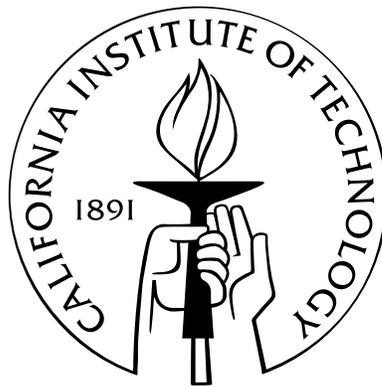


Hidden Dimensions in Protein Evolution: Stability, Mutational Robustness, and Evolvability

Thesis by
Jesse D. Bloom

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy



California Institute of Technology
Pasadena, California

2007
(Defended May 17, 2007)

© 2007

Jesse D. Bloom

All Rights Reserved

Acknowledgements

Graduate school has been a tremendously intellectually exciting time. At the top of the list of reasons why is my advisor, Frances Arnold. If you had told me five years ago that I would have spent my time in graduate school doing research in a Chemical Engineering lab, I would have thought you were crazy! But Frances has integrated the study of proteins, evolution, and engineering together into the fascinating area of evolutionary protein engineering, and I feel lucky to have stumbled into a lab pursuing research in such a fertile area. Frances has encouraged my theoretical/computational bent while reminding me to stay grounded in experiments, and she has encouraged my interests in some of the more obscure areas of evolution while keeping me in touch with the practical concerns of engineering. Above all, she has been tremendously encouraging, and has pushed me to extend myself into new areas. Because Frances is constantly challenging me to defend my ideas and think about what I am doing, I will be extremely well prepared to eventually make the transition to being a professor myself. When I do, one of my goals will be to emulate her skills as a manager and mentor.

The second person who has profoundly shaped my graduate path is Christoph Adami. When I arrived at Caltech, I knew almost nothing about evolution. It was Chris's Evolution and Biocomplexity class that opened my eyes to this entire field. Chris has some of the most creative and (I hope he won't mind me saying this) crazy ideas in science, and it is great to talk to someone who is constantly bridging entire scientific disciplines. The last few years at Caltech I have drifted out of Chris's lab and into the wet lab, but he remains a huge influence. He is one of the friendliest people I know, and it is wonderful to always be able to stop by his lab and chat for a few hours.

Claus Wilke has provided much of the "theoretical side" of my graduate education. Claus is an incredibly careful and methodical scientist, and he is always willing to take the

time to go over things carefully with me. He is really the one who has taught me how to do theory, walking me through the process of transferring an idea into a mathematical equation. Like Chris, Claus is a biologist who is an expert in areas that most biologists know little about, and I am grateful to have picked up a little of his expertise. In general I have been spoiled by Claus, because any time I have a question I simply send him an e-mail and back comes a carefully reasoned answer. I hope that this situation might continue for decades, as I'm sure that Claus will remain both a collaborator and a mentor.

Claus was responsible for introducing me to Alpan Raval, who might as well be a wizard when it comes to math. I still remember the first time that I drove to Claremont to talk to Alpan about a mathematical problem that Claus and I had been puzzling over for days. Alpan went to the blackboard and essentially solved the problem in real time. Since then I've gotten the opportunity to work with Alpan on several more problems, and it is always the same story: He comes up with elegant approaches that never even would have occurred to me. The selfish side of me hopes that he at least stays at Keck for a few more years so I can keep stopping by to have him solve my math problems.

I also owe a debt of gratitude to my undergraduate advisor Susan Lindquist and her postdoc Gunter Kramer. They helped me get started on a project related to role of molecular chaperones in protein evolution, which I still hope to finish at some point. They have been patient in explaining to me the intricacies of a new field. In addition, Sue remains a valuable mentor who has been very supportive. She is a phenomenally creative scientist, and I am still just beginning to appreciate the significance of some of the things I learned as an undergraduate in her lab many years ago.

I would also like to thank the other members of my thesis committee, Steve Mayo, Doug Rees, and Paul Sternberg. Steve has given me a number of pieces of important advice, and even let me mess around with the ORBIT program. Doug has a very deep perspective on protein thermodynamics, and I've benefited a lot from my conversations with him.

One of the most enjoyable parts of graduate school has been getting to work with my team of undergraduates, Zhongyi Lu, Matt Glassman, David Chen, Sy Labthavikul, and Ophelia Venturelli. In particular, Zhongyi has done work more befitting of a graduate student; I never could have finished the epic P450 evolution experiments (Chapter 5) without his diligence. Matt has been very stimulating to work with, and is going to have a very

bright future in science since he is already coming up with outstanding original ideas. David and Sy both did very productive SURF projects which directly helped with publications, and Ophelia was upbeat, enthusiastic, and creative; she will definitely prosper in her own PhD at Caltech.

People in the Arnold lab have been great coworkers. At the top of the list is Chris Otey, who spent countless hours sharing his accumulated knowledge about how to do P450 experiments. Chris's generosity with his time was essential in getting my work up and running. Matt Peters also played a huge role in getting me started on my P450 work, and a decade from now when he is CEO of the next Chevron (will he be a Republican then?), I will tell people that we used to spend hours rating the merits of the 2004 Democratic presidential candidates. Chris Snow is a relatively recent addition to the Arnold lab, but he has already taught me a lot about computer modeling and helped me with some pretty slick molecular graphics. Allan Drummond and I led somewhat parallel careers in the Adami and Arnold labs, and we had countless great discussion and brainstorming sessions. His ideas and input were always stimulating. Philip Romero spent his first quarter working with me and helping with ideas that turned into Chapter 6 of this thesis. And Marco Landwehr was an entertaining and helpful office-mate who (as he will always remind me) synthesized all of that 12-para-nitrophenoxydodecanoic acid.

There are lots of other people in the Arnold lab who helped me at various times, among them Jorge Rodriguez, Rudi Fasan, Mike Chen, Jeff Endelman, Joff Silberg, Michelle Meyer, Marco Landwehr, Andrew Sawayama, Geethani Bandara, and Cheryl Nakashima. I am grateful to all of these people and everyone else in the Arnold lab for simply making my day-to-day existence so enjoyable.

I also had lots of important scientific interactions with graduate students from other labs. In particular, Christina Vizcarra and Eric Zollars from the Mayo lab were always fun to chat with. Titus Brown gave me lots of good advice, the best of which was to take the time to learn Python.

In addition to the science, my years in graduate school have been exciting (sometimes a bit much so...) for many other reasons. My roommates Alex Gagnon, Ryan Ogliore, and Kenji Sasaki were always there to laugh, talk, encourage, or (in the case of Alex and Kenji) just ask to borrow my car. Overall, we carved out a pretty nice niche for ourselves up in

Sierra Madre, and made Los Angeles a very livable place. This livability was helped by many other friends, especially Dan Feldman, Pete Huskey, Mary Laura Lind, Daven Henze, Amy Eastwood, Alan Hampton, Janie Shelton, and Philip Gable. I'd also like to thank my running coach Evae Silva and all of Team Run With Us for giving me something other than research to think about, but also being understanding when the commitments of finishing my PhD caused a (hopefully temporary) decline in my appearances at practice.

My parents have of course been extremely supportive, and the older I get the more I appreciate them. My dad has been very encouraging, and has taken an active interest in my recent research and given me some excellent advice. I may yet end up being a bit of a virologist. Meanwhile, even as I have researched its obscure details, my mom has fought to preserve the teaching of evolution itself (and so much more) in the Bitterroot Valley high schools. I've tried to make the Introduction to this thesis intelligible for her. I'm sure that this wasn't necessary for my brother Seth, who I'm still hoping will get me a plaque like the one on Andrew's desk: "Dr. Jesse D. Bloom, smarter than most, but not his brother."

Finally, I would like to thank my girlfriend Sarah Thomas. We met almost exactly a year before I'm due to defend my thesis, and the time since then has been my best in Los Angeles. She has been supportive even as I spent a good chunk of the last year running around the country doing interviews. Because she works way out at the Getty Villa, I've written much of this thesis in Mogan's Coffee Shop near Sunset and the Pacific Coast Highway. There I have learned that when people from Malibu ask, "are you working on a script?" they are referring to a movie, not an interpreted computer program. I imagine I will get to learn other such lessons, because Sarah has helped prompt me to stay in Los Angeles a few more years to do postdoctoral research at Caltech.

Abstract

Proteins are evolvable in the sense that they are readily able to acquire new or improved functions through the process of mutation and selection. Here, I examine what properties influence the ability of proteins to evolve new functions. I show that proteins with similar biochemical properties can differ substantially in their capacities to withstand mutations and evolve new functions. Specifically, more stable proteins are both more mutationally robust and more evolvable, due to improved tolerance for mutations. This fact can be exploited in protein engineering. I then show how evolutionary theory can be modified to describe how a protein's mutational robustness changes during the normal course of neutral genetic drift. One of the main theoretical predictions is that proteins evolving in larger populations will gain excess stability and mutational robustness, a prediction which I confirm experimentally. Finally, I turn to the question of how neutral genetic drift can alter "promiscuous" protein functions that are not under selection. I show that promiscuous functions can change significantly during genetic drift, a phenomenon that may aid in the evolution of beneficial new functions. Overall, this work establishes two mechanisms whereby initially neutral mutations can influence the course of future evolution.

Contents

Acknowledgements	iii
Abstract	vii
1 Introduction	1
2 Thermodynamic Prediction of Protein Neutrality	6
2.1 Abstract	6
2.2 Introduction	6
2.3 Results	7
2.4 Discussion	15
2.5 Methods	17
2.6 Acknowledgements	21
3 Protein Stability Promotes Evolvability	22
3.1 Abstract	22
3.2 Introduction	22
3.3 Results	23
3.4 Discussion	30
3.5 Methods	33
3.6 Acknowledgments	52
4 Thermodynamics of Neutral Protein Evolution	53
4.1 Abstract	53
4.2 Introduction	54
4.3 Results	56

4.4	Discussion	68
4.5	Materials and Methods	72
4.6	Appendix	74
4.7	Acknowledgements	76
5	Evolution Favors Mutational Robustness in Sufficiently Large Populations	77
5.1	Abstract	77
5.2	Background	78
5.3	Results and Discussion	79
5.4	Conclusions	87
5.5	Methods	89
5.6	Mathematical Appendix	99
5.7	Acknowledgments	115
6	Neutral Genetic Drift Can Aid Functional Protein Evolution	116
6.1	Abstract	116
6.2	Background	117
6.3	Results and Discussion	118
6.4	Conclusions	128
6.5	Methods	130
6.6	Acknowledgements	136
	Bibliography	137

Chapter 1

Introduction

Proteins are the molecular workhorses of biology, carrying out a tremendous range of essential biochemical functions. The existence of proteins that are exquisitely tuned to perform such diverse tasks is a testament to the creative power of natural evolution. It is also a source of admiration among bioengineers, who seek to mimic evolution by tailoring proteins for a wide variety of medical and industrial applications. An understanding of the processes by which proteins evolve to perform new functions is therefore of interest to both biologists and engineers.

One of the most fascinating overarching questions about protein evolution was posed over 40 years ago by the great chemist Linus Pauling and his postdoctoral fellow Emile Zuckerkandl in research they performed at the California Institute of Technology. Working at the time when it was first becoming feasible to obtain the amino acid sequences of proteins, Pauling and Zuckerkandl assembled the sequences of hemoglobin proteins from a range of different species. They compared the protein sequences with an eye towards determining the molecular changes that underpinned the evolutionary divergence of these species. Their analysis showed that the hemoglobin sequences had accumulated many mutations since each pair of species had diverged. But although it was already well known (in part from Pauling's earlier work on sickle-cell anemia [1, 2]) that even a single amino acid mutation could dramatically alter a protein's function, the number of accumulated mutations in hemoglobin seemed more reflective of the amount of elapsed evolutionary time than any measure of functional alteration. Summarizing their research, Pauling and Zuckerkandl wrote [3],

Perhaps the most important consideration is the following. There is no

reason to expect that the extent of functional change in a polypeptide chain is proportional to the number of amino acid substitutions in the chain. Many such substitutions may lead to relatively little functional change, whereas at other times the replacement of one single amino acid residue by another may lead to a radical functional change. *Of course, the two aspects are not unrelated, since the functional effect of a given single substitution will frequently depend on the presence or absence of a number of other substitutions.*

This last italicized sentence (emphasis added) highlights what is one of the most important issues both for understanding the natural evolution of proteins and for modifying these molecules in engineering applications. In the absence of a dependence of the effect of one mutation on the presence of other mutations, then the evolution or engineering of a new or improved protein property can be viewed as a simple hill-climbing exercise, with each successive beneficial mutation moving a protein further up the path towards some desired objective. But if the impact of a mutation depends upon whether other mutations are present, then the situation becomes much more complicated. In the particular case emphasized by Pauling and Zuckerkandl, whether a mutation is beneficial depends on the presence of other mutations that themselves have no substantial effect. This type of dependence means that evolutionary optimization cannot occur by natural selection for beneficial mutations alone, since selectively favored “uphill” steps may only be possible after several “sideways” steps caused by the random occurrence and spread of mutations that are not favored by selection. In the title of this thesis, selectively neutral mutations that can enable future beneficial mutations are described as causing changes in “hidden dimensions” in protein evolution. This phrase is a reference to the fact that the effects of these mutations are hidden to direct selection for protein function, but nonetheless have important effects that are revealed by later beneficial mutations.

The goal of this thesis is to elucidate some of these “hidden dimensions” by identifying properties which are affected by selectively neutral mutations in a way that enables later beneficial mutations. Little effort is made to develop a detailed description of the biophysical interactions within specific sets of such mutations, as would be done by a modern molecular modeling force field. Instead, the focus is on changes in protein properties that tend to broadly affect the impact of future mutations. Chapters 2 and 3 use a combination of

modeling and experiments to show that one such property is protein stability. The basic idea is that a protein's stability is under selection only insofar as it must achieve some minimal threshold in order to allow the protein to reliably fold and perform its biological function. A protein has substantial latitude to further increase its stability above the threshold, but provides no direct fitness benefit by doing so. However, if a protein does achieve extra stability beyond the minimal required threshold, it can then tolerate a subsequent destabilizing mutation without any adverse consequences. In these terms, the "interaction" between several mutations can be highly generic, since it is simply due to their cumulative effect on protein stability. The situation is analogous to hanging several small weights from a thread. If the thread is strong enough to support one weight, but snaps under the force of two, then in a sense the effect of one weight on the thread depends on its "interaction" with the other weight. But if one recognizes that the thread is simply responding to the cumulative downward force, then it is clear that the "interaction" between the weights can be understood simply in terms of the sum of the downward forces they impose. Chapters 2 and 3 show that this is an apt analogy for describing the effects of mutations on two different enzymes, TEM1 β -lactamase and cytochrome P450 BM3. In particular, they demonstrate that mutations that make a protein more stable improve its tolerance for subsequent mutations, in effect drawing a link between the biophysical property of protein stability and the evolutionary property of mutational robustness.

