentials with the pairwise energy approximation for its computational manageability and because even the most sophisticated models at best only moderately predict mutational effects (Potapov *et al.*, 2009). Furthermore, this model was sufficient because I did not require exact amino acid sequences that can be folded in the native structure; that is a demanding task even when more computationally exhaustive methods are used. Instead, the objective was to simulate plausible evolutionary dynamics, and I have shown that the modelling framework is sufficient for this purpose. In addition, the models used here assume selection acting only on protein stability whereas natural proteins are subject to additional functional and structural constraints. A recent approach was presented by de la Paz *et al.* (2020) using multiple sequence alignments of natural protein families (>1,000 sequences) to estimate global epistatic contributions. The approach reproduces empirical and theoretical phenomena and is a promising tool for improving our understanding of protein evolution.

To conclude, I have found that epistasis alters the dynamics of how proteins evolve. It is therefore important to model epistatic interactions when the objective is to gain intuition

and develop a deeper understanding of how protein sequences change over time. However, with regards to inference of selective pressures, the data presented here suggests that explicit modelling of epistasis might not be of paramount importance. Instead, accounting for the phenomenological outcomes of epistasis, in allowing for more diversity in among-site amino acid preferences (Rodrigue and Lartillot, 2017; Tamuri *et al.*, 2014) and/or accounting for temporal fluctuations in substitution rates (Jones *et al.*, 2017; Murrell *et al.*, 2015), offer a promising avenue for the future development of inference models.

2.5 Methods

2.5.1 Natural protein alignments

Three globular, monomeric proteins were used throughout this study with PDB codes 1qhw, 1pek, and 2ppn. The 1qhw structure is from a purple acid phosphatase protein extracted from rat bone and is likely involved in bone resorption (Lindqvist *et al.*, 1999). The 2ppn protein is a peptidyl-prolyl cis-trans isomerase extracted from human cells which facilities the folding of other proteins (Szep *et al.*, 2009). The 1pek protein is a proteinase K used in protein digestion. The structure was extracted from *Engypdontium album* (Betzel *et al.*, 1993). The three protein structures differ in important ways. First, I included the 1qhw protein for consistency since it is the only protein to have previously been used in this modelling framework. I included the 2ppn protein because of its smaller size (it is approximately a third of the length of the other two proteins) and, more importantly, because it has been shown to fold following the two-state folding (Jackson, 1998) and therefore does not violate one of the core thermodynamic model assumption. Lastly, I selected the 1pek protein because, while it is comparable in length to the 1qhw protein, it is a more densely packed protein. The average number of contacts per site was 8.39 for the 1pek protein compared to 7.5 for the 1qhw protein (and 6.9 for the 2ppn structure).

For each of the three proteins I created a multiple sequence alignment of orthologous gene sequences using MUSCLE (Sievers *et al.*, 2011). Protein sequences were chosen if there were no insertions or deletions since that will likely imply changes in the protein structure which are not accounted for in the modeling framework. The accession numbers for the gene sequences are reported in table 2.5. The 1qhw and 2ppn alignments included gene sequences from fourteen taxa, whereas the 1pek alignment was made up of twelve sequences. The length of the 1qhw, 2ppn, and 1pek alignments were 300, 107 and 279

codon sites respectively.

Table 2.5: NCBI Accession numbers for sequences used to create the three natural protein alignments (1qhw, 2ppn, 1pek).

1qhw	**2ppn**	**1pek**
NP_001075457.1	M80199.1	XM_003713956.1
ELK28734.1	U09386.1	XM_016786789.1
DAA35014.1	NM_204330.1	AM412313.1
ELR59971.1	NM_001252190.3	X14688.1
XP_020749961.1	AF483488.1	XM_018327475.1
DAA35015.1	BT021075.1	XM_006967830.1
NM_001256558.1	KY474593.1	EF362571.1
XP_021499432.1	KY474591.1	AF104385.1
XM_022526904.1	KY474590.1	AJ427459.1
XM_008048505.2	BC059689.1	HM635906.2
M76110.1	BT075719.1	XM_014693831.1
CR457078.1	NM_001139669.1	M73795.1
NM_001284443.1	NM_001103022.1	
XP_005078573.1	BT082974.1	

For each protein alignment I inferred a phylogenetic tree using IQ-TREE (Nguyen *et al.*, 2014) with ModelFinder (Kalyaanamoorthy *et al.*, 2017) and ultrafast bootstrapping (Minh *et al.*, 2013). Maximum likelihood estimates yielded a wide range of tree lengths (table 2.1) which allowed us to investigate how the relationship between model assumptions and substitution rate was affected by tree length.

Following the protocol outlined in Sydykova *et al.* (2018). I calculate relative solvent accessibility (RSA) and weighted contact number (WCN) for all sites in each of the protein structures. Relative solvent accessibility (RSA) was calculated as

$$RSA = ASA/maxASA \tag{2.2}$$

where ASA is the accessible surface area calculated using DSSP (Kabsch and Sander, 1983), and $maxASA$ is the maximum accessible surface area as measured by Tien *et al.* (2013). WCN is calculated as $\sum_{j \neq j} 1/r_{ij}^2$ where r_{ij} is the distance between the geometric centres of the side chains of residues occupying sites i and j.

2.5.2 Mutation-Selection (MutSel)

The evolutionary process, for all the simulation models, was based on the mutation-selection (MutSel) framework (Halpern and Bruno, 1998). This framework was introduced in detail in section 1.3.1.

When generating protein-specific alignments, I used the nucleotide frequencies π_j and κ values estimated from the corresponding real alignment under inference model M3($k = 3$) (table 2.1). All models assume that selection acts on the final protein product. The models therefore assign all synonymous codons the same fitness.

2.5.3 C-series site-independent model (C-SI)

Under C-SI, amino acid fitness values were approximated from the C-series empirical frequency profiles (Quang *et al.*, 2008), commonly used in phylogenetic inference. The C-series model capture among site variation in amino acid preferences (and hence frequencies) by assuming that a site belongs to one of twenty different frequency profiles. In the MutSel framework, the frequency of amino acid a is related to its fitness f_a by the following relationship

$$\pi_a \propto \pi_a^{(0)} \exp(2N_e f_a) \tag{2.3}$$

where $\pi_a^{(0)}$ is the stationary frequency in the absence of selection pressure (dos Reis, 2015). I use this to convert each of the twenty C20 frequency profiles to twenty fitness vectors. Note that the amino acid frequencies in the absence of selection pressures, $\pi_a^{(0)}$, reflect underlying biases in the mutation process. The stationary frequency of a codon (or nucleotide triple ijk) is proportional to $\pi_i \pi_j \pi_k$. Then, $\pi_a^{(0)}$ is calculated as the sum of the stationary frequencies of synonymous codons corresponding to amino acid a. Because the three proteins studied here had different mutational parameters (table 2.1), the C20 profiles translated to twenty protein-specific fitness landscapes. When generating alignments under C-SI, each site was randomly assigned one of the twenty protein-specific fitness vectors. As such, the C-SI model assumes that sites evolve independently and are identically distributed.

2.5.4 Stability-informed models (S-SI and S-SD)

Alternatively, the stability-informed models (S-SI and S-SD) define fitness as the proportion of correctly folded proteins at thermodynamic equilibrium, which is a nonlinear function of the protein's folding stability.

Epistasis refers to the dependence of the fitness effect of a mutation on the background genetic sequence. To account for epistasis within the MutSel framework, each site was assigned a vector of amino acid fitness values $F^h(S) = \langle f_1^h(S), ..., f_{20}^h(S) \rangle$ where $f_a^h(S)$ is the fitness of the protein calculated as described in section 1.3.2, given amino acid a at site h and background sequence S. Throughout the evolution of the protein, all site-specific fitness vectors were recalculated following a nonsynonymous substitution somewhere in the protein.

To assess if and how epistasis influences substitution rates, I developed an analogous stability-informed site-independent model (S-SI) where epistatic effects on folding stability were marginalized such that the fitness landscape at a site, F^h, is independent of the background sequence and is therefore constant across time. To allow for a direct comparison between alignments generated with and without epistasis, I used the S-SD simulations to estimate the independent fitness landscapes, F^h (figure 2.2). In other words, let $\{S_1, ..., S_N\}$ be the extant sequences in an S-SD simulated alignment, where N is the number of taxa. I calculated f_a^h as the average fitness value for amino acid a over sequences $\{S_1, ..., S_N\}$:

$$f_a^h = (1/N) \sum_{t=1}^{N} f_a^h(S_t) \tag{2.4}$$

The average fitness values were used to specify the independent site-specific fitness vectors, F^h, under S-SI.

2.5.5 Scaling branch lengths

In order for branch lengths to have the desired interpretation as the mean number of single nucleotide substitution per codon site, the substitution rates must be rescaled. For -SI generated alignments, I rescaled the rate matrices in the conventional way by dividing all site-specific rate matricies Q^h by the mean expected rate of change:

$$(1/L) \sum_{h=1}^{L} \sum_{x=1}^{61} -\pi_x q_{xx}^h \tag{2.5}$$

where L is the number of sites and $q_{xx}^h = -\sum_{y \neq x} q_{xy}^h$ (Jones *et al.*, 2017). Alternatively, to obtain the appropriate scaling factor for the S-SD alignments, I ran the simulation for 1000 substitutions using the Gillespie algorithm (Gillespie, 1977). I recorded the overall time T required for 1000 substitutions to occur by summing over the waiting times between

substitutions,

$$T = \sum_{t=0}^{1000} \sum_{h=1}^{L} \tau_t^h \tag{2.6}$$

where τ^h is the waiting time until the next substitution event at site h which is exponentially distributed with mean $1/q_{xx}^h$. Branch lengths, b, were then rescaled such that $b = n(T/1000)$. I validated the scaling methods by comparing the inferred branch lengths from the simulated alignments to the true generating branch lengths (mean tree lengths from each set of simulations are reported in table 2.2).

2.5.6 Sampling high fit sequences

To avoid nonequilibrium behavior, each of the protein-specific simulations were initiated at amino acid sequences with fitness values > 0.99 given the respective protein structure. However, sequence space is immense, and most sequences have a fitness of zero. I developed and used the following algorithm to explore sequence space to find sequences with high fitness:

Algorithm 1: Algorithm for exploring sequence space and finding sequence with high fitness values

```
Start at randomly generated amino acid sequence S;
while fitness < 0.99 do
    calculate the site-specific fitness landscape at all sites F^1(S), ..., F^L(S);
    if a single step uphill move is possible then
        randomly choose the next substitution from the set of single amino acid
         changes that will increase fitness;
    else
        randomly choose 20 sites and substitute them to the fittest amino acid at
         that site;
    end
end
```

2.5.7 Expected substitution rate dN/dS calculations

The evolutionary rate at a site is commonly defined as the ratio of nonsynonymous to synonymous substitutions rates (N^h/S^h) normalized by the ratio of nonsynonymous to synonymous mutations rates (N_{mut}^h/S_{mut}^h). Assuming selection acting at the protein-level such that synonymous codons have the same fitness value, the rate of fixation of a

synonymous mutation will be equal to its underlying mutation rate, $S^h = S^h_{mut}$. Therefore, the expected substitution ratio simplifies to $dN^h/dS^h = N^h/N^h_{mut}$. In the traditional MutSel framework ($i.e.$, assuming site-independence as done in simulation models C-SI and S-SI), the evolutionary rate at a site, dN^h/dS^h, can be calculated directly from the site-specific fitness coefficients and the protein-specific mutation rates.

$$dN^h/dS^h = \frac{N^h}{N^h_{mut}} = \frac{\sum_x \sum_{y \in \mathcal{N}_x} \pi^h_x q^h_{xy}}{\sum_x \sum_{y \in \mathcal{N}_x} \pi^h_x \mu_{xy}} \tag{2.7}$$

where $\mathcal{N}_x$ is the set of codons that are nonsynonymous to codon x and differ by a single nucleotide, q^h_{xy} is the substitution rate from codon x to codon y calculated using equation (1.3), μ_{xy} is the mutation rate calculated under the HKY85 process (Hasegawa *et al.*, 1985), and π^h_x is the stationary frequency for codon x at site h. I note that dos Reis (2015) presented an alternative way of calculating dN^h/dS^h where the nonsynonymous mutation rate, N^h_{mut}, was calculated in reference to the neutral stationary frequencies $\pi^{(0)}_x$. While the interpretation of the dN^h/dS^h values differ (as discussed in Jones *et al.* (2017)), I found that both formulations resulted in highly comparable rate values (Pearson correlation coefficient = 0.99, p-value < 0.001; figure 2.12). The dN^h/dS^h rates reported in the chapter were calculated using equation (2.7).

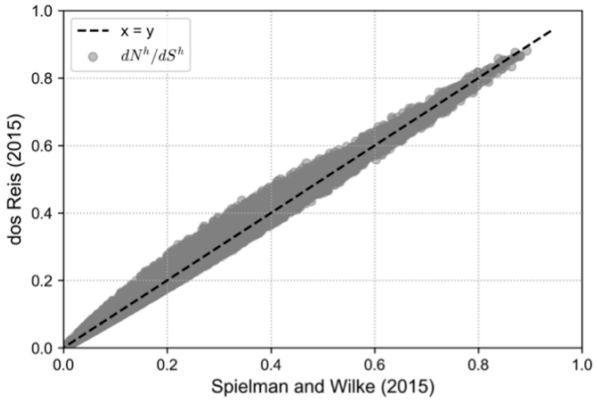


Figure 2.12: Correlation between different ways of calculating site-specific substitution rates, dN^h/dS^h. The dos Reis (2015) formulation measures the nonsynonymous mutation rate in reference to the neutral stationary frequencies $\pi_x^{(0)}$ based on mutational biases only. The Spielman and Wilke (2015) formulation measures the nonsynonymous mutation rate in reference to the site-specific stationary frequencies π_x^h in the presence of selection pressure and mutational biases. See Jones *et al.* (2017) for further discussion on the interpretation of both formulations. The Pearson correlation coefficient between both formulations was 0.99, p-value< 0.001.

When epistatic dependencies between sites are modelled within the MutSel framework, the average substitution rate at a site can in principle be calculated as

$$dN^h/dS^h = \frac{\sum\limits_S N^h(S)}{\sum\limits_S N^h_{mut}(S)} \tag{2.8}$$

where the sum is over all possible background sequences S. However, the number of possible sequences is very large, 20^L where L is the length of the protein. While dN^h/dS^h averaged over all 20^L background sequences is the theoretical rate expectation, it is impossible to calculate (because of the large number of sequences) and likely does not reflect the rates for real proteins. Instead, for S-SD simulations, I define the evolutionary rate at a site as the mean substitution rate observed throughout the evolution of a protein over a defined length of time averaged over all generated alignments. Specifically, for each S-SD alignment i (for $i = 1, ..., 50$) I approximate the rate at a site (dN_i^h/dS_i^h) by summing over the extant sequences $\{S_1, ..., S_N\}_i$.

2.5.8 Assessing robustness to sample size

Given the enormity of sequence space, it is unclear that any sampling, no matter how extensive, could characterize the entire fitness landscape. Since the evolution of natural proteins billions of years ago, even natural proteins have not adequately sampled their respective sequence space and are evolving on a small, localized portion of sequence space. To understand how epistasis influences protein evolution concerning rates of substitution, I consider comparisons between rate estimates from the S-SD and S-SI models in the same local-neighbourhood of sequence space. This avoided the difficulty of comparing behaviour in different regions of sequence space.

Specifically, to calculate the expected rate at a site, I approximate the rate as the average over the extant sequences observed in each S-SD simulated alignment. The extant sequences provide a sample of the local neighbourhood. To address the robustness of the results to a more extensive sampling of sequences in the local space, I compared the expected rate dN_i^h/dS_i^h considering all extant sequences from each alignment i to the rate dN_{ij}^h/dS_{ij}^h calculated by leaving out the j^{th} sequence. Then, I calculated the bias and mean squared error (MSE) as:

$$bias_i^h = 1/N \sum_{j=1}^{N} dN_{ij}^h/dS_{ij}^h - dN_i^h/dS_i^h \tag{2.9}$$

$$MSE_i^h = 1/N \sum_{j=1}^{N} [dN_{ij}^h/dS_{ij}^h - dN_i^h/dS_i^h]^2 \tag{2.10}$$

where N is the number of taxa per alignment. The distributions of the average bias and average MSE, for all three proteins, suggest that calculating expected rates based on the extant sequences does not systematically bias estimates, and has little impact on the expected rate values (figure 2.13). However, note that the bias and MSE are slightly higher for the 1pek simulations. This is likely due to two reasons: (1) the 1pek tree is deeper than the 1qhw and 2ppn trees (tree length = 13.88, 4.93, 8.04 for 1pek, 1qhw, and 2ppn) which means that the local neighbourhood is larger for the 1pek simulations, and (2) the number of taxa in the 1pek alignment (N= 12) is smaller than the number of sequences in the 1qhw and 2ppn alignments (N = 14). The larger local neighbourhood in conjunction with the smaller sampling likely lead to the increase in bias and MSE observed. Importantly, however, the bias and MSE are nonetheless minor which suggests that calculating rates as

the average over the extant sequences has minimal consequences on rate expectations.

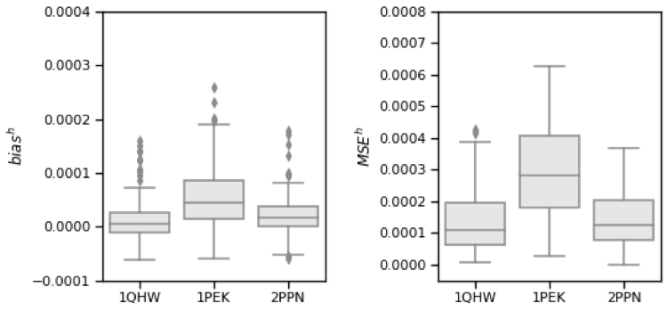


Figure 2.13: Assessing the average bias and average mean squared error in expected site-specific rates (dN^h/dS^h)

2.5.9 Thermodynamic model of protein folding

I used the stability-informed model described in section 1.3.2 to estimate fitness. The set of alternative structures used to estimate the mean ($\bar{E}$) and variance (ΔE^2) of energies in the unfolded states are listed in table 2.6.

Table 2.6: PDB codes for protein structures used to calculate average free energy, $\bar{E}$, and variance ΔE^2, for a sequence in the unfolded configurations. Structures were taken from (Goldstein and Pollock, 2017).

1cnz	1gyh	1jix	1m4l	1nsz	1pby	1t5j
1dmh	1hz4	1jj2	1mkf	1o4s	1pfk	1t5o
1e19	1i4w	1jkm	1moq	1o7j	1qo0	1to6
1ek6	1iom	1jl5	1mty	1o88	1qop	1uby
1esd	1ir6	1jub	1n00	1oc7	1rkd	1umd
1ga6	1jfb	1kwf	1nbf	1odm	1sbp	1v6s
1gwu	1jil	1l5o	1nd6	1ojj	1svm	1wch
1wer	1wkr	1woh	2bbv	2mas	3sil	

2.5.10 Maximum likelihood inference of selection pressure

2.5.10.1 M-series models

The M-series models assumes a time-reversible, stationary, continuous-time Markov chain where the instantaneous substitution rate matrix A defines the rate of substitution between codon x and y as

$$a_{xy} \propto \begin{cases} 0, & \text{if } x \text{ and } y \text{ differ by more than one nucleotide.} \\ \pi_j, & \text{if } x \text{ and } y \text{ differ by a synonymous transversion.} \\ \kappa\pi_j, & \text{if } x \text{ and } y \text{ differ by a synonymous transition.} \\ \omega\pi_j, & \text{if } x \text{ and } y \text{ differ by a nonsynonymous transversion.} \\ \omega\kappa\pi_j, & \text{if } x \text{ and } y \text{ differ by a nonsynonymous transition.} \end{cases} \tag{2.11}$$

κ is the transition to transversion rate ratio, π_j is the stationary frequency of the target nucleotide j, and ω is the nonsynonymous to synonymous rate ratio. This describes MG (Muse and Gaut, 1994) parameterization of M0, the simplest M-series model, with a single rate parameter estimated for all sites in the alignment. To account for variation in selection pressure across sites, M3(k) extends M0 by allowing for c discrete number of rate categories, each with a rate parameter ω_c and corresponding proportion of sites p_c. M0 is analogous to M3($k = 1$). The M3(k) versus M3($k + 1$) likelihood ratio test was used to determine the appropriate number of rate categories for each alignment.

2.5.10.2 CLM3

To test for variation in substitution rate across time, I used the covarion-like CLM3 as implemented by Jones *et al.* (2017) which assumes that the substitution process switches over time between one with an $\omega = \omega_1$ and another with $\omega = \omega_2$. The switching and substitution processes can be modeled as a two-dimensional Markov chain (X,Y) where X is the current codon and Y indicates the substitution process, 1 or 2. Ordering the possible states as (1,1),(2,1),..., the rate matrix is

$$A = \frac{1}{r_1}\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} + \frac{\delta}{r_2}\begin{pmatrix} -p_2 I & p_2 I \\ p_1 I & -p_1 I \end{pmatrix} \tag{2.12}$$

where A_1 and A_2 are the substitution rate matrices constructed using equation (2.11) with ω_1 and ω_2 respectively; p_1 and p_2 are the expected proportion of time a site evolves under the respective ω, I is the identity matrix, and δ denotes the rate of change between selection regimes. r_1 and r_2 are scaling parameters such that time is measured as the expected number of single nucleotide changes per codon site and δ is the expected number

of switches per unit time. The model contrast M3($k = 2$) versus CLM3 provides a likelihood ratio test for evidence of switching between rate categories ω_1 and ω_2 across the tree.

2.5.10.3 BUSTED

The branch-site unrestricted statistical test for episodic diversification, BUSTED (Murrell *et al.*, 2015), is based on the BS-REL framework (Kosakovsky Pond *et al.*, 2011) allowing for variations in rates across sites and branches. Specifically, BUSTED estimates three rate categories ($\omega_1 \leq \omega_2 \leq \omega_3$) where at each branch in the tree, a site belongs to one of the three ω categories. The model also estimates proportions p_1 and p_2 ($p_3 = 1–p_1–p_2$) shared across sites. If there is evidence for positive selection ($\omega_3 > 1$), then a likelihood ratio test of BUSTED with ω_3 constrained to be < 1 against an unconstrained BUSTED is conducted.

2.6 Code and Data Availability

Real and simulated alignments, as well as the code used to generate, analyze, and plot have been uploaded to GitHub (https://github.com/noory3/Consequences-of-stability-induced-epistasis.)

CHAPTER 3

TRAJECTORIES OF AMINO ACID PROPENSITIES UNDER STABILITY-MEDIATED EPISTASIS

This work was submitted to the journal *Molecular Biology and Evolution*, and was done in collaboration with Edward Susko, Andrew Roger, and Joseph Bielawski.

3.1 Abstract

Epistasis between residues significantly impacts protein evolution. The propensity of a resident amino acid can increase because of replacements at other sites—a nonadaptive phenomenon referred to as the *evolutionary Stokes shift*. Alternatively, decreases in propensities have been interpreted as evidence of adaptations. I show that propensities can decrease under nonadaptive stability-constrained evolution, a phenomenon I call *evolutionary anti-Stokes shifts*. Using extensive simulations based on three natural protein structures, I detect evolutionary Stokes shifts following approximately 50% of substitutions, and anti-Stokes shifts in the remaining substitutions. Therefore, nonadaptive evolution can lead to positive and negative shifts in propensities, and hence their detection is not conclusive evidence of adaptation. Nevertheless, two phenomena emerge from nonadaptive evolution (1) the magnitudes and frequencies of Stokes and anti-Stokes shifts tend to be balanced, and (2) epistasis leads to a significant negative autocorrelation in propensity changes, thereby limiting the severity of evolutionary shifts. Analyses of one-step propensity changes following the acceptance of stabilizing substitutions indicate

that they increase mutational tolerance such that site-specific landscapes are more uniform, causing a decrease in resident amino acid propensities. In contrast, destabilizing substitutions result in more rugged landscapes and tend to increase resident amino acid propensities. In summary, the results from this chapter characterize propensity trajectories under nonadaptive stability-constrained evolution, against which evidence of adaptations should be calibrated.

3.2 Introduction

Amino acid interactions within a protein are a fundamental form of epistasis. Interactions between sites occur because of functional, structural, or stability constraints (Ortlund *et al.*, 2007; Pollock *et al.*, 2012; Gong *et al.*, 2013). It has become evident that accounting for epistasis between sites is critical for explaining various properties observed in natural sequences (de la Paz *et al.*, 2020; Goldstein and Pollock, 2017). Here, I focus on stability-constraints by modelling protein evolution based on thermodynamic principles. This modeling framework reproduces realistic evolutionary dynamics with regards to protein stability values (Goldstein, 2011), evolutionary rates (Youssef *et al.*, 2020), temporal and spatial patterns of rate heterogeneity (Goldstein and Pollock, 2016), and convergence rates (Goldstein *et al.*, 2015). In this chapter, I explore long term shifts in amino acid preferences due to nonadaptive stability-constraints.

Under nonadaptive evolution, a protein evolves on a fixed fitness landscape with no changes in environment or function (Wright, 1932). Natural selection maintains the protein near a peak on its landscape with equilibrium dynamics shaped by mutation, drift, and selection. At equilibrium, most mutations are deleterious, while a small proportion is beneficial. The higher fixation probability of the fewer but more advantageous mutations is balanced by a lower fixation probability of the more frequent yet disadvantageous mutations. As a result, the proportion of deleterious and beneficial substitutions ($i.e.$, fixed mutations) are equal (Goldstein, 2013; Cherry, 1998). This scenario contrasts with the dynamics under adaptive evolution. Novel protein function or environment lead to shifts in the fitness landscape, rendering the current state suboptimal. Subsequent fixations that increase fitness transiently inflate substitution rates, a characteristic of adaptive episodes (dos Reis, 2015; Jones *et al.*, 2017).

Since its origin, the strictly neutral model of protein evolution is often treated as

the null scenario that must be rejected prior to postulating adaptive evolution (Kimura, 1968, 1991; Duret, 2008). Equilibrium dynamics under stability-constrained models are consistent with neutral theory (Goldstein, 2011). Using a stability-constrained model, Goldstein (2011) showed that marginal protein stability emerges from the balance between mutation, drift, and selection, challenging the notion that evolution actively selects for it (DePristo *et al.*, 2005). Nonadaptive epistatic models predict various characteristics in natural proteins—such as marginal stability (Goldstein, 2011) and differences in mutational tolerance across sites (Youssef *et al.*, 2020)—highlighting that adaptive evolution need not be invoked to explain their presence.

Using a nonadaptive stability-constrained model, Pollock *et al.* (2012) observed that the preference for a newly substituted amino acid tends to increase over time due to substitutions at other protein sites. They referred to this phenomenon as the 'evolutionary Stokes shift'. Shah *et al.* (2015) performed extensive *in silico* stability-constrained evolution and observed that substitutions are often contingent on prior substitutions that increased their probability of fixation and entrenched by subsequent replacements at other positions. In contrast with these theoretical predictions, experimental results report that amino acid preferences are often conserved over long evolutionary time scales (Starr *et al.*, 2018; Risso *et al.*, 2015; Doud *et al.*, 2015; Haddox *et al.*, 2018; Ashenberg *et al.*, 2013). Furthermore, evidence of decreases in propensities is emerging (Popova *et al.*, 2019; Stolyarova *et al.*, 2020). Decreases in preferences have been interpreted as evidence of adaptive evolution to a changing environment. Specifically, Popova *et al.* (2019) state that epistatic constraints "cannot lead to a systematic reduction in fitness of the incumbent alleles", naming this phenomena *senescence*.

Faced with seemingly conflicting observations, it is unclear if there are general patterns in how amino acid propensities shift during evolution. Do resident amino acid preferences increase, decrease, or remain conserved? And to what extent are these dynamics shaped by nonadaptive processes? Using extensive simulations under a stability-constrained model, I apply two quantitative metrics to evaluate long term propensity shifts. I observe that all three trajectories emerge from nonadaptive dynamics at mutation-drift-selection equilibrium. Importantly, resident amino acid preferences can decrease merely due to epistatic constraints. Building on previous work (Pollock *et al.*, 2012), I refer to this as the evolutionary anti-Stokes shift. I also observe a significant negative autocorrelation in

propensity changes in the stability simulations, suggesting that epistasis tends to conserve amino acid preferences rather than alter them.

Lastly, I characterize the underlying mechanisms that cause propensities to fluctuate. Following a stabilizing substitution, most site-specific landscapes become more uniform and hence more mutationaly tolerant. This leads to decreases in the propensities of the resident amino acids, since other residues may occupy each site with little or no detriment to fitness. In contrast, destabilizing substitutions tend to induce more restrictive site-specific landscapes, limiting potential substitutions and increasing the propensity for the resident amino acid. Importantly, these phenomena emerge from a nonadaptive model of sequence evolution with constraints on protein stability.

3.3 Results

I use a thermodynamic model of protein evolution and equate fitness to the probability of correct folding, a function of protein stability (ΔG). I assume no changes in structure or function so that the global fitness landscape (the mapping between sequence and fitness) remains constant. Nonetheless, this modelling framework accounts for epistasis by assigning site-specific fitness landscapes dependent on the background sequence ($f^h(S) = \{f_1^h(S), ..., f_{20}^h(S)\}$ for a given site h and background sequence S). Amino acids that confer higher fitness values (improve stability) will more frequently occupy the site and have higher expected frequencies ($i.e.$ propensities). In this way, the frequency of an amino acid is related to its fitness—the formal relationship is provided by equation (3.1). Frequency landscapes are also site-specific and context-dependent ($\pi^h(S) = \{\pi_1^h(S), ..., \pi_{20}^h(S)\}$). Note, that the fittest amino acid may not have the highest frequency. This occurs when a suboptimal amino acid has many codon aliases. The high number of synonymous codons and/or mutational bias can increase the residue's frequency despite its lower fitness.