Chapter 3 shows that high stability also increases a protein's "evolvability," which is defined as the probability that a random mutation improves the protein's performance in one or more specified functions. Stability promotes evolvability by the same basic mechanism by which it increases mutational robustness. Mutations that improve some desired aspect of a protein's biochemical function are usually detrimental to its stability. High stability allows a protein to withstand the destabilizing effects of these functionally beneficial mutations without any adverse consequences, thereby increasing its overall evolvability. In Chapter 3 I show how an understanding of this hidden dimension of stability can be exploited to engineer cytochrome P450 proteins with new enzymatic activities.

Chapter 4 turns to the question of how protein stability will change during the normal course of evolution, when proteins are bombarded with mostly deleterious or neutral mutations. This chapter derives a number of mathematical results that augment the normal

equations of population genetics with detailed information about the specific biophysical constraints that govern protein evolution. Chapter 4 also predicts that the amount of stability and mutational robustness possessed by natural proteins should depend on the size of the population in which they evolve (specifically, on whether this population is mostly monomorphic or highly polymorphic). Chapter 5 confirms this prediction with a series of epic laboratory evolution experiments on cytochrome P450 proteins. This chapter provides the first experimental evidence of how a population's mutational robustness can depend on its size.

Finally, Chapter 6 turns to a slightly different mechanism by which selectively neutral mutations can aid in the evolution of proteins. In this chapter, I describe measurements performed on a set of proteins that have all evolved under the same constant selection criterion. I show that while the function of these proteins that was under selection has been preserved above the selection threshold, a number of other “promiscuous” functions have changed substantially. This provides a second mechanism by which selectively neutral mutations can aid in functional evolution: they can create diversity in properties that are not currently under selection, poising the proteins to readily evolve new functions should selection “ask new questions” at some point in the future.

Overall, this thesis establishes two important mechanisms by which selectively neutral mutations can aid in future functional evolution. In the first mechanism, a neutral mutation increases a protein's stability, thereby improving its tolerance for subsequent mutations, some of which may confer new or improved functions. In the second mechanism, neutral mutations enhance a promiscuous protein function, allowing the protein to more easily undergo adaptive evolution should a change in selection pressures make the promiscuous function beneficial at some point in the future. Both of these mechanisms couple selectively neutral and functionally important mutations, and so confirm Pauling and Zuckerkandl's contention that these two modes of sequence change are crucially linked during evolution.

Caveat about protein stability

The basic approach of viewing proteins as evolving under a stability constraint proves to be highly successful in explaining experimental results presented in this thesis. This view

is certainly a substantial improvement over any of the previous perspectives that have been taken in quantitatively analyzing the molecular evolution of proteins. But it is important to remember that this view is still a severe approximation of the many complex evolutionary requirements faced by proteins. In an evolutionary sense, proteins can “break” for a variety of reasons: they can lose the ability to remain stably folded, become stuck in kinetic folding traps, aggregate, or undergo proteolysis. In fact, the experimental measurements on P450 proteins in Chapters 3 and 5 actually are of kinetic rather than thermodynamic stability (since the denaturation is irreversible). The saving grace of the approach in this thesis is that the propensities of proteins to meet a mutational doom from many different factors are correlated to stability, since mutations that destabilize a protein also tend to decrease its kinetic stability [4], increase its tendency to aggregate [5], and increase the likelihood that it is proteolyzed [6]. Therefore, the term “protein stability” as it is used in the overall view presented in this thesis is probably actually serving as a proxy for a number of other properties. In an evolutionary analysis, where one is concerned with the average properties of many mutations, such imprecision is acceptable (indeed, it is not clear how to build a more precise picture). However, modern biophysics certainly allows for more detailed characterization of proteins. In fact, I suggest that a major arena for further work is a more exhaustive characterization of the specific biophysical problems that tend to be caused by the types of random mutations that proteins experience during their evolution.

Chapter 2

Thermodynamic Prediction of Protein Neutrality

A version of this chapter has been published as [7].

2.1 Abstract

We present a simple theory that uses thermodynamic parameters to predict the probability that a protein will retain function after one or more random amino acid substitutions. Our theory predicts that for large numbers of substitutions, the probability that a protein retains function will decline exponentially with the number of substitutions, with the severity of this decline determined by the protein's structure. Our theory also predicts that a protein can gain extra robustness to the first few substitutions by increasing its thermodynamic stability. We validate our theory with simulations on lattice protein models and by showing that it quantitatively predicts previously published experimental measurements on subtilisin and our own measurements on variants of TEM1 β -lactamase. Our work unifies observations about the clustering of functional proteins in sequence space, and provides a basis for interpreting the response of proteins to substitutions in protein engineering applications.

2.2 Introduction

The ability to predict a protein's tolerance to amino acid substitutions is of fundamental importance in understanding natural protein evolution, developing protein engineering strategies, and understanding the basis of genetic diseases. Computational and experimen-

tal studies have identified a bewildering array of factors that affect a protein’s tolerance to substitutions. Simulations have highlighted the contributions of both protein stability and structure by showing that more stable proteins have a higher fraction of folded mutants [8, 9, 10, 11], and that some structures are encoded by more sequences than others [12, 13, 14]. Experiments have demonstrated that proteins can be extremely tolerant of a single substitution; for example, 84% of single-residue mutants of T4 lysozyme [15] and 65% of single-residue mutants of *lac* repressor [16] mutants were scored as functional. For multiple substitutions, the fraction of functional proteins decreases roughly exponentially with the number of substitutions, although the severity of this decline varies among proteins [17, 18, 19]. Protein mutagenesis experiments have also underscored the contribution of protein stability to mutational tolerance by finding “global suppressor” substitutions that buffer a protein against otherwise deleterious substitutions by increasing its stability [20, 21].

We unify these diverse experimental and computational results into a simple framework for predicting a protein’s tolerance for substitutions. A fundamental measure of this tolerance is the fraction of proteins retaining biochemical function after a single random substitution, often called the neutrality [22]. We extend this concept to multiple substitutions by defining the m -neutrality as the fraction of functional proteins among all sequences that differ from the wildtype sequence at m residues. We show that a protein’s m -neutrality can be accurately predicted from measurable thermodynamic parameters, and that these predictions capture the contributions of both stability and structure to determining a protein’s tolerance for substitutions.

2.3 Results

2.3.1 Thermodynamic Framework for Predicting Neutrality

A protein’s m -neutrality is defined with respect to the retention of biochemical function, and so implicitly depends on the stringency of the particular process that selects for function. Nonetheless, we can unambiguously define an upper limit on a protein’s m -neutrality as the fraction of mutants that still fold to the wildtype structure because mutants that fail to fold will generally also fail to function. For many proteins, this upper limit should closely approximate the actual m -neutrality, since mutagenesis studies suggest that most

functionally disruptive substitutions disrupt the structure rather than specifically affecting functional residues [20, 23, 24].

The native structure is thermodynamically stable [25], with typical free energies of folding (ΔG_f) between -5 and -15 kcal/mol [26]. A mutant sequence will still fold to the wildtype structure so long as its stability in that structure meets some threshold. We call the extra stability of the native structure beyond this minimal threshold $\Delta G_f^{\text{extra}}$, and note that functional proteins always have $\Delta G_f^{\text{extra}} \leq 0$. A protein's m -neutrality is therefore the fraction of sequences with m substitutions that still meet the stability threshold.

A substitution causes a stability change of

$$\Delta\Delta G = \Delta G_f^{\text{mut}} - \Delta G_f^{\text{wt}}$$

where ΔG_f^{wt} and ΔG_f^{mut} are the wildtype and mutant protein stabilities. Substitutions tend to be destabilizing: although there are no large collections of $\Delta\Delta G$ measurements for truly random substitutions, in a likely biased collection of more than 2,000 measured $\Delta\Delta G$ values for single-residue substitutions [27], the mean is 0.9 kcal/mol and the values at the 10th and 90th percentiles are -1.0 and 3.2 .

The thermodynamic effects of most substitutions are independent and additive [28, 29], meaning that if the stability changes due to two different single substitutions are $\Delta\Delta G^a$ and $\Delta\Delta G^b$, then the stability change due to both substitutions is $\Delta\Delta G^a + \Delta\Delta G^b$. With this additivity assumption, if we know the probability distribution $p_1(\Delta\Delta G^1)$ that a single random substitution causes a stability change $\Delta\Delta G^1$, we can compute the probability $p_m(\Delta\Delta G^m)$ that m substitutions will cause a stability change of $\Delta\Delta G^m$.

Denote the m -neutrality as $P_f(m)$. When we approximate $P_f(m)$ as the fraction of proteins that continue to meet the stability threshold after m substitutions, we have

$$P_f(m) = \int_{-\infty}^{-\Delta G_f^{\text{extra}}} p_m(\Delta\Delta G^m) d(\Delta\Delta G^m).$$

This formula predicts a protein's m -neutrality from knowledge of its extra stability and the distribution of $\Delta\Delta G$ values for all possible single substitutions.

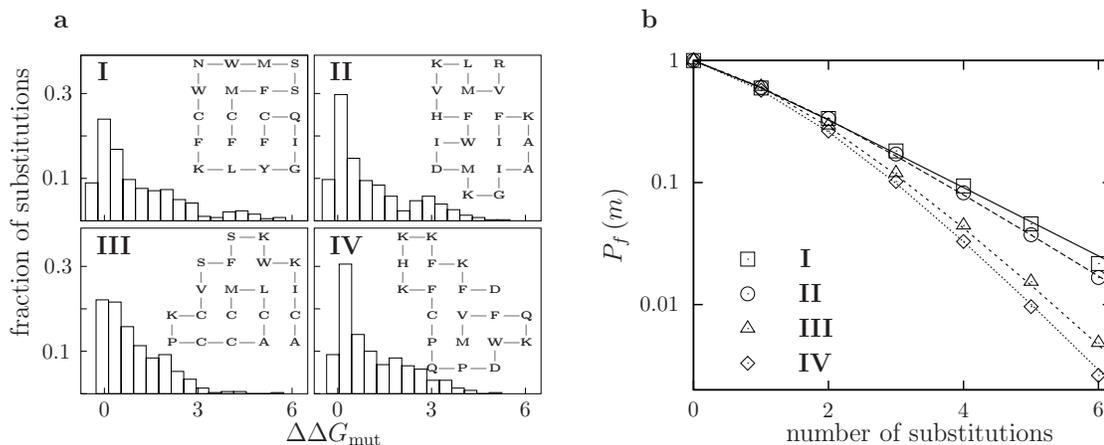


Figure 2.1: Lattice proteins with different structures but the same stability ($\Delta G_f = -1.0$) converge to different exponential declines in m -neutrality. (a) The distributions of $\Delta\Delta G$ for all 380 single amino acid substitutions to the inset lattice proteins. (b) The measured (symbols) and predicted (lines) m -neutralities for the four proteins. A sequence is functional if it folds to the original native structure with $\Delta G_f < 0.0$.

2.3.2 Lattice Proteins Support Predictions

We used lattice protein simulations to test our m -neutrality predictions. Lattice proteins are highly simplified models of proteins that provide a useful tool for studying protein folding [30, 31, 32, 33] and evolution [34, 35] (some example lattice proteins are shown in Figure 2.1). Like real proteins, our lattice proteins exist in a vast conformational space (4.2×10^7 possible conformations), yet we can exactly compute ΔG_f in just fractions of a second. We considered a mutant lattice protein to be functional if it still stably folded to the wildtype structure, and so these simulations directly tested our theory under the assumption that the retention of fold is equivalent to the retention of structure. We carried out our simulations by evolving lattice proteins with various stabilities and structures, and then computing their m -neutralities by sampling mutants with random amino acid substitutions.

Our theory accurately predicted the m -neutralities of all of the lattice proteins we tested. Lattice proteins with different structures have different m -neutralities, even when they have the same ΔG_f (Figure 2.1). The 1-neutralities of proteins with different structures and the same ΔG_f look similar, but for larger values of m some proteins clearly show higher m -neutralities than others. For large m , the m -neutralities of all of the proteins converge to a

simple exponential of the form

$$P_f(m) \propto \nu^m$$

where ν varies among proteins and gives the severity of the exponential decline of m -neutrality with m . The underlying reason for this exponential decline is clear: after several substitutions, the distribution of ΔG_f among the remaining functional sequences reaches a steady state, and so each new substitution pushes the same fraction of proteins beyond the stability threshold. Interestingly, although $P_f(m=1)$ is similar for all of the proteins in Figure 2.1, the factors that give rise to the different ν parameters are present in the distribution of single mutant $\Delta\Delta G$ values, since it is used to predict the m -neutralities for all values of m .

Figure 2.2 shows the m -neutralities of proteins with the same structure but different stabilities. After several substitutions, all of the proteins converge to the same value of ν , suggesting that ν is a generic property of a protein's structure and does not depend on its particular sequence or stability. On the other hand, the response of a protein to the first few substitutions depends strongly on its stability, with more stable proteins exhibiting higher initial m -neutrality. The high initial m -neutrality of stable proteins is readily rationalized in terms of the thermodynamic model: substitutions tend to disrupt a protein's structure by pushing its stability below the minimal threshold, but proteins with an extra stability cushion are buffered against the first few substitutions. Proteins that sit on the very margin of the minimal stability threshold exhibit lower 1-neutrality than is predicted by an exponential decline because these proteins are less stable than the average functional protein, and so surviving sequences will tend to be more stable than the wildtype sequence and so exhibit a higher tolerance for the next substitution.

2.3.3 Real Proteins Support Predictions

Our theory makes two main predictions: first, that the decline in m -neutrality can be predicted from the $\Delta\Delta G$ values for single amino acid substitutions, and second, that among proteins with the same structure, more stable variants will have higher m -neutralities. In order to test these predictions on real proteins, we needed a method for measuring the fractions of functional proteins at various levels of amino acid substitutions. Actually

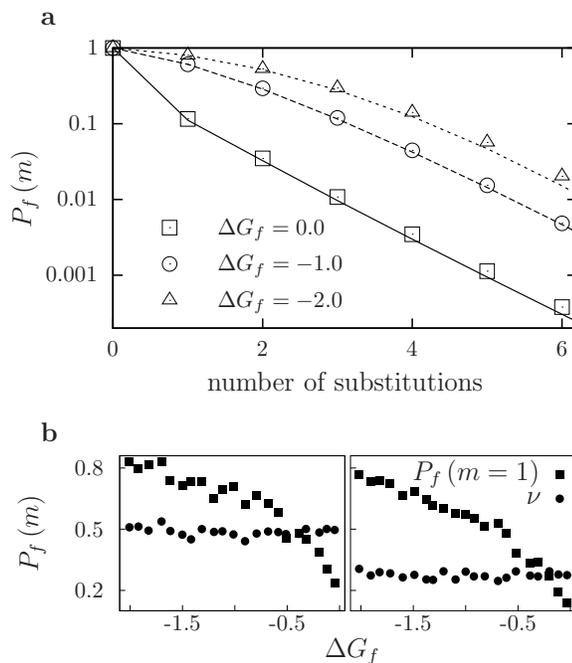


Figure 2.2: Lattice proteins with the same structure but different stabilities have different 1-neutralities but converge to the same exponential decline in m -neutrality. **(a)** Predicted (lines) and measured (symbols) m -neutralities for proteins with different stabilities and the same structure (**III** in Figure 2.1). **(b)** Measured values of the 1-neutralities (squares) and exponential m -neutrality decline parameter ν (circles) for proteins with different stabilities but the same structures (the plots at left and right are for structures **I** and **IV** from Figure 2.1 respectively). The sequences of the proteins with different stabilities are highly diverged, with average pairwise sequence identities of 15% and 41% for the structures at left and right respectively.

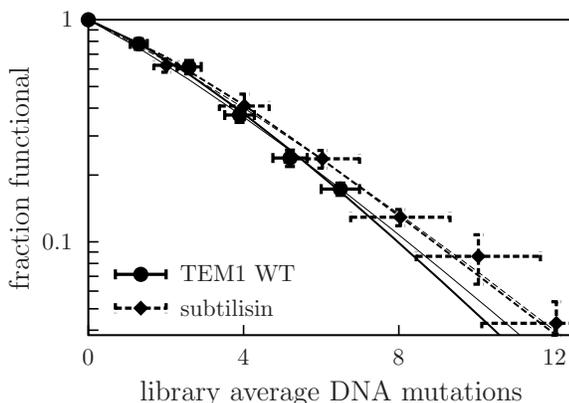


Figure 2.3: Theoretical predictions match measured neutralities in mutant libraries of subtilisin (dashed lines) and TEM1 β -lactamase (solid lines) genes. Thick lines show predictions made using the PoPMuSiC potential and thin lines show predictions made using the FOLDEF potential. The measurements are from Table 2.2, normalized by the values from the control unmutated library.

creating many sequences with specific numbers of amino acid substitutions is extremely difficult, but it is relatively easy to use error-prone PCR to make libraries of mutant genes with randomly distributed nucleotide mutations. The fraction of functional proteins \mathcal{F} encoded by the genes in a mutant gene library can easily be related to the distribution of $\Delta\Delta G$ values for single amino acid substitutions by

$$\mathcal{F} = \sum_{m_{\text{nt}}=0}^{\infty} f(m_{\text{nt}}) \times \hat{P}_f(m_{\text{nt}})$$

where $f(m_{\text{nt}})$ is the fraction of genes with m_{nt} nucleotide mutations in the mutant library and $\hat{P}_f(m_{\text{nt}})$ is the fraction of functional proteins encoded by genes with m_{nt} nucleotide mutations, computed as for $P_f(m)$ except that we now use $\hat{p}_1(\Delta\Delta G^1)$, the probability that a single nucleotide mutation made according to the particular protocol used causes a stability change of $\Delta\Delta G^1$.

We considered data from mutant gene libraries of subtilisin and variants of TEM1 β -lactamase. Shafikhani and coworkers [17] created libraries of mutagenized subtilisin genes using multiple rounds of error-prone PCR such that each round introduced the same average number of nucleotide mutations, and screened the resulting genes for their ability to encode active secreted proteases in *Bacillus subtilis*. We used a similar procedure to create mutant gene libraries of the wildtype and the thermostable M182T variant [36] of TEM1

Base pairs sequenced	22,800
Total mutations	172
Mutation frequency (%)	0.75 ± 0.06
Mutations per gene	6.5 ± 0.5
Mutation types (%)	
A \rightarrow T, T \rightarrow A	22
A \rightarrow C, T \rightarrow G	9
A \rightarrow G, T \rightarrow C	42
G \rightarrow A, C \rightarrow T	20
G \rightarrow C, C \rightarrow G	1
G \rightarrow T, C \rightarrow A	3
frameshift	3

Table 2.1: Mutation frequencies for TEM1 β -lactamase mutagenesis determined by sequencing 20 unselected clones each from the wildtype and M182T error-prone PCR round 5 libraries. Standard errors are calculated assuming Poisson counting statistics.

β -lactamase, and screened genes from these libraries in *Escherichia coli* for their ability to confer resistance to the antibiotic ampicillin. Table 2.1 shows the frequency of nucleotide mutations introduced by our mutagenesis protocol, while Table 2.2 shows how the fraction of genes in the library that conferred ampicillin resistance decreased with increasing average numbers of nucleotide mutations.

In order to test the ability of our theory to predict the fractions of functional proteins in these libraries, we also needed a method for calculating the $\Delta\Delta G$ values for single amino acid substitutions. We used two existing computational potentials, the database-derived PoPMuSiC potential of Gilis and Rooman [37, 38] and the empirical FOLDEF potential of Serrano and coworkers [39], and corrected for the fact that some amino acid substitutions are more likely than others. The only remaining unknown parameter is the extra stability of the protein, $\Delta G_f^{\text{extra}}$, which cannot be directly measured because we do not know the minimal stability required for the protein to function. However, $\Delta G_f^{\text{extra}}$ only influences the initial behavior of the m -neutrality curve and does not affect the exponential decline parameter ν (as shown in Figure 2.2), so we can fit $\Delta G_f^{\text{extra}}$ to the experimental measurements and still test the ability of the theory to predict the severity of the decline in m -neutrality.

Figure 2.3 shows that the measured fractions functional for subtilisin and wildtype TEM1 β -lactamase are in good agreement with the predictions made with the PoPMuSiC and FOLDEF potentials. Subtilisin and TEM1 β -lactamase exhibit similar declines in m -neutrality, which is expected because both have similar structures (the same CATH

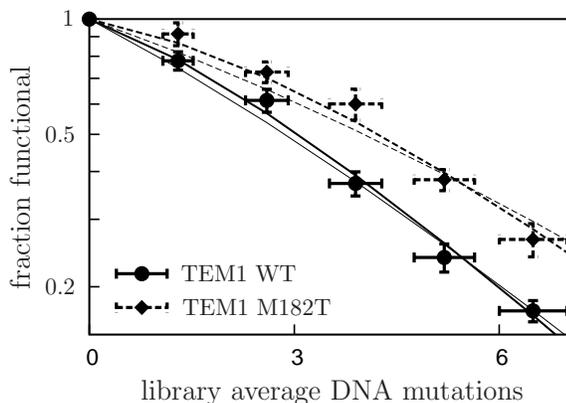


Figure 2.4: The more stable M182T variant of TEM1 β -lactamase (dashed lines) exhibits enhanced neutrality relative to wildtype (solid lines) as predicted by the theory. Thick lines show predictions made using the PoPMuSiC potential and thin lines show predictions made using the FOLDEF potential. The measurements are from Table 2.2, normalized by the values from the control unmutated library.

Round	$\langle m_{nt} \rangle$	WT	M182T
0	0.0 \pm 0.0	0.76 \pm 0.03	0.74 \pm 0.04
1	1.3 \pm 0.2	0.59 \pm 0.03	0.68 \pm 0.03
2	2.6 \pm 0.3	0.47 \pm 0.03	0.54 \pm 0.02
3	3.9 \pm 0.4	0.28 \pm 0.02	0.45 \pm 0.04
4	5.2 \pm 0.4	0.18 \pm 0.01	0.28 \pm 0.01
5	6.5 \pm 0.5	0.13 \pm 0.01	0.20 \pm 0.02

Table 2.2: Probabilities that genes from mutant libraries of wildtype and the thermostable M182T variant of TEM1 β -lactamase encode functional proteins. The table shows the number of rounds of error-prone PCR, the average number of nucleotide mutations per gene, and the fractions of mutated genes that confer ampicillin resistance on *E. coli*. Values are shown \pm their standard errors.

classification as $\alpha\beta$ 3-layer ($\alpha\beta\alpha$) sandwiches [40]), and our lattice protein simulations suggest that proteins with the same structure should have the same exponential decline in m -neutrality. Computations on proteins with markedly different structures predicted different declines in m -neutralities (data not shown), but no experimental data is yet available for these proteins.

The second major prediction of our theory is that among proteins with the same structure, more stable variants will exhibit higher initial m -neutralities, but converge to same exponential decline parameter ν . Our measurements on wildtype and the M182T variant of TEM1 β -lactamase allowed us to test this prediction, since the M182T variant is 2.7 kcal/mol more stable than wildtype [36], yet should have the same structure since it differs by only a single amino acid substitution. Figure 2.4 shows the measured fractions functional for wildtype and the M182T variant, as well as the theoretical predictions made using both the PoPMuSiC and FOLDEF potentials. The M182T variant exhibits enhanced initial m -neutrality as predicted by the theory, and once again the predictions made with both potentials are in good agreement with the experimental measurements.

2.4 Discussion

We have presented a theory for calculating the probability that a protein will remain functional after random amino acid substitutions, and have confirmed the main theoretical predictions with simulations and experiments. Our theory naturally separates a protein’s m -neutrality into components due to structure and stability. The eventual severity of the exponential decline in m -neutrality with the number of substitutions is a property of a protein’s structure. On the other hand, a protein can increase its tolerance for the first few mutations by increasing its stability, in effect allowing it to “take a few hits” before it is pushed into the inevitable structurally determined exponential decline in m -neutrality. This increased tolerance to mutations due to extra stability is probably also the underlying reason for the existence of global suppressor mutations [20, 21] that buffer proteins against otherwise deleterious mutations by increasing their thermodynamic stability.

Two simplifying assumptions underlie our model. The first is that mutations are additive, an assumption that is clearly not strictly true since protein residues do interact.

Mutations are most likely to be non-additive if the mutated residues are in direct contact in a protein’s structure [28, 29]. Since proteins are large, two randomly chosen residues will only rarely contact each other, and so although the additivity assumption is violated for some specific mutations, it is accurate when averaged over all possible mutations. The second assumption is that mutations affect function only through their effects on stability. This assumption ignores some effects of mutating residues that are directly involved in a protein’s function. Therefore, for proteins with a high fraction of functional residues, our theory only provides an upper bound on m -neutrality. However, our theory’s remarkable success in describing the m -neutralities of both subtilisin and the TEM1 β -lactamase variants suggests that this assumption is also accurate for most proteins.

Our theory provides a quantitative rationale for earlier work with lattice proteins on the organization of functional proteins in sequence space. Bornberg-Bauer and Chan [9] proposed that proteins are located in superfunnels in sequence space with the most stable sequence having the most neutral neighbors; others have reported that folded proteins surround highly stable prototype sequences in sequence space [41, 10, 11], and Shakhnovich and coworkers [8] showed that proteins with a large energy gap between the lowest and second lowest energy conformations are stabilized against mutations. We provide a clear explanation: more stable proteins are able to tolerate more of the possible mutations before unfolding, and so a higher fraction of their neighboring sequences fold.

In addition to these stability-based effects, different protein structures have different inherent designabilities, with more sequences folding into some structures than others [12, 42, 43]. Proteins with more designable structures might be expected to show a milder exponential decline in m -neutrality (a larger value of ν), since their structures occupy a larger fraction of sequence space. The structural neutrality given by ν therefore provides a quantitative measure of designability that can be estimated with current computational techniques.

Our theory suggests a more nuanced approach to experimentally analyzing protein neutralities than has been applied in the past. Loeb and coworkers [18] have performed a careful and thorough analysis of the neutralities of several proteins or regions of proteins under the assumption of a strict exponential decline in m -neutrality. However, our work suggests that a protein’s m -neutrality can deviate from a strict exponential for the first few substitutions

if the protein has a large amount of extra stability, as is seen for the M182T variant of TEM1 β -lactamase. Therefore, it is important to examine whether natural proteins have evolved stability above that required for function in their natural environments in order to provide them with additional robustness [22] to the first few amino acid substitutions.

Our theory also has applications in protein engineering. Directed evolution involves screening libraries of mutant proteins for new or improved functions [44]. Each round of directed evolution typically introduces only one or two amino acid substitutions because the rapid decline in m -neutrality means that larger numbers of substitutions will yield libraries of mostly unfolded proteins. Our model suggests that using highly stable parents for directed evolution should increase the fraction of folded mutants at a given level of substitutions. It also provides a method for predicting which structures will respond better to large numbers of substitutions.

2.5 Methods

2.5.1 Convolution of Mutational Effects

The distribution $p_m(x)$ was calculated as

$$p_m(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} [F(k)]^m \exp(-ikx) dk$$

where

$$F(k) = \int_{-\infty}^{\infty} p_1(y) \exp(iky) dy$$

is the characteristic function of $p_1(x)$. For the numerical calculations, $p_1(x)$ was constructed by binning all single mutant $\Delta\Delta G$ values with a bin size of 0.01 and representing it with a list padded by a number of zeroes equal to $m - 1$ times the number of bins. Each element of the fast-Fourier transform (FFT) of this list was raised to the m power, and p_m was recovered through an inverse FFT.

2.5.2 Lattice Proteins

We used a two-dimensional lattice protein model with proteins composed of $N = 20$ monomers of 20 types representing the natural amino acids, and allowed to occupy any

of the 41,889,578 conformations corresponding to all length 20 self-avoiding walks [45]. The sum over these conformations was tractable because they correspond to only 910,972 unique contact sets. The energy $E(\mathcal{C})$ of conformation \mathcal{C} is

$$E(\mathcal{C}) = \sum_{i=1}^N \sum_{j=1}^{i-2} C_{ij}(\mathcal{C}) \times \epsilon(\mathcal{A}_i, \mathcal{A}_j),$$

where $C_{ij}(\mathcal{C})$ is one if residues i and j are nearest neighbors in conformation \mathcal{C} and zero otherwise, and $\epsilon(\mathcal{A}_i, \mathcal{A}_j)$ is the interaction energy between residue types \mathcal{A}_i and \mathcal{A}_j , given by [46] (Table 5). The energies are in units of $k_B T$ where $T = 1.0$ for all simulations, corresponding to room temperature.