An evolutionary Stokes shift is a phenomenon whereby the propensity (π_a^h) for a resident amino acid increases due to substitutions at other positions (Pollock *et al.*, 2012). Here, I define propensity as the equilibrium frequency given a fixed background sequence,

$$\pi_a^h(S) = \pi_a^{(0)} e^{2N_e f_a^h(S)} / \sum_x \pi_x^{(0)} e^{2N_e f_x^h(S)} \tag{3.1}$$

where N_e is the effective population size and $\pi_a^{(0)}$ are the neutral stationary frequencies (dos Reis, 2015). In previous work (Pollock *et al.*, 2012; Goldstein and Pollock, 2017), propensities represented thermodynamic preferences and effectively assumed no mutational biases. This can be accommodated in this formulation by assuming no mutational biases such that $\pi_a^{(0)}$ is uniform ($\pi_a^{(0)}$ = 1/20). However, the simulations presented here are based on three proteins with unequal nucleotide frequencies and transition/transversion rate biases. I account for these by estimating protein-specific $\pi_a^{(0)}$ (see Methods section for details; figure 3.1). Nevertheless, the results remained consistent under both definitions of propensity. Unless otherwise stated, I use equilibrium frequencies to measure amino acid propensities.

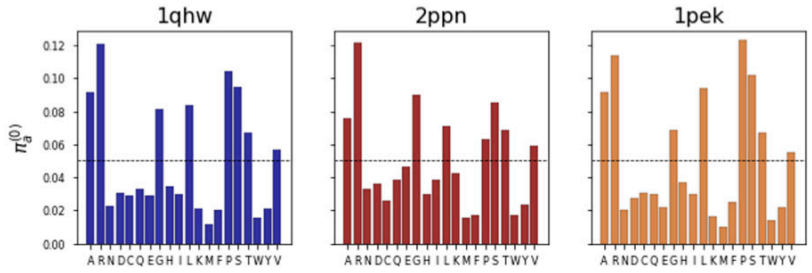


Figure 3.1: The expected amino acid frequencies in the absence of selection but accounting for underlying mutational biases. The dotted line represents the expected frequency values in the absence of mutational biases and assuming all amino acids have the same number of codon aliases (=1/20).

Results are based on 500 protein-specific simulations for three proteins. The proteins differ in structure (PDB codes 1qhw, 2ppn, and 1pek), function (a phosphatase, an isomerase, and a proteinase), and length (300, 107, and 279 amino acids; see Methods for more detail). I ran each simulation for 500 substitutions with $N_e = 100$. On average, sequences diverged at 43% of sites within a simulation. Increasing the number of substitutions or the effective population size did not alter the results. See section 3.5.3 for a discussion assessing the robustness of the results from this chapter to different simulation settings.

3.3.1 Increases, decreases, and conservation of preferences under non-adaptive evolution

Throughout the simulations, and in natural protein evolution (Risso *et al.*, 2015; Gong *et al.*, 2013; Ashenberg *et al.*, 2013), the preference for an amino acid changes over time.

In natural proteins, these variations can occur because of adaptive or nonadaptive processes. By contrast, variations in sites' fitness and propensity landscapes in the simulations are due to stability-induced epistasis and are not adaptive. Examples of these propensity dynamics are shown in figure 3.2. The propensity for aspartic acid (D), the resident amino acid at site 232, changes as substitutions occur at other protein sites (figure 3.2A). The site experiences an evolutionary Stokes shift where its propensity increases over time. Alternatively, at site 72, the propensity for the resident amino acid proline (P) decreases (figure 3.2B), while the propensity for glutamine (Q), at site 88, was conserved (figure 3.2C). All three trajectories emerged at mutation-drift-selection equilibrium and in the absence of adaptive changes.

Shifts in amino acid propensities are not directly observable in natural proteins. However, Popova *et al.* (2019) suggested that shifted propensities alter replacement rates, producing a detectable signal in protein alignments. An amino acid's replacement rate is inversely related to its propensity: if the propensity for the resident amino acid is high, then its replacement rate will be low, and vice versa. Therefore, in addition to the amino acid propensities, I calculated the expected replacement rate as the sum of transition rates to neighbouring sequences that differ from the current sequence at the site of interest. Figure 3.2D and 3.2E confirm the predicted effect on replacement rates. At site 232, the increase in propensity ($i.e.$, evolutionary Stokes shift) is accompanied by a decrease in the replacement rate (figure 3.2A&D). Similarly, the decrease in resident amino acid propensity at site 72 ($i.e.$, evolutionary anti-Stokes shift), is accompanied by an increase in replacement rate (figure 3.2B&E). Therefore, both increases and decreases in replacement rates can occur because of nonadaptive evolutionary Stokes and anti-Stokes shifts.

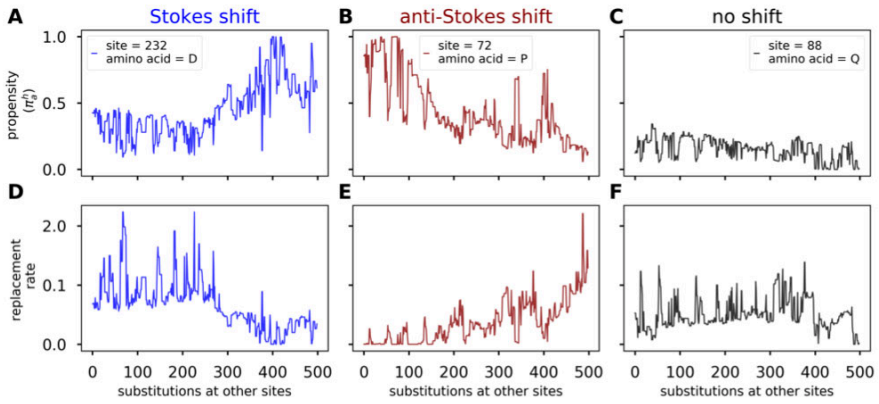


Figure 3.2: Trajectories of amino acid preferences under nonadaptive evolution. (A,D) site 232 undergoes an evolutionary Stokes shift where the propensity for the resident amino acid, aspartic acid (one letter code D), increases over time. (B,E) site 72 undergoes an anti-Stokes shift where the propensity for the resident amino acid, proline (one letter code P), decreases over time. (C,F) The propensity for the resident amino acid glutamine (one letter code Q) at site 88 remains conserved. (A,B,C) plot the propensity of the resident amino acids as substitutions occur at other positions in the protein. (D, E, F) show the expected replacement rates. Results are from a simulation of the 1pek protein.

3.3.2 A balance in the occurrence of evolutionary Stokes and anti-Stokes shifts

The previous results demonstrate that propensity shifts can occur under nonadaptive evolution. However, it remains unclear whether shifts are widespread or rare. To address this, I developed two metrics to quantify trends in propensities. The metrics are described in detail in the Methods section and illustrated in figure 3.3A. Briefly, metric M_{SLR} is the **S**lope of the **L**inear **R**egression where the covariate x is time (measured in substitutions) and the response y is the propensity of the resident amino acid. In defining the evolutionary Stokes shift, Pollock *et al.* (2012) state that "the inherent propensity for [an] amino acid at that position will be, on average, higher than it was when the substitution occurred". I, therefore, defined the metric M_{AMI}, consistent with this definition, calculated as the **A**verage propensity of an amino acid while it is resident **M**inus its **I**nitial propensity. Values of M_{SLR} and $\mathrm{M}_{AMI} > 0$ indicate an evolutionary Stokes shift, while values < 0 suggest an evolutionary anti-Stokes shift.

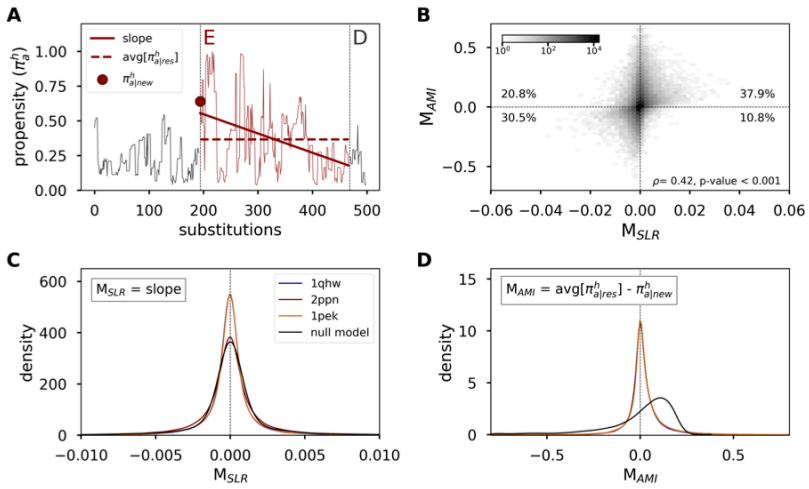


Figure 3.3: Description and analysis of metrics used to estimate propensity shifts. (A) Plotted is an example trajectory observed at site 82 of the 1pek protein. The site accepts two substitutions (vertical dotted lines) and the resident amino acid changes from D→E→D. Consider the dynamics following the acceptance of amino acid E. The first metric (M_{SLR}) is the **S**lope of the **L**inear **R**egression where x is the number of substitutions and y is the propensity of the resident amino acid a at site h (π_a^h) calculated over $i \leq x \leq j$; i is the substitution where amino acid a first occupies the site and j is the last substitution. The second metric M_{AMI} is the **A**verage propensity of an amino acid while it is resident (avg[$\pi_{a|res}^h$]) **M**inus its **I**nitial propensity ($\pi_{a|new}^h$). Metrics values > 0 indicate evolutionary Stokes shifts and values < 0 indicate evolutionary anti-Stokes shifts. (B) Hexbin plot showing the relationship between M_{SLR} and M_{AMI}. The shade of each hexbin represents the number of points per hexbin. Reported are the relative percentage of points within each quadrant across all simulations. (C, D) The distribution of M_{SLR} and M_{AMI} respectively estimated from 500 simulations for each of three proteins (1qhw, 2ppn, and 1pek), and the distributions based on a null model where propensities changed randomly over time.

Estimates from M_{SLR} suggest that both Stokes and anti-Stokes shifts occurred with similar frequencies (figure 3.3C, table 3.1). Alternatively, estimates from M_{AMI} suggest an excess of Stokes compared to anti-Stokes shifts (figure 3.3D, table 3.1). Why do percentages differ under M_{SLR} and M_{AMI}? To investigate this, I developed a null model by randomly sampling propensities from the empirical distribution of resident amino acid propensities observed throughout the simulations. In these null model simulations, M_{SLR} estimated equal percentages for both evolutionary shifts (table 3.1). However, M_{AMI} estimated a higher percentage of Stokes shifts (65.4%) compared to anti-Stokes shifts (34.6%), with an excess of Stokes shifts that is greater than in the stability simulations. The higher

occurrence of Stokes shifts is, therefore, unlikely to be a consequence of stability-mediated epistasis. Rather, it possibly reflects a statistical artifact associated with M_{AMI}.

Table 3.1: Percentage of Stokes and anti-Stokes shifts from the stability simulations are consistent with random fluctuations in propensities. Results are based on 500 protein-specific simulations (1qhw, 2ppn, and 1pek), and a null model where propensities changed randomly over time.

	% anti-Stokes		**% Stokes**	
	$\mathbf{M}_{SLR}$	$\mathbf{M}_{AMI}$	$\mathbf{M}_{SLR}$	$\mathbf{M}_{AMI}$
1qhw	51.8	42.4	48.2	57.6
2ppn	51.0	41.0	49.0	59.0
1pek	50.9	40.8	49.1	59.2
Null model	49.0	34.6	51.0	65.4

We can understand the cause of this kind of artifact by evaluating the distribution of amino acid propensities. Propensities are often less than 0.5 when an amino acid is first substituted, and in most cases, they remain low (figure 3.4A). The distribution of an average of sampled propensities from such a distribution will not be the same as the distribution of a single propensity (figure 3.4A compared to 3.4B). I, therefore, hypothesize that the asymmetry in the propensity distribution is leading to higher estimates of Stokes shifts under M_{AMI}. To test this hypothesis, I developed two additional null models. The first null model samples propensities from the normal distribution N(0,0.1). The second model samples from the uniform distribution U(0,1). When the propensity distribution was symmetric, both metrics estimated equal proportions of Stokes and anti-Stokes shifts. This demonstrates that M_{AMI} is sensitive to the shape of the propensity distribution and will estimate an excess of Stokes shift if the distribution is asymmetric.

An important question remains: Why are propensities so often less than 0.5 in the stability simulation? Substitutions tend to occur within a "neutral zone" where the original and newly substituted amino acids have similar fitness contributions, and therefore similar propensity values (Goldstein and Pollock, 2017). This is evident from the higher correlation between propensities of the original and newly substituted amino acids than the correlation between the original amino acid and other residues (figure 3.4C). Since all 20 amino acid propensities must sum to one, and the propensities for the original and newly substituted

amino acids must be similar, they are likely to be ≤ 0.5.

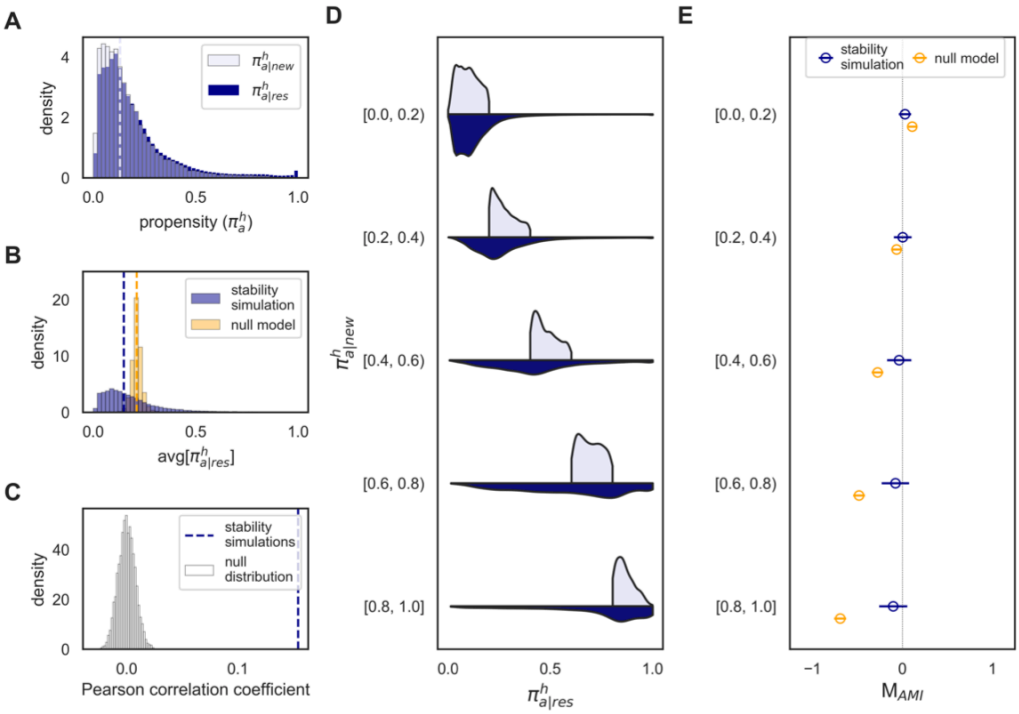


Figure 3.4: Stability-mediated epistasis conserves amino acid preferences. (A) Empirical distribution of initial ($\pi^h_{a|new}$) and resident ($\pi^h_{a|res}$) amino acid propensities observed during simulations of the 1qhw protein. Dotted line represents the median $\pi^h_{a|new}$ value. (B) Distribution of the average propensity of an amino acid while it is resident in the stability simulation (blue distribution) and the null model where propensities varied randomly over time (yellow distribution). Lines represent the median value from the respective distribution. (C) Pearson correlation between propensities of previously resident amino acids (π^h_{old}) and newly accepted residues (π^h_{new}) observed in the simulations (blue line) compared to a null distribution (grey distribution). The null distribution was obtained by estimating the correlation between the propensity of the previous amino acid (π^h_{old}) and the propensity of a randomly sampled residue given the same site and background sequence. This was repeated 10,000 times. (D) Violin plots showing the distributions of $\pi^h_{a|res}$ (dark blue) given that $\pi^h_{a|new}$ (light blue) was within a specific range. (E) The mean and standard deviation for M_{AMI} estimates within each $\pi^h_{a|new}$ range.

3.3.3 Stability-mediated epistasis conserves, rather than alters, amino acid propensities

The percentages of Stokes and anti-Stokes shifts from the random null model were similar to those from the stability simulations under M_{SLR} (Table 3.1). However, estimates under

M_{AMI} differed and the distributions of values were markedly different (figure 3.3D). The median M_{AMI} was an order of magnitude higher in the random model (5e-2) compared to the stability simulation (7e-3). In the null model, average propensities were often higher than initial propensities, whereas average and initial values were approximately equal in the stability simulations (compare median lines in figure 3.4A & 3.4B). Furthermore, the distribution of resident amino acid propensities in the stability model tended to closely match the distribution of initial propensities (figure 3.4D). Taken together, these results suggest that propensities were more conserved in the stability simulations compared to the null expectations.

Indeed, in the stability simulations, an amino acid having a high initial propensity is likely to continue enjoying high propensity throughout its residency, and low initial propensities often remain low (figure 3.4D). When initial propensities were between 0.0 and 0.2, there were fewer instances of Stokes shift in the stability simulation (65%) than the null model (98%), leading to a lower average M_{AMI} value (figure 3.4E). In contrast, when initial propensities were high, between 0.8 and 1.0, there were fewer instances of anti-Stokes shifts in the stability simulation (75%) than the null model (100%), leading to a higher average M_{AMI} value (figure 3.4E). Across all initial propensity ranges, fewer instances of either Stokes or anti-Stokes shifts occur under stability-constrained evolution, as compared to the unconstrained null. This further supports that propensities are more conserved in the stability simulations.

Stability-mediated epistasis conserves propensities through a significant negative first-order autocorrelation in propensity changes (autocorrelation averaged across sites were between -0.22, -0.24, and -0.21 for the 1qhw, 1pek, and 2ppn proteins). In other words, increases in propensity tend to be followed by decreases (and vice versa) leading to lower variability in propensities in the stability model compared to the null expectation. While these results suggest that stability-mediated epistasis frequently conserves amino acid propensities, there will be instances where propensities shift considerably over time at some sites. Importantly, however, nonadaptive dynamics will be balanced in the frequencies and magnitudes of Stokes and anti-Stokes shifts.

3.3.4 The dynamics of evolutionary Stokes and anti-Stokes shifts are comparable under nonadaptive evolution

The current metrics cannot distinguish between different underlying propensity dynamics. For example, the metrics estimate similar values for the following scenarios: (1) a rapid increase (or decrease) in amino acid propensity followed by a longer period where the propensity remains high (or low), and (2) a more gradual increase (or decrease) in propensity over time. It may be the case that evolutionary Stokes shifts occur soon after a substitution, while evolutionary anti-Stokes shifts are more gradual. To quantify whether propensity changes accelerated or decelerated, I compared the absolute value of each metric calculated over the first half of the amino acid residency ($M1_X$) and the estimate over the second half ($M2_X$), where X is either SLR or AMI . Specifically, I calculated ($|M2_X|$ - $|M1_X|$) / T_{res} where T_{res} is the amino acid residency time (measured in number of substitutions). I found no significant differences in the average rates of change between Stokes and anti-Stokes shifts (Welch's t-test, P-values > 0.05, table 3.2).

Table 3.2: Differences in average rate of change between substitutions experiencing evolutionary Stokes and anti-Stokes shifts. The rate of change is calulated as $|M2_X|$ - $|M1_X|$) / T_{res}. Reported are the p-values based on Welch's t-test. Substitutions are classified as undergoing an evolutionary Stokes (or anti-Stokes shift) if the corresponding metric value was greater than (or less than) zero.

	1qhw		**1pek**		**2ppn**	
	Difference in means	P-value	Difference in means	P-value	Difference in means	P-Value
$X = SLR$	-1.7e-6	0.55	1.3e-6	0.58	-1.8e-6	0.59
$X = AMI$	-1.3e-6	0.62	2.3e-6	0.31	-2.2e-6	0.41

Another dynamic might be missed with the current metrics: do physicochemically similar amino acids experience similar shifts in propensities? Goldstein and Pollock (2017) observed that when a site experiences an evolutionary Stokes shift, not only does the propensity for the resident amino acid increase but so does the propensity for physicochemically similar residues. For example, if V becomes newly resident at a site, then the propensity for it and similar amino acids ($e.g.$ L) will increase. Is the same behaviour expected for anti-Stokes shifts? To address this, I grouped amino acids that tend to interchange rapidly and that have similar chemical properties: [AST], [C], [DE], [FY], [GN], [HQ], [IV], [KR], [LM], [P], [W] (Susko and Roger, 2007). Rather than

evaluating the propensity dynamics for an individual amino acid, I tracked the dynamics of amino acid groups and applied the metrics to the summed group propensities. If evolutionary anti-Stokes shifts only affect individual amino acids, I expect %anti-Stokes in the binned analysis to be less than the %anti-Stokes for individual amino acids. However, the estimated percentage from the binned and individual amino acid analyses were similar (figure 3.5). Overall, these results suggest that both evolutionary shifts tend to induce comparable dynamics for similar amino acids.

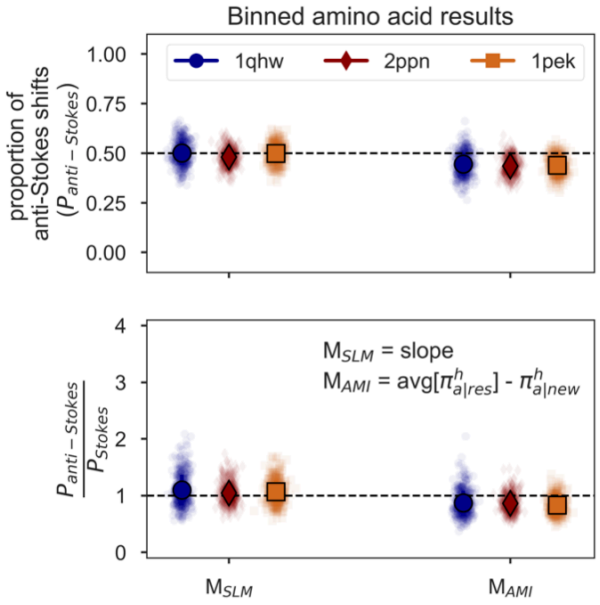


Figure 3.5: Percentages of Stokes and anti-Stokes shifts based on binned analyses. Amino acids were grouped as: AST, C, DE, FY, GN, HQ, IV, KR, LM, P, W (Susko and Roger, 2007). Bins represent amino acids which tend to interchange rapidly and have similar chemical properties. Evolutionary shifts were calculated based on the sum of propensities for all amino acids in a specific bin. (A) Approximately half of substitutions are followed by evolutionary anti-Stokes shifts based on metrics M_{SLR} and M_{AMI}. (B) Evolutionary Stokes and anti-Stokes shifts occur at similar frequencies (%anti-Stokes / %Stokes $\approx$ 1)

3.3.5 Evolutionary Stokes and anti-Stokes shifts both occur at exposed and buried sites

A site's location in a protein influences its evolutionary dynamics. For globular proteins, surface residues are usually involved with protein function ($e.g.$, binding affinity, enzymatic activity) with a preference for hydrophilic residues, while buried sites prefer hydrophobic

residues and evolve slower (Yeh *et al.*, 2014; Shahmoradi *et al.*, 2014; Marcos and Echave, 2015; Echave *et al.*, 2015). Two measures of a site's location in the protein are relative solvent accessibility (RSA) and weighted contact number (WCN). Both RSA and WCN correlate significantly with substitution rates in natural (Yeh *et al.*, 2014; Shahmoradi *et al.*, 2014; Marcos and Echave, 2015) and simulated proteins (Youssef *et al.*, 2020). Exposed sites have higher substitution rates, higher RSA, and lower WCN than buried sites. In line with these observations, I found a negative correlation between average residency time and RSA and a positive correlation with WCN (figure 3.6).

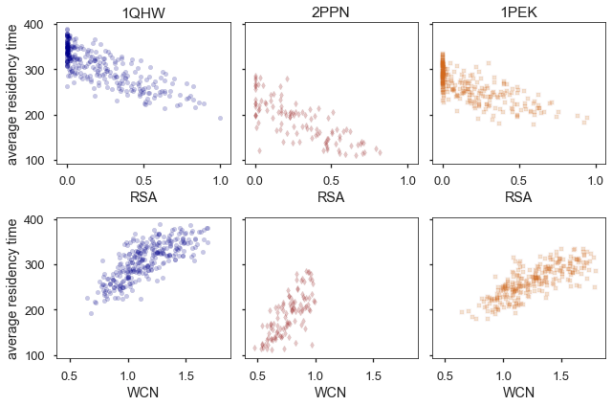


Figure 3.6: Relationship between average amino acid residency time and location in the protein. Plotted are the correlations with relative solvent accessibility (RSA, top row), and weighted contact number (WCN, bottom row) for three proteins (1qhw, 2ppn, 1pek).

Popova *et al.* (2019) recently suggested that buried sites are more likely to undergo evolutionary Stokes shifts, while exposed sites are more prone to decreases in propensities. I assessed sites' susceptibility to evolutionary Stokes and anti-Stokes shifts by examining the relationship between the metrics and location in the protein (figure 3.7). The average values of M_{SLR} at exposed and buried sites were not significantly different (Welch's t-test, P-values > 0.05 for all proteins; table 3.3). While the average M_{AMI} values were significantly higher at buried compared to exposed sites (Welch's t-test, P-values < 0.001 for all proteins; table 3.3), the effect sizes were minor (6e-3, 1e-2, 8e-3 for the 1qhw, 1pek, and 2ppn proteins respectively). Therefore, evolutionary Stokes and anti-Stokes shifts are not associated with the locations of sites in a protein. This conclusion is consistent with experimental results in the HIV envelope protein where sites with shifted propensities were observed across the protein (Haddox *et al.*, 2018).

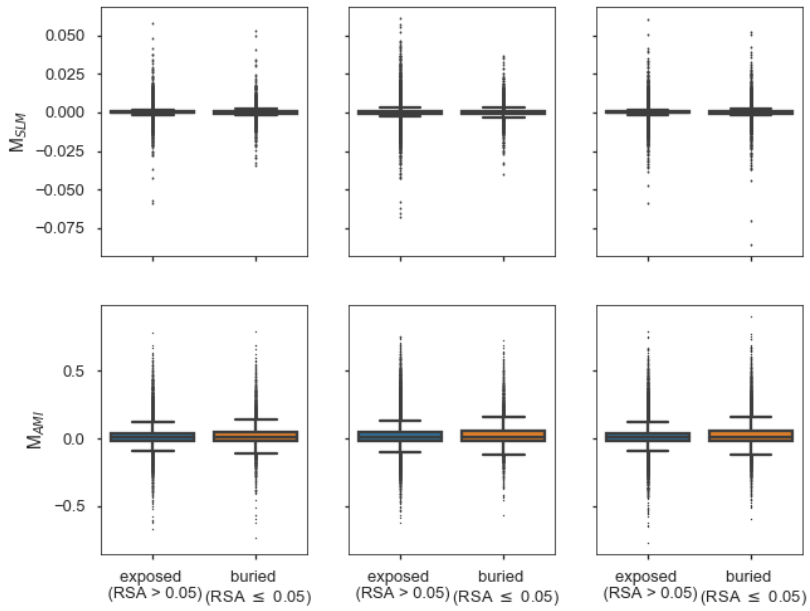


Figure 3.7: Evolutionary shifts in propensities occur with similar frequency and magnitude at exposed and buried sites. Sites are considered exposed if their relative solvent accessibility (RSA) is > 0.05, and are considered buried if $RSA \leq 0.05$. The columns report the results from simulations of the 1qhw, 2ppn, and 1pek proteins, respectively.

Table 3.3: Differences in the average metric value based on position in the protein (exposed versus buried sites). Reported P-values are based on Welch's t-test. Null hypothesis is that both exposed and buried sites have identical mean values. A site is considered buried if its relative solvent accessibility (RSA) ≤ 0.05, and is exposed if $RSA > 0.05$.

	1qhw		**1pek**		**2ppn**	
	Difference in means	P-value	Difference in means	P-value	Difference in means	P-Value
$\mathbf{M}_{SLR}$	-4.2e-5	0.16	-7.5e-5	0.005	7.5e-5	0.063
$\mathbf{M}_{AMI}$	-0.006	<0.001	-0.010	<0.001	-0.008	<0.001

While the previous result suggests that all sites are equally susceptible to undergoing evolutionary Stokes or anti-Stokes shifts, it remains unclear if the entailed dynamics are comparable. I was interested in assessing whether location in the protein might influence the rate of propensity changes. For example, a deleterious substitution at a surface site might be compensated for by adjustments at a small number of interacting sites, leading to

a rapid evolutionary shift. Alternatively, a deleterious substitution at a highly connected site might require more adjustments at other positions, and, therefore, the propensity shift may be gradual. However, I found that the average rates of change were not significantly different at buried and exposed sites (all p-values were > 0.17, Welch's t-test; table 3.4).

Table 3.4: Differences in average rate of change between substitutions based on position in the protein (exposed versus buried sites). The rate of change is calulated as $|M2_X|$ - $|M1_X|$) / T_{res}. Reported are the p-values based on Welch's t-test. Substitutions. Null hypothesis is that both exposed and buried sites have identical mean values. A site is considered buried if its relative solvent accessibility (RSA) ≤ 0.05, and is exposed if $RSA > 0.05$.

	1qhw		**1pek**		**2ppn**	
	Difference in means	P-value	Difference in means	P-value	Difference in means	P-Value
$X = SLR$	3.1e-6	0.39	1.0e-6	0.72	-4.9e-6	0.35
$X = AMI$	-1.2e-7	0.97	3.8e-6	0.17	-2.3e-7	0.58

3.3.6 Stabilizing substitutions increase resident amino acid propensities while destabilizing substitutions decrease them

I have shown that long term shifts in amino acid preferences can occur because of non-adaptive stability-mediated epistasis. Next, I turn to the underlying mechanisms that cause changes in propensities after a single substitution. Following a substitution, the fitness and propensity landscapes at most sites in the protein will change because of epistasis. Important questions about how substitutions alter propensities remain unanswered: Do substitutions tend to favourably impact some sites (by increasing their resident amino acid propensities) while simultaneously disadvantaging other sites (by decreasing their resident amino acid propensities)? Or does a substitution impact propensities similarly across sites? I found that the effect of substitution on resident amino acid propensities is unbalanced. Substitutions either favourably (or disfavorably) impact most sites by increasing (or decreasing) their resident amino acid propensity (figure 3.8A). Stabilizing substitutions ($\Delta\Delta G < 0$) were associated with decreases in propensities of resident amino acids at most sites while destabilizing substitutions ($\Delta\Delta G > 0$) caused propensities to increase (figure 3.8B).