The stability with which a protein folds to a target conformation \mathcal{C}_t is

$$\Delta G_f(\mathcal{C}_t) = E(\mathcal{C}_t) + T \ln \{Q(T) - \exp[-E(\mathcal{C}_t)/T]\}$$

where $Q(T)$ is the partition sum

$$Q(T) = \sum_{\{\mathcal{C}_i\}} \exp[-E(\mathcal{C}_i)/T].$$

We considered a mutant functional if $\Delta G_f(\mathcal{C}_t) < 0.0$, and nonfunctional otherwise.

To generate lattice proteins, \mathcal{C}_t was randomly chosen from the subset of conformations with unique contact sets. For each \mathcal{C}_t , an adaptive walk was begun with a random starting sequence, with each step of the walk choosing the first point mutant with a better $\Delta G_f(\mathcal{C}_t)$. The adaptive walk was terminated after a sequence was found with $\Delta G_f(\mathcal{C}_t) \leq -2.0$, 500 steps were taken, or no improved mutants were found. If 200 random walks failed to generate a sequence with $\Delta G_f(\mathcal{C}_t) \leq -2.0$, a new \mathcal{C}_t was chosen. A clonal population of the final sequence from the random walk was then subjected to 2.5×10^5 generations of neutral evolution with a population size of 100 and a per residue mutation rate of 5×10^{-5} , with all sequences having $\Delta G_f(\mathcal{C}_t) \leq -2.0$ assigned a fitness of one and all other sequences a fitness of zero. Clonal populations of the most abundant sequence from this evolution were used to begin evolutionary runs to generate sequences with $\Delta G_f(\mathcal{C}_t)$ equal to the values indicated in Figure 2.2. Neutral evolution was performed for each target stability, now assigning a

fitness of one to any sequence that met the target stability and a fitness of zero to any other sequence. After the 2.5×10^5 generations, we selected the first sequence generated with a stability within 0.025 of the target stability, and used this sequence for the m -neutrality analysis. Lattice protein m -neutralities were computed by sampling all mutants for $m \leq 2$ or 5×10^5 randomly generated mutants for $m > 2$. In Figure 2.2, ν was computed as $\sqrt{P_f(m=6)/P_f(m=4)}$.

2.5.3 Measured Neutralities of Subtilisin and TEM1 β -lactamase

The measured neutralities for subtilisin were those from population 6B in Table 2 of [17], normalized by the fraction of functional clones in the control library.

The 861 bp genes for wildtype and M182T TEM1 β -lactamase were a kind gift from Brian Shoichet [36] and were subcloned into the pMON 1A2 plasmid [47] with SacI and HindIII using PCR primers 5'-GCGGCGGAGCTC TGAGTATTCAACATTTCCGT GTCGC-3' and 5'-GCGGCGAAGCTTTTACCAATG CTTAATCAGTGAGGCAC-3' and sequenced. For the round zero library, unmutated gene was cut directly from the plasmid. Each successive round of error-prone PCR used 3 ng of SacI/HindIII digest of gene from the previous round in 100 μ l reactions containing 0.5 μ M of each of the above primers, 7 mM MgCl₂, 75 μ M MnCl₂, 200 μ M of dATP and dGTP, 500 μ M of dTTP and dGTP, 1X Applied Biosystems PCR buffer without MgCl₂, and 5 U of Applied Biosystem *Taq* DNA polymerase. The PCR conditions were 95°C for 5 minutes, and then 14 cycles of 30 s each at 95°C, 50°C, and 72°C. The number of doublings per round was determined to be approximately ten by quantifying the DNA versus a marker on an agarose gel.

The SacI/HindIII digested genes were ligated into the pMON plasmid with T4 Quick DNA Ligase in 20 μ l reactions containing 50 ng each of gene and digested pMON plasmid, and 10 μ l of the ligation reactions were transformed into 100 μ l of XL1Blue Supercompetent cells from Stratagene. Transformed cells were plated on LB-agar plates containing 10 μ g/ml of kanamycin (selective only for plasmid) and on LB-agar plates containing 10 μ g/ml of both kanamycin and 20 μ g/ml of ampicillin (selective for both plasmid and active β -lactamase gene) at a density that gave 100-300 colonies per unselected plate. The fraction functional was computed as the average of at least five pairs of selected/unselected plates.

The mutation frequency in the round five library was determined by sequencing the first 570 bp of twenty genes each from the unselected wildtype and M182T plates with the sequencing primer 5'-GGTCGATGTTTGGATGTTATGGAGC-3'. No biases in the locations of mutations were observed. The wildtype and M182T genes were mutated under identical conditions, and the sequencing found the same mutation frequency in the round five library for both ($0.77 \pm 0.08\%$ for wildtype and $0.74 \pm 0.08\%$ for M182T). For better statistics, the sequencing results for both libraries were combined to give the data in Table 2.1. The per round mutation frequency was calculated assuming that each round of error-prone PCR introduced the same average number of mutations. This gives a per round mutation frequency of $0.15 \pm 0.03\%$. To confirm this assumption, we sequenced ten unselected clones from both of the round one libraries, and found mutation frequencies of $0.16 \pm 0.05\%$ for wildtype and $0.19 \pm 0.06\%$ for M182T. Standard errors were computed assuming Poisson sampling statistics.

2.5.4 Neutrality predictions

The single residue mutant $\Delta\Delta G$ values were estimated with the web version of the PoPMuSiC program [38] available at <http://babylone.ulb.ac.be/popmusic/> and the FOLDEF program [39] available at <http://fold-x.embl-heidelberg.de:1100/cgi-bin/main.cgi> using PDB structures IIAV and 1BTL for subtilisin and TEM1 β -lactamase. The probabilities $\mathcal{P}(Y | X)$ that a single nucleotide mutation to a gene would change base X to base Y were computed from the nucleotide frequencies of the gene sequences and the probabilities $\mathcal{P}(X, Y)$ that a random mutation was from base X to base Y given for subtilisin in Table 1 of [17] and for TEM1 β -lactamase in Table 2.1 of this work. We constructed the probability distribution $\hat{p}_1(\Delta\Delta G^1)$ for the effects of single nucleotide mutations by constructing a list of $\Delta\Delta G$ values by including $1000 \times L \times \mathcal{P}(Y | X)$ entries for the $\Delta\Delta G$ value associated with each mutation from X to Y , where L is the gene length, assigning synonymous mutations $\Delta\Delta G = 0$ and nonsense and frameshift mutations $\Delta\Delta G = 25$ kcal/mol, ignoring mutations for which no $\Delta\Delta G$ value could be computed, and assigning any mutation with $\Delta\Delta G > 25$ a value of $\Delta\Delta G = 25$. The values of $\hat{P}_f(m_{\text{nt}})$ were computed by convolutions of $\hat{p}_1(\Delta\Delta G^1)$ as described above.

The probability $f(m_{\text{nt}}; r)$ of finding a sequence with m_{nt} nucleotide mutations in a

library with an average of $\langle m_{\text{nt}} \rangle$ mutations created by r rounds of error-prone PCR is

$$f(m_{\text{nt}}; r) = (1 + \lambda)^{-N} \sum_{k=0}^N \binom{N}{k} \lambda^k \frac{(kx)^{m_{\text{nt}}} e^{-kx}}{m_{\text{nt}}!}$$

where n_{cycles} is the number of PCR cycles per round, λ is the PCR efficiency, $N = n_{\text{cycles}} \times r$, and $x = \frac{\langle m_{\text{nt}} \rangle (1 + \lambda)}{N \lambda}$ [48]. For subtilisin $n_{\text{cycles}} = 13$ and $\lambda = 0.77$ and for the TEM1 $n_{\text{cycles}} = 14$ and $\lambda = 0.71$. For these parameter values, $f(m_{\text{nt}}; r)$ is nearly Poisson.

The values of $\Delta G_f^{\text{extra}}$ for the predictions were computed by a least squares fitting to the experimental measurements. The $\Delta G_f^{\text{extra}}$ values differed for the PoPMuSiC and FOLDEF predictions because the $\Delta \Delta G$ values from these methods are scaled differently. For the PoPMuSiC predictions, $\Delta G_f^{\text{extra}}$ was -2.5 , -2.5 , and -3.9 kcal/mol for subtilisin, wildtype TEM1 β -lactamase, and M182T TEM1 β -lactamase. For the FOLDEF predictions, the values were -7.3 , -5.7 , and -12.0 .

2.6 Acknowledgements

We thank Brian Shoichet for providing us with genes for the TEM1 β -lactamase variants, and Titus Brown for programming assistance. We thank Michelle Meyer, Steve Mayo, Zheng-Gang Wang, and Eric Zollars for helpful advice and discussions. J.D.B. is supported by a Howard Hughes Medical Institute predoctoral fellowship. D.A.D. is supported by the National Institutes of Health, National Research Service Award 5 T32 MH19138 from the National Institute of Mental Health. C.A. and C.O.W. were supported in part by the National Science Foundation under grant DEB-9981387.

Chapter 3

Protein Stability Promotes Evolvability

A version of this chapter has been published as [49].

3.1 Abstract

The biophysical properties that enable proteins to so readily evolve to perform diverse biochemical tasks are largely unknown. Here we show that a protein's capacity to evolve is enhanced by the mutational robustness conferred by extra stability. We use simulations with model lattice proteins to demonstrate how extra stability increases evolvability by allowing a protein to accept a wider range of beneficial mutations while still folding to its native structure. We confirm this view experimentally by mutating marginally stable and thermostable variants of cytochrome P450 BM3. Mutants of the stabilized parent were more likely to exhibit new or improved functions. Only the stabilized P450 parent could tolerate the highly destabilizing mutations needed to confer novel activities such as hydroxylating the anti-inflammatory drug naproxen. Our work establishes a crucial link between protein stability and evolution. We show that we can exploit this link to discover new protein functions, and we suggest how natural evolution might do the same.

3.2 Introduction

Biological systems are evolvable in the sense that mutation and selection are able to create new or improved phenotypes. A major biological question is how a system's physical

properties influence its capacity to evolve [50, 51]. In recent years, understanding the determinants of evolvability has also become an important practical concern, as researchers increasingly use evolution to engineer everything from proteins [52] to designs for civil engineering structures [53].

Proteins are one of the simplest and best examples of evolvable biological systems, since they possess biochemical functions that can be altered with just a few mutations [54]. One property that has been broadly hypothesized to contribute to evolvability is robustness to mutations [55], and proteins are often quite mutationally robust, with over half of the single mutants of many proteins retaining their native functions [15, 16, 18, 7]. Because proteins usually must fold in order to function, and because mutated proteins generally adopt the original native structure if they fold at all [56, 57], retention of the basic native structure is a normally a prerequisite for the acquisition of new functions. Extra thermodynamic stability makes a protein’s native structure and function more robust to random mutations by increasing the fraction of mutants that continue to possess the minimal stability required to fold [7, 58].

Here we investigate how stability affects a protein’s evolvability by using controlled experiments to measure the fraction of a protein’s mutants that exhibit new or improved function. We first use a simple computational model to establish a framework for understanding the relationships among protein stability, mutational robustness, and evolvability. We then validate this framework with experiments on members of the biochemically important cytochrome P450 enzyme family, and describe specific examples that illustrate the biophysical basis of the connection between protein stability and evolvability. Finally, we discuss the implications of our work for understanding natural protein evolution and designing better protein engineering strategies.

3.3 Results

3.3.1 Simulations with Model Lattice Proteins

We used a simple conceptual framework [7] for understanding the relationship between protein stability and evolution. The premise is that evolution selects for a protein’s biochemical function rather than its stability. However, since a protein’s function typically

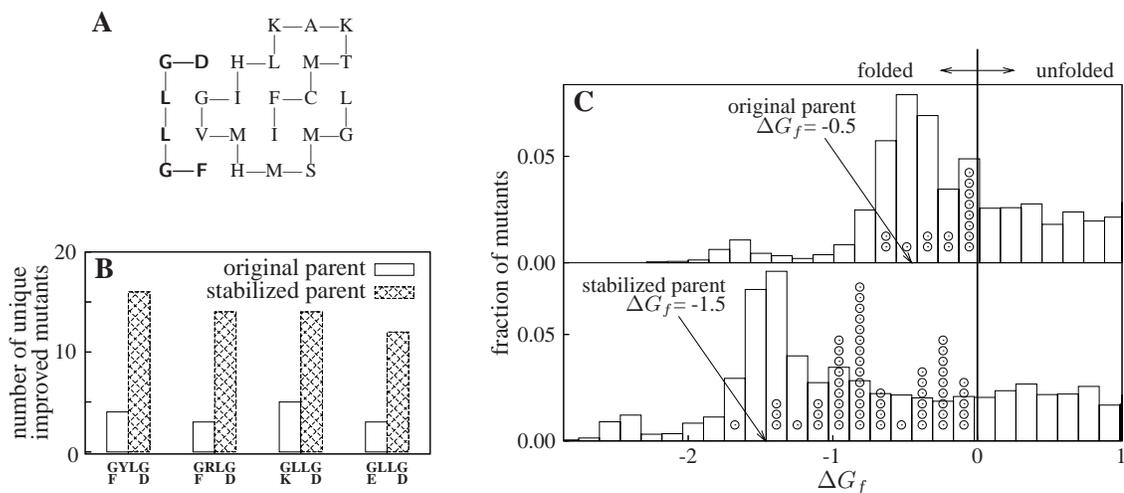


Figure 3.1: Increased stability enhances evolvability of a model lattice protein. (A) The original model protein (right) that had been evolved to bind to a rigid ligand (left in bold). (B) Mutants of a stabilized model protein were more likely than mutants of the original model protein to show improved binding to the four new ligands shown below the bars. The bars show the number of mutants out of 1500 screened that bound the new ligand with at least twice the affinity of the parent. (C) The stabilized model protein was more evolvable because more of its destabilized but improved mutants satisfied the minimal stability cutoff. The bars show the distribution of stabilities among all mutants in the libraries, while the circles show the stabilities of the improved mutants.

depends on its ability to fold to a thermodynamically stable native structure [25], stability is still constrained during evolution. Specifically, we imagine that a protein must fold to its native structure with some minimal stability in order to remain folded at physiological conditions. If a protein fails to meet this minimal stability threshold, then it will neither fold nor function. If a protein does fold with at least the minimal required stability, then evolution selects for a protein’s function and is indifferent to the amount of extra stability it possesses. Most proteins, however, will still be marginally stable since highly stable sequences are rare [41].