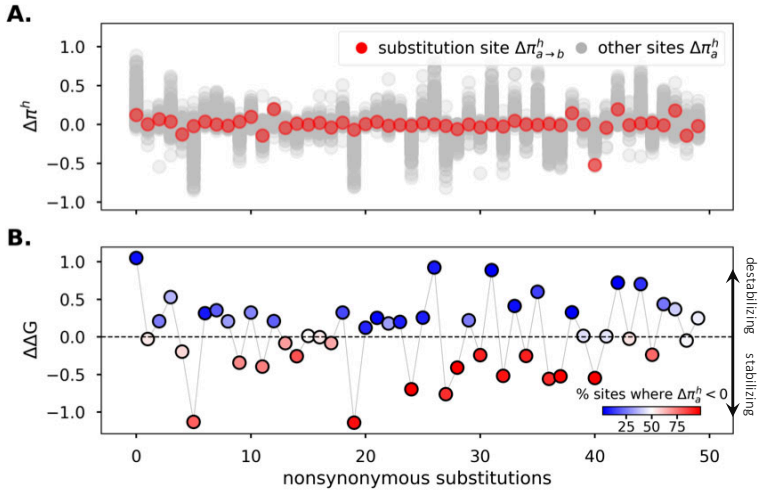


Figure 3.8: Stabilizing substitutions reduce resident amino acid propensities while destabilizing substitutions often increase propensities. (A) Stability-mediated epistasis between sites results in changes in resident amino acid propensities as substitutions accrue. Following an amino acid replacement at one position in the protein, so that the sequence changes from $S_x \rightarrow S_{x+1}$, the propensity of the resident amino acids at all sites will change. The grey dots are the changes in the propensities of the resident amino acids at each site following a substitution, $\Delta\pi^h_a = \pi^h_a(S_{x+1}) - \pi^h_a(S_x)$. The red dots are the change in the propensity of the resident amino acid at the substitution site, and therefore a change in the amino acid from $a \rightarrow b$ ($\Delta\pi^h_{a\to b} = \pi^h_b(S_{x+1}) - \pi^h_a(S_x)$). (B) Stabilizing substitutions ($\Delta\Delta G < 0$) decrease resident amino acid propensities at most sites. In contrast, destabilizing substitutions ($\Delta\Delta G > 0$) result in a lower percentage of sites where $\Delta\pi^h_a < 0$.

To illustrate the effect, consider the dynamics following a stabilizing substitution from $S_1 \rightarrow S_2$ (figure 3.9). I focus on site 145 as an example of the site-specific dynamics. The uphill move from S_1 to S_2 flattened the fitness landscape at site 145. Given that sequence S_2 has greater stability, a destabilizing mutation has a smaller fitness effect relative to the same mutation in the less stable S_1 sequence. How does the change in the fitness landscape relate to variations in propensities? Since a higher number of amino acids can now occupy the site with little or no detriment to protein fitness, the propensity landscape will similarly become more uniform (figure 3.9C). Amino acids like R, N, and P that had low propensity in the context of sequence S_1, are more likely given the "stability-buffered" sequence S_2 (figure 3.9C). Since propensities must sum to one, the increase in the propensity of some amino acids ($e.g.$, R, N, and P) will cause a decrease in the propensity of the resident

amino acid (K in this example; figure 3.9). The opposite trends are evident following the fixation of a destabilizing mutation (figure 3.9D). The fitness (figure 3.9E) and propensity (figure 3.9F) landscapes became less uniform, with fewer amino acids having non-zero propensities, and an increase in resident amino acid propensities.

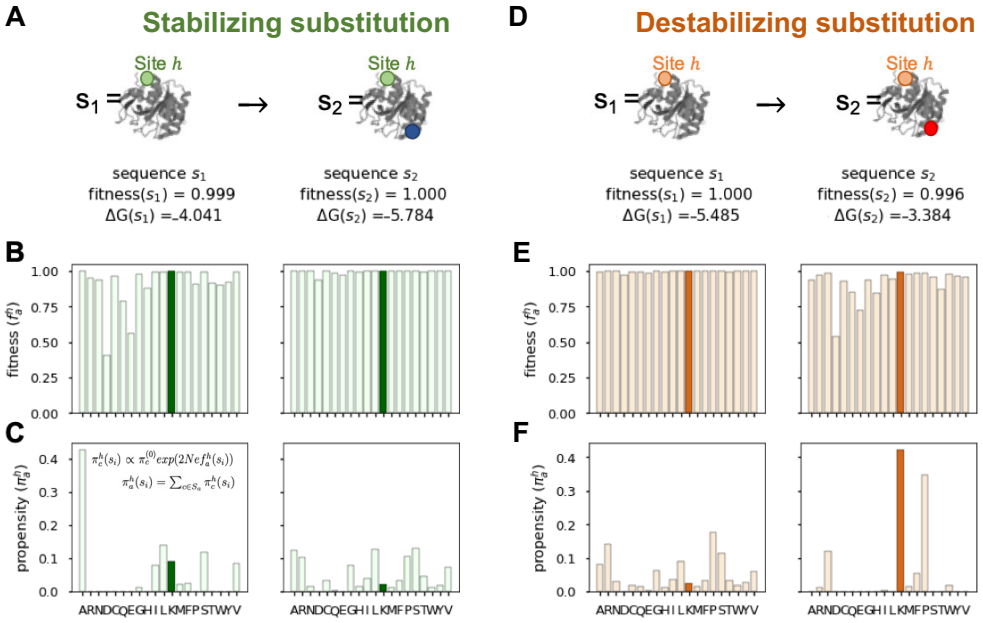


Figure 3.9: Epistatic dynamics following the fixations of stabilizing (A,B,C) and destabilizing (D,E,F) substitutions. (A) Let S_1 be the initial protein sequence, and S_2 be the sequence following the acceptance of a stabilizing substitution (blue dot). Given the stability-buffered sequence S_2, deleterious mutations which would not have been fixed in S_1 are now more likely to be fixed (e.g. R, N,P). The fitness landscape (B) and propensity (C) landscapes at a non-substituted site 145 becomes more uniform. The fitness and propensity of the resident amino acid is shown in dark green. (D, E, F) are the respective plots following the fixation of a destabilizing substitution (red dot). The fitness and propensity landscapes at the non-substituted site become less uniform. These landscapes were observed in simulations of the 1pek protein.

To quantify the effect across all sites, I measured landscape uniformity using Shannon entropy $\mathrm{H}^h(S)$ (see Methods section for detail). Entropy is highest when the landscape is uniform ($i.e.$, all amino acids have equal frequencies) and is at a minimum (= 0) when only one amino acid has a non-zero propensity. Note that the uniformity of fitness and propensity landscapes are highly correlated (figure 3.10). The fitness landscape describes the fitness of nearby sequences, while propensity landscapes consider how frequently nearby sequences

are explored. I, therefore, report the entropy of the propensity landscapes, although I expect similar results based on fitness landscapes. As expected, at higher stability values (lower ΔG), the landscapes were more uniform compared to at lower stability values (figure 3.11A, Spearman correlation coefficients < -0.98 for all proteins, P-values < 0.001).

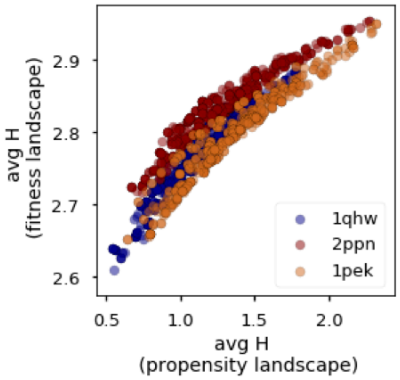
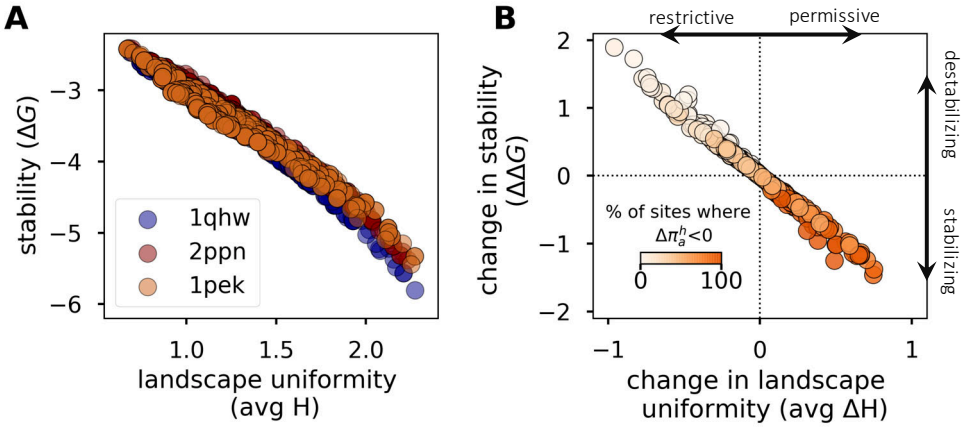


Figure 3.10: Relationship between the Shannon entropy of a propensity landscape compared with the entropy of the fitness landscape. Reported are the average entropy values over all sites given a particular background sequence from a single simulation trial for each of three protein structures (1qhw, 2ppn, and 1pek).

Next, I assessed how substitutions alter landscape uniformity. A change from a uniform to a rugged landscape (with a small number of amino acids having non-zero propensities), will result in a negative $\Delta \mathrm{H}^h$. In contrast, a positive $\Delta \mathrm{H}^h$ indicates an increase in landscape uniformity. I considered a substitution as permissive if, on average, it increased landscape uniformity across sites ($i.e.$, a positive average ΔH). A restrictive substitution is one where following its acceptance, the landscapes at most sites permit fewer amino acids ($i.e.$, a negative average ΔH). The stability effect of a substitution ($\Delta\Delta G$) is strongly correlated with its influence on landscape uniformity (figure 3.11B, Spearman correlation coefficient -0.99, P-value < 0.001). Consistent with the results in figure 3.9, stabilizing substitutions provide a stability-buffered background so that slightly destabilizing mutations are more likely to be fixed, expanding the space of potential evolutionary paths (figure 3.11C). In contrast, destabilizing substitutions were restrictive, limiting potential evolutionary trajectories (figure 3.11C).

These results are consistent with evolutionary dynamics on saturating fitness functions (Cherry, 1998; Goldstein, 2013). On a saturating fitness curve, mutational effects decrease

with increasing fitness. At a higher point on the fitness curve, site-specific fitness landscapes will be more uniform. Since more amino acids may be tolerated at the site, the propensity for the resident amino acid will decrease. Therefore, stabilizing substitutions entail decreases in the propensities of resident amino acids due to the general flattening of the site-specific fitness landscapes. Alternatively, destabilizing substitutions tend to increase resident amino acid propensities because some mutations become selectively inviable and have low fixation probabilities.



C

Substitution Type	Percentage of nonsynonymous substitutions		
	1qhw	1pek	2ppn
Stabilizing and permissive	48.4	49.1	48.8
Stabilizing and restrictive	0.21	0.23	0.47
Destabilizing and permissive	0.25	0.17	0.39
Destabilizing and restrictive	51.1	50.5	50.3

Figure 3.11: Stabilizing substitutions are permissive and destabilizing substitutions are restrictive. (A) The relationship between protein stability (ΔG) and landscape uniformity, measured as the entropy of the propensity landscape averaged over all sites in the protein (avg H). (B) The relationship between the stability effect of a substitutions ($\Delta\Delta G$) and the resulting average change in landscape uniformity (avg ΔH). Color bar represents the percentage of sites for which the propensity for the resident amino acid decreased ($\Delta\pi_a^h < 0$). Positive avg ΔH values imply that, on average, the landscapes became more uniform. Therefore, the substitution is deemed permissive. Negative avg ΔH are indicative of restrictive substitutions. Plotted results are based on a single simulation of the 1pek protein. (C) The percentages of different types of substitutions for each of three proteins (1qhw, 2ppn, and 1pek). Percentages are calculated from 500 protein-specific simulations

3.4 Discussion

In this chapter, I have examined the evolutionary dynamics of proteins under nonadaptive stability-constraints. I found that as proteins become more stable adverse fitness effects of mutations diminish, thereby expanding the space of potential evolutionary trajectories. It has been suggested that highly stable proteins may be more adaptable to new functions since they are more likely, than less stable proteins, to accept destabilizing yet functionally beneficial mutations (Schreiber *et al.*, 1994; Nagatani *et al.*, 2007; Wang *et al.*, 2002; DePristo *et al.*, 2005). I suggest that highly stable proteins, all other things being equal, may also be more adaptable because they are more apt to explore neighbouring regions of sequence space. Nevertheless, it is important to note that selection on other properties of proteins, such as their expression level and the cost of translation error (Drummond *et al.*, 2005), can also influence their evolution. Therefore, the relationship between evolvability and stability of proteins is likely to reflect the complex interplay of multiple factors.

As more (or fewer) mutations become accessible, the propensity for the resident amino acid at a site will change. Stabilizing substitutions expand evolutionary paths and, in doing so, decrease resident amino acid propensities (figure 3.8 & 3.11). By contrast, destabilizing substitutions limit accessible trajectories and increase resident amino acid propensities. At mutation-selection-drift equilibrium, the proportion of stabilizing and destabilizing substitutions should be equal (Cherry, 1998; Goldstein, 2011). As such, a balance is expected in the proportion of increases and decreases in propensities. This balance may arise if (1) for any given site the number of propensity increases is the same as the number of decreases; (2) some sites undergo systematic increases in propensities, an evolutionary Stokes shift, while others undergo systematic decreases in propensities, an anti-Stokes shift; or (3) some combination of these two phenomena. The results presented here favour the latter scenario since there is no tendency for sites to experience Stokes versus anti-Stokes shifts and fluctuations in propensities are negatively autocorrelated. While, in some instances, the propensity at a site may drift upwards with time, the dynamics at equilibrium are such that there will be an approximately equal number of sites experiencing the opposite trend.

Propensity shifts cannot be directly observed in natural proteins. Instead variation in replacement rates over time may be inferred as a proxy: increases in resident amino acid propensity lead to decreases in replacement rates, and decreases in propensity lead to

higher replacement rates (Popova *et al.*, 2019; Stolyarova *et al.*, 2020; Gelbart and Stern, 2020). Analysis of natural protein alignments often reveal a balance in the number of rate accelerating and decelerating sites. For example, across five mitochondrial genes 21/28 sites showed evidence of replacement rate decreases/increases (Stolyarova *et al.*, 2020), and 137/134 sites across nine proteins in HIV and SIV (Gelbart and Stern, 2020). Popova *et al.* (2019) analysed four influenza A protein alignments and found that the ratios of replacement rate decreases/increases were 2/0, 0/0, 4/12, and 5/8 for the H1, N2, H3, and N2 proteins, respectively. The balance in rate increases and decreases in these datasets is suggestive of nonadaptive processes. Nonetheless, the excess of rate increases in the H3 protein could be evidence of adaptive herd immunity (Popova *et al.*, 2019). Since nonadaptive processes can shift propensities, it is important to calibrate our evidence of adaptations with nonadaptive signals. Future work assessing the dynamics of propensity shifts under adaptive evolution is warranted.

I defined an evolutionary Stokes shift as an increase in the propensity of a resident amino acid at a site (Pollock *et al.*, 2012). This can be thought of as a 'site-level' evolutionary Stokes shift. Goldstein and Pollock (2017) later described a 'sequence-level' Stokes shift where they break down the stability of the sequence into two components: the pairwise energetic contribution from all interactions with a resident amino acid at a focal site, and the contribution from all other interactions in the sequence. The sequence-level Stokes shift is then described in terms of a pull towards regions in sequence space where the energetic contribution of a resident amino acid is high. The prediction is: if the resident amino acid at a focal site is highly stabilizing, then the stability contribution from the background sequence will be low. Since a higher number of background sequences will satisfy lower stability contributions, the evolutionary process will remain in such regions of sequence space for more extended periods. Note, however, that the choice of focal site is arbitrary. Therefore, as the stabilizing contributions of the resident amino acid at a particular site increases, it alleviates the need for the resident amino acid at another site to be high. As such, the sequence-level Stokes shift hints at the existence of a site-level anti-Stokes shift, since the lower stability contributions from the background sequence imply that some non-focal sites may be free to have lower stability contributions (and thus lower propensities). Here, I make explicit that such decreases in preference do occur.

Evolutionary Stokes and anti-Stokes shifts relate to changes in the propensities of

resident amino acids due to nonadaptive stability-mediated epistatic effects. Alternatively, contingency and entrenchment are phenomena describing changes in the relative fixation probabilities of substitutions (Shah *et al.*, 2015) which can arise by adaptive or nonadaptive means. While these phenomena (entrenchment $\approx$ evolutionary Stokes shift, and contingency $\approx$ evolutionary anti-Stokes shift) have been used interchangeably (de la Paz *et al.*, 2020), they are related yet distinct. It is evident that a site-level evolutionary Stokes shift may lead to the entrenchment of a resident amino acid at a site making it less likely to revert over time. I suggest that a site-level anti-Stokes shift could similarly contribute to the phenomenon of contingency. A particular evolutionary history that decreases the propensity of the resident amino acid concurrently entails increases in propensities of non-resident amino acids, thereby increasing their probabilities of fixation. This includes cases where the resident and a non-resident amino acid enter the nearly neutral zone, and the increased probability of their substitution at such times is consistent with the concept of contingency in Shah *et al.* (2015). As such, substitutions may be contingent on other changes occuring in the protein. However, the original conception of contingency was related to external chance events (Gould, 1991). Therefore, evolutionary Stokes and anti-Stokes shifts may produce signals consistent with entrenchment and contingency. Nevertheless, contingency and entrenchment may arise by other means and are not limited to variations in propensities due to stability constraints.

An advantage of thermodynamic stability models is that they provide plausible nonadaptive null models for protein evolution (Goldstein, 2011; Pollock *et al.*, 2012; Goldstein and Pollock, 2017). They have been used to critically assess adaptationist claims about the trade-offs between protein function and stability (Taverna and Goldstein, 2002; Goldstein, 2011), and protein function and foldability (Govindarajan and Goldstein, 1996). "Despite the seduction of adaptive rationalizations", to quote one of the original authors of this model, "neutral evolutionary dynamics remains the null model that must first be rejected" (Goldstein, 2011). The demonstration that amino acid propensities may decrease over time in the absence of external environmental changes does not preclude that environmental shifts could render resident amino acids less favourable. Rather the results presented here demonstrate that decreases in propensities are expected to occur in the absence of external changes, and therefore that their mere occurrence should not, on their own, be taken as conclusive evidence of adaptations.

3.5 Methods

3.5.1 Descriptions of natural proteins

I simulated the evolution of three proteins with PDB codes 1qhw, 2ppn, and 1pek. The proteins differ in structure, function, length, and contact density. The 1qhw protein is a phosphatase, the 1pek protein is a proteinase, and the 2ppn protein is an isomerase. The 1qhw protein has 300 amino acids, 1pek is made of 297 amino acids, and the 2ppn protein comprises 107 residues. The 1pek protein was the most densely packed with an average number of contacts per site of 8.4 compared to 7.5 for the 1qhw protein and 6.9 for the 2ppn protein. During the simulations, I used the nucleotide frequencies (π_n) and transition/transversion rate (κ) parameters estimated from multiple sequence alignments for the corresponding protein used in Youssef *et al.* (2020) (Chapter 1 in this thesis). The mutation parameters ($\kappa, \pi_A, \pi_C, \pi_G, \pi_T$) were set equal to (4.37, 0.21, 0.32, 0.28, 0.20) for the 1qhw protein; (0.90, 0.19, 0.35, 0.56, 0.21) for the 1pek protein; and (2.50, 0.27, 0.24, 0.29, 0.19) for the 2ppn protein.

3.5.2 Evolutionary model

For simulations I used the MutSel model (section 1.3.1) with fitness values estimated from the stability-infromed framework (section 1.3.2). Sequence space is vast and randomly sampling sequences rarely produces viable proteins. Starting the simulation in such dire conditions leads to evolution randomly drifting between low fit sequences. Therefore, I used the algorithm outlined in section 2.5.6 to obtain protein sequences with fitness values ≥ 0.99 given the corresponding protein structure. Then, I evolved the equilibrated sequence for 500 substitutions while keeping track of the site-specific fitness landscapes at all sites. The reported results are based on the post-equilibration phase. I generated 500 protein-specific replicates for each protein. Unless otherwise stated, I assumed $N_e = 100$.

3.5.3 Assessing robustness of results to simulation settings

In order to assess the robustness of the results to different simulation settings, I conducted additional simulations using the 1qhw protein. The additional simulations included: (1) allowing for a longer equilibration phase; (2) increasing the number of substitutions per simulation; (3) increasing the effective population size; (4) increasing both the effective population size and the lenght of the simulation; and (5) looking at changes in propensities

as defined in Pollock *et al.* (2012). In all the additional experiments, the percentage of Stokes and anti-Stokes shifts were consistent with the results reported above (table 3.5).

Table 3.5: Assessing robustness of results to simulation settings. Reported are the percentage of substitutions where the metric (M_{SLR} or M_{AMI}) values was negative. P-values as based on the Binomial test where the null hypothesis assumes an equal percentage of positive and negative values. Results based on simulations of the 1qhw protein.

Simulation	Number of trials	Number of windows	$\%M_{SLR}$ <0	P-value	$\%\ M_{AMI}$ <0	P-value
N_e = 1e2	500	36,934	51.8	<0.001	42.4	<0.001
N_e = 1e6	50	3,659	50.0	1.0	40.8	<0.001
Thermodynamic propensity	500	36,934	51.2	<0.001	42.2	<0.001
Longer equilibration phase	50	4,268	49.3	0.40	39.6	<0.001
# subs = 5000	50	88,229	51.1	<0.001	34.3	<0.001
N_e = 1e6 and # subs = 5000	50	85,643	51.2	<0.001	34.8	<0.001

For all results described in this chapter, I only considered the dynamics when a residue was accepted and subsequently replaced within the time-frame of the simulation, and where the amino acid was resident for at least ten substitutions. However, I repeated the analyses with the inclusion of partial windows (where for example an amino acid is accepted during the simulation but the simulation ends prior to its replacement) which revealed similar results with respect to the proportion of evolutionary Stokes and anti-Stokes shifts.

3.5.4 Amino acid propensities

Suppose that for a simulation trial we observe $S_1 \rightarrow S_2 \rightarrow ... \rightarrow S_{500}$ where the S_x's are the codon sequences realized during the simulations, and S_x and S_{x+1} differ by a single nucleotide substitution (synonymous or nonsynonymous). Given a sequence $S = [c^1, \ldots, c^L]$, I can calculate the fitness of amino acid a at site h holding the rest of the sequence constant, $f_a^h(S) = f(c^1, \ldots, c^{h-1}, c^h(a), c^{h+1}, \ldots, c^L)$ where $c(a)$ is a codon encoding amino acid a. The fitness landscape is then $f^h(S) = \{f_1^h(S), \ldots, f_{20}^h(S)\}$. I use the fitness values to calculate the amino acid stationary frequencies using (3.1). I calculate $\pi_a^{(0)}$ as the sum over the neutral stationary frequencies for synonymous codons for each amino acid. The neutral frequency for a codon made up of nucleotide triplet ijk will be

proportional to $\pi_i \pi_j \pi_k$. For the mutation-selection model, the stationary frequency of a sequence S having codons $c^1, \ldots c^L$ is proportional to

$$\prod_h \pi^{(0)}_{c^h} \exp[f(S)]. \tag{3.2}$$

It follows that the (marginal) probability of amino acid a at site h is given by (3.1).

3.5.5 Description of metrics used to quantify evolutionary Stokes and anti-Stokes shifts

I defined two metrics to quantify shifts in propensities. First, let the residence time of an amino acid (T_{res}) be the time period between i and j, where i is the substitution when amino acid a first occupies the site and j is the last substitution. The first metric is the **S**lope of the **L**inear **R**egression over T_{res} where the covariate x is time (measured in substitutions) and the response y is the propensity of the resident amino acid a at site h (π^h_a). I refer to this metric as M_{SLR}. The second metric M_{AMI} is the difference in the **A**verage propensity of an amino acid while it is resident (avg[$\pi^h_{a|res}$]) **M**inus its **I**nitial propensity ($\pi^h_{a|new}$). Metrics values greater than 0 are suggestive of an evolutionary Stokes shift and values less than 0 are indicative of evolutionary anti-Stokes shifts. Figure 3.3 provides a visual representation of the metrics.

3.5.6 Quantifying the uniformity of a landscape

I used the Shannon entropy of a propensity landscape as a measure of its uniformity. I calculate entropy as

$$\mathrm{H}^h(S) = -\sum_a \pi^h_a(S) \ln \pi^h_a(S) \tag{3.3}$$

where $\pi^h_a(S)$ is the propensity of amino acid a at site h given background sequence S. The entropy is maximized when all amino acids are equally likely, and is minimized (= 0) when only a single amino acid is observed. To determine how the landscapes change in response to changes in the background protein sequence, I compared the entropy before and after the substitution

$$\Delta\mathrm{H}^h = \mathrm{H}^h(S_{x+1}) - \mathrm{H}^h(S_x) \tag{3.4}$$

I classified a substitution as permissive if the average ΔH across all sites was positive, and restrictive if the average ΔH was negative.

3.5.7 The rate of amino acid replacement

I calculated the rate of leaving the resident amino acid at a site h as the sum of the transition rates over all sequences that differ from the current sequence by a single nucleotide and have a different amino acid at site h.

3.5.8 Null model

I developed a null model in order to examine the dynamics of propensity shifts in the absence of the temporal effects of epistasis. I sampled 10,000 window sizes ($i.e.$, residency times) from the empirical distribution observed in the stability simulations. For each window, I randomly sampled propensity values from the empirical propensity distribution plotted in figure 3.4A. Then I estimated M_{SLR} and M_{AMI} for each window. The distributions of M_{SLR} and M_{AMI} are plotted in figure 3.3C and D, and the estimated percentages of Stokes and anti-Stokes based on each measure are reported in table 3.1.

3.5.9 Code availability

All code used to simulate, analyze, and plot data has been uploaded and is freely available from https://github.com/noory3/antiStokes_shifts.

CHAPTER 4

SHIFTS IN AMINO ACID PREFERENCES AS PROTEINS EVOLVE: A SYNTHESIS OF EXPERIMENTAL AND THEORETICAL WORK

This work was submitted to the journal *Protein Science*, and was done in collaboration with Edward Susko, Andrew Roger, and Joseph Bielawski.

4.1 Abstract

Amino acid preferences vary across sites and time. While variation across sites is widely accepted, the extent and frequency of temporal shifts are contended. Our understanding of the underlying drivers changing amino acid preferences is incomplete: To what extent are temporal shifts driven by adaptive versus nonadaptive evolutionary processes? I review phenomena that cause preferences to vary (e.g., evolutionary Stokes shift, contingency, entrenchment) and clarify how these phenomena differ. Then, to determine the extent and prevalence of shifted preferences, I review experimental and theoretical studies. Analyses of natural sequence alignments often detect decreases in homoplasy (convergence and reversions) rates, and variation in replacement rates with time. Such signals are consistent with temporally changing preferences. For example, as proteins diverge their set of preferred amino acids will likely differ, leading to lower homoplasy rates. While approaches inferring shifts in preferences from patterns in natural alignments are valuable, they are

indirect since multiple mechanisms (adaptive and nonadaptive) could have led to the observed signal. Alternatively, site-directed mutagenesis experiments allow for a more direct assessment of shifted preferences. They corroborate evidence from multiple sequence alignments, revealing that the preference for an amino acid at a site varies depending on the background sequence. However, shifts in preferences are usually minor in magnitude and sites with significantly shifted preferences are low in frequency. Nevertheless, the small yet consistent perturbations in preferences as proteins evolve could jeopardize the accuracy of inference procedures, which assume constant preferences. I conclude by discussing if and how such shifts in preferences influence widely used time-homogenous inference procedures and potential ways to mitigate their effects.

4.2 Introduction

Protein evolution is complex, leaving confounding signals in natural sequences. An evolutionary biologist interested in understanding the evolutionary history of a population, species, or protein must investigate these patterns and decipher their likely causes: Is the observed signal evidence of adaptive evolution, or could it have arisen by nonadaptive processes? To address these questions, we must first have a rigorous understanding of the patterns emerging under the interplay of random genetic drift and selective pressure to maintain protein function, but in the absence of adaptive processes. To this end, I review nonadaptive evolutionary phenomena and their identifiable footprint in natural sequences.

The space of possible protein sequences is vast. For an average-sized protein of length 300, the number of possible sequences (20^{300}) exceeds the number of atoms in the observable universe (10^{82}). This combinatorial explosion prohibits our ability to fully characterize the sequence-to-sequence (S2S) fitness landscape on which a protein evolves. A more tractable approach is to define the fitness landscape at an individual site in the protein. The *site-specific fitness landscape* is fully defined by a vector of 20 describing the fitness of the mutant protein created by placing each amino acid at the site given a particular background sequence S, where $f^h(S) = \{f_1^h(S), \ldots, f_{20}^h(S)\}$ defines the fitness landscape at a site h (Bazykin, 2015). From fitness landscapes, we can estimate *site-specific propensity landscapes*. Propensity can be defined as the expected frequency with which an amino acid occurs at a site (e.g., Chapter 3), or the fraction of sequences at thermodynamic equilibrium carrying that particular mutation (Pollock *et al.*, 2012). The

propensity for an amino acid is related to its fitness by

$$\pi_a^h(S) = \pi_a^{(0)} e^{2N_e f_a^h(S)} / \sum_x \pi_x^{(0)} e^{2N_e f_x^h(S)} \tag{4.1}$$

where N_e is the effective population size and $\pi_a^{(0)}$ is the expected frequency of amino acid a in the absence of selection (dos Reis, 2015). In this review, I use the more general term *site-specific preference landscape* to describe the relative preferences for amino acids, estimated from any of the above definitions. Preference landscapes are often normalized so that the sum of all amino acid preferences is equal to one and are usually represented using a heatmap (Bazykin, 2015), a sequence-logo plot (Bloom, 2015), or a barplot (Jones *et al.*, 2017) (figure 4.1).