This conceptual framework formed the basis for simulations with lattice proteins. Lattice proteins are highly simplified protein models that are useful tools for studying protein folding and evolution [34, 35]. Our lattice proteins were chains of 20 amino acids that fold on a two-dimensional lattice, with the energy of each conformation equal to the sum of the pairwise interactions between nonbonded amino acids [46]. Each lattice protein can occupy any of 41,889,578 possible conformations, and by summing over all of these conformations we

could exactly determine the partition function and free energy of folding (ΔG_f). We set a minimal stability threshold for our lattice proteins by requiring them to fold to the original native structure with a stability of $\Delta G_f \leq 0$ (in no case did we observe a protein that stably folded to a new structure), which is equivalent to requiring the native structure to have a lower free energy than the ensemble of all non-native conformations. For those proteins that stably folded, we measured function as the binding energy of the folded protein to a small rigid ligand [45], as shown in Figure 3.1A. Our model therefore recapitulated the essential requirements imposed on real proteins of simultaneously folding and performing a biochemical task.

We first evolved a model protein to stably fold and strongly bind a ligand (Figure 3.1A). This evolved protein had a stability of $\Delta G_f = -0.5$, meaning that it was only marginally stable as is typical for real proteins [26]. We then simulated the process of directed evolution with two rounds of random mutagenesis by error-prone PCR and screening to identify a stabilized variant of our model protein ($\Delta G_f = -1.5$) that contained three amino acid substitutions and exhibited the same ligand binding energy as the original protein. To examine the evolvabilities of the original and stabilized model proteins, we computationally simulated screening libraries of 1,500 randomly mutated sequences for mutants that bound to new ligands with at least twice the affinity of the parent proteins. For all four new ligands we examined, the parent proteins bound the new ligand with equal affinity, yet each time the mutant library from the stabilized parent produced over twice as many unique improved mutants (Figure 3.1B).

Figure 3.1C shows why the stabilized model protein was more evolvable. The mutants in both libraries exhibited similar changes in stability ($\Delta\Delta G$ values), but the extra stability of the stabilized protein meant that a larger fraction of its mutants continued to fold (46% versus 35% among all mutants with at least one mutation), confirming previous findings that more stable lattice proteins are more robust to mutations [7, 9]. The improved mutants tended to be destabilized, so were more frequent in the library from the stabilized parent. Although the more stable parent had less than a 50% increase in the fraction of mutants that folded, it had nearly four times more improved mutants (56 versus 15). The fact that extra stability increases the number of improved mutants much more than it increases the number of mutants that retain parental function indicates that improved mutants tended

to be more destabilized than the typical folded mutant.

3.3.2 Experiments on Cytochrome P450 BM3 Variants

To experimentally test the effect of stability on the evolvability of real proteins, we randomly mutated two variants of a cytochrome P450 BM3 (also known as CYP102A1) heme domain peroxxygenase [59] and screened for mutants with new or improved activity on five substrates. The cytochrome P450 superfamily contains members involved in important biochemical processes such as drug metabolism and steroid biosynthesis [60, 61]. P450 BM3 catalyzes sub-terminal hydroxylation of medium- and long-chain fatty acids [62]. The 21B3 variant is a laboratory-evolved version of the P450 BM3 heme domain that efficiently hydroxylates 12-*p*-nitrophenoxycarboxylic acid (12-pNCA, structure shown in Figure 3.3), utilizing hydrogen peroxide in place of the NADPH cofactor and oxygen [59]. The 5H6 variant was created by laboratory evolution of 21B3, selecting for mutants that were more thermostable while retaining activity on 12-pNCA [63]. We quantified the stabilities of the enzymes by the temperature (T_{50}) at which half of the protein irreversibly denatured after a 10 minute incubation. Because denaturation is irreversible, these T_{50} values are not equilibrium thermodynamic measurements, and so cannot be directly related to ΔG_f . However, the T_{50} values were highly correlated with the stability to irreversible denaturation by urea, supporting the notion that they reflect universal aspects of protein stability rather than unique characteristics of the process of irreversible thermal denaturation (Figure 3.12 and related discussion in the Methods section). As measured by the T_{50} values, the 21B3 enzyme is only marginally stable ($T_{50} = 47^\circ\text{C}$), while 5H6 is much more stable ($T_{50} = 62^\circ\text{C}$) (melting curves are shown in Figure 3.7). The 5H6 enzyme differs from 21B3 at eight residues (out of 464 total). Both variants displayed nearly the same activity (measured as total turnovers) on 12-pNCA and all other substrates examined in this work.

We created mutant libraries of both 21B3 and 5H6 using error-prone PCR. The libraries were generated under identical conditions and had the same distributions of mutations (Figure 3.2). The overall mutation rate was 4.5 ± 0.3 nucleotide mutations per gene (Table 3.1). We examined 522 mutants from each library for retention of folding, as assayed by the characteristic Soret band at 450 nm in the carbon monoxide binding difference spectrum [66]. As expected, mutants of the stabilized 5H6 protein were more likely than those of the 21B3

Base pairs sequenced	58,719
Total mutations	182
Mutation frequency (%)	0.31 ± 0.02
Avg. mutations per gene	4.5 ± 0.3
% synonymous mutations	28
% nonsynonymous mutations	63
% frameshift/nonsense mutations	9
Mutation types (%)	
A \rightarrow T, T \rightarrow A	25
A \rightarrow C, T \rightarrow G	6
A \rightarrow G, T \rightarrow C	53
G \rightarrow A, C \rightarrow T	10
G \rightarrow C, C \rightarrow G	0
G \rightarrow T, C \rightarrow A	3
frameshift	3

Table 3.1: Mutation frequencies in error-prone PCR libraries. Statistics are for all 41 randomly chosen mutants. Standard errors are calculated assuming Poisson counting statistics.

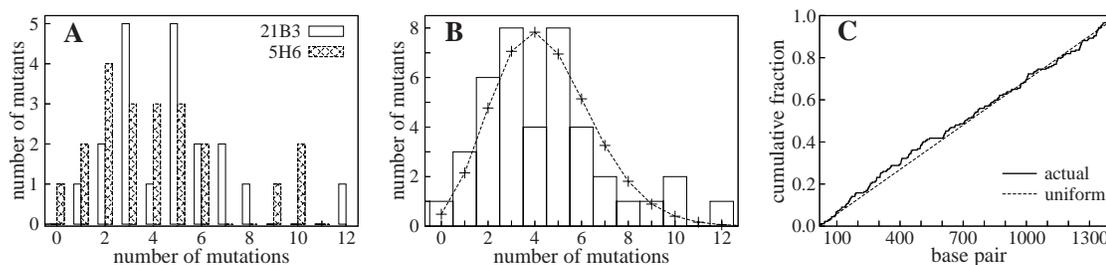


Figure 3.2: Distribution of mutations in the two P450 error-prone PCR libraries. (A) The distribution of mutations among 20 randomly chosen 21B3 mutants and 21 randomly chosen 5H6 mutants. The distributions are statistically indistinguishable ($P = 0.84$). (B) The distribution of mutations among all 41 sequenced mutants is consistent with the theoretical prediction for an error-prone PCR library (lines) [48, 64] ($P = 0.11$). (C) The mutations are uniformly distributed along the gene ($P = 0.66$). The lines show the cumulative fraction of mutations that occur at or before that position in the gene. All P -values are from Kolmogorov-Smirnov tests [65], and represent the probability that the samples or theoretical curves would differ by at least this much if they were generated by the same underlying distribution.

Threshold	Pr (F 21B3)	Pr (F 5H6)	Pr (A F ; 21B3)	Pr (A F ; 5H6)
0.10	0.44	0.69	0.98	0.96
0.25	0.39	0.66	0.98	0.96
0.50	0.33	0.61	0.96	0.94
0.75	0.29	0.57	0.92	0.93
0.90	0.24	0.51	0.93	0.90

Table 3.2: Retention of folding and function. The probabilities that 21B3 and 5H6 mutants fold (F), and that those folded mutants are active (A) on 12-pNCA at different thresholds for folding and activity. The probabilities are determined from the six plates of each parent shown in Figure 3.8. The folding and activity status of each mutant is determined relative to the four parents on the plate. If the CO binding difference spectrum $A_{447} - A_{490}$ is greater than the threshold times the parental median, the mutant is considered folded. Folded mutants are classified as active if the 12-pNCA A_{398} value is greater than the threshold times the parental median.

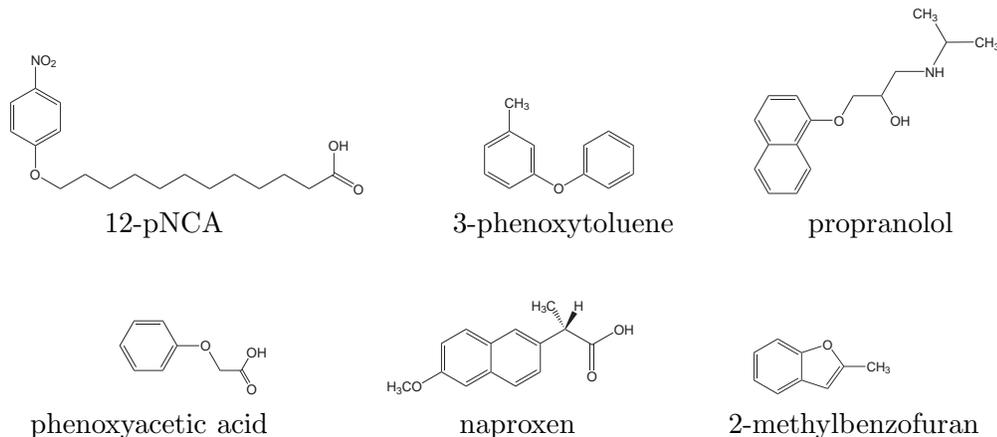


Figure 3.3: The substrates on which the P450 mutants were screened for activity.

protein to fold (61% of 5H6 mutants contained at least half the folded protein of the parent versus 33% for 21B3, raw data are shown in Figure 3.8 and Table 3.2). Most of these folded mutants retained at least half the parental activity on 12-pNCA (94% and 96% for 5H6 and 21B3), indicating that mutations that disrupted parental function generally did so by preventing the formation of properly folded protein. This confirms the experimental findings of [7] that more stable proteins are more robust to mutations.

We examined the evolvability of the 21B3 and 5H6 enzymes by screening for mutants that hydroxylated any of five new substrates: the anti-inflammatory drug naproxen, 3-phenoxytoluene, phenoxyacetic acid, the beta-adrenergic receptor blocking agent propranolol, and 2-methylbenzofuran (structures shown in Figure 3.3). We screened for hydrox-

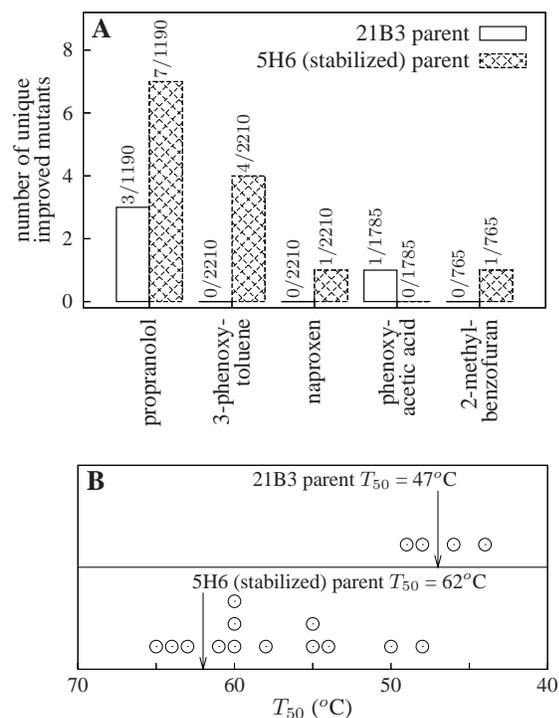


Figure 3.4: Increased stability enhances evolvability of the P450 BM3 heme domain. (A) The stable 5H6 protein yielded more mutants with new or improved activity than the marginally stable 21B3 protein. The counts above the bars give the number of improved mutants out of the total number of mutants screened. (B) Some of the improved 5H6 mutants were greatly destabilized relative to the parent protein, while the stabilities of the improved 21B3 mutants clustered around those of the parent protein (circles show T_{50} values for improved mutants).

ylation activity using the 4-aminoantipyrene (4-AAP) assay to measure the total amount of product after completion of the reaction [67], and determined that neither 21B3 nor 5H6 had detectable activity on the first three substrates, both had equal weak activity on propranolol, and 21B3 had trace activity on 2-methylbenzofuran (Table 3.3). We used consistent quantitative criteria to identify mutants that had either acquired new activity or improved by more than 50% over the parental level in the 4-AAP assay. We screened 8,160 mutant-substrate pairs for each parent. From these, we identified 13 improved mutants of 5H6 and 4 improved mutants of 21B3 (Figure 3.4A, Table 3.3). All the improved mutants had unique protein sequences (given in Table 3.5). Thus, we found over three times more improved mutants in the 5H6 library than in the 21B3 library.

To assess the importance of stability in conferring enhanced evolvability on the 5H6 pro-

tein, we measured the stabilities of all improved mutants (melting curves are in Figure 3.7). Figure 3.4B shows that none of the improved 21B3 mutants was destabilized by more than 3°C, but that the thermostable 5H6 parent produced improved mutants that were destabilized by as much as 14°C. We identified specific beneficial but destabilizing substitutions that could be made only in the stabilized parent. For example, neither 21B3 nor 5H6 exhibited activity on naproxen, presumably because the negatively charged naproxen molecule does not enter the hydrophobic P450 BM3 substrate binding pocket. Mutating leucine 75 in the substrate binding pocket to arginine allowed 5H6 to hydroxylate naproxen by providing a compensating positive charge for the naproxen molecule (Figure 3.5). However, burial of this arginine residue in the hydrophobic binding pocket substantially destabilized the 5H6 mutant ($\Delta T_{50} = -8^\circ\text{C}$). When we made the same substitution to 21B3, we could only recover inactive and improperly folded protein (as indicated by a carbon monoxide difference spectrum peak at 420 nm [68] as shown in Figure 3.9). The F275S substitution (located 12Å from the substrate [69]) is another example of a beneficial substitution which could be made only in the stabilized parent. This substitution conferred 3-phenoxytoluene activity on the 5H6 parent, but decreased the T_{50} by 7°C. When we made this substitution in 21B3 we again could not recover any folded protein (Figure 3.10). In contrast, the F205L substitution (located near the substrate binding pocket [69]) found in a 21B3 mutant with improved activity on propranolol did not have a substantial effect on stability, and slightly improved the activity of both 21B3 and 5H6 when introduced into those sequences (Figure 3.11).