Proteins evolve with various biophysical and evolutionary constraints on their structures and functions. Such selective constraints manifest as differences in preference landscapes among sites and across time. Spatial, or among-site, variability has been extensively studied revealing commonly observed patterns (Echave *et al.*, 2016). Buried sites often prefer hydrophobic residues, while surface sites have a higher affinity for hydrophilic amino acids. In addition, preference landscapes at surface sites are usually more uniform, with many residues having similar preferences, than at buried sites, where only a small number of amino acids have high preferences (Youssef *et al.*, 2020). Failing to account for such spatial variability can jeopardize the accuracy of inference procedures. As a result, various inference methodologies accommodate differences in frequency profiles across sites (e.g., Wang *et al.* (2008)). Temporal, or across-time, variability in preference landscapes is comparatively less understood. This has lead to the interpretation of temporal rate shifts as evidence of adaptive evolution (Yang and Nielsen, 2002; Zhang *et al.*, 2005); however, the role of nonadaptive processes, such as epistatic interactions between sites, in changing preferences and rates is gaining appreciation (Pollock *et al.*, 2012; Goldstein and Pollock, 2017; Shah *et al.*, 2015).

I begin by reviewing various nonadaptive phenomena that give rise to temporal shifts in preferences. Then, I discuss evidence for shifted preferences gleaned through analyses of natural sequence alignments. The observed levels of convergence rates, reversion rates, and replacement rates are broadly consistent with nonadaptive evolution. However, this evidence is inferential and indirect—other mechanisms which we may not yet appreciate

may be the ultimate causes of such signals. To more directly quantify the magnitude and prevalence of shifted landscapes, I discuss results from site-directed mutagenesis experiments. The conclusion from these datasets is that amino acid preferences shift over time. However, nonadaptive shifts are usually minor in magnitude and low in frequency. Nevertheless, such minor yet consistent perturbation in preference landscapes lead to variations in rates across time. I end by discussing the consequence these shifts might have on widely-used inference procedures and potential ways to mitigate their effects.

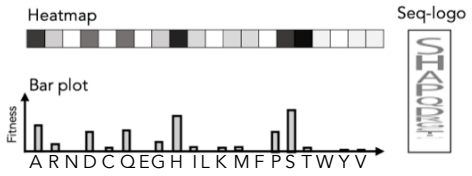


Figure 4.1: Different representations of site-specific preferences. In the heatmap representation darker shades imply higher preference. In the barplot, bar height represents the preference for the respective amino acid. In the sequence-logo (seq-logo) representation, the size of the letter represents its preference relative to other amino acids.

4.3 Causes of nonadaptive shifts in preferences

Protein evolution is commonly viewed as a walk in sequence space directed by natural selection, drift, and mutations. This was intuitively summarized by John Maynard Smith, where he used a word game as an analogy of protein evolution (Maynard Smith, 1970). Starting with a meaningful word, the objective is to, at each turn, change one letter to yield a different meaningful word. His example trajectory was WORD $\rightarrow$ WORE $\rightarrow$ GORE $\rightarrow$ GONE $\rightarrow$ GENE. Meaning, in this case, is defined as any English word and is therefore binary (a word is either meaningful or not). Despite its simplicity, Maynard Smith's word game analogy illuminates various salient evolutionary dynamics (figure 4.2). Relevant to this review, I will use it to illustrate how adaptive and nonadaptive processes can lead to similar dynamics for site-specific landscapes.

Analogous to a site-specific landscape, let us define a position-specific landscape as a 26-element vector for each letter in the English alphabet. Each letter is assigned a value of zero if it does not produce a meaningful word in the context of the characters present at the other positions and is assigned a value of one otherwise. A change in the

background sequence from -ORE to -ONE will cause a shift in the first-position fitness landscape. Letters (such as W) that produced meaningful words in the previous background (e.g.,WORE) are no longer meaningful in a new background (e.g., WONE). Similarly, letters that were nonviable may become permissible (e.g., DORE versus DONE). In this way, the position-specific landscape is dependent on the background sequence. In proteins, site-specific preference landscapes undergo similar dynamics—such context-dependence is referred to as *epistasis*.

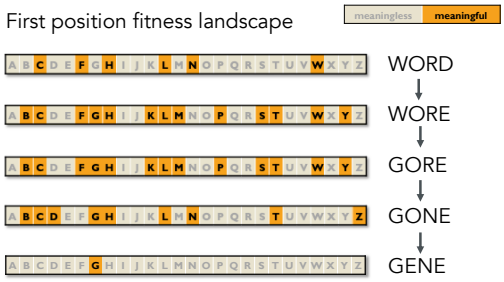


Figure 4.2: Depicting epistatic dynamics using Maynard Smith's (1970) word game analogy of protein evolution. The fitness landscape at the first letter position changes as letters at other positions change. Fitness is binary: a word is either meaningful or not. The provided trajectory is from Maynard Smith (1970). These dynamics are akin to epistatic dynamics in protein evolution where site-specific fitness landscapes depend on the residues present at other sites in the protein.

It is important to differentiate between shifts in S2S landscapes and shifts in site-specific landscapes. A change in the protein's environment or function will lead to a shift in the ordering of preferred sequences and hence a shift in the S2S landscape. Such a shift is analogous to a change in the definition of a meaningful word (e.g., Spanish rather than English words are considered meaningful). The evolutionary response to a shift in the S2S landscape is often considered adaptive with an excess of beneficial substitutions compared to neutral or deleterious fixations. Alternatively, site-specific fitness landscapes can change solely due to epistasis in the absence of any external change. In this scenario, the proportions of beneficial and deleterious (fixed by random genetic drift) substitutions remain equal at equilibrium (Goldstein, 2013). As such, changes in site-specific landscapes are often considered nonadaptive when the S2S landscape is unchanged. Here, I will refer to *adaptive shifts* as changes in site-specific fitness landscapes in conjunction with

a shift in the S2S landscape. Alternatively, *nonadaptive shifts* constitute changes in site-specific landscapes caused by the interplay of mutations, drift, and selection on a fixed S2S landscape.

For most proteins, a prerequisite to proper biological functioning is correct folding into a native structure in which the protein is sufficiently stable. As such, many authors have investigated the level of nonadaptive preference shifts by modeling stability-mediated epistasis and found that amino acid preferences changed over time (Pollock *et al.*, 2012; Youssef *et al.*, 2020; Shah *et al.*, 2015). In particular, Pollock *et al.* (2012) observed a tendency for the preference for a newly substituted amino acid to increase through adjustments at other sites in the protein. They refer to this as an *evolutionary Stokes shift*, analogous to the spectroscopy effect known as the Stokes shift where a molecule receives a quantum of energy, moves to a higher energy state, and adjusts to the new state by emitting a smaller quantum of energy than was first absorbed. More recently, evidence for the opposite trend, where the preference for the resident amino acid decreases over time, was observed (Chapter 3). This phenomena was dubbed as the *evolutionary anti-Stokes shift*. Using a different stability model, Shah *et al.* (2015) observed similar trends where substitutions were often *entrenched*, becoming increasingly deleterious to revert over time, and were usually *contingent* on prior substitutions that increased their fixation probability.

While entrenchment and evolutionary Stokes shifts have been used interchangeably (Rodrigue and Lartillot, 2017; Bastolla *et al.*, 2017; Teufel *et al.*, 2018), they are related yet distinct phenomena. Briefly, a substitution may be entrenched "by-any-means" (adaptive or nonadaptive); whereas an evolutionary Stokes shift refers to the increase in preference of a residue by nonadaptive stability-mediated effects. An evolutionary Stokes shift may lead to an entrenched allele; however, not all entrenched alleles are conserved because of an evolutionary Stokes shift. Similarly, the notion of contingency and evolutionary anti-Stokes shifts are related yet not synonymous.

To illustrate their differences, consider an adaptive episode where a protein was evolving in the context of environment A when an external change occurs (environment B) with a shift in the S2S landscape and accompanying changes in the site-specific landscapes. Let us consider the dynamics at a focal site. In environment A, amino acid alanine (one-letter code A) was the most preferred residue at a site (figure 4.3). In environment B, the site's preferences change such that valine (one-letter code V) is now

the most preferred residue. Positive selection will likely lead to the fixation of beneficial substitutions that increase fitness, resulting in the fixation of the newly favoured amino acid V. The substitution to V is therefore contingent on the environmental change that increased its favourability. Once on (or near) the new landscape peak, mutations away from amino acid V will be purged by purifying selection. The beneficial effects of subsequent mutations at other sites may depend on the presence of V as part of the genetic background. As such, substitution away from V may become increasingly deleterious, leading to its entrenchment. In this way, a residue may be contingent and subsequently entrenched through an adaptive process.

Alternatively, a substitution may be contingent on or become entrenched by nonadaptive processes. Suppose that, instead of an environmental change, a mutation is fixed by drift at another site in the protein, changing the preference landscape at the site of interest. Such a shift in the landscape could increase the preference for alanine (an evolutionary Stokes shift) or decrease it (an evolutionary anti-Stokes shift). Given an increase in the preference for A, substitutions away from A are unlikely to be fixed leading to its conservation, or entrenchment. Alternatively, if the landscape shift resulted in a decrease in the preference of the resident amino acid such that another amino acid is the fittest at the site (e.g., V, figure 4.3), then the subsequent fixation of V is contingent on the change in the background sequence. These examples offer snap-shots of different preference landscapes. In natural protein evolution, these processes are dynamic and gradual over long periods (Pollock *et al.*, 2012).

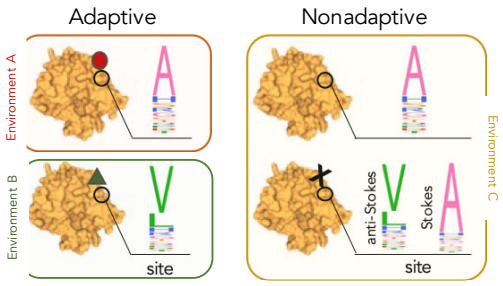


Figure 4.3: Adaptive evolution often causes substantial shifts in amino acid preferences. For example, suppose a change occurs in the protein's environment (e.g.,a change in an interacting protein or a ligand; depicted by the red circle or green triangle), then the landscape shifts from having a strong preference for amino acid alanine (one-letter code A), to strongly preferring valine (one-letter code V). Nonadaptive evolution can also cause shifts in preferences. Following a substitution at another position in the protein (depicted with an X), the fitness landscape at a focal site could increase the preference for A or could change the ordering of amino acid preferences such that V is the most preferred residue. Evolutionary Stokes and anti-Stokes shifts are gradual phenomena that could in the long run lead to these example landscapes.

Shifted preferences can have significant consequences for protein evolution. Dobzhansky-Muller incompatibilities, where a mutation is neutral (or beneficial) in one protein but is pathogenic in a homologous protein, highlight the potential significance of shifted preferences on speciation (Kondrashov *et al.*, 2002). Furthermore, entrenched substitutions play a significant role in maintaining molecular complexes (Hochberg *et al.*, 2020). It is, therefore, crucial to understand the drivers of shifted preferences. My aim in this review is an attempt to quantify the magnitudes and frequencies of nonadaptive shifts in preferences.

4.4 Evidence of preference shifts from multiple sequence alignments

A challenge with estimating shifts in preferences is that they are not directly observable in extant sequences. However, models which permit variation in site-specific preferences make explicit predictions that can be validated or refuted by patterns in natural alignments. Analysis of natural proteins often reveals evidence for temporal variation in replacement rates and homoplasy rates (reversions, convergence, and parallelism). Are these patterns explainable by nonadaptive processes, or are they the result of adaptive evolution? As reviewed below, in most instances, the observed patterns are consistent with predictions

from nonadaptive epistatic models.

4.4.1 Convergence rates

Convergence refers to the evolutionary phenomenon whereby similar traits emerge independently in multiple lineages. Convergence may occur at the phenotypic level, such as the origins of wings in bats and birds (Stern, 2013), or echolocation in bats and toothed whales (Parker *et al.*, 2013). Phenotypic convergence is commonly viewed as evidence of adaptations of different lineages to similar environmental challenges (Mcgee and Wainwright, 2013). Alternatively, molecular convergence, the emergence of identical states (nucleotide, codon, or amino acid) in two independent lineages, is not convincing evidence of adaptation since they could happen by chance owing to the limited number of possible states at a site. Independent changes from the same ancestral state to the same derived state, are convergent substitutions that transpired in parallel (figure 4.4).

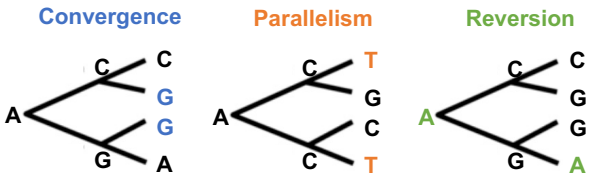


Figure 4.4: Examples of molecular homoplasy. Convergence refers to substitutions at independent lineages from different ancestral states to the same derived state. Parallelism refers to independent substitutions from the same ancestral state to the same derived state. Reversion refers to a change from a derived state back to an ancestral state.

Evidence of convergent substitutions abounds (Parker *et al.*, 2013; Thomas and Hahn, 2015; Zou and Zhang, 2015b; Goldstein *et al.*, 2015; Zou and Zhang, 2015a; Mendes *et al.*, 2016). An adaptive explanation would suggest that convergent substitutions are due to similar selection pressures in different taxa. For example, Parker *et al.* (2013) compared 22 mammalian genome sequences (composed of 2,326 orthologous proteins) and reported rampant levels of convergent substitutions. They concluded that adaptive molecular convergence is widespread and explains the independent evolution of echolocation in bats and whales. However, their conclusions were challenged by two subsequent studies which reanalyzed their (and additional) data and found that convergence levels between bats and toothed whales are no greater than the levels of molecular convergence between bats and cows (Zou and Zhang, 2015b; Thomas and Hahn, 2015). These studies highlight

that rigorous assessments of the prevalence of adaptive convergence require properly formulated null models. Such null models allow us to assess whether it is necessary to invoke adaptive processes to explain the observed patterns of substitution.

The simplest models for sequence evolution assume equal substitution rates between states. When applied to amino acids, it is referred to as the Poisson model which assumes that all amino acids have the same fitness effect so that site-specific landscapes are uniform at all sites and are constant across time. The Poisson model predicts a relatively constant and low level of convergence rates as proteins diverge. However, evidence of convergence in natural datasets often exceeds the levels of convergences predicted by the Poisson model, and the level of convergence in natural alignments usually decreases as sequences diverge. Therefore, using the Poisson model as a null model, one might inaccurately reject the null in favour of an adaptive explanation. However, models which account for differences in rates of exchange among amino acids, e.g.,WAG (Whelan and Goldman, 2001), and models that allow for variability across sites, e.g.,MutSel (Halpern and Bruno, 1998), predict higher levels of convergence than the Poisson model, and declining levels with time. Nonetheless, rates of convergence inferred from natural alignments exceed the levels predicted by these heterogeneous models. A limitation of these models is that they do not account for epistasis. Independently, Goldstein *et al.* (2015) and Zou and Zhang (2015a) showed that accounting for epistatic interactions leads to patterns and levels of convergence rates in line with observations in natural data. Their work highlights that understanding substitutions patterns under epistatic models are imperative for accurately detecting adaptive evolution. In box 1, I review two datasets with declining convergence rates. In both datasets, the observed patterns are consistent with nonadaptive epistatic dynamics.

Why do convergence levels decrease over time under epistatic models? To illustrate this, let us again consider Maynard Smith's word game analogy. The first-position fitness landscapes are more similar when the background sequences have fewer differences (for example, consider the first position landscapes given background sequences -ORD and -ORE; figure 4.2). As more differences accumulate (e.g., -ORD and -ENE), the first-position landscapes become more dissimilar (Usmanova *et al.*, 2015). Similarly, in protein evolution, as sequences diverge the amino acid preference landscapes accumulate more differences. Nevertheless, structural or functional constraints may limit variability in

amino acid preferences across diverged proteins. The extent to which such restrictions limit variability in preferences, however, is still unknown.

Box 1 | Convergence rate: evidence of adaptations or expected under nonadaptive evolution?

13 orthologous mitochondrial proteins from 629 vertebrate mitochondrial genomes

- Goldstein *et al.,* 2015 observed declining levels of convergence rates with time in vertebrate mitochondrial proteins.
- To dissect if the levels of convergence are evidence of adaptive evolution or are explainable by nonadaptive convergences, they simulated data under two substitution models: the WAG (Whelan and Goldman 2001) which is site- and time-homogenous but allows for difference in rates of exchange across sites; and a stability-mediated epistatic model (Pollock *et al.,* 2012) which accounts for differences among sites and across time.
- They found that the levels of convergence rates in the mitochondrial proteins were highly compatible with the levels expected under a nonadaptive epistatic model.

5,935 orthologous proteins from 12 fruit fly species

- Zou and Zhang (2015a) report a large amount of variability in convergence rates across the different pairs of orthologous proteins. Convergence rates were higher in recently diverged proteins and declined with evolutionary distance.
- To determine if the convergence levels are due to adaptive or non-adaptive process, they developed various evolutionary models and compared the expected rates to those observed in the natural proteins.
- The simplest model estimates gene-wide equilibrium amino acid frequencies which are constant across sites and time. Based on this model, the observed number of convergences were significantly higher than the null expectation.
- They developed two additional substitution models both of which account for variation across sites by either grouping sites into classes with similar amino acid frequencies, or by assigning site-specific equilibrium frequencies. Under both these site-heterogenous models, observed convergence rates were significantly lower than predicted.
- Lastly, using simulations they showed that the lower rates of convergence in the empirical data compared to the site-heterogenous null models is likely due to epistatic interactions.
- In conclusion, they found that the observed amounts of convergence is explainable by nonadaptive models which account for site- and time- heterogenous process.

4.4.2 Reversion rates

Reversion describes a return to an ancestral state during evolution (figure 4.4). Molecular reversions are common in natural sequences (Rokas and Carroll, 2008; Breen *et al.*, 2012; Naumenko *et al.*, 2012). More than a century ago, Muller (1918, 1939) hypothesized that epistasis causes reversion rates to decrease with time. McCandlish *et al.* (2016) proved that involvement in at least one epistatic interaction is sufficient to cause decreases in reversion rates and that in the absence of epistasis reversions rates are constant through time.

Naumenko *et al.* (2012) analysed two datasets of genome-wide alignments from vertebrates (7,967 genes from 9 species) and insects (8,477 genes from 8 species). In both datasets, they observed decreases in reversion rates as sequences diverged, consistent with expectations under epistatic models (McCandlish *et al.*, 2016). Epistasis can lead to diminishing rates of reversion through (1) a nonadaptive increase in fitness for the derived residue (i.e., an evolutionary Stokes shift), or (2) a nonadaptive decrease in the fitness of the replaced residue (Naumenko *et al.*, 2012). Naumenko *et al.* (2012) argued that the second effect is stronger and that "negative epistatic interaction with currently absent amino acids" is responsible for most of the observed declines in reversion rates.

4.4.3 Replacement rates

Another signal commonly observed in natural alignments is changes in replacement rates over time, or heterotachy. Various adaptive and nonadaptive mechanisms can produce this signal. For example, evolution on a static site-specific fitness landscape, in the absence of both epistatic and adaptive processes, can lead to heterotachy (Jones *et al.*, 2017). On a static landscape, a chance fixation to a suboptimal amino acid is followed by a period of positive selection restoring the site to its optimal state, a process referred to as *nonadaptive shifting balance* (Jones *et al.*, 2017). Alternatively, heterotachy can also be caused by changes in site-specific fitness landscapes because of epistasis. Changes at other positions can lead to a more uniform fitness landscape having higher substitution rates, or a more rugged landscape with fewer opportunities for change (Gong *et al.*, 2013). Lastly, heterotachy may also occur because of changes in the S2S landscape congruent with an adaptive episode—the shift in the S2S landscape is often followed by a period of high substitution rates as the protein adapts to the new conditions (dos Reis, 2015; Jones *et al.*, 2017). Given the diversity of processes that can lead to heterotachy, accurate interpretation of the mechanisms at play in natural data is challenging.

Can heterotachy from adaptive versus nonadaptive evolution be distinguished? Two studies have recently suggested that nonadaptive and adaptive processes cause idiosyncratic variations in replacement rates (Popova *et al.*, 2019; Stolyarova *et al.*, 2020). They hypothesized that epistasis causes a reduction in replacement rate with time, while adaptive evolution leads to increases in rates. The reason, they suggest, is that adaptive shifts in preferences often render the current state suboptimal for the new conditions. Positive selection will restore equilibrium through the subsequent fixations of beneficial substitutions, leading to an increase in substitution rate following the landscape shift. In contrast, nonadaptive evolutionary Stokes shifts increase the favorability of the resident amino acid. Such an increase in favourability leads to declining rates of replacement. However, the existence of an evolutionary anti-Stokes shift—where decreases in resident amino acid favorability lead to increases in replacement rates—challenges this claim (Chapter 3).

In this way, both adaptive and nonadaptive processes may lead to an increase in replacement rates over time. Nevertheless, I hypothesize that heterotachy caused by adaptive and nonadaptive processes can be differentiated. In the absence of adaptations, a balance is expected in the frequency and magnitude of both evolutionary Stokes and

anti-Stokes shifts (Chapter 3). This balance suggests that under nonadaptive evolution, the proportion of sites that experience increases in replacement rates should be approximately equal to the proportion experiencing a decrease in rate. Alternatively, adaptive shifts will lead to an excess of sites with increased rates compared to the proportion of sites for which replacement rates decreased. This is akin to the expectations of the proportions of beneficial and deleterious substitutions under adaptive and nonadaptive processes. Under nonadaptive evolution, a balance exists in the proportions of beneficial and deleterious substitutions. However, following an adaptive change, the proportion of beneficial substitutions exceeds that of deleterious substitutions (dos Reis, 2015; Jones *et al.*, 2017). While the dynamics of landscape shifts under adaptive evolution are yet to be thoroughly investigated, I suspect that adaptive episodes will analogously lead to an excess in the proportion of sites undergoing increases in substitution rates relative to the proportion of rate decreasing sites.

I summarise the results from three recent studies investigating changes in replacement rates in table 4.1. In the reported datasets, the number of rate accelerating or decelerating sites is comparable—except for the hemagglutinin H3 subtype protein where a higher number of accelerating sites was observed (12 accelerating sites and only four decelerating sites). The sites with the largest increase in replacement rates were experimentally shown to affect antigenic properties (Popova *et al.*, 2019). The observed increase in rates in the H3 protein may be a true signal of adaptive evolution. Nevertheless, the similar numbers of accelerating and decelerating sites in all other proteins are in line with the expectations from nonadaptive epistatic models (Chapter 3). The results presented in table 4.1 are from a relatively narrow range of proteins, making it difficult to draw general conclusions. Future work establishing the differences and similarities in variability in replacement rates due to adaptive versus nonadaptive processes is warranted.

Table 4.1: Number of rate accelerating sites is often equal to the number of rate decelerating sites inline with expectations from nonadaptive epistatic models.

Reference	Dataset	Rate increases	Rate decreases	Total num. of alleles
Popova *et al.* (2019)	H1 proteins from 1,613 strains	0	2	83
	N1 proteins from 2,015 strains	0	0	82
	H3 proteins from 1,832 strains	12	4	117
	N2 proteins from 1,996 strains	8	5	93
Stolyarova *et al.* (2020)	Five mitochondrial genes across 3,557 metazoan species	28	21	42,637
Gelbart and Stern (2020)	Nine proteins across 126 HIV-1/SIV strains	134	137	5,902

4.5 Experimental evidence of shifts in preferences

While the patterns discussed above—decreases in homoplasy rates with divergence levels, and patterns of heterotachy—are consistent with temporally varying preferences, they could have arisen by nonepistatic mechanisms. For example, inaccurate tree inference could lead to diminishing rates of convergence (Mendes *et al.*, 2016), or nonadaptive shifting balance can lead to the observed heterotachy (Jones *et al.*, 2017). A more direct approach for inferring preference shifts driven by epistasis is to compare mutational effects across background sequences. If variations in preferences due to epistasis are minor, then a mutation should have a similar phenotypic effect regardless of the background sequence. Alternatively, if preferences depend heavily on sequence-context, then mutational effects will vary across different background sequences. Until recently, experimental methods were restricted in the number of mutations they can introduce (Fowler and Fields, 2014). Most studies performed one of three types of pairwise amino acid replacements (figure 4.5): (1) *Forward mutations* by replacing the residue in an ancestral protein with a derived state; (2) *Backward mutations* which introduce an ancestral state into an extant protein; and (3) *Exchange mutations* by replacing the resident amino acid in one protein with the resident residue in an orthologous protein.

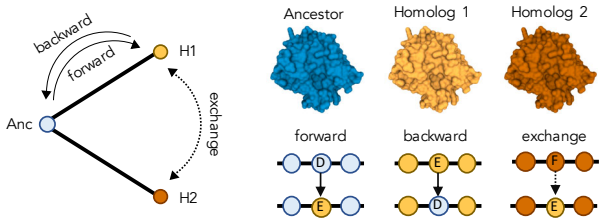


Figure 4.5: Diagram representing the different mutation experiments. Forward substitutions place a derived amino acid into the context of an ancestral (Anc) sequence. Backward substitutions place an ancestral amino acid in the context of an extant sequence. Exchange substitutions refer to changing the resident amino acid in one homolog (e.g.,E in H1) with the resident amino acid in another (e.g.,F in H2). Forward and backward substitutions are shown in solid lines. Exchange substitutions are shown in dashed lines.

4.5.1 Effects on stability

Protein stability is a holistic property determined by all residues. A stabilizing mutation in one sequence may be destabilizing in another. To investigate the dependence of the stability effect of a mutation on the background protein sequence, Ashenberg *et al.* (2013) introduced the same mutations into a series of diverged homologs of the influenza nucleoprotein (NP). Specifically, they separately introduced six mutations (I186V, V239M, L259S, A280V, H334N, G384R) into four NP homologs (Brisbane/2007, Aichi/1968, California/2009, bat/2009). The level of sequence divergence relative to the Brisbane/2007 sequence is 8% with Aichi/1968, 10% with California/2009, and 28% with bat/2009. They observed that stability effects of mutations were conserved across background sequences: only a single mutation induced a substantial shift in stability effects (A280V). The substitution from A $\rightarrow$ V at site 280 was stabilizing in the context of the Brisbane/2007, Aichi/1968, and California/2009 sequences, but was destabilizing in bat/2009 NP. Analysis of their data revealed that the standard deviation in mutational effects on melting temperature across background sequences was on average 0.86°C. Furthermore, the stability effects of mutations in the context of different homologous proteins were significantly correlated. However, correlations decreased as sequence divergence increased: the correlation in stability effects of mutations between Brisbane/2007 and Aichi/1968 (8% sequence divergence) was 0.90, falling to 0.89 in California/2009 (10%sequence divergence), and 0.82 in bat/2009 (28% sequence divergence).

To assess how stability effects of mutations change over time, Risso *et al.* (2015) performed forward and backward substitutions between extant and ancestral reconstructions of

thioredoxin proteins. Specifically, they assayed stability effects in the context of the extant *E. coli* protein and a resurrected protein present in the last bacterial common ancestor (LBCA). These proteins differ at 44% of sites. They introduced 21 mutations of the types E $\leftrightarrow$ D, I $\leftrightarrow$ V into both background sequences and assayed their effect on stability. The stability effects in the LBCA and *E. coli* thioredoxin proteins were strongly correlated (Pearson correlation coefficient of 0.89). Only 2 of the 21 mutations were stabilizing in one protein and destabilizing in the other. In general, stability effects were within the range of ± 1 kcal/mol. These results suggest that stability effects among biochemically similar amino acids (E and D, V and I) are conserved over long evolutionary time scales (approximately 4 billion years). To investigate the generalizability of this observation to biochemically dissimilar mutations, Risso *et al.* (2015) introduced L $\leftrightarrow$ K mutations across a series of ancestral thioredoxin proteins, and T $\leftrightarrow$ M mutations across ancestral β-lactamases. Variability in stability effects was more pronounced in the L $\leftrightarrow$ K and T $\leftrightarrow$ M mutations than in the E $\leftrightarrow$ D and V$\leftrightarrow$ I mutations. Nevertheless, the most energetically preferred amino acid at a site remained the same in the extant and ancestral proteins.

The experimental studies reviewed above investigated the stability effects of a limited number of mutations. Alternatively, simulations of stability-constrained evolution allow for a more comprehensive assessment of stability effects across a wide range of background sequences (Pollock *et al.*, 2012; Shah *et al.*, 2015). Shah *et al.* (2015) simulated the evolution of the lysine-arginine-ornithine-binding periplasmic protein (argT) using the force-field approach FoldX to estimate stability. They performed forward and backward mutations *in silico* and assayed the stability effects across all background sequences. They observed that variability in stability effects was common in frequency, yet minor in magnitudes. On average, stability effects were within 0.8 kcal/mol. In summary, theoretical and experimental investigations reveal that stability effects of mutations are conserved across background sequences, consistent with the expectation that fitness effects are often nearly neutral at mutation-selection-drift equilibrium (Cherry, 1998; Goldstein, 2011).

4.5.2 Effects on function

The previous results suggest that stability effects of mutations are conserved across diverged sequences. Are functional effects of mutations similarly conserved, or is protein function highly attuned to the background sequence such that functional effects of mutations differ substantially across background sequences?

Lunzer *et al.* (2010) were amongst the first to investigate the functional effects of mutations in orthologous proteins. They individually introduced 168 mutations into the wild-type *E. coli* isopropyl malate dehydrogenase (IMDH) protein and assayed their impact on enzyme performance (k_{cat}/K_m). At each site, they performed exchange mutations with the resident amino acids present in the *P. aeruginosa* IMDH homolog. The vast majority of single mutant enzymes (104/168) performed similarly to the wild-type IMDH proteins, suggesting that functional effects of mutations are conserved.

Emlaw *et al.* (2020) compared the effects of mutations on single-channel conductance using human muscle-type acetylcholine receptor (AChR) and an ancestral AChR (the AChR present in the last common ancestor between humans and cartilaginous fish). The proteins differed at 36% of sites. At two sites where the resident amino acids differed between the two proteins (sites 2 and 6), they performed backward substitutions, placing the ancestral amino acids into the human AChR (mutations G2T and F6S). They also performed forward substitutions, placing the derived amino acids into the ancestral sequence (mutations T2G and S6F). Lastly, they introduced the double mutants into both the extant and ancestral proteins. Analysis of their data revealed high concordance between the effects of the studied mutations in the different background sequences (Pearson correlation was 0.90).

Starr *et al.* (2018) performed forward and backward replacements between a heat shock protein 90 (Hsp90) ATPase domain present in modern *Saccharomyces cerevisiae* (ScHsp90) and a deep eukaryotic ancestor (ancAmoHsp90, the common ancestor of Amorphea). In particular, their analysis focused on the N-terminal domain (NTD). The ancestral and extant NTDs differ at 60 of 221 sites (27% sequence divergence). They individually introduced each ancestral amino acid into the extant ScHsp90 protein and each derived state into ancAmoHsp90. Then, they estimated the fitness of yeast cells carrying the mutant proteins by measuring the change in the ratio of a mutant to wildtype frequency over time. Approximately 48% of derived states reduced fitness when placed in the context of the ancestral NTD, 32% were neutral, and 20% were beneficial. When placed in the modern NTD, 92% of ancestral amino acids were deleterious, 7% were neutral, and 1% were beneficial. Across all mutations studied, 77% had different impacts on fitness depending on the background sequence. However, the effects of most mutations were minor: the average selection coefficient was -0.02, and -0.01 for backward and forward substitution, respectively. Note that the relatively small selection coefficient does

not imply that epistasis plays a minor role in protein evolution. Even a relatively small degree of nonadditivity in the effects of mutations can have a considerable impact on evolutionary processes (Shah *et al.*, 2015).

4.5.3 Quantifying the frequency and magnitude of shifts in preferences using Deep Mutational Scanning (DMS)

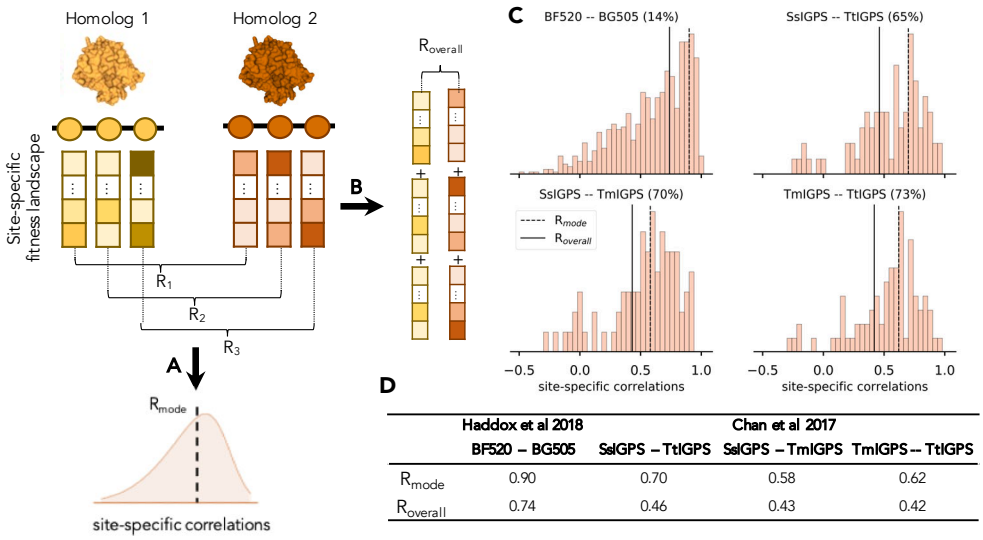
The previously discussed studies were limited to a small number of mutations. However, recent advancements, known collectively as deep mutational scanning (DMS), allow us to estimate the fitness effect of all single amino acid mutations at many (or all) sites in a protein (Fowler and Fields, 2014; Hietpas *et al.*, 2011). First, a single-mutant library of proteins is created. The mutants are then subjected to a selection or screen in which the frequency of each genotype in the library is measured using deep sequencing. Fitness can then be estimated from the frequency measures. One approach is to evaluate a mutant's frequency relative to the wildtype over time as a measure of fitness (Starr *et al.*, 2018). Others have used the relative frequency of a mutant pre- and post-selection as a measure of the mutant's fitness (Bloom, 2015). More sophisticated Bayesian approaches which correct for low sequencing depth have also been developed (see Bloom (2015) for a detailed description of models used to analyse DMS data and software implementations). While DMS approaches are a powerful tool for assessing the extent of shifts in amino acid preferences, the level of experimental noise is often high. Site-specific landscapes estimated from replicate experiments can have correlation coefficients as low as 0.59 (Haddox *et al.*, 2018; Doud and Bloom, 2016).

Despite its recency and potential limitations, DMS methodologies have been used to estimate site-specific fitness landscapes in many proteins in various organisms. Livesey and Marsh (2020), report on the results from 31 publicly available DMS datasets: 13 from human proteins, 9 from bacterial proteins, 5 from yeast proteins, and 4 viral proteins. However, only four studies have applied DMS to homologous proteins (Doud *et al.*, 2015; Lee *et al.*, 2018; Haddox *et al.*, 2018; Chan *et al.*, 2017). Six datasets from these four studies are available to compare site-specific preferences across different background sequences (table 4.2). Three studies were carried out in viruses (Haddox *et al.*, 2018; Lee *et al.*, 2018; Doud *et al.*, 2015). The fourth study (Chan *et al.*, 2017) compared site-specific fitness landscapes in orthologous indole-3-glycerol phosphate synthase (IGPS) proteins present in the archaeon *Sulfolobus solfataricus* (ssIGPS) and in two bacteria:

Thermotoga maritima (TmIGPS) and *Thermus thermophilus* (TtIGPS). Collectively, the studies compare site-specific landscapes across sequences with as little as 6% and up to 73% sequence divergence.

There are broadly two ways of comparing site-specific landscapes across different sequences. The first approach is to calculate correlation coefficients between landscapes. This has been done in two ways (figure 4.6): (A) calculate the landscape correlation at homologous sites, and report the mode of the correlation coefficient distribution (R_{mode}; figure 4.6A); or (B) concatenate all landscapes and estimate a single overall correlation coefficient ($R_{overall}$; figure 4.6B). Chan *et al.* (2017) used the first approach and found that site-specific landscapes were significantly correlated (with modes ranging from 0.62 and 0.72; table 4.2). Alternatively, Bloom and colleagues report the overall correlation from the second approach: $R_{overall}$ ranged from 0.36 to 0.72 (Doud *et al.*, 2015; Haddox *et al.*, 2018; Lee *et al.*, 2018). It is currently unclear if both approaches lead to similar correlation estimates and hence similar biological conclusions.

To compare the two correlation approaches, R_{mode} and $R_{overall}$, I reanalyzed the datasets from Chan *et al.* (2017) and Haddox *et al.* (2018) using both methods. Note that Haddox *et al.* (2018) conducted three replicate experiments for each homologous protein (BF520 and BG505). It is valuable to obtain the across-replicate average landscapes prior to obtaining correlations (see box 2 for more details). I report the correlations between site-specific landscapes given the different background sequences in figure 4.6C and 4.6D. It is clear from this analysis that R_{mode} and $R_{overall}$ can differ; specifically, $R_{mode} > R_{overall}$ in the four datasets. The largest difference is observed in the TmIGPS – TtIGPS comparison where R_{mode} and $R_{overall}$ differ by 0.20. Because fitness profiles can be expected to vary substantially over sites, conditioning on a site by reporting site-specific correlations may be more statistically robust and is more informative regarding the dynamics at a site. For example, it is evident from the site-specific correlation distributions that most landscapes correlate strongly ($R > 0.5$). However, some sites have landscapes that are negatively correlated. A negative correlation of the preference landscape given different genetic background indicates substantially shifted amino acid preferences, which might suggest different functional or structural constraints in the different proteins. For these reasons, the R_{mode} approach may be preferable for comparing the correlations between site-specific landscapes given different background sequences.



	Haddox et al 2018	Chan et al 2017		
	BF520 – BG505	SsIGPS – TtIGPS	SsIGPS – TmIGPS	TmIGPS -- TtIGPS
R_{mode}	0.90	0.70	0.58	0.62
$R_{overall}$	0.74	0.46	0.43	0.42

Figure 4.6: Different approaches for comparing correlations between site-specific landscapes across different background sequences. The first approach (A) estimates the correlation between landscapes at homologous sites given different background sequences and reports the mode of distribution (R_{mode}). The second approach (B) concatenates all site-specific landscapes and estimates an overall correlation value ($R_{overall}$). (C) Distribution of site-specific correlation values from four deep mutational scanning experiments. The BF520 - BG505 dataset is from Haddox *et al.* (2018). The remaining datasets are from Chan *et al.* (2017). Percentages in parentheses are the percent sequence divergence between the two proteins. (D) Reports the values of R_{mode} and $R_{overall}$ from the four datasets. Reported correlations for the BF520 - BG505 dataset are from site-specific fitness landscapes averaged over replicate experiments.

Box 2 | Higher correlations when averaging over replicate experiments.

Deep mutational scans can display high levels of experimental noise. Therefore, triplicate experiments are usually conducted for a given protein. Let P and Q be the true site-specific (or concatenated) fitness landscapes given different background sequences. Then let P^r and Q^r be the fitness landscapes estimated from a DMS experiments r and s such that $P^r = P + e^r$ and $Q^s = Q + d^s$, where e^r and d^s are the measurement errors. If these are uncorrelated then,

$$\mathrm{Cov}(P^r, Q^s) = \mathrm{Cov}(P, Q)$$

but, $\mathrm{Var}(P^r) = \mathrm{Var}(P) + \mathrm{Var}(e^r)$ and $\mathrm{Var}(Q^s) = \mathrm{Var}(Q) + \mathrm{Var}(d^s)$. Thus, the correlation for a replicate pair (r, s) is

$$\begin{aligned}
\mathrm{Corr}(P^r, Q^s) &= \frac{\mathrm{Cov}(P^r, Q^s)}{\sqrt{\mathrm{Var}(P) + \mathrm{Var}(e^r)} * \sqrt{\mathrm{Var}(Q) + \mathrm{Var}(d^s)}} \\
&= \frac{\mathrm{Cov}(P, Q)}{\sqrt{\mathrm{Var}(P) + \mathrm{Var}(e^r)} * \sqrt{\mathrm{Var}(Q) + \mathrm{Var}(d^s)}} \\
&\leq \frac{\mathrm{Cov}(P, Q)}{\sqrt{\mathrm{Var}(P)} * \sqrt{\mathrm{Var}(Q)}} \\
&= \mathrm{Corr}(P, Q)
\end{aligned}$$

This shows that correlations estimate from replicate pairs of experiments will be less than or equal to the true correlation. The above argument also holds for across-replicate averaged landscapes $\bar{P}$ and $\bar{Q}$. However,

$$\mathrm{Var}(\bar{P}) = \mathrm{Var}(P) + \mathrm{Var}(e^r)/3$$
$$\mathrm{Var}(\bar{Q}) = \mathrm{Var}(Q) + \mathrm{Var}(d^s)/3$$

Therefore, the denominator term, causing the underestimation, is smaller for averaged landscapes. For example, consider the Haddox et al., 2018 study where they performed DMS triplicate experiments for envelope proteins present in HIV stains BF520 and BG505. The correlations between replicate experiments, $\mathrm{Cov}(P^r, Q^s)$ were less than 0.58. However, the correlation between across-replicate average landscapes was 0.74.

In summary, it is valuable to average prior to obtaining correlations. If errors in approximating the landscapes are uncorrelated, the covariance does not change by averaging but the variance contributions due to errors in approximation are reduced giving a better approximation to the correlation of interest that one would obtain had there been no variation over replicates.

In order to accurately detect shifts in preferences using DMS data we must account for high amounts of experimental noise. Therefore, a second approach for quantifying shifts in amino acid preferences compares the distance between two landscapes using the Jensen-Shannon distance metric (Doud *et al.*, 2015) (see box 3 for detailed discussion). The distance is equal to zero when amino acid preferences are identical, and is one if the preferences are dissimilar. The distance approach accounts for the level of variability in site-specific landscapes due to experimental noise by estimating the average root-mean-square distance within replicate experiments (RMSD_{within}). The distance between site-specific landscapes in homologs is similarly calculated ($\mathrm{RMSD}_{between}$). The magnitude of shift at a site ($\mathrm{RMSD}_{corrected}$) is then calculated as the difference between $\mathrm{RMSD}_{between}$ and RMSD_{within}. At each site $\mathrm{RMSD}_{corrected}$ provides a measure of the magnitude of the shift in preference while calibrating for experimental noise. Furthermore, the $\mathrm{RMSD}_{corrected}$

approach can be used to quantify the prevalence of significantly shifted sites. To do this, a null distribution of $\text{RMSD}_{corrected}$ values is generated through an exact permutation test by reassigning site-specific landscapes among the two protein groups. If preferences have not shifted significantly between the two proteins, then the true distribution of $\text{RMSD}_{corrected}$ values should be similar to the null distribution. This method can be used to identify sites for which the null hypothesis of no shifts is rejected. Note that permutation tests can be conservative because they construct a null distribution from data that may instead support the alternative hypothesis. As such, this approach may be susceptible to high false negative rates.

Box 3 | Quantifying shifts in DMS datasets while controlling for experimental noise

Deep mutational scanning methodologies offer a powerful tool for assessing the effect of mutations. However, the level of experimental noise may be problematic; correlations between identical replicates can be as low as 0.59, matching correlation coefficients observed across different background sequences (Haddox *et al.,* 2018; Doud and Bloom, 2016). Therefore, quantifying the extent and prevalence of shifts in preferences must be calibrated to the observed levels of experimental noise.

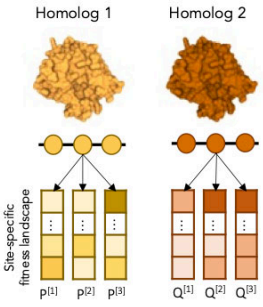
Doud *et al.,* 2015 used the Jensen-Shannon divergence (JSD) to quantify the level of similarity (or dissimilarity) between the fitness landscapes at homologous sites given different background sequences. Let P and Q be the site-specific fitness landscapes at a site given the background sequences H1 and H2, respectively. Then

$$\text{JSD}(P\,||Q) = \frac{1}{2}\,\text{D}(P\,||A) + \frac{1}{2}\,\text{D}(Q\,||A)$$

$$\text{where } A = \frac{1}{2}(P + Q) \text{ is the average fitness landscape}$$

$$\text{and } \text{D}(P||A) = \sum_i P_i \log(P_i / Q_i) \text{ is the Kullback-Leibler divergence}$$

Let $d(P, Q) = \sqrt{JSD(P||Q)}$, such that d(P,Q) is a metric of the distance between landscapes P and Q. The utility of d(P,Q) is that it is symmetric and ranges from 0 (identical distributions) to 1 (dissimilar distributions).



Replicate experiments yield different landscape estimates. To quantify the level of variability within replicates, calculate the average root-mean-squared distance at a site 'within' replicate experiments

$$\text{RMSD}_{\text{within}} = \frac{1}{2}\sqrt{\frac{1}{n_P}\sum_{r,s\,\in\,N_P} d(P^r,\,P^s)^2} + \frac{1}{2}\sqrt{\frac{1}{n_Q}\sum_{r,s\,\in\,N_Q} d(Q^r,\,Q^s)^2}$$

where P^r is the estimated landscape at a site in replicate r, N_P is the set of nonredundant pairwise comparisons within replicates (for example, given three replicate experiments, $N_P = \{(1,2), (1,3), (2,3)\}$), n_P is the number of comparisons, and the respective definitions for Q^r, N_Q, and n_Q. Then, calculate the root-mean-square distance between landscapes in different background sequences

$$\text{RMSD}_{\text{between}} = \sqrt{\frac{1}{n_{P,Q}}\sum_{r,s\,\in\,N_{P,Q}} d(P^r,\,Q^s)^2}$$

where $N_{P,Q}$ is the set of nonredundant pairwise comparisons between replicates (for example, given three replicate experiments for each background sequence, $N_{P,Q} = \{(1,1), (1,2), (1,3), (2,1)\ldots, (3,3)\}$), and $n_{P,Q}$ is the number of comparisons. The magnitude of preference change after correcting for site-specific noise is calculated as

$$\text{RMSD}_{\text{corrected}} = \text{RMSD}_{\text{between}} - \text{RMSD}_{\text{within}}$$

Doud *et al.* (2015) performed DMS on two homologs of influenza A virus (IAV) nucleoproteins (NP) in the H1N1 and H3N2 viral strains. The proteins differed at 6%

of sites. Using the $\text{RMSD}_{corrected}$ approach, they found that only a modest fraction of sites exhibited significant shifts in amino acid preferences: at a false discovery rate of 0.05, 14 of 497 sites (2.8%) showed evidence of significantly shifted preferences. Haddox *et al.* (2018) used the same method to quantify the magnitude and prevalence of shifted preferences between homologous HIV envelope (env) proteins that differ at approximately 14% of sites. Only 30 of the 659 sites (4.6%) showed evidence of significantly shifted preferences (at an FDR of 0.01). Lee *et al.* (2018) performed a similar analysis between homologous hemagglutinin (HA) proteins present in influenza viruses H1N1 and H3N2. The proteins were highly diverged, having 58% sequence divergence. The number of significantly shifted sites was not reported. However, it is evident from the distribution of $\text{RMSD}_{corrected}$ (figure 7C in Lee *et al.* (2018)) that a large number of sites had significantly shifted preferences. Also, the magnitude of the shifts was more pronounced than in other DMS studies. For example, the largest $\text{RMSD}_{corrected}$ reported in Doud *et al.* (2015) was 0.45, whereas $\text{RMSD}_{corrected}$ values were as high as ≈ 0.8 between the hemagglutinin homologs.

Table 4.2: Site-specific preference landscapes estimated across diverged background sequences are positively correlated. Listed are the Pearson correlations between landscapes within replicate experiments, and correlations between landscapes estimated in different background sequences. Prevalence is estimated from the $\text{RMSD}_{corrected}$ approach.

Ref	**Organism**	**Protein**	**Comparison**	**Seq length**[a] **(# sites**[b]**)**	**% div**	**Correlation between**	**Correlation within**	**Prevalence**
Doud et al 2015	IAV	NP	H1N1 – H3N2	497 (497)	6%	0.78[c]	0.83[c]	2.8% (FDR of 0.05)
Haddox et al 2018	HIV	env	BF520 – BG505	836 (659)	14%	0.57 – 0.58[d]	0.59 – 0.78[e]	4.6% (FDR of 0.01)
Lee et al 2018	IAV	HA	H1N1 – H3N2	566 (566)	58%	0.36 – 0.47[d]	0.69 – 0.82[e]	–
	S. Solfataricus (Ss)		SsIGPS – TtIGPS	271 (80)	65%	0.72[f]		
Chan et al 2017	*T. thermophilus* (Tt)	IGPS	SsIGPS – TmIGPS	267 (80)	70%	0.62[f]	0.94[f]	–
	T. maritima (Tm)		TmIGPS – TtIGPS	277 (80)	73%	0.62[f]		

[a] Pairwise alignable sites.
[b] Number of mutated sites.
[c] $\text{R}_{overall}$ between replicate-averaged site-specific landscapes. Within replicate correlations are based on comparison with site-specific landscape estimat from a previous study (Bloom, 2014a).
[d] Range of $\text{R}_{overall}$ over all replicate pairs between homologs.
[e] Range of $\text{R}_{overall}$ over all replicate pairs within homologs.
[f] R_{mode}

A challenge with assessing shifts in preferences using the correlation approaches is that, while it is clear that correlations between landscapes inferred from homologs are lower than correlations from biological replicates, it is unclear if the observed decreases are significant. As such, the $\text{RMSD}_{corrected}$ approach is more preferable for detecting significantly shifted preferences. Nonetheless, a limitation of the $\text{RMSD}_{corrected}$ approach is that it cannot distinguish between instances where the order of amino acid preferences has changed versus cases where there is an intensification (or relaxation) of selection between sequences. An example of this is provided in figure 4.7. Amino acid alanine (one-letter code A) is the most prefered residue at site 512 in both homologs of the env protein (Haddox *et al.*, 2018). However, site 512 is more mutationally tolerant in the context of the BG505 sequence versus the BF520 background. Conversely, at site 288, there is a clear shift in the ordering of amino acids. Despite having different shifted dynamics, the $\text{RMSD}_{corrected}$ approach estimates a similar degree of shift at sites 288 and 512. Alternatively, the Pearson correlation between landscapes is substantially lower for site 288 (figure 4.7), highlighting that the correlation approach might be more suitable for identifying sites having different preferred amino acids given different background sequences.

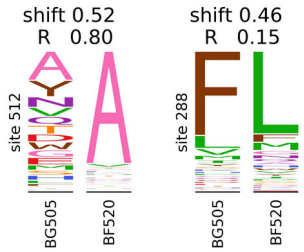


Figure 4.7: Correlation approach is better at identifying a reordering of amino acid preferences compared to the $\text{RMSD}_{corrected}$ approach. Site-specific preference landscapes in homologous envelope proteins in HIV strains BG505 and BF520. Shown are the across-replicate average preference landscapes at a site. The reported shift is the $\text{RMSD}_{corrected}$ values. The reported R value is the Pearson correlation coefficients between site-specific preference landscapes. Data obtained from Haddox *et al.* (2018) under the Creative Common Attribution license.

Deep mutational scanning is a promising tool for quantifying the magnitude and prevalence of shifted amino acid preferences. In addition to the analyses discussed above, data from DMS can be used to assess multiple additional questions: How often is a

substitution deleterious in one protein but beneficial in another? How often does the most preferred amino acid at a site differ across background sequences? How often are the detected shifts due to a reordering of the preferred amino acid versus a relaxation (or intensification) of selection pressure? Answers to these questions can be illuminated using data from DMS.

4.6 Limitations

The studies reviewed above suggest that temporal variability in amino acid preferences is usually minor in magnitude and low in frequency. However, each of the methods used for inferring preference shifts has potential limitations. Detecting variations in rates of homoplasy or replacement rates in natural alignments are indirect ways of assessing preference shifts. While theoretical models suggest that epistasis could result in the observed signal, other mechanisms may also be at play (Mendes *et al.*, 2016). Alternatively, deep mutational scanning approaches allow for a more direct assessment of site-specific preferences in different background sequences. These approaches offer snap-shots of preference landscapes in the context of different sequences but tell us little about the trend of change over time. For example, we cannot use current DMS data to assess if changes in preference are abrupt or gradual. Nevertheless, comparing preference landscapes between ancestral and extant proteins (as done in Starr *et al.* (2018)) to track how preferences change over time is valuable for understanding trends in preference shifts.

Deep mutational scanning data currently available to assess shifts in preferences is limited. There are only four studies that compared preference landscapes between homologous protein sequences (table 4.2), and only one of which compares orthologous bacterial and archeal proteins (Chan *et al.*, 2017). The remaining three studies were conducted in viruses, specifically RNA viruses. The high mutation rates in RNA viruses may have selected for loosely packed protein structures which buffer the deleterious effects of mutations (Tokuriki *et al.*, 2008). This would suggest that the low levels of mutational effects observed in these experimental settings may not generalize to non-viral proteins. This has led to concerns regarding the utility of viral DMS data in more generally assessing levels of preference shifts (Pollock and Goldstein, 2014). However, results from Chan *et al.* (2017) corroborate that drastic shifts in preference landscapes are usually rare in non-viral proteins even at high levels of sequence divergences. Furthermore, Ferrada (2019)

curated a dataset of 124 pairs of homologous protein (sequence divergences ranged from 0-100%) and computationally estimated site-specific landscapes using FoldX. Using the $\text{RMSD}_{corrected}$ approach, they observed that the number of sites with significantly shifted preferences increases with sequence divergence. Nevertheless, even at 100% sequence divergence on average less than 30% of sites had significantly shifted preferences. This study only modelled the effects of stability. Additional functional and structural constraints on natural proteins may further limit the amount of variability in preferences.

4.7 Consequences of shifts for time-homogenous evolutionary models

One way of deducing information about evolutionary processes is to analyze multiple sequence alignments with a quantitative model of sequence evolution. Two widely used classes of evolutionary models are phylogenetic models used to infer relationships between taxa and ω models used to estimate selection intensity. Inference procedures for either class of models often assume that the evolutionary process is identical across sites and constant through time. Specifically, most models assume (1) independent evolution across sites, (2) time-homogeneous substitution processes, and (3) a common vector of stationary frequencies; assumptions that are all violated in the presence of epistasis.

Various amendments have been applied to allow for heterogeneity (spatial and temporal) in the evolutionary process in both phylogenetic and ω models. However, due to the difficulty in tractably modelling co-dependencies among sites, models are limited in the extent of heterogeneity they can account for. In practice, inference procedures model among-site heterogeneity through a mixture model with different substitution processes as classes in the mixture, and can allow for temporal changes in the substitution process at prespecified branches along the tree (Yang and Nielsen, 2002; Yang *et al.*, 2005; Zhang *et al.*, 2005), or using a covarion-like process (Galtier, 2001; Guindon *et al.*, 2004; Jones *et al.*, 2020). More recently there has been a push towards using experimentally informed evolutionary models where site-specific substitution processes are informed by data from DMS (Bloom, 2014b,a; Hilton and Bloom, 2018). While these models offer improved likelihood scores over more traditional approaches, they are limited in applicability to the currently small number of proteins for which DMS data is available.

While the challenges associated with allowing for temporal and spatial heterogeneity

place a high barrier for their widespread incorporation into inference procedures, it is nonetheless of paramount importance to understand how they may bias our inferences. To this end, recent studies have advocated for the use of models of protein evolution with plausible levels of spatial and temporal heterogeneity as a tool for assessing the accuracy of inference in the face of realistic levels of heterogeneity (Spielman and Wilke, 2015; Jones *et al.*, 2017, 2018, 2020; Youssef *et al.*, 2020). Simulations of stability-informed models recapitulate levels of both spatial and temporal heterogeneity present in real data (Youssef *et al.*, 2020). They are therefore a powerful tool for assessing inference accuracy. To this end, sequences are first generated under a stability-constrained evolutionary model. The simulated sequences are then analyzed using traditional inference procedures. The true parameter values, predicted from the generating model, are then compared to the inferred parameters to assess inference accuracy.

Using the procedure outlined above, it is evident that traditional ω models underestimated levels of among-site heterogeneity; ω models estimated only 2-4 rate classes when a much richer distribution of rate classes (> 100) is present in the true generating process (Youssef *et al.*, 2020). Nevertheless, the inferred rates corresponded to the most common substitution rates across sites. Furthermore, inclusion of a covarion-like component in the substitution model, allowing rates at sites to vary over time, fit the data significantly better. These results suggest that ω models need not explicitly include epistatic interactions for reasonable inference of selection pressure when averaging over time and sites, and that allowing for a covarion-like component seems to capture temporal heterogeneity in rates arising due to epistasis (Youssef *et al.*, 2020).

The procedure outlined above has not yet been implemented to assess the sensitivity of phylogenetic inference to extensive and persistent levels of heterogeneity due to non-adaptive stability-constrained epistasis. However, the literature assessing the accuracy of phylogenetic inference in the face of temporal and spatial heterogeneity "by-any-means" is vast. Simulations show that ignoring temporal heterogeneity can induce systematic errors in phylogenetic inference, including topological and branch length inaccuracies (Magee *et al.*, 2020; Nasrallah *et al.*, 2011; Kolaczkowski and Thornton, 2008, 2009; Whelan, 2008). However, it remains unclear if the level of heterogeneity arising from nonadaptive epistatic processes is substantial enough to similarly bias our phylogenetic inferences.

In contrast with the relatively minor changes in preferences over time, differences

in amino acid preferences among sites is substantial (Echave *et al.*, 2016). Models that accommodate among-site heterogeneity fit the data significantly better than site-homogeneous models (Hilton and Bloom, 2018). This leads to the question: How can site-specific fitness profiles be estimated? There are currently two approaches for obtaining site-specific fitness landscapes: (1) they can be statistically inferred from large multiple sequence alignments (e.g., Rodrigue and Lartillot (2017)), or (2) experimentally obtained from deep mutational scans (e.g., Hilton and Bloom (2018)).

A new approach, informed by developments in the field of systems biology, might be worth exploring. Various computational variant effect predictors (VEPs) have recently been developed to predict the effects of mutations in a given protein sequence, often for clinical applications. In a recent study, Livesey and Marsh (2020) compared the performance of 46 different computational VEPs to data obtained from DMS. These VEPs rely on various structural, evolutionary, and biophysical features (see Livesey and Marsh (2020) for details of the different VEPs). The best performing VEP was DeepSequence (Riesselman *et al.*, 2018), an unsupervised machine learning approach. DeepSequence had an average correlation coefficient between predicted and observed (DMS) landscapes equal to 0.43 across all human proteins and 0.46 across all non-human proteins. While these correlation coefficients are low, it is relevant to note that the average Pearson correlation between different DMS studies on the same protein is only 0.66 (Livesey and Marsh, 2020), and correlations between replicate experiments can be as low as 0.59 (Haddox *et al.*, 2018; Doud and Bloom, 2016). A noteworthy limitation of the DeepSequence method is that it necessitates the availability of large multiple sequence alignments. For proteins where a large alignment is not available, other VEPs that rely on structural or biophysical features, such as DEOGEN2 (Raimondi *et al.*, 2017) and SNAP2 (Hecht *et al.*, 2016), may be preferable. Note that while DEOGEN2 and SNAP2 are supervised approaches, with potential limitations related to overfitting of the training dataset, they performed well against DMS datasets from viral, eukaryotic, and bacterial proteins.