3.4 Discussion

We have shown that more stable proteins are more evolvable because they are better able to tolerate functionally beneficial but destabilizing mutations. Our work touches on the relationship between protein stability and function, which has historically been a subject of considerable confusion. Despite repeated speculation to the contrary [70, 71, 72], high stability and function are not inherently incompatible, since a wealth of experiments have demonstrated that proteins can be dramatically stabilized without sacrificing their biological functions [73, 74, 75, 76]. But protein stability and function often appear to trade off at the

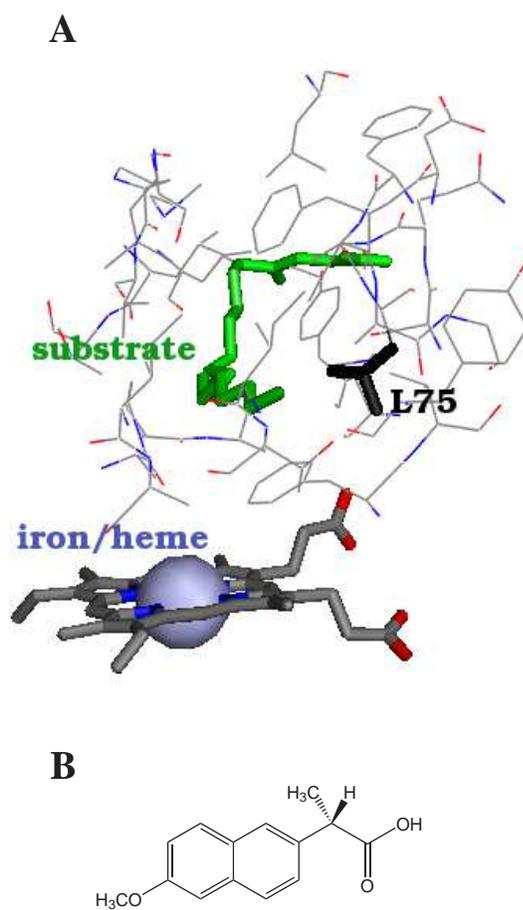


Figure 3.5: The functionally innovative but destabilizing L75R mutation can only be tolerated by the stabilized parent. (A) Leucine 75 is positioned close to the substrate in the hydrophobic P450 BM3 substrate binding pocket [69]. Mutating L75 to a positively charged arginine conferred naproxen activity on the stabilized 5H6 parent but disrupted the proper folding of the marginally stable 21B3 parent. (B) The anti-inflammatory drug naproxen, which contains a negatively charged carboxylic acid group.

level of individual mutations. This apparent tradeoff is at least partly due to the simple fact that most randomly chosen mutations (functionally beneficial or not) are destabilizing [77, 78, 79, 27]. In addition, residues in a protein's active site often must satisfy functional constraints (such as maintaining buried charges or cavities in a protein's interior) that make them poorly optimized for stability [80, 81, 82]. Therefore, mutating active-site residues often enhances stability at the cost of function [80, 81, 82], and likewise acquiring new functions can require destabilizing mutations (as is the case for our L75R mutation in P450, which confers activity on naproxen by burying a positive charge). However it remains unclear whether active-site constraints intensify the tradeoff between stability and functional evolution, since a seemingly opposite argument can be made that mutations to an active site that is already poorly optimized for stability should be less destabilizing than typical mutations (they could even enhance stability if, for example, they confer function on smaller substrates by reducing the size of a cavity in a protein's interior). If functionally innovative mutations tend to be more destabilizing than random mutations, then extra protein stability should enhance the rate of functional innovation more than it enhances the mutational robustness of the native function. In our lattice protein simulations, extra stability increased the rate of functional innovation by nearly 400% while it increased mutational robustness by only 50%; however, we feel our lattice model is too crude to confidently extrapolate conclusions involving residue-level properties to real proteins. In our P450 experiments, extra stability also increased the number of functionally improved mutants (from 4 to 13) more than it increased mutational robustness (by a factor of 1.8); however, here the statistics are too poor to conclude that functional innovation is improved significantly more than mutational robustness. Therefore, in our minds it remains unclear whether extra protein stability promotes evolvability simply by improving the tolerance to all mutations (some of which happen to be functionally beneficial), or whether the effect is further amplified by a tendency for functionally innovative mutations to be especially destabilizing.

In either case, our work argues quite generally that extra stability will enhance evolvability. Although it is clearly possible to stabilize proteins without interfering with their functions [73, 74, 75, 76], proteins tend to be only marginally more stable than is required by their environment [26]. This marginal stability is probably due to the fact that natural selection does not directly favor extra stability in the face of predominantly destabilizing

mutations, causing stability to drift towards the minimum evolutionary requirement [41, 83]. Naturally evolving proteins must therefore wait for functionally neutral mutations to stabilize the structure in order to counterbalance the effects of other destabilizing but functionally beneficial mutations [36]. In this sense, a protein’s stability represents a hidden dimension in evolution: extra stability is neutral with respect to selection for protein function, but it can be crucial in allowing a protein to tolerate mutations that confer beneficial phenotypes. We have shown that protein engineering by directed evolution is more effective if direct selection for extra stability is used to increase a protein’s evolvability. The extent to which natural evolution might also select for evolvability has been the subject of much recent theoretical speculation [51, 55, 84]. We suggest that one possible mechanism by which natural evolution could increase evolvability would be to stabilize proteins undergoing adaptive evolution or provide systems to buffer the effects of destabilizing protein mutations.

3.5 Methods

3.5.1 Model lattice protein simulations

A model lattice protein was represented as a chain of $N = 20$ monomers of 20 types corresponding to the natural amino acids. A model protein could occupy any of the 41,889,578 conformations (representing 910,972 unique contact sets) corresponding to all length 20 self-avoiding walks on a two-dimensional lattice. A conformation \mathcal{C} had an energy of

$$E(\mathcal{C}) = \sum_{i=1}^N \sum_{j=1}^{i-2} C_{ij}(\mathcal{C}) \times \epsilon(\mathcal{A}_i, \mathcal{A}_j),$$

where $C_{ij}(\mathcal{C})$ is one if residues i and j are nearest neighbors in conformation \mathcal{C} and zero otherwise, and $\epsilon(\mathcal{A}_i, \mathcal{A}_j)$ is the interaction energy between residue types \mathcal{A}_i and \mathcal{A}_j , given by Table 5 of reference [46]. Energies are in units of $k_B T$, and $T = 1.0$ for all simulations.

A model protein folds to a target conformation \mathcal{C}_t with stability

$$\Delta G_f(\mathcal{C}_t) = E(\mathcal{C}_t) + T \ln \{Q(T) - \exp[-E(\mathcal{C}_t)/T]\}$$

where $Q(T)$ is the partition sum

$$Q(T) = \sum_{\{\mathcal{C}_i\}} \exp[-E(\mathcal{C}_i)/T].$$

Protein were considered stably folded to \mathcal{C}_i if and only if $\Delta G_f(\mathcal{C}_i) \leq 0.0$.

A model protein's activity was represented by binding to a small rigid ligand, much as in [45]. If a model protein did not stably fold to a unique conformation, it was considered inactive. If it did stably fold to a unique conformation, then we computed the binding energy of the ligand as the sum of the pairwise interactions between the ligand and protein residues, searching over all possible rotational and translational positions of the ligand. The binding affinity of the ligand to the folded conformation is given by the association constant K_a , calculated as the exponential of the negative binding energy.

To create the model protein shown in Figure 3.1A, we first chose the target conformation of the protein, and then performed an adaptive walk from a random starting sequence until we found a sequence that folded to the conformation. We evolved this sequence for 10,000 generations with a population size of 10 and a mutation rate of 5×10^{-4} per residue per generation, with the fitness of a model protein proportional to its binding affinity to the ligand shown in Figure 3.1A. To create a stable variant of this model protein, we simulated using error-prone PCR to make a library of 1,500 mutant sequences with the mutations distributed according to the error-prone PCR distribution [48, 85] with an average of 1.5 mutations per protein, 16 PCR cycles, and a PCR efficiency of 0.45. We performed two rounds of this error-prone PCR, selecting the most stable sequence that still bound to the ligand with the same binding energy as the original model protein at the end of each round. This procedure yielded the stabilized model protein with the sequence IFFMTKIKFHIGVMHMSMGL. We then simulated creating error-prone PCR libraries of both the original and stabilized model proteins using the same procedure as above. We screened each of the four new ligands shown in the legend of Figure 3.1B on one library from each parent, and recorded the number of mutants that bound to the new ligand with a binding affinity at least two-fold higher than that for the parent. These counts are the data shown by the bars in Figure 3.1B.

3.5.2 P450 mutant libraries.

We used error-prone PCR to create mutant libraries of the marginally stable 21B3 [59] and the thermostable 5H6 [63] variants of the cytochrome P450 BM3 heme domain. The template DNA was the appropriate gene cloned into the pCWori [86] plasmid as described in references [59] and [63]. We confirmed the sequences of the 21B3 and 5H6 genes by sequencing them with the primers *pCWori_for* (5'-GAAACAGGATCCATCGATGCTTAGGAGGTCAT-3'), *pCWori_rev* (5'-GCGTATCACGAGGCCCTTTCGTCTTCAAGC-3'), and *pCWori_mid_rev* (5'-CCAGCTTGTGGCCAACCCGAC-3'). The sequences matched those that were reported [59, 63], with 21B3 containing ten amino acid substitutions relative to the wildtype P450 BM3 heme domain (I58V, F87A, H100R, F107L, A135S, M145V, N239H, S274T, K434E, and V446I in the numbering scheme where residue one is the threonine after the cleaved N-terminal methionine) and 5H6 containing eight amino acid substitutions relative to 21B3 (L52I, S106R, V145M, A184V, L324I, V340M, I366V, and E442K) as well as the removal of one histidine from the C-terminal His tag.

The error-prone PCR reactions for the two parents were carried out in parallel using identical conditions to ensure the same mutation rate for both. The reactions were 100 μ l and contained 13 ng of template plasmid (corresponding to 3 ng of gene), 0.5 μ M of the oligonucleotide primers (*pCWori_for* and *pCWori_rev_clone*, 5'-GCTCATGTTTGACAGCTTATCATCG-3'), 200 μ M dATP and dGTP, 500 μ M dTTP and dCTP, 7 mM MgCl₂, 200 μ M MnCl₂, 1X Applied Biosystems PCR Buffer, and 5 units of *Taq*. PCR conditions were 95°C for 5 min. followed by 16 cycles of 30 s at 95°C, 30 s at 51°C, and 60 s at 72°C. Gel electrophoresis versus a known standard indicated that this yielded PCR product at a concentration of \approx 12 ng/ μ l, for a PCR efficiency of $\lambda = 0.45$. The PCR products were cloned into pCWori with *Bam*HI and *Eco*RI, electroporated into a catalase-free strain of *E. coli* [86], and plated on LB plates containing 100 μ g/ml ampicillin. Transformation of a control ligation with no insert indicated that the background rate of plasmid self ligation was less than 1%.

To measure the mutation rates, we randomly selected twenty 21B3 clones and twenty-one 5H6 clones for sequencing with primers *pCWori_for* and *pCWori_rev*, allowing us to read each gene from bp 18 to bp 1436. The 21B3 clones contained a total of 95 nucleotide mutations in the 28,380 sequenced base pairs, with 28 synonymous mutations, 60 nonsynonymous mutations, and 7 mutations leading to premature truncation of the gene (frameshift or non-

sense mutations). The 5H6 clones contained a total of 87 mutations in the 29,799 sequenced base pairs, with 23 synonymous mutations, 55 nonsynonymous mutations, and 9 mutations leading to premature truncation of the gene. The distributions of mutations in the two libraries were statistically indistinguishable (Figure 3.2A). After confirming that the mutation rates in the two libraries were indistinguishable, we combined the sequencing results for further analysis (Table 3.1). Figure 3.2B shows that the distribution of mutations is consistent with the theoretical distribution for error-prone PCR [48, 64], which gives the probability that a mutant in a library with an average of $\langle m_{\text{nt}} \rangle$ mutations per gene has m_{nt} nucleotide mutations as

$$\Pr(m_{\text{nt}}) = (1 + \lambda)^{-n} \sum_{k=0}^n \binom{n}{k} \lambda^k \frac{(kx)^{m_{\text{nt}}} e^{-kx}}{m_{\text{nt}}!}, \quad (3.1)$$

where n is the number of PCR cycles, λ is the PCR efficiency, and $x = \frac{\langle m_{\text{nt}} \rangle (1 + \lambda)}{n\lambda}$. We also confirmed that the mutations were distributed uniformly along the gene sequence (Figure 3.2C). If each position in the gene is equally likely to be mutated, then among 41 sequenced clones, 156.3 positions should be mutated once, 9.7 positions should be mutated twice, and 0.4 positions should be mutated three times. This is in good agreement with the observed values of 148, 13, and 1.

3.5.3 Screening for improved mutants

Single mutant colonies were picked from transformation plates with sterile toothpicks and used to inoculate sterile 1 ml deep-well plates (Falcon) with each well containing 400 μl of liquid LB with 100 $\mu\text{g}/\text{ml}$ ampicillin. As controls, wells A1, A2, A3, and A4 were always inoculated with the parent (21B3 or 5H6); wells A5, A6, A7, and A8 were always inoculated with a negative control (the pCWori plasmid lacking a P450 gene); and well H12 was not inoculated. Each of the remaining 87 wells contained a different mutant. These deep-well plates were grown for 20-24 hours in a humidified shaker (Kuhner ISF-1-W) at 215 or 225 rpm, 30°C, and 80% relative humidity. To express the proteins, a pipetting robot (Beckman Multimek 96) was used to transfer 100 μl per well from these LB deep well plates to 2 ml deep-well plates (Falcon) containing 400 μl of terrific broth (TB) supplemented with 100 $\mu\text{g}/\text{ml}$ ampicillin, 0.5 mM δ -aminolevulinic acid, and 0.2 mM IPTG. These TB deep-well

plates were also grown for 20-24 hours in the humidified shaker at 215-225 rpm, 30°C, and 80% relative humidity. After this growth, the cells were harvested by centrifugation at 6,100 g for 10 min at 4°C, and stored at -20°C. The LB deep-well plates were stored at 4°C so that improved mutants could be streaked from the plates.

To perform assays, the pelleted cells were resuspended in 600 μ l of lysis buffer (100 mM EPPS pH 8.2 with 0.5 mg/ml lysozyme and 2 U/ml deoxyribonuclease) per well with the pipetting robot. The plates were incubated for one hour at 37°C, and then centrifuged at 6,100 g for 10 min at 4°C to pellet cell debris. The pipetting robot was used to add 80 μ l per well of this clarified lysate to 96-well microtiter plates. To assay for folded protein, high-throughput carbon monoxide (CO) binding difference spectra were measured as described in [66] with the modifications that 80 μ l of clarified lysate in the EPPS buffer was used, and that the heme was reduced by the addition of 20 μ l of 0.1 M sodium hydrosulfite in 1.3 M potassium phosphate, pH 8.0. Blank spectra were read prior to binding CO with a Spectra Max Plus 384 plate reader (Molecular Devices) at every 10 nm from 400 to 500 nm, as well as at the points 447 nm (the absorbance peak for both 21B3 and 5H6) and 490 nm. After 5-10 minutes of incubation with CO, the absorbance readings were read at these points.