Site-specific fitness landscapes can be estimated from VEPs and used to inform evolutionary models. For example, site-specific frequency landscapes can be estimated from the site-specific fitness landscapes and provided to phylogenetic models, similar to the phylogenetic application of DMS data (Bloom, 2014b; Hilton and Bloom, 2018). Alternatively, fitness values can be used directly in models of sequence evolution to specify

the rates of substitutions between codons or amino acids. Bloom (2014a) proposed two heuristic approaches of converting site-specific fitness landscapes to fixation probabilities. These approaches were first developed in the context of DMS data but can be used to estimate fixation probabilities from landscapes predicted from VEPs.

While we do not yet have a complete understanding of the degree of temporal shifts in most proteins, the reviewed studies suggest that they are usually minor in magnitude and low in frequency. These consistent yet minor perturbations in preferences can lead to variation in rates across time; however, most inference models assume constant preferences. Evidence is emerging highlighting the value of accounting for temporal heterogeneity in inference procedures using a covarion-like process (e.g., Lu and Guindon (2014); Jones *et al.* (2020)). Therefore, allowing for temporal variability in addition to allowing preferences to vary across sites might lead to better models of protein evolution.

4.8 Conclusions

From the foregoing, it is clear that nonadaptive processes can alter site-specific amino acid preferences. Experimental studies suggest that at high sequence divergence levels only a small proportion of sites experience significantly shifted preferences (Doud *et al.*, 2015; Haddox *et al.*, 2018; Lee *et al.*, 2018; Chan *et al.*, 2017). Extensive computational studies (Ferrada, 2019; Shah *et al.*, 2015) corroborate this conclusion. Furthermore, pairwise amino acid exchange mutations between highly divergent sequences often have only minor differential effects on fitness (Starr *et al.*, 2018), function (Lunzer *et al.*, 2010; Emlaw *et al.*, 2020), and stability (Ashenberg *et al.*, 2013; Risso *et al.*, 2015). Together these results suggest that preferences at most sites vary slightly but are usually conserved over long evolutionary time scales. Nevertheless, the frequent, but small, changes in amino acid preferences leave an identifiable footprint in natural sequences: decreases in convergence rates (Goldstein *et al.*, 2015; Zou and Zhang, 2015a), reversion rates (Naumenko *et al.*, 2012; McCandlish *et al.*, 2016), and variation in replacement rates (Popova *et al.*, 2019; Stolyarova *et al.*, 2020; Gelbart and Stern, 2020) with time. While explicitly including epistatic interactions between all sites is computationally prohibitive, allowing for temporal variations in substitution processes (using a covarion-like process) and differences in preferences across sites (determined computationally or experimentally) are tractable ways of phenomenologically accounting for epistasis in inference models. Lastly, mutational

effects which appear inconsequential in experimental or computational settings may be exacerbated in nature. Further investigations into how nonadaptive processes alter evolutionary dynamics will be important, not only to better understand how proteins evolve but also to better identify adaptive episodes when they occur in natural proteins.

CHAPTER 5

DIFFERENCES IN EPISTATIC RESPONSE TO DESTABILIZING SUBSTITUTIONS ACROSS AND WITHIN PROTEINS

This work was done in collaboration with Scott McCain, Edward Susko, and Joseph Bielawski.

5.1 Abstract

Protein structures have significant implications for sequence evolution. Highly designable structures tend to have high contact densities and often evolve faster than less densely packed structures. On a finer scale, a site's location in the protein is strongly associated with its substitution rate, with buried sites often having lower rates than exposed sites. While rate dynamics have been extensively studied, it is currently unclear if the dynamics of recovery from a destabilizing substitution differ across proteins, and/or among sites within a protein. I perform extensive stability-informed simulations on six protein structures (PDB codes 2ppn, 1pek, 1qhw, 5jq3, 6vxx, 6nb6), which reveal that recovery dynamics differ across and within protein structures. In particular, I investigate the relationship between various features of protein structure and recovery time, measured as the number of compensatory substitutions that occur prior to protein stability returning to, or exceeding, the equilibrium stability value. In line with rate expectations, I observed that proteins with higher contact densities tend to recover more rapidly than proteins with lower contact densities. With

regards to within-protein dynamics, destabilizations at buried sites required a longer recovery time than destabilizations at exposed sites. This phenomenon is explicable by three underlying effects: (1) buried sites tend to have less uniform fitness landscapes; (2) buried sites have lower evolutionary rates; and (3) the fitness landscapes at buried sites are less robust to changes in the background sequence. Therefore, destabilizations at buried sites were often more severe and required a larger number of compensatory substitutions to recover stability than destabilizations at exposed sites. Overall, the results presented here provide evidence that protein structure plays an influential role in shaping the evolutionary response to destabilizing substitutions.

5.2 Introduction

Evolution, through the forces of natural selection, genetic drift, and mutation, has led to a diversity of proteins with different structures and functions. Until recently, the prevailing view has been that functional constraints constitute the predominant selective pressures throughout protein evolution. For example, amongst the five principles governing molecular evolution, Kimura and Ohta state that "[f]unctionally less important molecules or parts of a molecule evolve (in terms of mutant substitutions) faster than more important ones" (Kimura and Ohta, 1974). Nevertheless, it has now become apparent that while functional pressures do exert a strong influence on rates, effect sizes are large only at a small number of functionally relevant sites (e.g., active sites). Conversely, over the last two decades, evidence has accumulated highlighting the importance of another, more general constraint on proteins: structural constraints (DePristo *et al.*, 2005; Bershtein *et al.*, 2006; Tokuriki and Tawfik, 2009; Bershtein *et al.*, 2017). Unlike functional requirements, protein structure affects all sites, having a global influence.

Two studies have systematically examined the structural underpinnings of differences in evolutionary rates among proteins: Bloom *et al.* (2006) investigated yeast proteins, and Zhou *et al.* (2008) extended the analyses to *Escherichia coli*, fruit fly, and human proteins. The consensus from these studies is that more densely packed proteins evolve faster than proteins with lower contact densities. Contact density is related to the designability of a structure (Wolynes, 1996; Shakhnovich, 1998; England and Shakhnovich, 2003; Bloom *et al.*, 2006). The general hypothesis is that if many sequences can fold into a given structure (i.e., it is highly designable), then most mutations will preserve sequence stability

within the folded structure. The higher tolerance to mutations of more designable structures leads to higher evolutionary rates (measured by the number of substitutions). Interestingly, however, other summaries of secondary structures of proteins (e.g., the fraction of helix, sheet, turn, or coil sites) do not significantly correlate with the rate of subsitutions (Bloom *et al.*, 2006; Zhou *et al.*, 2008).

In addition to the differences across proteins, sites within the same protein typically exhibit differences in evolutionary dynamics. The among-site rate variability observed in natural sequence alignments is, to a large extent, driven by global structural constraints on proteins (see Echave *et al.* (2016) for a review). For example, sites on the protein surface tend to be more mutationally tolerant (i.e., can accept different amino acid mutations with little fitness effects) and have higher substitution rates than sites towards the core of the protein (Shahmoradi *et al.*, 2014; Yeh *et al.*, 2014; Echave *et al.*, 2015; Marcos and Echave, 2015; Nisthal *et al.*, 2019). This observation is explicable in terms of stability constraints: buried positions play a more substantial role in folding and maintaining adequate stability than exposed sites. Therefore, mutations at buried positions are often destabilizing and are less likely to be fixed than mutations at exposed sites. Further, the evolutionary dynamics at exposed locations are typically more dependent on the residues occupying other positions in the protein (i.e., the background protein sequence) (Youssef *et al.*, 2020). The collective effect is that rates at buried sites tend to be lower and more robust to different background sequences as compared to the rates at exposed positions.

Proteins are marginally stable, teetering on the verge of unfolding (Taverna and Goldstein, 2002; Williams *et al.*, 2006b; Goldstein, 2011). Therefore, destabilizing substitutions are usually purged by purifying selection. Nevertheless, there are instances where their fixation is inevitable. They may become fixed in a population due to random genetic drift, or a destabilizing residue might be essential for protein function—functional residues often compromise stability, resulting in a trade-off between functionality and stability (Tokuriki *et al.*, 2008; Miller, 2017). It is therefore valuable to understand the response to destabilizations across and within protein structures. Furthermore, a deeper understanding of the response to destabilization can help inform drug design and therapeutic interventions. For example, the mode of inhibition for six anti-viral drugs is through binding and destabilization of the Ebola virus spike protein (Ren *et al.*, 2018). The destabilization leads to premature uncoupling of the two spike subunits, preventing viral entry into host cells.

However, viruses evolve rapidly and are notorious for their ability to escape drug interventions. Therefore, characterizing the evolutionary response to destabilizations is crucial for treating viral infections where evolution is rapid and resistance to drugs is common. Do protein structures, or positions within a protein, respond differently to destabilizing substitutions? Is the ability to restore stability facilitated by some structural properties? Are destabilizations at some sites more challenging to compensate for? Addressing these questions is the central objective of this chapter, leading to important insights into protein evolution and the response dynamics that transpire after a destabilizing substitution.

To investigate how proteins and sites within proteins respond to destabilizations, I simulated protein evolution using a stability-informed model. In recent years, advances in biophysical models of protein evolution, in particular with selection for protein stability, have provided insights into evolutionary dynamics that are otherwise unobservable in natural proteins (Wylie and Shakhnovich, 2011; Pollock *et al.*, 2012; Shah *et al.*, 2015; Goldstein and Pollock, 2016; Echave, 2019; Youssef *et al.*, 2020). These models are grounded in the formalisms of both thermodynamics and population genetics theory. Predictions from stability-informed models often recapitulate trends present in empirical data, with regards to stability values (Goldstein, 2011), magnitude of stability effects of mutations (Shah *et al.*, 2015; Wylie and Shakhnovich, 2011), substitutions rates (Youssef *et al.*, 2020), and levels and trends in convergence rates (Goldstein *et al.*, 2015). Further, such models intrinsically account for structural constraints and reproduce correlations between structural descriptors (e.g., relative solvent accessibility, RSA, and weighted contact number; WCN) and evolutionary rates, even in the absence of functional constraints (Youssef *et al.*, 2020).

After ensuring that the evolutionary process was at mutation-drift-selection equilibrium, I introduced and held constant a destabilizing substitution. Then, I tracked subsequent evolution. In particular, I investigated differences in recovery times, measured as the number of substitutions after a destabilization until protein stability returns to, or exceeds, the equilibrium stability values. Recovery times varied across protein structures and among sites within the same protein. Protein properties that are known to correlate with evolutionary rates were also significantly associated with recovery times. At the protein level, differences in recovery times were generally explained by contact density,

with other structural features having little influence on recovery times. On a site level, variability in recovery times depended considerably on location in the tertiary protein structure: destabilizations at buried sites required on average significantly longer for recovery than destabilizations at exposed positions.

5.3 Results

The space of possible amino acid sequences is vast. Most sequences do not produce stable proteins for a given structure. The algorithm from Youssef *et al.* (2020) (Chapter 1 in this thesis) was used to obtain 50 unique sequences that are stable in a given protein structure (figure 5.1A). Each sequence was evolved for 200 substitutions to ensure the process was at mutation-drift-selection equilibrium. Following this *equilibration phase*, the sequences were evolved for 300 additional substitutions while keeping a record of site-specific fitness landscapes at all sites and given all background sequences. The purpose of this phase (hereafter, called *pre-intervention*) was to (i) characterize the dynamics and properties of sequences at equilibrium; and (ii) characterize the distribution of fitness effects at each site, including which state is the most destabilizing. The simulations are computationally expensive since each sequence must be thread through the set of alternative structures (see section 1.3.2 for details). Therefore, an intervention was not introduced at all sites. Instead, a subset of target sites was selected *a priori* (approximately 50 per protein) by ordering sites based on weighted contact number and sampling at uniform intervals (see Methods for details). At each target site, an *intervention* was introduced by individually fixing the most destabilizing mutation. Lastly, during the *recovery phase*, the destabilized sequences were evolved for 30 further substitutions with the constraint that no changes can occur at the target site. This constraint was enforced to prevent reversions and to allow for an assessment of the dynamics of compensatory substitutions.

For each simulation, recovery time was calculated as the number of substitutions post-intervention that occurred prior to restoring, or exceeding, the mean stability value in the pre-intervention phase. For example, consider the stability trajectories at two intervention sites 51 and 23 in the 2ppn protein (figure 5.1B). After the intervention at site 51, seven substitutions at other sites were fixed prior to the protein returning to its equilibrium stability value. In contrast, the destabilization at site 23 was adjusted for (i.e., equilibrium stability was restored) after 29 substitutions at other positions.

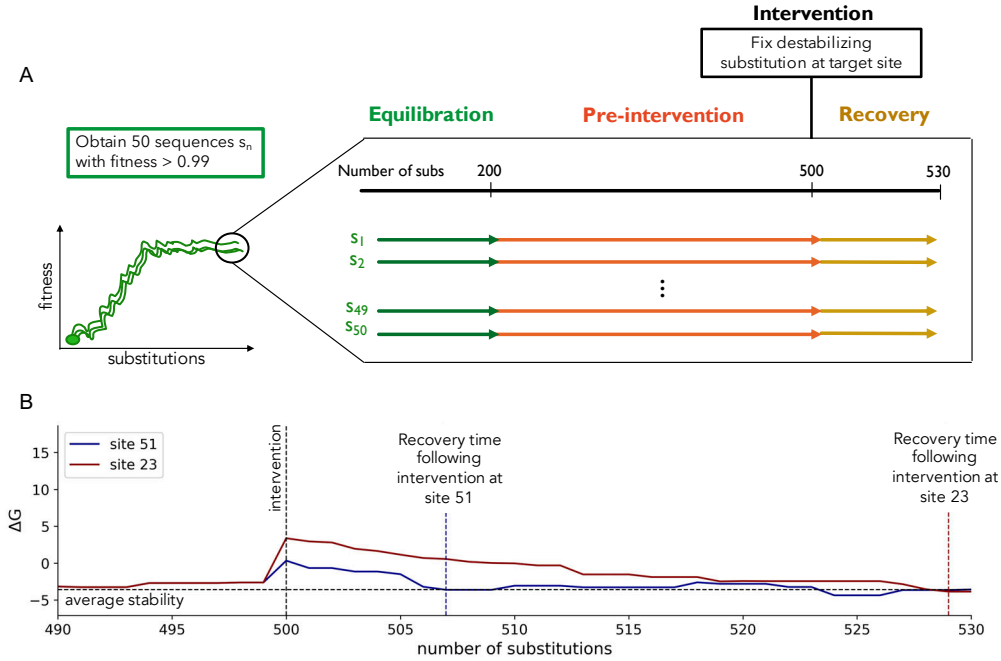


Figure 5.1: Methods outline. For each protein structure, 50 unique sequences with fitness greater than 0.99 were obtained. During the *equilibration phase*, each sequence was evolved for 200 substitutions to ensure mutation-drift-selection equilibrium was reached. Then, the sequences were evolved for 300 subsequent substitutions during the *pre-intervention phase*. To investigate the response to destabilizing substitutions, an *intervention* was introduced at a set of target sites (approximately 50 sites in each protein). The intervention constituted the fixation of the most destabilizing mutation at the site, with the constraint that subsequent substitutions cannot occur at the target site. During the *recovery phase*, the proteins adjusted to the destabilization through compensatory substitutions at other positions in the protein. (B) Example stability trajectories from a simulation of the 2ppn protein. Plotted are trajectories following interventions at sites 51 (blue) and 23 (red). The horizontal dotted line represented the average stability value in the pre-intervention phase. The vertical dotted lines represent recovery times. *Recovery time* was calculated as the number of substitutions that occurred post-intervention prior to the sequence returning to, or exceeding, the average stability value in the pre-intervention phase.

5.3.1 Structural classification of proteins

The method outlined above was applied to six protein structures (PDB codes: 1qhw, 2ppn, 1pek, 5jq3, 6vxx, 6nb6). The structures used here inevitably represent a minute fraction of all possible protein structures. The computational cost associated with the simulations prohibits its replication across a large number of protein structures. Nevertheless, the

selected proteins differed in length, function, and structural features representing a diverse range of structures (table 5.1). The 2ppn protein was the smallest, with only 107 amino acid sites. In contrast, the longest protein (6nb6) was composed of 1052 amino acid sites. Three of the selected structures were of globular proteins (2ppn, 1pek, and 1qhw). The remaining three were spike proteins (5jq3, 6nb6, and 6vxx), having globular and transmembrane domains. The advantage of the selected protein structures is that (i) they cover a wide range of protein lengths, (ii) simulations of the globular protein can be directly compared to previous studies (e.g., Pollock *et al.* (2012), and the simulations in chapter 3), and (iii) they expand the set of protein structures for which this stability-informed modelling framework has been applied.

For each protein, sites were classified into four secondary structure classes (helix, sheet, turn, and coil) using the DSSP software (Kabsch and Sander, 1983). Sites were also classified based on their location in the tertiary protein structure: buried sites were those with relative solvent accessibility (RSA) ≤ 0.05; exposed positions were those with $RSA > 0.05$. The percentages of sites within each secondary and tertiary structure category are summarized in table 5.1. Furthermore, for each protein, contact density was calculated as the average number of contacts per site. The 1pek and 1qhw protein were the most densely packed, having the highest contact densities and percentage of buried sites.

Table 5.1: Structural classification of proteins.

Protein	**Function**	**Organism**	**Len**	**Contact density**	**%Buried**	**Secondary Structure (% sites)**			
						Helix	**Sheet**	**Turn**	**Coil**
2ppn	isomerase	Human	107	6.90	21.5	10.3	38.3	25.2	26.2
1pek	proteinase	Fungus	279	8.39	43.0	24.0	21.5	27.2	27.2
1qhw	phosphatase	Rat	300	7.51	41.0	19.0	24.3	24.3	32.3
5jq3	spike	Ebola virus	384	6.32	21.1	16.4	27.6	21.4	34.6
6vxx	spike	SARS-COV2	972	6.76	29.9	19.4	33.2	20.0	27.4
6nb6	spike	SARS-COV	1052	6.59	18.1	18.4	33.8	19.4	28.3

Given the differences in structural features among the proteins, I was interested in assessing if the equilibrium properties of these proteins differ. For a given protein, stability values remained relatively constant during the pre-intervention phase, indicative of mutation-drift-selection equilibrium. The average stability value was -3.43 kcal/mol for the 2ppn protein; -3.71 kcal/mol for 1pek; -3.63 kcal/mol for 1qhw; -3.27 kcal/mol for 5jq3; -3.34 kcal/mol for 6vxx; and -3.32 kcal/mol for 6nb6 (figure 5.2A). The distribution of stability effects of all possible single-step mutations ($\Delta\Delta$G) was similar across protein

structures (figure 5.2B). As expected, the vast majority of mutations were destabilizing. The average $\Delta\Delta$G for all mutations ranged from 2.21 kcal/mol, for the 2ppn protein, to 2.53 kcal/mol, for the 6nb6 protein. Nevertheless, the distributions of fixed mutations (i.e., substitutions) were centred around zero for all proteins with a balance in the number of stabilizing and destabilizing substitutions (figure 5.2C). Such a balance is expected at mutation-drift-selection equilibrium (Goldstein, 2011; Jones *et al.*, 2017).

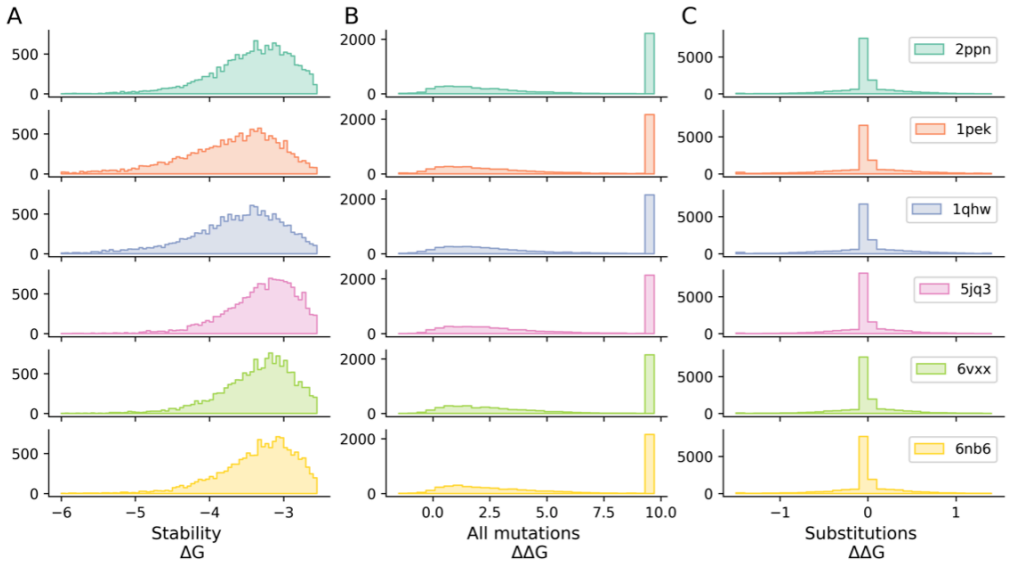


Figure 5.2: Distributions of equilibrium properties from simulations of six protein structures. Each row corresponds to a different protein (2ppn, 1pek, 1qhw, 5jq3, 6vxx, 6nb6). Proteins are ordered with respect to length. (A) Distributions of stability values of all sequences observed in the pre-intervention phase. (B) Distributions of all single-step mutations during the pre-intervention phase. Distributions were truncated such that $\Delta\Delta$G $<$ 10kcal/mol is set equal to 10 kcal/mol. (C) Distributions of all accepted substitutions during the pre-intervention phase. Results are based on 50 protein-specific simulations, each run for 300 substitutions.

5.3.2 Differences in mean recovery time across proteins

To investigate differences in recovery dynamics across the protein structures, I estimated the mean recovery time for a protein as the average across all intervention sites. Figure 5.3A reports the relationship between structural features (protein length, contact density, % buried, % helix, % sheet, % coil, and % turn sites) and average recovery times. Of the structural descriptors, only contact density and the percentage of buried sites were significantly associated with mean recovery times. The Pearson correlation between

contact density and mean recovery time was -0.97 (p-value < 0.001, figure 5.3B). The correlation between the percentage of buried sites and mean recovery time was -0.93 (p-value $= 0.001$, figure 5.3C). Contact density and the fraction of buried sites are both measures of the packing density of a protein; they were significantly correlated with each other (Pearson correlation 0.88, p-value $= 0.019$, figure 5.3D).

While the protein-level analyses presented here are based on only six protein structures, interventions were introduced at approximately 50 target sites per protein. These simulations, therefore, allow for a much more granular exploration of the recovery dynamics between sites *within* proteins.

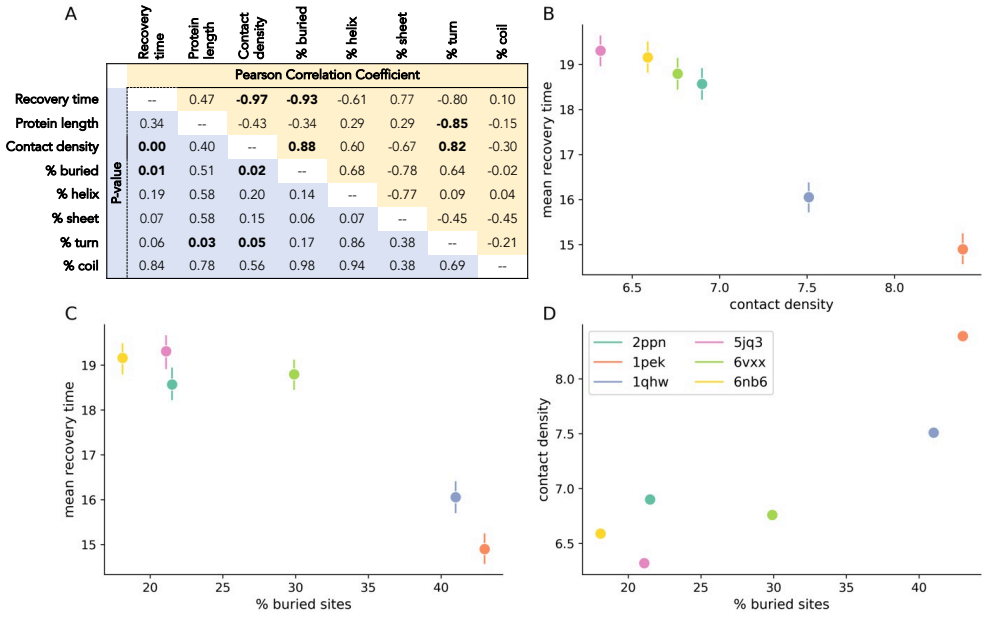


Figure 5.3: Relationship between mean time to recovery and structural features. (A) The upper triangle (shown in yellow) reports the Pearson correlation coefficients between protein properties and mean recovery times. The lower triangle (shown in blue) reports the corresponding p-values. Significant correlations (p-values ≤ 0.05) are bolded. (B) The relationship between mean recovery time and contact density. (C) The relationship between mean recovery time and percentage of buried sites. Plotted are the across-site average recovery times with bars representing the 95% confidence interval. (D) The relationship between contact density and percentage of buried sites.

5.3.3 Differences in mean recovery time across sites

The above results suggest that protein structure can influence the time to recovery following destabilizing substitutions. Nevertheless, an important question remains: Does response

to intervention differ depending on the location of the destabilization *within* the protein? There were 54 target sites for the 2ppn protein; 47 for 1pek; 50 for 1qhw; 48 for 5jq3; 49 for 6vxx; and 53 for 6nb6 (see section 5.5.3 for details and a list of target sites). At each target site, an intervention was introduced by fixing the most destabilizing substitutions (prohibiting further change at the site). This process was repeated across 50 different background sequences per protein for a total of (54×50 =) 2700 simulations of the 2ppn protein; 2350 simulations for 1pek; 2500 simulations for 1qhw; 2400 simulations for 5jq3; 2450 simulations for 6vxx; and 2650 simulations for 6nb6.

Response to destabilizations differed across sites (figure 5.4). For example, in the 2ppn protein, recovery times were longer for interventions at site 58 than at site 41 (mean recovery time was 26.1 substitutions for site 58 and 11.9 substitutions for site 41; figure 5.4A). To assess the dynamics across sites more rigorously, an intervention was performed at all 107 sites in the 2ppn protein and repeated across 50 different background sequences. Figure 5.4B shows the mean recovery times mapped onto the 2ppn protein structure. Qualitative assessment reveals that sites towards the core of the protein have higher recovery times than exposed sites. Furthermore, destabilizations at β-sheet sites appear to have longer recovery times than destabilizations at other secondary structures. For the 2ppn protein, mean recovery times ranged from approximately 12 to 26 substitutions (figure 5.4C). Similar levels of among site variability in recovery times were observed across the protein structures (figure 5.4C).

What explains these differences in response times? Site-specific structural properties have well-documented impacts on various evolutionary properties. For example, protein mutagenesis experiments reveal that β strands (i.e., sheet sites) are on average less mutationally tolerant than α helices (i.e., helix sites); and that turn sites are the most tolerant to mutations (Guo *et al.*, 2004). As a consequence, turn sites tend to have higher replacement rates than sheet sites (Goldman *et al.*, 1998). In addition, exposure to solvent (buried versus exposed sites) tends to have a significant impact on the evolutionary rates at sites; exposed sites are often more tolerant of mutations and have higher substitution rates than buried sites (Shahmoradi *et al.*, 2014; Yeh *et al.*, 2014; Echave *et al.*, 2015; Marcos and Echave, 2015). I, therefore, hypothesized that differences in structural features among sites may impact the ability of the protein to adjust to the destabilization.

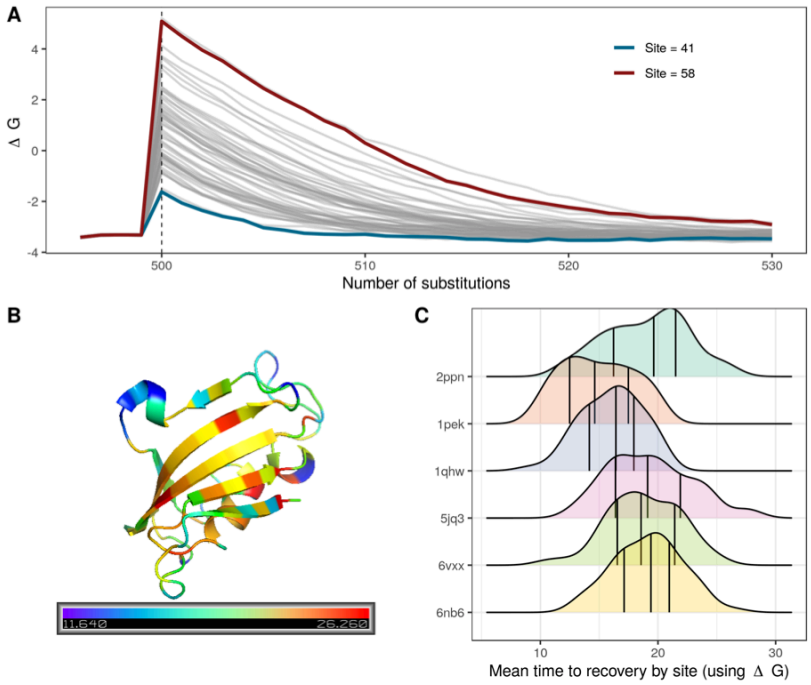


Figure 5.4: Variability in mean recovery times across sites. (A) Average stability trajectories across sites in the 2ppn protein. Each site is represented by a line and stability values are averaged over fifty different background sequences. Sites 41 and 58 are highlighted in blue and red, respectively. Site 58 required, on average, the longest to recover while site 41 had the quickest recovery time. (B) Mean recovery times mapped onto the 2ppn protein structure. (C) Distributions of site-specific mean recovery times across all protein structures. Vertical lines represent the first, second, and third quantiles. Figure was made using R packages ggridges (Wilke, 2020) and ggplot2 (Wickham, 2016).