Mutants were screened for the retention of activity on 12-*p*-nitrophenoxycarboxylic acid (12-pNCA, Figure 3.3) using a slightly modified version of the method described in [59]. A 6X stock of 12-pNCA was prepared by combining 3.6 ml of 4.17 mM 12-pNCA in DMSO with 6.4 ml of 100 mM EPPS (pH 8.2) immediately before use. Twenty μ l of this 6X 12-pNCA stock was added to the 80 μ l of clarified lysate in each well of the microtiter plate, and the plate reader was used to mix the plate and blanked at 398 nm. To initiate the reaction, 20 μ l of a 6X stock of hydrogen peroxide (6X stock was 24 mM H₂O₂ in 100 mM EPPS pH 8.2, made immediately before use) was added to each well and mixed with the plate reader. The amount of final product was quantified by reading the absorbance at 398 nm after 20-30 minutes.

Mutants were screened for the acquisition of new or improved activity using the 4-aminoantipyrene (4-AAP) assay, which detects phenol-like compounds [67], produced by either direct hydroxylation of an aromatic ring or as dealkylation products after hydroxylation of a carbon that is ether bonded to an aromatic ring. This assay was used to screen for activity on five substrates (Figure 3.3) at the following final concentrations: 3-phenoxytoluene

(10 mM), naproxen (10 mM), phenoxyacetic acid (25 mM), 2-methylbenzofuran (5 mM), and propranolol (5 mM). Stocks with 6X concentrations of these substrates were made immediately before use by dissolving the substrate in equal volumes of DMSO and acetone and then adding 100 mM EPPS (pH 8.2) so that the final concentrations of DMSO and acetone in the 6X stocks were each 6%. Twenty μl of 6X substrate stock was added to the 80 μl of clarified lysate in each well of the microtiter plate, and 20 μl of the 6X hydrogen peroxide stock was added to initiate the reaction. The microtiter plates were mixed with the plate reader, and the reactions were allowed to run for 1.5 to 2 hours at room temperature before being quenched by the addition of 120 μl of 4 M urea in 0.1 M sodium hydroxide. A pipetting robot was used to add and mix 36 μl of 0.6% 4-AAP, and the plate reader was zeroed at 500 nm. The robot was then used to add and mix 36 μl of 0.6% potassium peroxodisulfate, and after 30 minutes the product was quantified by reading the absorbance at 500 nm.

3.5.4 Stability and Verification Assays

Protein was expressed for stability and verification assays by growing the cells in flasks with 200-300 ml of TB supplemented with 100 $\mu\text{g}/\text{ml}$ of ampicillin at 30 °C and 215 rpm. When the cells reached an optical density between 0.8 and 1.2 at 600 nm, they were induced by adding IPTG to a final concentration of 0.4 mM, as well as δ -aminolevulinic acid to a final concentration of 0.5 mM. The cells were then grown at 30 °C and 215 rpm for an additional 20-24 hours. A few mutants did not express well in these conditions, and so were grown at the milder conditions of 28°C and 180 rpm. The cells were harvested by centrifugation at 6,000-8,000 g and 4 °C for 10 min, and the pellets stored at -20 °C. Prior to use, pellets from 100 ml of culture were resuspended in 4 ml of 100 mM EPPS (pH 8.2) and lysed by sonication. Cell debris was pelleted by centrifugation at 6,000-8,000 g and 4 °C for 10-15 min. The clarified lysates were passed through PD-10 desalting columns (Amersham Biosciences) to remove small molecules that might appear as background in the 4-AAP assay. The P450 protein was quantified from the CO binding difference spectrum (extinction coefficient of 91 $\text{mM}^{-1} \text{cm}^{-1}$ for the $A_{447} - A_{490}$) and the concentration of P450 was adjusted to 5 μM by the addition of more 100 mM EPPS (pH 8.2).

For verification assays, the 5 μM clarified lysate was pipetted into 96-well microtiter

substrate	21B3				5H6			
	protein	activity	m _{aa}	T ₅₀	protein	activity	m _{aa}	T ₅₀
propranolol (1,190 screened)	parent	0.07 ± 0.02	NA	47°C	parent	0.08 ± 0.02	NA	62°C
	14C10	0.29 ± 0.05	3	44°C	27G8	0.15 ± 0.01	1	65°C
	27B2	0.15 ± 0.03	3	49°C	27G12	0.25 ± 0.03	7	55°C
	31B12	0.14 ± 0.02	2	48°C	30B10	0.19 ± 0.02	6	58°C
					32F7	0.15 ± 0.02	4	55°C
					36G11	0.22 ± 0.02	3	64°C
					37F4	0.21 ± 0.02	2	63°C
					38F11	0.16 ± 0.02	3	60°C
3-phenoxytoluene (2,210 screened)	parent	none	NA	47°C	parent	none	NA	62°C
					20D1	0.05 ± 0.02	1	60°C
					23E4	0.30 ± 0.03	5	48°C
					28B5	0.05 ± 0.03	3	60°C
					29G8	0.05 ± 0.02	2	50°C
naproxen (2,210 screened)	parent	none	NA	47°C	parent	none	NA	62°C
					13C9	0.16 ± 0.02	1	54°C
phenoxyacetic acid (1,785 screened)	parent	none	NA	47°C	parent	none	NA	62°C
	38F7	0.30 ± 0.04	3	46°C				
2-methylbenzofuran (765 screened)	parent	0.07 ± 0.02	NA	47°C	parent	none	NA	62°C
					32H1	0.11 ± 0.02	1	61°C

Table 3.3: Summary of improved P450 mutants. The leftmost column gives the total number of mutants of each parent screened on that substrate. Subsequent columns give the activity, number of amino acid substitutions (m_{aa}), and stabilities as measured by the T_{50} values. Mutants are named according to the plate and well in which they were found, and sequences are given in Table 3.5. Activities represent the median \pm the standard deviation of the total product formed per well of 80 μ l of 5 μ M protein, as measured by the A_{500} in the 4-AAP assay (raw data is in Figure 3.6), and are indicated as “none” when indistinguishable from the background. Using an extinction coefficient of 4,800 $M^{-1}cm^{-1}$ for the 4-AAP/phenol complex [87], each unit of A_{500} corresponds to ≈ 80 nmol of product from the ≈ 0.4 nmol of protein per well.

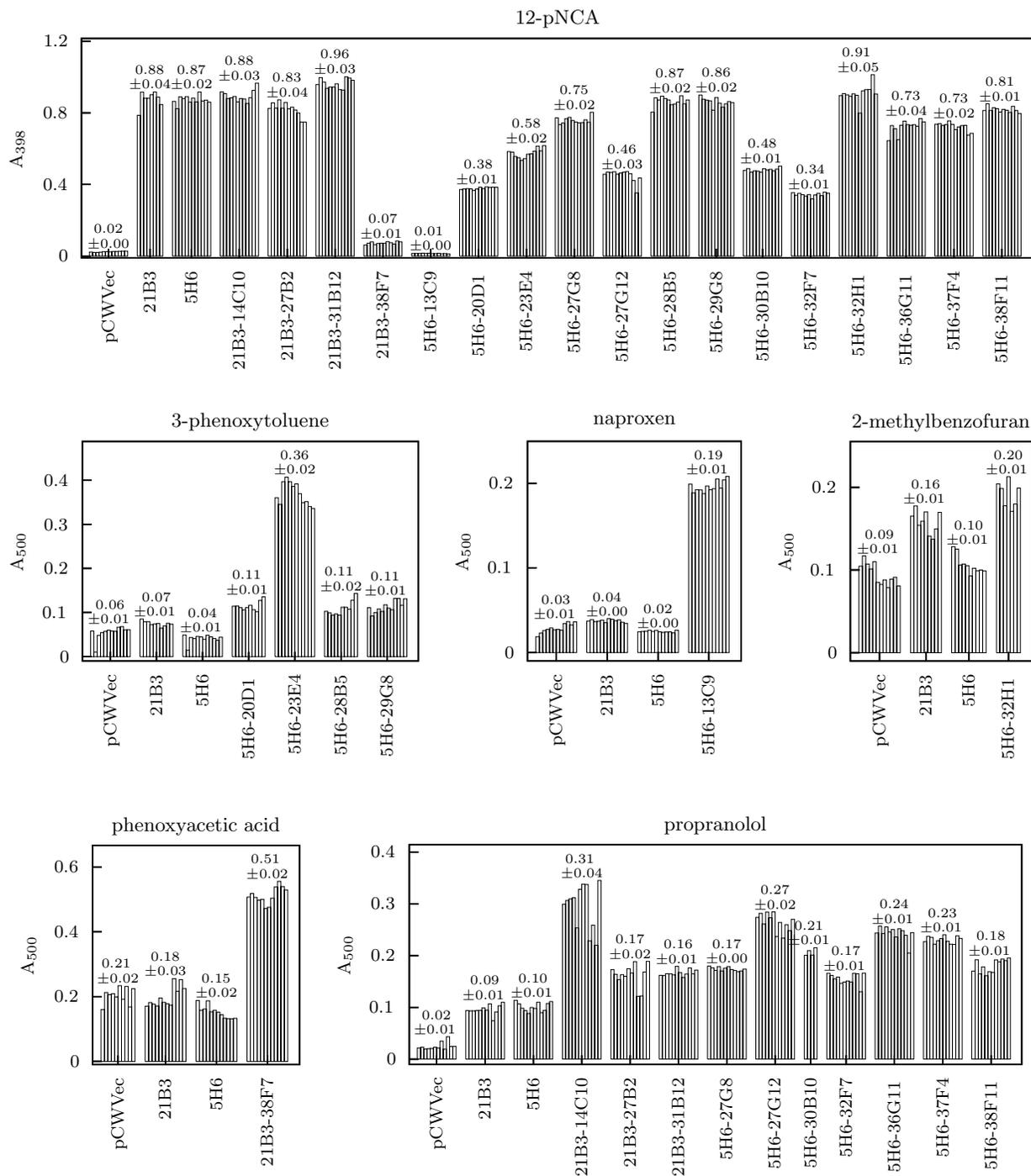


Figure 3.6: Activities of parents and improved mutants. Measurements are for all proteins on 12-pNCA and the substrate on which the mutants were improved as, measured in the verification assays. The top number above each readings indicates the median reading. The second number indicates the standard deviation of the readings.

plates (typically a full row of 12 wells for each mutant, although some samples had fewer wells due to air bubbles or a limited protein supply) and 12-pNCA and 4-AAP assays were performed as described above.

For stability assays, 150 μl of the 5 μM clarified lysate was added to rows of 96-well PCR plates (Bio-Rad). The PCR plates were heated to different temperatures using the gradient method of a PCR cycler (MJ Research, PTC-200) for 10 minutes, then cooled to 4°C. The PCR plates were centrifuged at 5,000-6,000 g for 5-10 minutes at 4°C to pellet denatured debris, and 80 μl of the supernatant was used for CO binding difference spectrum assays. The temperature at which half of the protein was denatured (T_{50}) was determined by fitting a sigmoidal curve to the percentage of remaining CO binding difference spectrum, as shown in Figure 3.7.

3.5.5 Retention of Folding and Activity on 12-pNCA

Our high-throughput screening allowed us to determine whether mutants retained the CO binding difference spectrum peak characteristic of folded P450 heme domains, and whether they retained the high activity on 12-pNCA of both the 21B3 and 5H6 parents. We collected CO binding difference spectra and 12-pNCA activity data for six plates each of 21B3 and 5H6 (522 mutants of each). The CO binding difference spectra and the 12-pNCA readings for these plates are shown in Figure 3.8. Since there was often some variation between plates, we classified the mutants as folded/unfolded and active/inactive relative to the parental controls on the same plate. These binary classifications require defining a threshold for the fraction of the parental reading the mutant must retain. Table 3.2 shows the fractions for different thresholds.

3.5.6 Identification of Mutants with New Activity

We sought to measure the frequencies with which 21B3 and 5H6 mutants acquired activity on 3-phenoxytoluene, naproxen, propranolol, 2-methylbenzofuran, and phenoxyacetic acid. In order to compare the frequencies for the 21B3 and 5H6 libraries, we developed consistent quantitative methods for identifying improved mutants to ensure that both libraries were treated identically. We first performed high-throughput screening of the mutants for activity using the 4-AAP assay. Overall we screened 8,160 mutants each of 21B3 and 5H6 (Table

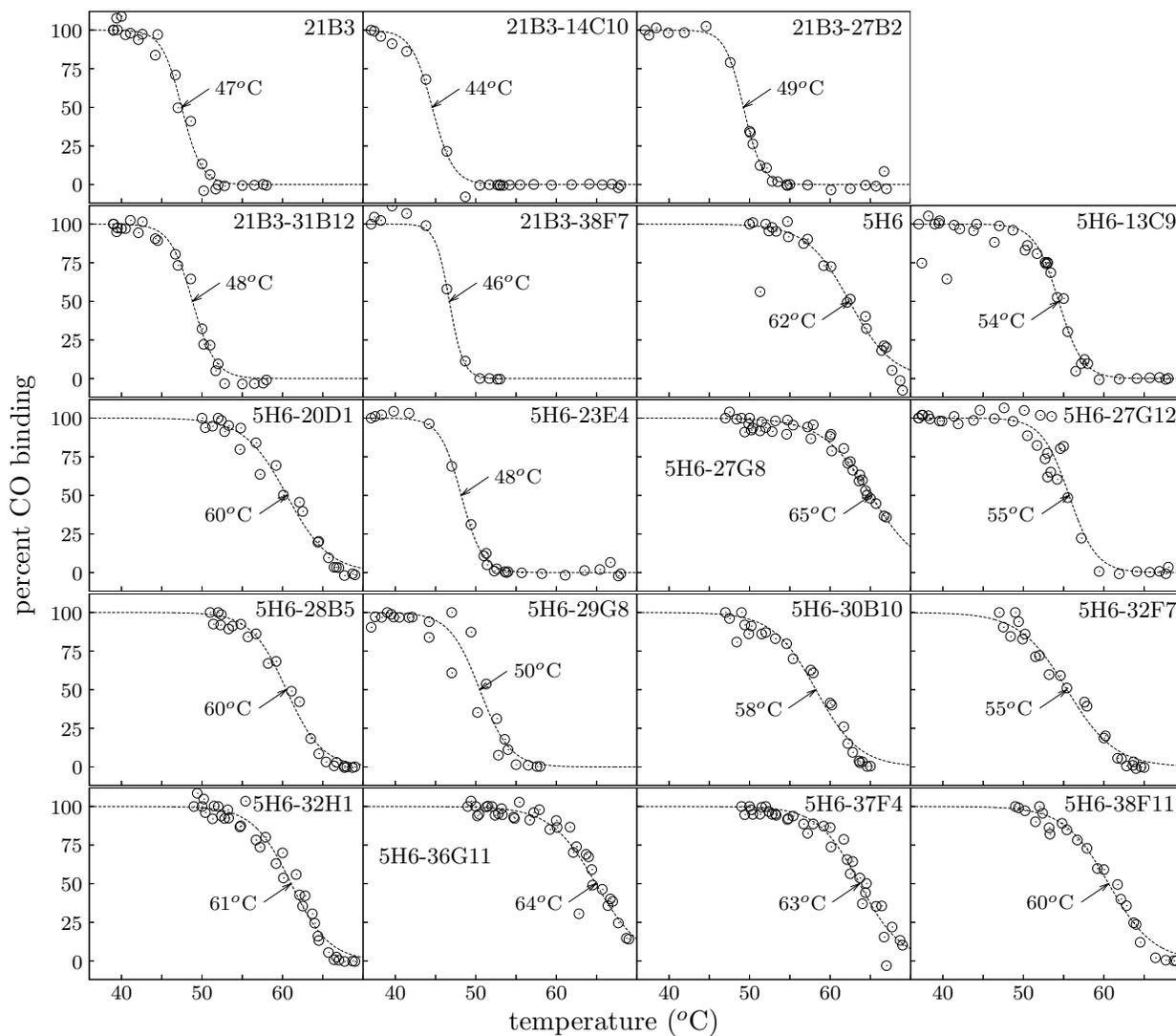


Figure 3.7: Stabilities of improved P450 mutants. Protein stabilities are measured by the percentage of protein that still folds after 10 minute incubations at the indicated temperatures. The T_{50} is the temperature at which 50% of the protein still folds.