To assess the impact of secondary and tertiary structural features on response dynamics, I compared the distributions of mean recovery times for sites belonging to different secondary structure classes (helix, sheet, turn, or coil; figure 5.5A). A one-way ANOVA test revealed that there was a statistically significant difference in recovery times between site classes in the 2ppn (p-value < 0.001), 1pek (p-value < 0.001), 1qhw (p-value < 0.001), and the 5jq3 (p-value $= 0.011$) proteins. The differences in recovery times between site classes in the 6nb6 and 6vxx proteins were not significant (p-values were 0.332 and 0.073, respectively). Tukey post-hoc analyses revealed that mean recovery times were significantly higher at sheet compared to turn sites in the 2ppn (p-value = 0.001), 1pek (p-value = 0.001), 1qhw (p-value = 0.004), and 5jq3 (p-values = 0.050) proteins. Furthermore,

differences in recovery times were significant between sheet and coil sites in the 2ppn (p-value = 0.025) and 5jq3 (p-value = 0.015) proteins; and between helix and turn sites in the 1pek protein (p-value = 0.020). In summary, secondary structure had some effect on mean recovery times. The largest discrepancy was observed between destabilizations at sheet and turn sites: interventions at sheet sites required the longest recovery times, while destabilizations at turn sites tended to recover more quickly.

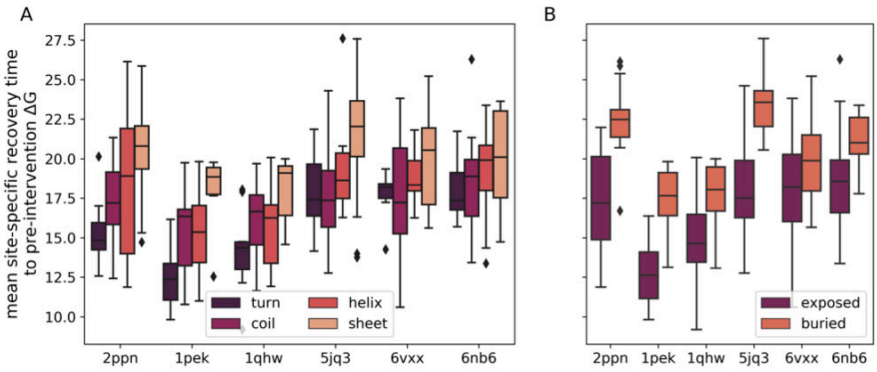
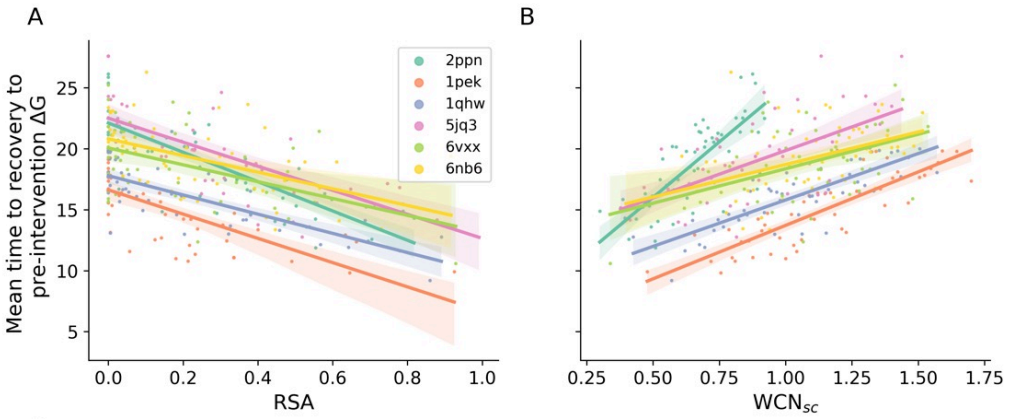


Figure 5.5: Distributions of mean recovery times for different site types. (A) Sites are categorized based on secondary structure (coil, sheet, turn, or helix). The average recovery times were significantly different across secondary structure classifications for the 2ppn, 1pek, 1qhw, and 5jq3 proteins (one-way ANOVA, p-values ≤ 0.05). Differences were not significant for the 6vxx and 6nb6 (p-values were 0.07 and 0.33, respectively). Post-hoc analyses revealed that mean recovery times were significantly different between turn and sheet sites in the 2ppn, 1pek, 1qhw, and 5jq3 proteins (Tukey HSD, p-values < 0.05). Coil and sheet sites had significantly different mean recovery times in the 2ppn, and 5jq3 proteins (Tukey HSD, p-values ≤ 0.025). Helix and turn sites had significantly different mean recovery times in the 1pek protein (Tukey HSD, p-value $= 0.02$)(B) Buried sites recovered significantly slower than exposed sites across all protein structures (Welch's t-test, all p-values ≤ 0.05).

In addition to secondary structure, the association between the location of a site in the tertiary structure and its evolutionary rates is often significant (Shahmoradi *et al.*, 2014; Yeh *et al.*, 2014; Echave *et al.*, 2015; Marcos and Echave, 2015). Relative solvent accessibility (RSA) and weighted contact number (WCN) are two commonly used measures for determining a site's location in the tertiary protein structure. The weighted contact number can be measured in two ways: with respect to the C_α of the amino acid (WCN_α); or with respect to the geometric centre of the side chain (WCN_{sc}). The latter measure is a better

predictor of rate variation among sites (Marcos and Echave, 2015). I, therefore, report on the results based on WCN_{sc} although similar conclusions are expected for WCN_{α}.

Destabilizations at exposed sites were more rapidly compensated (through adjustments at other sites) than interventions at buried sites (figure 5.5B; Welch's t-test, p-value $<$ 0.05 for all proteins). Buried and exposed sites are often classified based on an RSA cut-off: buried sites having $RSA \leq 0.05$; exposed sites having $RSA > 0.05$. Based on this site 58 ($RSA = 0.03$) in the 2ppn protein was classified as a buried site. However, it was less densely packed than other buried sites, with a relatively low WCN_{sc} equal to 0.89. As such, site 58 was an outlier having the lowest mean recovery time among the buried sites in the 2ppn protein (figure 5.5B). To avoid classifications based on arbitrary cut-offs, I looked at the association between location in the tertiary structure more closely by correlating mean recovery times with RSA (figure 5.6A) and WCN_{sc} (figure 5.6B). Mean recovery times were significantly associated with both structural descriptors; a significant negative relationship was observed between RSA and mean recovery times across all protein structures (Pearson correlations < -0.474, all p-values ≤ 0.001, figure 5.6C); while a significant positive relationship was present between WCN_{sc} and mean recovery times (Pearson correlations > 0.447, all p-values ≤ 0.001).



C

	2ppn	1pek	1qhw	5jq3	6vxx	6nb6
RSA vs mean time to recovery	-0.822	-0.720	-0.775	-0.728	-0.474	-0.512
WCN$_{sc}$ vs mean time to recovery	0.774	0.847	0.802	0.590	0.457	0.447

All p-values ≤ 0.001

Figure 5.6: Relationship between mean recovery times and a site's location in the protein measured by relative solvent accessibility (RSA; A) and weighted contact number (WCN_{sc}; B). (C) Pearson correlation coefficients for the different protein structures. All p-values were ≤ 0.001.

To assess whether both RSA and WCN_{sc} had significant and independent impacts on recovery times, I performed a multiple linear regression to predict mean recovery times based on RSA, WCN_{sc}, and protein (table 5.2). These variables explained approximately 60% of the variance in recovery times. Importantly, the results demonstrate that both RSA and WCN_{sc} were significant predictors of mean recovery times.

Table 5.2: Multiple regression of mean recovery time on RSA (Å^2), WCN_{sc} (Å^{-2}) and protein.

	coef	**std err**	**p-value**
Intercept	12.3	1.08	< 0.001
Protein-1qhw	1.78	0.46	< 0.001
Protein-2ppn	6.06	0.55	< 0.001
Protein-5jq3	6.05	0.47	< 0.001
Protein-6nb6	4.86	0.45	< 0.001
Protein-6vxx	4.14	0.46	< 0.001
RSA	-6.45	0.85	< 0.001
WCN_{sc}	3.26	0.81	< 0.001

5.3.4 Why buried sites take longer to recover than exposed sites

What explains the longer recovery times at buried sites compared to exposed sites? I hypothesize that there are at least three contributors to this observation: (i) landscape entropy, (ii) differences in evolutionary rates, and (iii) epistatic sensitivity. The consequence of each is described in detail below.

Exposed sites are often more tolerant of mutations than buried sites such that a higher number of residues may occupy the site with little to no fitness effects (Youssef *et al.*, 2020; Nisthal *et al.*, 2019). Therefore, exposed positions will tend to have more uniform site-specific fitness landscapes than buried sites and, on average, have higher substitution rates. To investigate if these observations are recapitulated within the simulation framework presented here, the uniformity of a landscape was calculated as the Shannon entropy at each site and given all background sequences as:

$$\mathrm{H}^h(S) = -\sum_a f_a^h(S) \ln f_a^h(S) \tag{5.1}$$

where $f_a^h(S)$ is the fitness of the sequence carrying amino acid a at site h in the context of the background sequence S. The entropy of a uniform landscape will be ≈ 3. Alternatively, if only a single amino acid is permissible at a site h, then H^h will be equal to zero. Consistent with previous observations, landscapes at exposed positions were significantly more tolerant of mutations (i.e., have more uniform fitness landscapes) as compared to buried sites (figure 5.7A, Welch's t-test all p-values < 0.001).

Since buried sites have less uniform fitness landscapes, the destabilizing substitutions fixed during the intervention tended to have more substantial fitness effects than at exposed sites. The stability effect of a mutation was significantly correlated with recovery times (Pearson correlation > 0.429, all p-values < 0.001). As an example, consider the fitness landscapes at buried site 85 (figure 5.9A) and exposed site 121 (figure 5.9B) of the 1qhw protein across different background sequences. Site 85 was the most densely packed target site (WCN_{sc} = 1.55) and site 121 was the least packed (WCN_{sc} = 0.42). The landscapes at buried site 85 are much less uniform than the landscapes at site 121. At the time of interventions (number of substitutions = 300), the most destabilizing amino acid (leucine, one-letter code L) was fixed at site 85 (figure 5.9A). The stability effect of the K to L substitution was highly destabilizing with a $\Delta\Delta$G = 7.33 kcal/mol, leading to a protein

with fitness equal to 0.001. In contrast, at site 121 the most destabilizing mutation (to phenylalanine, one-letter code F) resulted in a comparatively minor destabilization, $\Delta\Delta$G = 2.32 kcal/mol with fitness equal to 0.84.

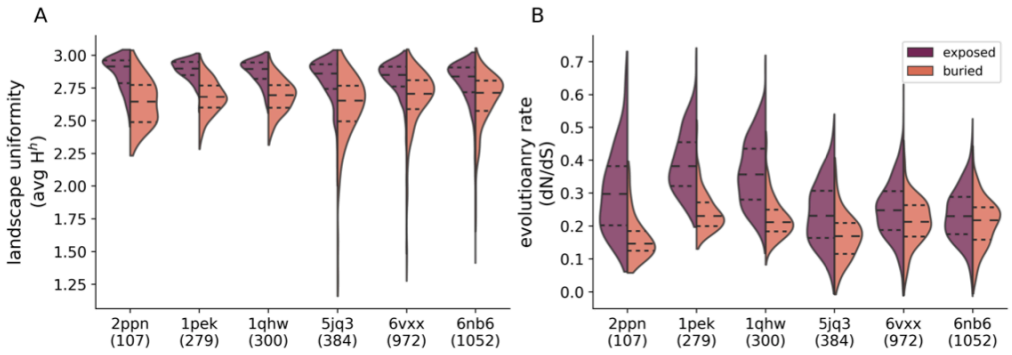


Figure 5.7: Differences in evolutionary dynamics between buried and exposed sites; surface sites are more mutationally tolerant and evolve quicker than buried sites. (A) Distributions of average landscape entropy (avg H) at a site, averaged over all background sequences observed in the pre-intervention phase. (B) Distributions of expected substitution rates (dN/dS). All p-values were < 0.001, Welch's t-test. Plotted results are from six protein structures (PDB code: 2ppn, 1pek, 1qhw, 5jq3, 6vxx, and 6nb6). Protein lengths are provided in parentheses below the PDB code.

All else being equal (i.e., constant effective population size, environment, and mutation rates), the fitness landscape at a site dictates its rate of substitution. Therefore, the observation that landscapes at exposed sites are more uniform than at buried sites suggests that they would have higher substitution rates. To illustrate this, consider, again, the fitness landscapes at buried site 85 and exposed site 121 (figure 5.9A and 5.9B). The relatively uniform fitness landscapes at site 121 indicate that stability effects of mutations will be less severe. As such, the rate of fixation of nonsynonymous mutations will be relatively high. Alternatively, at buried site 85, amino acid K is selectively preferred over all other residues. Since most nonsynonymous mutations will decrease fitness, the substitution rate at the site will be low, reflective of purifying selection.

To explore the relationship between recovery times and substitution rates, I estimated the evolutionary rate at a site by calculating the expected nonsynonymous to synonymous substitution rate ratio (dN/dS; see Methods for details). In line with empirical observations, buried sites had significantly lower evolutionary rates than exposed sites in all proteins (Yeh *et al.*, 2014; Shahmoradi *et al.*, 2014; Marcos and Echave, 2015) (figure

5.7B). However, the effect sizes, and strength of the correlation, tended to decrease as protein length increased. Furthermore, I observed a significant association between mean recovery times and site-specific substitution rates: the Pearson correlation coefficients were $-0.923, -0.861, -0.887, -0.805, -0.688,$ and -0.688 for the 2ppn, 1pek, 1qhw, 5jq3, 6vxx, and 6nb6 proteins respectively (all p-values < 0.001; figure 5.8).

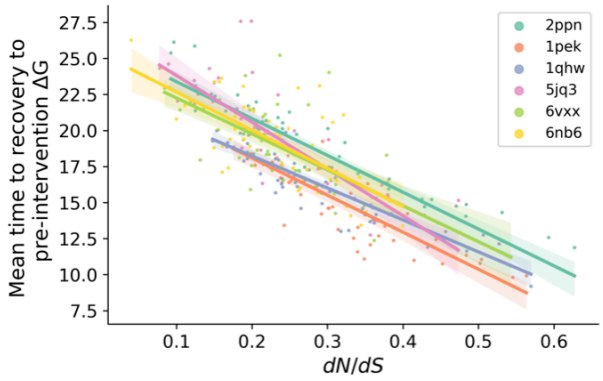


Figure 5.8: Relationship between mean recovery times and substitution rates (dN/dS) across six protein structures (2ppn, 1pek, 1qhw, 5jq3, 6vxx, 6nb6).

Substitution rates at exposed sites tend to be more variable than rates at buried sites (Youssef *et al.*, 2020). Since site-specific fitness landscapes underlie substitution rates, this observation suggests that landscapes at exposed sites are more variable depending on the background protein sequence than the landscapes of buried sites. As a proxy for the variability in the fitness landscapes at a site, I measured the number of unique amino acids that produced the fittest protein. For example, lysine (one-letter code K) was the most preferred amino acid at site 85 regardless of the background sequence (figure 5.9B). In contrast, the fittest amino acid at site 121 varied depending on the background sequence (figure 5.9C). A similar trend was observed across all buried and exposed sites: the fittest amino acid at buried sites is less dependent on the background sequence compared to exposed sites (figure 5.9D).

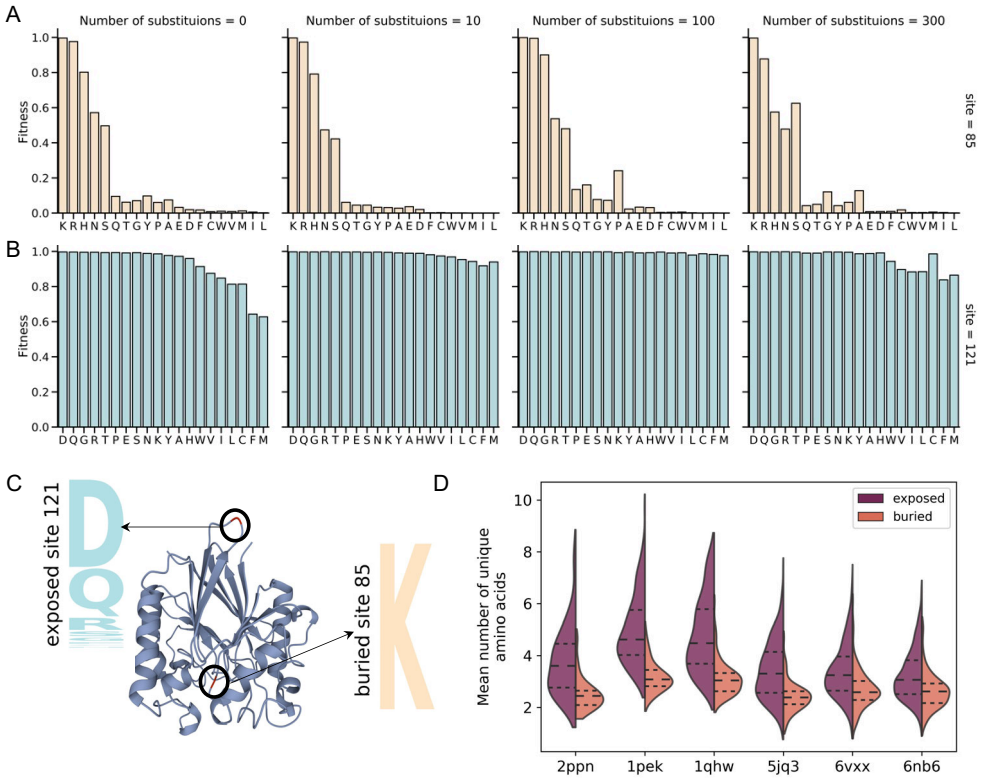


Figure 5.9: Higher variability in fittest amino acids at exposed sites. Fitness landscapes at buried site 85 (A) and exposed site 121 (B) of the 1qhw protein given the background sequences observed at the start of the pre-intervention phase, and after 10, 100, and 300 substitutions. (C) Logo plots representing the variability in the most preferred residue across background sequences at exposed site 121 and buried site 85 of the 1qhw structure. The letter size is proportional to the number of background sequences where the respective amino acid was the fittest. (D) The number of unique amino acids that were most preferred at a site was significantly higher at exposed than buried sites.

5.4 Discussion

The work presented in this chapter highlights the association between protein structure and recovery dynamics. Across various protein-level structural features, contact density had the strongest effect on recovery times: more densely packed proteins tended to recover more quickly to destabilizations than loosely packed proteins. An important caveat to this analysis is that the simulations were based on only six protein structures, limiting

the ability to derive general conclusions regarding the influence of protein structure on recovery dynamics. Nevertheless, the observation that more densely packed proteins recover more quickly follows from previous conclusions based on the relationship between contact density and evolutionary rates (Bloom *et al.*, 2006; Zhou *et al.*, 2008). Structures with higher contact densities are more designable (i.e., a greater number of amino acid sequences can fold stably into the given structure)(Wolynes, 1996; Shakhnovich, 1998). The observation that more densely packed structures (e.g., 1pek and 1qhw) were able to recover stability after a comparatively smaller number of compensatory substitutions (i.e., have shorter recovery times) than less packed structures (6nb6 and 5jq3) follows from previous claims that more sequences can fold into densely packed structures (Wolynes, 1996; Shakhnovich, 1998; England and Shakhnovich, 2003; Bloom *et al.*, 2006).

The simulations presented here allow for a more rigorous investigation of the dynamics of recovery across sites within a protein. Destabilizations at exposed sites were more easily compensated for than destabilizations at buried sites. Three underlying phenomena explain this observation. First, exposed sites tended to have more uniform fitness landscapes (Youssef *et al.*, 2020; Nisthal *et al.*, 2019), a pattern that is recapitulated in the stability-informed simulations. A consequence of this is that interventions constituted less severe destabilizations at exposed compared to buried sites. As such, fewer adjustments were required in order to restore stability.

Second, it has previously been observed that buried sites tend to be less sensitive to epistasis (Youssef *et al.*, 2020), where epistatic sensitivity was measured as the variability in the site-specific dN/dS rate across different background sequences. Rates at buried sites were less variable given different background sequences compared to the rates at exposed positions. Within the stability-informed framework presented here, (with fixed effective population size, environment, and mutation rates) variations in dN/dS rates are a direct consequence of varying site-specific fitness landscapes. Therefore, that exposed sites have higher variability in dN/dS suggests that fitness landscapes at exposed sites may be more influenced by the background sequence than landscapes at buried sites. Indeed, I observed that the number of unique amino acids that produced the protein with the highest fitness was more variable (i.e., more dependent on the background protein sequence) at exposed sites compared to buried sites. This is perhaps expected since buried sites have a higher number of interactions. Therefore, the fitness landscapes at a buried site must be

compatible with the residues present at the many neighbouring sites. In contrast, exposed sites have fewer contacts, and a substitution at one of the few interacting sites might more easily dictate the fitness landscape.

Lastly, exposed sites tend to have higher substitution rates than buried sites (Shahmoradi *et al.*, 2014; Yeh *et al.*, 2014; Echave *et al.*, 2015; Marcos and Echave, 2015). This observation is also recapitulated within the modelling framework. However, it is not directly apparent why evolutionary rate should influence recovery time since the substitution process was constrained to prevent further substitutions at the intervention site. In other words, recovering stability does not directly depend on the evolutionary rate at the target site. The association between recovery times and evolutionary rates can be understood through the dynamics at neighbouring sites. Exposed (or buried) sites tend to neighbour other exposed (or buried) sites, which also have more uniform (or peaky) landscapes and evolve quickly (or slowly). Therefore, a destabilizing intervention at a buried site requires longer for the protein to recover stability since the sites neighbouring the intervention site are less likely to substitute away from the current resident amino acids.

By looking across both protein-level (e.g., contact density and the fraction of buried sites) and site-level (e.g., buried and exposed sites) structural descriptors an interesting phenomenon emerges: structural features that influence evolutionary rates also impact recovery times. More specifically, features that correlate positively (or negatively) with mean recovery times have an inverse relationship with evolutionary rates. More densely-packed proteins tend to evolve at higher substitution rates, and tend to have shorter recovery times than less densely packed proteins. Turn sites are more mutationally tolerant and have higher evolutionary rates than sheet sites (Guo *et al.*, 2004; Goldman *et al.*, 1998). In line with this observation, I found that destabilizations at turn sites were more quickly adjusted for than destabilizations at sheet sites. Furthermore, destabilization at buried sites (which tend to have lower substitution rates) required more adjustments at other positions than destabilizations at exposed sites in order to recoup equilibrium stability values. Therefore, there appears to be a general relationship between evolutionary rates and recovery dynamics both across and within proteins.

Protein evolution is a stochastic process. Destabilizing substitutions may be fixed in populations by random genetic drift, as a result of a trade-off between stability and functionality (Tokuriki *et al.*, 2008; Miller, 2017), or as a consequence of targeted drug

interventions (Ren *et al.*, 2018). Therefore, characterizing general associations of response to destabilization with phenotypic features such as protein structure is essential to understanding the process of sequence evolution both within and across proteins. Overall, the findings presented here suggest that the structures of proteins and sites within a protein may differ substantially in their recovery dynamics.

5.5 Methods

5.5.1 Model of sequence evolution

To examine the relationship between recovery time and structural features, I simulated sequence evolution using six protein structures (PDB codes: 2ppn, 1pek, 1qhw, 5jq3, 6vxx, and 6nb6). I modelled the evolutionary process as a continuous-time Markov chain with site-specific fitness landscapes according to the MutSel model (Halpern and Bruno, 1998) with fitness values determined from the stability-informed framework. See sections 1.3.1 and 1.3.2 for details regarding the MutSel and stability-informed models.

I assumed a fixed effective population size with $N_e = 100$, equal nucleotide frequencies ($\pi_A = \pi_C = \pi_T = \pi_G = 1/4$) and a transition-transversion rate $\kappa = 2$. Given each protein structure, 50 simulations were performed starting at different initial sequences. The simulations were run for 500 substitutions. To ensure that the process has reached mutation-drift-selection equilibrium, I report on results after an initial equilibration phase of 200 substitutions.

5.5.2 Secondary structure (helix, sheet, coil, turn), relative solvent accessibility (RSA), and weighted contact number (WCN)

Site-specific structural properties were estimated following the protocol in Sydykova *et al.* (2018). The xssp web server (https://www3.cmbi.umcn.nl/xssp/) was used to estimate secondary structure and to calculate the solvent accessible surface area (ASA) for each site. Following Bloom *et al.* (2006), sites were grouped into four secondary structure types based on their DSSP class (Kabsch and Sander, 1983): helix (class H), sheet (class E), turn (class S and T), and coil (class B, G, I). Relative solvent accessibility (RSA) was calculated as

$$RSA = ASA/maxASA \tag{5.2}$$

where $maxASA$ is the maximum accessible surface area as measured by Tien *et al.* (2013).

A site's weighted contact number was calculated as:

$$WCN_x = \sum_{j \neq i} \frac{1}{d^2_{x_i,x_j}} \tag{5.3}$$

The distance $d^2_{x_i x_j}$ can be calculated as either the distance between the C_α for sites i and j (then $x = \alpha$), or as the distance between the geometric center of the side-chain (then $x = sc$). Weighted contact number based on side-chain contact density was a better predictor of rate variation among sites (Marcos and Echave, 2015). Therefore, the results presented here are based on WCN_{sc}. The results remained consistent when WCN_α was used.

5.5.3 Identifying target sites

The simulation procedure outlined above is computationally expensive, limiting the ability to evaluate the recovery dynamics at all sites. Instead, a subset of sites was selected *a priori* by ordering sites based on WCN_{sc} and sampling at fixed intervals for a total of approximately 50 target sites per protein. Note that the sampling interval depends on protein length. For example, for the 2ppn protein (length = 107) a site was sampled every 2nd interval for a total of 54 sites. Whereas, for the 6nb6 protein (length = 1052) a site was sampled every 20th interval for a total of 53 sites. The target sites for each protein are provided in table 5.3. After the intervention, the sequences were evolved for 30 further substitutions. Then, recovery time was calculated as the number of substitutions after the intervention that occurred before the sequence returned to, or exceeded, the mean pre-intervention stability value.

Table 5.3: Target sites selected for intervention. Sites were ordered based on WCN_{sc} and sampled at uniform intervals. The intervals were determined such that approximately 50 target sites were obtained per protein.

Protein (length)	**Sampling interval**	**Total # of sites**	**Site number**
2ppn (107aa)	2nd	54	0, 3, 4, 6, 7, 8, 9, 12, 13, 14, 20, 21, 22, 23, 24, 25, 28, 30, 33, 35, 36, 37, 38, 40, 41, 48, 51, 52, 54, 55, 56, 57, 58, 59, 62, 64, 70, 71, 74, 75, 78, 82, 85, 87, 88, 91, 93, 95, 97, 99, 100, 102, 103, 106
1pek (279aa)	6th	47	2, 3, 8, 16, 17, 24, 32, 38, 44, 62, 66, 68, 74, 92, 94, 102, 103, 106, 107, 111, 114, 128, 137, 142, 143, 147, 155, 158, 162, 165, 166, 182, 186, 195, 199, 202, 213, 214, 224, 226, 232, 242, 246, 254, 260, 265, 266
1qhw (300aa)	6th	50	1, 4, 18, 19, 29, 33, 35, 39, 51, 53, 56, 59, 80, 85, 92, 93, 96, 101, 104, 112, 114, 121, 130, 133, 134, 141, 146, 157, 160, 163, 164, 165, 173, 186, 193, 194, 204, 213, 214, 224, 231, 245, 247, 250, 253, 269, 276, 285, 294, 295
5jq3 (384aa)	8th	48	0, 6, 20, 25, 29, 43, 71, 75, 90, 96, 107, 110, 112, 137, 146, 150, 174, 178, 179, 187, 188, 195, 207, 214, 223, 224, 225, 235, 238, 241, 246, 249, 250, 251, 270, 276, 294, 298, 317, 331, 334, 335, 342, 351, 366, 369, 370, 382
6vxx (972aa)	20th	49	6, 50, 65, 72, 85, 91, 96, 105, 185, 262, 283, 292, 310, 314, 315, 323, 335, 359, 369, 387, 393, 425, 448, 465, 471, 477, 491, 504, 526, 531, 534, 535, 554, 565, 585, 634, 640, 664, 697, 719, 725, 729, 823, 849, 874, 901, 917, 956, 971
6nb6 (1052aa)	20th	53	2, 5, 27, 29, 31, 36, 65, 102, 118, 130, 134, 155, 158, 178, 228, 236, 296, 298, 336, 353, 420, 433, 439, 444, 480, 489, 492, 529, 543, 559, 580, 590, 609, 619, 622, 680, 703, 716, 720, 774, 782, 787, 812, 828, 830, 831, 884, 886, 914, 948, 997, 1001, 1051

5.5.4 Site-specific dN/dS

To calculate the expected rate at a site we must consider all possible background sequences since epistatic interactions with other sites can change the site's rate of evolution. The substitution rate dN/dS can, in principle, be calculated as the rate of nonsynonymous substitutions (N) normalized by the rate of nonsynonymous mutations (N_{mut}) given all

sequences S

$$dN/dS = \frac{\sum_S N}{\sum_S N_{mut}} \tag{5.4}$$

where

$$N = \sum_x \sum_{y \in \mathcal{N}_x} \pi_x q_{xy} \tag{5.5}$$

$$N^h_{mut} = \sum_x \sum_{y \in \mathcal{N}_x} \pi^h_x \mu_{xy} \tag{5.6}$$

$\mathcal{N}_x$ is the set of codons that are nonsynonymous to codon x and differ by a single nucleotide, q_{xy} is the substitution rate from x to y calculated using equation (1.3), μ_{xy} is the mutation rate calculated based on the HKY85 model, and π_x is the stationary frequency for codon x at site h. However, the space of possible sequences is vast (20^L where L is the length of the protein), prohibiting our ability to get an exact estimate of dN/dS. Instead, equation (5.4) was summed over all sequences observed in the pre-intervention phase for a given protein (a total of 300×50 sequences per protein).

CHAPTER 6

DISCUSSION

> Wind back the tape of life, and let it play again. Would the replay ever yield anything like the history that we know?
> —Stephen Jay Gould. *Wonderful Life*.

In the decades since Darwin's *Origin*, our understanding of the evolutionary process has grown immensely. Nevertheless, we cannot predict with absolute accuracy which new variants of proteins will emerge (Agor and Özaltın, 2018) or map out the exact trajectories in sequence space leading to extant proteins (Sailer and Harms, 2017). These shortcomings are not because of a superficial understanding of the evolutionary process. Rather, evolution is stochastic in nature, traversing an immense sequence space, with varying outcomes expected even under identical start conditions (Gould, 1991). Our best bet is to embrace a phenomenological approach, characterising emerging patterns and dynamics, and identifying phenotypic features of proteins that are associated with particular phenomena.

All models, in biology and elsewhere, make simplifying assumptions. When nature violates these assumptions, the inferences derived from such models may be biased. It is, therefore, crucial to understand how violations of model assumptions may be impacting our evolutionary inference. Throughout this dissertation, I used a stability-informed modelling framework to constrain the evolutionary process. A central assumption in this framework is that natural selection is acting solely on protein stability, a gross oversimplification of the constraints governing protein evolution. Yet, this modelling framework recapitulates numerous patterns present in natural sequences: (i) the marginal stability of proteins (Goldstein, 2011); (ii) the levels and patterns of convergence rates (Goldstein *et al.*, 2015);

(iii) substitution rates (Youssef *et al.*, 2020); (iv) the relationship between a site's location in the protein and its evolutionary rate (Youssef *et al.*, 2020); and (v) the distribution of stability effects of mutations (Goldstein, 2013). For these reasons, I believe that this modelling framework is appropriate for investigating the questions addressed in this dissertation, and, more generally, for developing a deeper understanding of the dynamics of protein evolution.

In addition to reproducing various empirically observed phenomena, the stability-informed modelling framework has the advantage that it implicitly accounts for epistatic interactions between sites. It is therefore useful for understanding how epistasis may influence the evolution of protein sequences. As discussed in Chapter 2, stability-mediated epistasis resulted in higher rates of substitution than the expected rates had sites evolved independently, an initially counter-intuitive phenomenon that can be understood by viewing the fitness landscape dynamics at a site. As substitutions accrue in the protein, the site-specific fitness landscape will change in response. As such, a site must constantly adjust for changes in the background protein sequence, resulting in dynamics reminiscent of adaptive Red-Queen regimes (Van Valen, 1973; Rodrigue and Lartillot, 2017).

In contrast with such a dynamic and interdependent view of the evolutionary process at a site, widely used inference procedures (e.g., ω-based models and phylogenetic models) often assume that sites evolve independently. It is all but impossible to forgo this assumption since the combinatorial explosion of accounting for all possible interactions would make these inference procedures intractable. Some have argued that such limitations may significantly impair the accuracy of inferred parameters (Pollock *et al.*, 2012). Others have asserted that temporal nonstationarity (e.g., changes in site-specific fitness landscapes over time) are relatively minor compared to among-site variability, arguing in support of the simplifying, albeit unrealistic, assumption of site-independence (Ashenberg *et al.*, 2013). I have shown in Chapter 2, that, with regards to the inference of substitution rates, site-independence models underestimated the amount of among-site rate variability. Nevertheless, they captured the most common rates across sites. As such, the results presented in Chapter 2, favour the interpretation that while temporal epistatic dynamics are unequivocal, their magnitudes are comparatively minor.

While epistatic models might not be essential for accurate inference, they are powerful tools for investigating evolutionary dynamics. They have revealed that proteins can adjust

for resident amino acids by increasing their propensities over time (Pollock *et al.*, 2012). This has led some to conclude that the detection of the counter dynamics—decreases in resident amino acid propensities—provide evidence of adaptive evolution (Stolyarova *et al.*, 2020; Popova *et al.*, 2019). In Chapter 3, I presented results that challenge this claim, showing that decreases in propensities can occur in the absence of adaptive evolution. Furthermore, I have shown that the proportion of sites for which the propensity of the resident amino acid increases is expected to be equal to the proportion where propensities decrease under non-adaptive evolution. Under adaptive circumstances (e.g., external environmental changes), I suspect that the proportion of sites experiencing decreases in resident amino acid propensities will exceed those for which the propensities increase. As such, site-specific dynamics may still be useful for revealing instances of adaptive regimes. This work highlights that characterizing general evolutionary dynamics using plausible stability-informed models is essential not only for understanding the process of sequence evolution but also for the accurate detection of molecular adaptations.

What constitutes a molecular adaptation? It is commonly defined as the evolutionary response that transpires following an environmental or functional change external to the protein (dos Reis, 2015; Jones *et al.*, 2017). Following the external change, the mapping of protein sequences to fitness values shifts such that the current protein state is suboptimal for the new conditions. However, site-specific fitness landscapes are constantly shifting in response to changes at other positions. In a way, the environment surrounding the site varies as amino acids change in the background sequence. The site must therefore evolve in response to that shift in the landscape. Can such site-level epistatic dynamics be differentiated from protein-level adaptations? In chapter 4, I reviewed theoretical and experimental work on site-specific shifts in fitness landscapes. The consensus from these studies is that changes in site-specific fitness landscapes due to epistatic interactions are often minor in magnitude. Over geological time scales (approximately four billion years of evolution), the most preferred residue at a site often remains conserved (Risso *et al.*, 2015). Alternatively, when faced with different environmental conditions, the set of prefered amino acids often differ drastically (Hietpas *et al.*, 2013). Furthermore, based on results from Chapter 2, I expect that the balance of sites for which propensities, and hence substitution rates, increase or decrease will no longer apply when a protein is subjected to novel environmental conditions. Therefore, while shifts in landscapes can occur due to

nonadaptive epistasis and adaptive evolution, the dynamics are likely to differ, allowing them to be distinguishable.

Lastly, as indicated by the strength of the association of structural protein features with substitution rates, the process of protein evolution—both across proteins and among sites within a protein—is underpinned by physical constraints. Yet little previous effort has been devoted to characterizing the impact of protein structure on the evolutionary response to destabilizing substitutions. The investigation of the relationship between response to destabilizing substitutions and protein structure presented in Chapter 5 reveals that both solvent accessibility and secondary structure lead to differences in recovery times. Destabilizations at buried sites, towards the core of the protein, required more compensatory substitutions to restore stability compared to destabilizations at surface sites. Similarly, destabilizations at sites within β-sheets required longer recovery times than destabilizations at turn sites. Overall, this work underscores the importance of protein structure in governing response dynamics.

Protein stability is but one of many constraints guiding the evolutionary process. Nevertheless, in most proteins (intrinsically disordered proteins excepted) it is an essential constraint that must be satisfied prior to proper biological functions. Taken together, the results presented in this dissertation highlight that stability requirements place a major constraint on the evolutionary process. The proteins investigated herein perform various biological functions (isomerase, protease, kinase, and spike proteins) that undoubtedly place additional constraints on their sequence evolution. Integrating both structural and functional constraints, see for example Echave (2019), will lead to improved models and may reveal further intricacies regarding the process of evolution.

BIBLIOGRAPHY

Agor, J. K. and Özaltın, O. Y. 2018. Vaccine Strain Selection. *Hum Vaccines Immunother*, 14(3): 678–683.

Ashenberg, O., Gong, L. I., and Bloom, J. D. 2013. Mutational effects on stability are largely conserved during protein evolution. *Proc Natl Acad Sci*, 110(52): 21071–21076.

Bastolla, U., Dehouck, Y., and Echave, J. 2017. What evolution tells us about protein physics, and protein physics tells us about evolution. *Curr Opin Struct Biol*, 42: 59–66.

Bateson, W. and Mendel, G. 1909. *Mendel's Principles of Heredity: A Defence, with a Translation of Mendel's Original Papers on Hybridisation*. Cambridge Library Collection - Darwin, Evolution and Genetics. Cambridge University Press.

Bazykin, G. A. 2015. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. *Biol Lett*, 11: 1–7.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D. S. 2006. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121): 929–932.

Bershtein, S., Serohijos, A. W., and Shakhnovich, E. I. 2017. Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. *Curr Opin Struct Biol*, 42: 31–40.

Betzel, C., Singh, T., Visanji, M., Peters, K., Fittkau, S., Saenger, W., and Wilson, K. 1993. Structure of the complex of proteinase k with a substrate analogue hexapeptide inhibitor at 2.2-a resolution. *J Biol Chem*, 268: 15854–15858.

Bloom, J. D. 2014a. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol Biol Evol*, 31(8): 1956–1978.

Bloom, J. D. 2014b. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol Biol Evol*, 31(10): 2753–2769.

Bloom, J. D. 2015. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinform*, 16: 168.

Bloom, J. D., Drummond, D. A., Arnold, F. H., and Wilke, C. O. 2006. Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol*, 23(9): 1751–1761.

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. 2012. Epistasis as the primary factor in molecular evolution. *Nature*, 490: 535–538.

Chan, Y. H., Venev, S. V., Zeldovich, K. B., and Matthews, C. R. 2017. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nat Comm*, 8: 14614.

Cherry, J. L. 1998. Should we expect substitution rate to depend on population size? *Genetics*, 150: 911–919.

Darwin, C. 1859. *On the Origin of Species By Means of Natural Selection*. John Murray, London.

Darwin, C. 1958. *The Autobiography of Charles Darwin, 1809-1882 Edited, with Appendix and notes, by his granddaughter Nora Barlow*. Collins, London.

de la Paz, J. A., Nartey, C., Yuvaraj, M., and Morcos, F. 2020. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc Natl Acad Sci*, 117(11): 5873–82.

De Vries, H. 1919. The present position of the mutation theory. *Nature*, 104(2610): 213–214.

DePristo, M. A., Weinreich, D. M., and Hartl, D. L. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet*, 6(9: 678–687.

dos Reis, M. 2015. How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the fisher–wright mutation–selection framework. *Biol Lett*, 11: 20141031.

Doud, M. B. and Bloom, J. D. 2016. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses*, 8(6): 155.

Doud, M. B., Ashenberg, O., and Bloom, J. D. 2015. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol Biol Evol*, 32: 2944–2960.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci*, 102: 14338–14343.

Duret, L. 2008. Neutral theory: The null hypothesis of molecular evolution. *Nat Education*, 1(1): 218.

Echave, J. 2019. Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity. *Mol Biol Evol*, 36(3): 613–620.

Echave, J., Jackson, E. L., and Wilke, C. O. 2015. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Phys Biol*, 12: 025002.

Echave, J., Spielman, S., and Wilke, C. O. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*, 17: 109–121.

Emlaw, J. R., Burkett, K. M., and Dacosta, C. J. 2020. Contingency between Historical Substitutions in the Acetylcholine Receptor Pore. *ACS Chem Neurosci*, 11(18): 2861–2868.

England, J. L. and Shakhnovich, E. I. 2003. Structural Determinant of Protein Designability. *Phys Rev Lett*, 90(21): 4.

Ferrada, E. 2019. The site-specific amino acid preferences of homologous proteins depend on sequence divergence. *Genome Biol Evol*, 11(1): 121–135.

Fisher, R. A. 1922. On the dominance ratio. *Proc R Soc Edinb*, 42: 321–341.

Fisher, R. A. 1958. *The genetical theory of natural selection.* Dover Publications, New York, second edition.

Fowler, D. M. and Fields, S. 2014. Deep mutational scanning: A new style of protein science. *Nat Methods*, 11(8): 801–807.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol*, 18(5): 866–873.

Gavrilets, S. 2003. *Evolution and speciation in a hyperspace: the roles of neutrality, selection, mutation, and random drift*. Oxford University Press, New York.

Gavrilets, S. 2004. *Fitness Landscapes and the Origin of Species (MPB-41)*. Princeton University Press, Oxfrod.

Gelbart, M. and Stern, A. 2020. Site-specific evolutionary rate shifts in hiv-1 and siv. *Viruses*, 12: 1312.

Gillespie, D. 1977. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*, 81(25): 2340–2361.

Goldman, N. and Yang, Z. H. 1994. Codon-based model of nucleotide substitution for protein-coding dna-sequences. *Mol Biol Evol*, 11: 725–736.

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

Goldstein, R. A. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins*, 79: 1396–1407.

Goldstein, R. A. 2013. Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability. *Genome Biol Evol*, 5: 1584–1593.

Goldstein, R. A. and Pollock, D. D. 2016. The tangled bank of amino acids. *Protein Sci*, 25: 1354–1362.

Goldstein, R. A. and Pollock, D. D. 2017. Sequence entropy of folding and the absolute rate of amino acid substitutions. *Nat Ecol Evol*, 1: 1923–1930.

Goldstein, R. A., Pollard, S. T., Shah, S. D., and Pollock, D. D. 2015. Nonadaptive amino acid convergence rates decrease over time. *Mol Biol Evol*, 32(6): 1373–1381.

Gong, L. I., Suchard, M. A., and Bloom, J. D. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2: e00631.

Gould, S. J. 1991. *Wonderful life: The Burgess shale and the nature of history*. WW. Norton and Company, New York.

Govindarajan, S. and Goldstein, R. A. 1996. Why are some protein structures so common? *Proc Natl Acad Sci*, 93: 3341–3345.

Guerois, R., Nielsen, J. E., and Serrano, L. 2002. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J Mol Biol*, 320: 369–387.

Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci*, 101: 12957–12962.

Guo, H. H., Choe, J., and Loeb, L. A. 2004. Protein tolerance to random amino acid change. *Proc Natl Acad Sci*, 101(25): 9205–9210.

Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J., and Bloom, J. D. 2018. Mapping mutational effects along the evolutionary landscape of hiv envelope. *eLife*, 7: e34420.

Haldane, J. B. 1927. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Math Proc Camb Philos Soc*, 23(7): 838–844.

Halpern, A. L. and Bruno, W. J. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*, 15: 910–917.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, 22: 160–17.

Hecht, M., Bromberg, Y., and Rost, B. 2016. Better prediction of functional effects for sequence variants From VarI-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease. *BMC Genomics*, 16(Suppl 8): 1–12.

Hietpas, R. T., Jensen, J. D., and Bolon, D. N. 2011. Experimental illumination of a fitness landscape. *Proc Natl Acad Sci*, 108(19): 7896–7901.

Hietpas, R. T., Bank, C., Jensen, J. D., and Bolon, D. N. 2013. Shifting fitness landscapes in response to altered environments. *Evolution*, 67(12): 3512–3522.

Hilton, S. K. and Bloom, J. D. 2018. Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence. *Virus Evol*, 4(2): vey033.

Hochberg, G. K., Liu, Y., Marklund, E. G., Metzger, B. P., Laganowsky, A., and Thornton, J. W. 2020. A hydrophobic ratchet entrenches molecular complexes. *Nature*, 588(7838): 503–508.

Huxley, J. 1942. *Evolution the Modern Synthesis*. Harper and Brothers, New York; London.

Jackson, S. 1998. How do small single-domain proteins fold? *Fold Des*, 3: 81–91.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2017. Shifting balance on a static mutation-selection landscape: a novel scenario of positive selection. *Mol Biol Evol*, 34: 391–407.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2018. Phenomenological load on model parameters can lead to false biological conclusions. *Mol Biol Evol*, 35: 1473–1488.

Jones, C. T., Youssef, N., Susko, E., and Bielawski, J. P. 2020. A phenotype-genotype codon model for detecting adaptive evolution. *Syst Biol*, 0: 1–17.

Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. In M. H. N, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York.

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

Kaffe-Abramovich, T. and Unger, R. 1998. A simple model for evolution of proteins towards the global minimum of free energy. *Fold and Des*, 3(5): 389–399.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. 2017. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nat methods*, 14: 587–589.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. *Genetics*, 47: 713–719.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature*, 217(5129): 624–626.

Kimura, M. 1991. The neutral theory of molecular evolution: A review of recent evidence. *Jpn J Genet*, 66: 367–386.

Kimura, M. and Ohta, T. 1974. On some principles governing molecular evolution. *Proc Natl Acad Sci*, pages 2848–2852.

Kolaczkowski, B. and Thornton, J. W. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol Biol Evol*, 25(6): 1054–1066.

Kolaczkowski, B. and Thornton, J. W. 2009. Long-branch attraction bias and inconsistency in bayesian phylogenetics. *PLoS ONE*, 4(12): e7891.

Kondrashov, A. S., Sunyaev, S., and Kondrashov, F. A. 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci*, 99(23): 14878–14883.

Kosakovsky Pond, S. L. and Frost, S. D. W. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*, 22: 1208–1222.

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D. W., Delport, W., and Scheffler, K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*, 28: 3033–3043.

Le, S. Q., Gascuel, O., and Lartillot, N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24: 2317–2323.

Lee, J. M., Huddleston, J., Doud, M. B., Hooper, K. A., Wu, N. C., Bedford, T., and Bloom, J. D. 2018. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proc Natl Acad Sci*, 115(35): E8276–E8285.

Letunic, I. and Bork, P. 2021. Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*, 1: 1–4.

Lewontin, R. 1970. The units of selection. *Annu Rev Ecol Evol Syst*, pages 1–18.

Lindqvist, Y., Johansson, E., Kaija, H., Vihko, P., and Schneider, G. 1999. Three-dimensional structure of a mammalian purple acid phosphatase at 2.2 a resolution with a mu-(hydr)oxo bridged di-iron center. *J Mol Biol*, 291: 135–147.

Livesey, B. J. and Marsh, J. A. 2020. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol Syst Biol*, 16(7): e9380.

Lu, A. and Guindon, S. 2014. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. *Mol Biol Evol*, 31(2): 484–495.

Lunzer, M., Brian Golding, G., and Dean, A. M. 2010. Pervasive cryptic epistasis in molecular evolution. *PLoS Genetics*, 6(10): 1–10.

Magee, A. F., Hilton, S. K., and DeWitt, W. S. 2020. Robustness of phylogenetic inference to model misspecification caused by pairwise epistasis. *bioRxiv*, page 2020.11.17.387365.

Marcos, M. L. and Echave, J. 2015. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ*, 3: e911.

Massingham, T. and Goldman, N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169(3): 1753–62.

Maynard Smith, J. 1970. Natural selection and the concept of a protein space. *Nature*, 225(5232): 563–564.

Mayr, E. and Provine, W. B. 1981. The Evolutionary Synthesis. *Bull AAAS*, 34(8): 17–32.

McCandlish, D. M., Shah, P., and Plotkin, J. B. 2016. Epistasis and the dynamics of reversion in molecular evolution. *Genetics*, 203(3): 1335–1351.

Mcgee, M. D. and Wainwright, P. C. 2013. Convergent evolution as a generator of phenotypic diversity in threespine stickleback. *Evolution*, 67(4): 1204–1208.

Mendes, F. K., Hahn, Y., and Hahn, M. W. 2016. Gene tree discordance can generate patterns of diminishing convergence over time. *Mol Biol Evol*, 33(12): 3299–3307.

Meyes, S. and vonHaeseler, A. 2003. Identifying site-specific substitution rates. *Mol Biol Evol*, 20(2): 182–189.

Miller, S. R. 2017. An appraisal of the enzyme stability-activity trade-off. *Evolution*, 71(7): 1876–1887.

Minh, B. Q., Nguyen, M. A., and von Haeseler, A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*, 30: 1188–1195.

Miyazawa, S. and Jernigan, R. 1985. Estimation of effective interresidue con- tact energies from protein crystal structures: quasi-chemical approximation. *Macromol*, 18: 534–552.

Mukherjee, S. 2016. *The Gene: An Intimate History*. Scribner, New York.

Muller, H. J. 1918. Genetic Variability, Twin Hybrids and Constant Hybrids, in a Case of Balanced Lethal Factors. *Genetics*, 3(5): 422–499.

Muller, H. J. 1939. Reversibility in Evolution Considered From the Standpoint of Genetics. *Biol Rev*, 14(3): 261–280.

Murrell, B., Wertheim, J., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*, 8: e1002764.

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D. P., Smith, D. M., Scheffler, K., and Kosakovsky Pond, S. L. 2015. Gene-wide identification of episodic selection. *Mol Biol Evol*, 32: 1365–1371.

Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol Biol Evol*, 11: 715–724.

Nagatani, R. A., Gonzalez, A., Shoichet, B. K., Brinen, L. S., and Babbitt, P. C. 2007. Stability for function trade-offs in the enolase superfamily "catalytic module". *Biochemistry*, 46: 6688–6695.

Nasrallah, C. A., Mathews, D. H., and Huelsenbeck, J. P. 2011. Quantifying the impact of dependent evolution among sites in phylogenetic inference. *Syst Biol*, 60(1): 60–73.

Naumenko, S. A., Kondrashov, A. S., and Bazykin, G. A. 2012. Fitness conferred by replaced amino acids declines with time. *Biol Lett*, 8(5): 825–828.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. 2014. Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, 32: 268–274.

Nisthal, A., Wang, C. Y., Ary, M. L., and Mayo, S. L. 2019. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc Natl Acad Sci*, 116(33): 16367–16377.

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., and Thornton, J. W. 2007. Crystal structure of an ancient protein: Evolution by conformational epistasis. *Science*, 317: 1544–1548.

Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E., and Rossiter, S. J. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, 502(7470): 228–231.

Pollock, D. D. and Goldstein, R. A. 2014. Strong evidence for protein epistasis, weak evidence against it. *Proc Natl Acad Sci*, 111(15): 2014.

Pollock, D. D., Thiltgen, G., and Goldstein, R. A. 2012. Amino acid coevolution induces an evolutionary Stokes shift. *Proc Natl Acad Sci*, pages E1352–E1359.

Popova, A. V., Safina, K. R., Ptushenko, V. V., Stolyarova, A. V., Favorov, A. V., Neverov, A. D., and Bazykin, G. A. 2019. Allele-specific nonstationarity in evolution of influenza a virus surface proteins. *Proc Natl Acad Sci*, 116: 21104–21112.

Potapov, V., Cohen, M., and Schreiber, G. 2009. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel*, 22(9): 553–560.

Quang, L. S., Gascuel, O., and Lartillot, N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24: 2317–2323.

Raimondi, D., Tanyalcin, I., FertCrossed, J. S. D., Gazzo, A., Orlando, G., Lenaerts, T., Rooman, M., and Vranken, W. 2017. DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res*, 45(W1): W201–W206.

Ren, J., Zhao, Y., Fry, E., and Stuart, D. I. 2018. Target identification and mode of action of four chemically divergent drugs against ebolavirus infection. *J Med Chem*, 61: 724–733.

Richmond, M. L. 2001. Women in the early history of genetics. William Bateson and the Newnham College Mendelians, 1900-1910. *Isis; an international review devoted to the history of science and its cultural influences*, 92: 55–90.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods*, 15(10): 816–822.

Risso, V. A., Manssour-Triedo, F., Delgado-Delgado, A., Arco, R., Barroso-delJesus, A., Ingles-Prieto, A., Godoy-Ruiz, R., Gavira, J. A., Gaucher, E. A., Ibarra-Molero, B., and Sanchez-Ruiz, J. M. 2015. Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol Biol Evol*, 32: 440–455.

Rodrigue, N. and Lartillot, N. 2014. Site-heterogeneous mutation-selection models within phylobayes-mpi package. *Bioinformatics*, 30: 1020–1021.

Rodrigue, N. and Lartillot, N. 2017. Detecting adaptation in protein-coding genes using a bayesian site-heterogeneous mutation-selection codon substitution model. *Mol Biol Evol*, 34(1): 204–214.

Rodrigue, N., Philippe, H., and Lartillot, N. 2010. Mutation-selection models for coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci*, 107: 4629–4634.

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. 2004. Protein structure prediction using rosetta. *Methods in enzymol*, 383: 66–93.

Rokas, A. and Carroll, S. 2008. Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol*, 25: 1943–1953.

Sailer, Z. R. and Harms, M. J. 2017. Molecular ensembles make evolution unpredictable. *Proc Natl Acad Sci*, 114(45): 11938–11943.

Schreiber, G., Buckle, A. M., and Fersht, A. R. 1994. Stability and function: two constraints in the evolution of barstar and other proteins. *Structure*, 2: 945–951.

Shah, P., McCandlish, D. M., and Plotkin, J. B. 2015. Contingency and entrenchment in protein evolution under purifying selection. *Proc Natl Acad Sci*, 112(25): E3226–E3235.

Shahmoradi, A., Sydykova, D. K., Spielman, S. J., Jackson, E. L., Dawson, E. T., Meyer, A. G., and Wilke, C. O. 2014. Predicting evolutionary site variability from structure in viral proteins: Buriedness, packing, flexibility, and design. *J Mol Evol*, 70: 130–142.

Shakhnovich, E. I. 1998. Protein design: a perspective from simple tractable models. *Fold and Des*, 3: R45–R58.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Sys Biol*, 7: 539.

Spielman, S. J. and Wilke, C. O. 2015. The relationship between dn/ds and scaled selection coefficients. *Mol Biol Evol*, 32: 1097–1108.

Starr, T. N. and Thornton, J. W. 2016. Epistasis in protein evolution. *Protein Sci*, 25: 1204—1218.

Starr, T. N., Flynn, J. N., Mishra, P., Bolon, D. N. A., and Thornton, J. W. 2018. Pervasive contingency and entrenchment in a billion years of hsp90 evolution. *Proc Natl Acad Sci*, 115(17): 4453–4458.

Steinberg, B. and Ostermeier, M. 2016. Environmental changes bridge evolutionary valleys. *Sci Adv*, 2: e1500921.

Stern, D. L. 2013. The genetic causes of convergent evolution. *Nat Rev Genet*, 14(11): 751–764.

Stoltzfus, A. and Cable, K. 2014. Mendelian-mutationism: The forgotten evolutionary synthesis. *J Hist Biol*, 47: 501–546.

Stolyarova, A., Nabieva, E., Ptushenko, V., Favorov, A. V., Popova, A., Neverov, A., and Bazykin, G. 2020. Senescence and entrenchment in evolution of amino acid sites. *Nat Commun*, 11: 4603.

Susko, E. and Roger, A. J. 2007. On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol*, 24(9): 2139–2150.

Sydykova, D. K., Jack, B. R., Spielman, S. J., and Wilke, C. O. 2018. Measuring evolutionary rates of proteins in a structural context. *F1000Research*, 6: 1845.

Szep, S., Park, S., Boder, E., Van Duyne, G., and Saven, J. G. 2009. Structural coupling between fkbp12 and buried water. *Proteins*, 74: 603–611.

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. 2012. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190: 1101–1115.

Tamuri, A. U., Goldman, N., and dos Reis, M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics*, 197: 257–271.

Taverna, D. M. and Goldstein, R. A. 2002. Why are proteins marginally stable? *Proteins*, 46: 105–109.

Teufel, A. I., Ritchie, A. M., Wilke, C. O., and Liberles, D. A. 2018. Using the mutation-selection framework to characterize selection on protein sequences. *Genes*, 9: 409.

Thomas, G. W. and Hahn, M. W. 2015. Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals. *Mol Biol Evol*, 32(5): 1232–1236.

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., and Wilke, C. O. 2013. Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE*, 8(11): e80635.

Tokuriki, N. and Tawfik, D. S. 2009. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*, 19(5): 596–604.

Tokuriki, N., Stricher, F., Serrano, L., and Tawfik, D. S. 2008. How protein stability and new functions trade off. *PLoS Comp Biol*, 4(2): 35–37.

Usmanova, D. R., Ferretti, L., Povolotskaya, I. S., Vlasov, P. K., and Kondrashov, F. A. 2015. A model of substitution trajectories in sequence space and long-term protein evolution. *Mol Biol Evol*, 32(2): 542–554.

Van Valen, L. 1973. A new evolutionary law. *Evol theory*, 1: 1–30.

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

Wang, X., Minasov, G., and Shoichet, B. K. 2002. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J Mol Biol*, 320: 85–95.

Whelan, S. 2008. The genetic code can cause systematic bias in simple phylogenetic models. *Philos Trans R Soc Lond B Biol Sci*, 363(1512): 4003–4011.

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

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilke, C. O. 2020. *ggridges: Ridgeline Plots in 'ggplot2'*. R package version 0.5.2.

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

Williams, P. D., Pollock, D. D., and Goldstein, R. A. 2006b. Functionality and the Evolution of Marginal Stability in Proteins: Inferences from Lattice Simulations. *Evol Bioinform*, 2: 91–101.

Wolynes, P. G. 1996. Symmetry and the energy landscapes of biomolecules. *Proc. Natl. Acad. Sci. USA*, 93: 14249–14255.

Wright, S. 1931. Evolution in mendelian populations. *Bull Math Biol*, 52(1-2): 241–295.

Wright, S. 1932. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*.

Wylie, C. S. and Shakhnovich, E. I. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc Natl Acad Sci*, 108(24): 9916–9921.

Yang, Z., Wong, W. S., and Nielsen, R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*, 22(4): 1107–1118.

Yang, Z. H. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19: 908–917.

Yang, Z. H. and Nielsen, R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, 25: 568–579.

Yang, Z. H., Nielsen, R., Goldman, N., and Pedersen, A. M. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155: 431–449.

Yeh, S.-W., Liu, J.-W., Yu, S.-H., Shih, C.-H., Hwang, J.-K., and Echave, J. 2014. Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. *Mol Biol Evol*, 31: 135–139.

Youssef, N., Susko, E., and Bielawski, J. P. 2020. Consequences of stability-induced epistasis for substitution rates. *Mol Biol Evol*, 37(11): 3131–3148.

Zhang, J., Nielsen, R., and Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, 22: 2472–2479.

Zhou, T., Drummond, D. A., and Wilke, C. O. 2008. Contact density affects protein evolutionary rate from bacteria to animals. *J Mol Biol*, 66(4): 395–404.

Zou, Z. and Zhang, J. 2015a. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol Biol Evol*, 32(8): 2085–2096.

Zou, Z. and Zhang, J. 2015b. No genome-wide protein sequence convergence for echolocation. *Mol Biol Evol*, 32(5): 1237–1241.