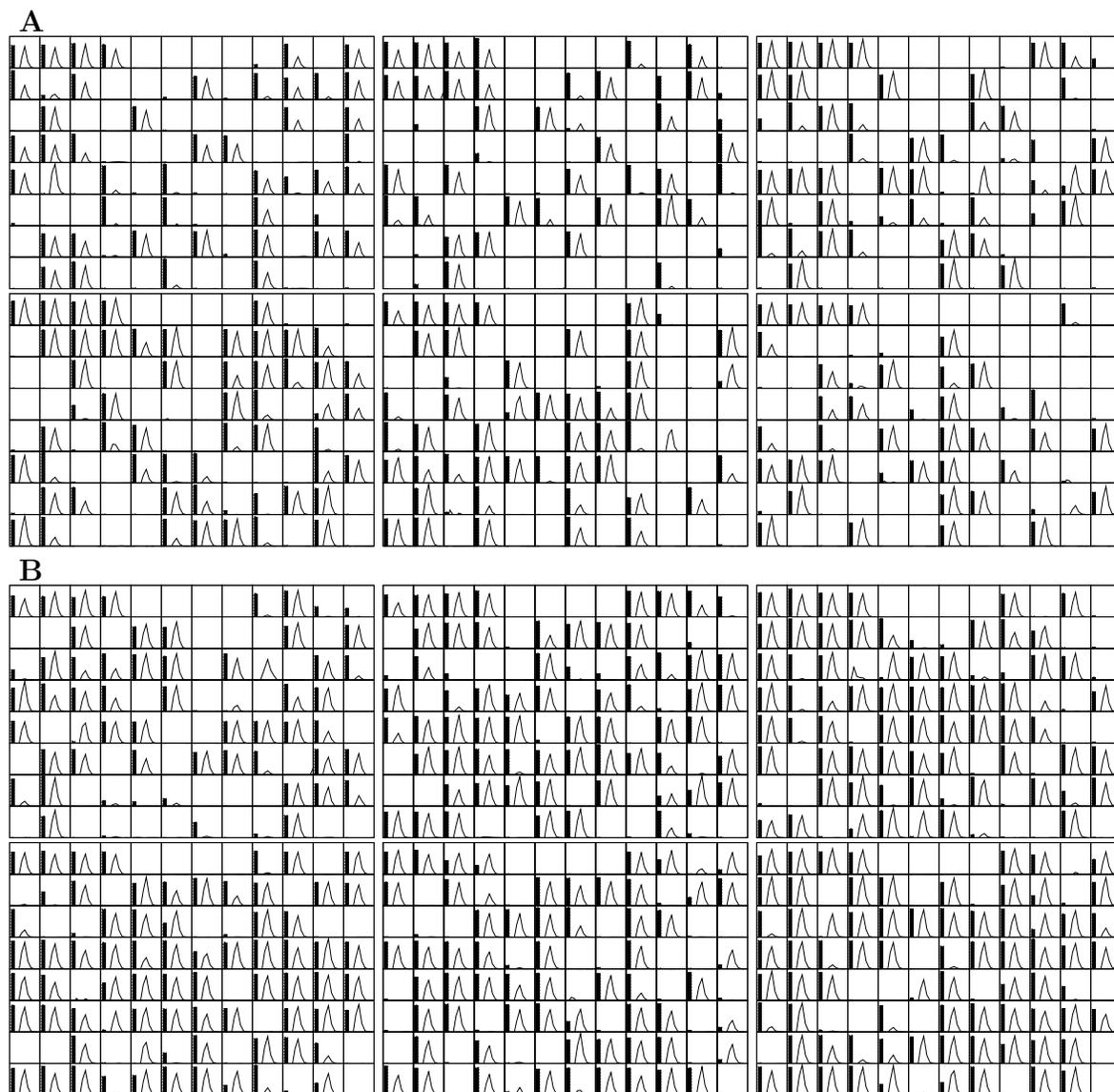


Figure 3.8: Folding and retention of function. (A) Six 96-well plates of 21B3 mutants. (B) Six 96-well plates of 5H6 mutants. Each graph corresponds to a single well, with the lines showing the CO binding difference spectrum from 400 to 500 nm and bars showing the activity on 12-pNCA (A_{398}). For each plate, the first four wells in the top row are the parent, the next four wells are cells containing a null vector, and the last well in the bottom row contains no cells. Axes are scaled so that the A_{490} is at zero and the maximum value is the highest reading for that plate.

substrate	parent	plates screened	mutants screened	mutants rescreened	mutants verified	mutants improved
3-phenoxytoluene	21B3	26	2,210	16	0	0
	5H6	26	2,210	21	4	4
naproxen	21B3	26	2,210	3	0	0
	5H6	26	2,210	8	1	1
propranolol	21B3	14	1,190	9	3	3
	5H6	14	1,190	10	7	7
phenoxyacetic acid	21B3	21	1,785	26	2	1
	5H6	21	1,785	24	1	0
2-methylbenzofuran	21B3	9	765	4	0	0
	5H6	9	765	16	2	1
total	21B3	96	8,160	58	5	4
	5H6	96	8,160	79	15	13

Table 3.4: Summary statistics for the screening of the 21B3 and 5H6 mutants for new activity. The table shows the numbers of mutants screened, the numbers selected for rescreening, the numbers expressed for verification assays, and the numbers that were determined to actually be improved.

3.4). Because of the variation between plates, each mutant was evaluated relative to the four parental and four null vector controls on its own plate. Because of problems with the pipetting robot, two wells in the microtiter plates (A12 and D10) consistently contained air bubbles, and so were disregarded, meaning that 85 mutants were screened for each plate. We identified candidates for improved mutants, then rescreened these candidates to eliminate false positives, and finally performed carefully controlled assays to verify improved mutants. The procedure for this process was as follows:

1. The A_{500} reading for each mutant was compared to the median parental reading and the median null control reading. If it was both 1.5 times greater and 0.01 greater than both medians, then the mutant was considered a candidate to be improved. In a few cases, candidates identified by these criteria were disregarded because they were obviously spurious (noticeable air bubble in the well, no folded protein as indicated by the CO binding difference spectrum for the well, or parental medians were abnormally low due to poor lysis in some wells). Candidates for improved mutants were streaked for single colonies on LB plates supplemented with 100 $\mu\text{g}/\text{ml}$ ampicillin from the 1 ml deep-well plates that had been stored at 4°C.
2. All candidate improved mutants were rescreened by growing new plates using the same

high-throughput screening procedure, but now growing an entire row (12 samples) in the 96-well plate for each mutant. The median reading for the candidate mutant was compared to the median readings for rows of parental and null vector samples. If wells were obviously spurious (air bubbles or poor lysis as indicated by CO binding difference spectrum), they were excluded from the calculations of the medians. If the mutant reading was 0.01 greater than both the parental and null vector samples, and if the difference between the mutant and null control reading was greater than 1.5 times the difference between the mutant and parental reading, then the mutant was considered to have passed the rescreen, and was analyzed on a verification plate.

3. Verification plates were used to gather high quality data, since all protein samples in these plates were adjusted to the same concentration (5 μ M). If the parental reading was significantly higher than the null control reading (indicating some parental activity), the median mutant reading minus the null vector reading was required to be 1.5 times greater than the median parent reading minus the null vector reading. If the parental reading was roughly equal to the null control reading, the median reading was required to be 1.5 times greater than both the parental and null control reading.

Figure 3.6 shows the activities for the parents and all improved mutants as measured on the verification plates. Table 3.4 summarizes the statistics for the different steps of this process. In the end, we identified 13 improved 5H6 mutants and 4 improved 21B3 mutants. These mutants were sequenced using the primers *pCWori_for*, *pCWori_rev*, and *pCWori_rev_clone* to identify the mutations. These mutations are summarized in Table 3.5.

3.5.7 Creation and analysis of site-directed mutants

We examined the effects of parental stability on some of the amino acid substitutions observed in our improved mutants by using site-directed mutagenesis to make these mutations in both the 21B3 and 5H6 parents. We constructed the mutants by PCR overlap extension mutagenesis [88] using the following primers (the induced amino acid substitutions are indicated in the primer names): *mutL75R_for* (5'-CTTAAGTCAAGCGCGTAAATTTGTACGTG-3'), *mutL75R_rev* (5'-CACGTACAAATTTACGCGCTTGACTTAAG-3'), *mutF205L_for* (5'-CAAGCGC CAGCTTCAAGAAGATATCAAGG-3'), *mutF205L_rev* (5'-CCTTGATATCTTCTTGAAGCTGGCGC

Mutant	Substrate	T_{50} ($^{\circ}$ C)	Amino acid substitutions	Synonymous mutations
21B3-14C10	propranolol	44	F205L (T617C, b); H239R (A719G, s); K306N (A921C, s)	T138C, C1255T
21B3-27B2	propranolol	49	S106R (A319C, b); F205S (T617C, b); E434A (A1304C, s)	T700C
21B3-31B12	propranolol	48	A82T (G247A, b, 4Å); A295S (G886T, s)	T243C
21B3-38F7	phenoxyacetic acid	46	N134D (A403G, s); F205L (T618A, b); E267R (G802A & A803G, b, 7Å)	none
5H6-13C9	naproxen	54	L75R (T227G, b, 6Å)	none
5H6-20D1	3-phenoxytoluene	60	A74E (C224A, b, 6Å)	A573G, A1305G
5H6-23E4	3-phenoxytoluene	48	V26A (T80C, b, 6Å); H92R (A278G, s); D151G (A455G, b); L188Q (T566A, b, 5Å); M237L (A712C, b)	A981G, A1014G
5H6-27G8	propranolol	65	F173L (T520C, b)	A318G, G1317A
5H6-27G12	propranolol	55	E35D (A108T, s); K98R (A296G, b); F173V (T520G, b); L437P (T1313C, b, 6Å); K451E (A1354G, s); P461S (C1384T); H468R (A1406G)	A1230G, A1305G
5H6-28B5	3-phenoxytoluene	60	M118V (A355G, s); I357T (T1073C, b, 10Å); F444L (T1333C, b)	T465G, T862C, T955C, T1318C
5H6-29G8	3-phenoxytoluene	50	W90R (T271A, s); F275S (T827C, b)	A1176G, T1227C
5H6-30B10	propranolol	58	P172A (C517G, s); F173I (T520A, b); D199G (A599G, s); D231E (T696A, s); L273V (T820G, b); T274S (A823T, b)	T259C, T399A
5H6-32F7	propranolol	55	S108G (A325G, s); D214Y (G643T, s); V366A (T1100C, s); I433F (A1300T, b)	A873G
5H6-32H1	2-methylbenzofuran	61	A61T (G184A, b)	none
5H6-36G11	propranolol	64	K113E (A340G, s); F173L (T520C, b); E430G (A1292T, s)	none
5H6-37F4	propranolol	63	F173L (T520C, b); D338N (G1015A, s)	A912G
5H6-38F11	propranolol	60	F173L (T520C, b); D217G (A653G, s); E430G (A1292G, s)	T445C, C501T, A798T

Table 3.5: Summary of the improved mutants and the mutations relative to the parental sequence. Substituted amino acids are labelled as surface (s) exposed (>20% exposed surface area) or buried (b). The distance to the substrate in the 1JPZ crystal structure [69] is indicated for all residues within 12Å of the substrate. Amino acid substitutions are labeled using the standard P450 numbering scheme where one is assigned to the threonine after the cleaved N-terminal methionine.

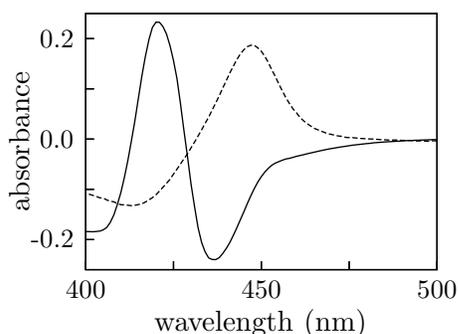


Figure 3.9: The L75R substitution is not tolerated by 21B3. The CO binding difference spectrum of 21B3-L75R (solid line) shows the peak at 420 nm characteristic of an improperly folded P450 sequence that binds heme but is catalytically inactive. The spectrum for the 21B3 parent is shown for comparison (dashed line).

TTG-3'), *mutF275S_for* (5'-GCGGTCTTTTAACATCTGCGCTGTATTTCTTAG-3'), *mutF275S_rev* (5'-CTAAGAAATACAGCGCAGATGTAAAAGACCGC-3'), *pCWori_for*, and *pCWori_rev*. All mutants were sequenced to ensure the genes contained only the desired mutations. Protein was expressed in flasks as for the improved mutants, and clarified lysate was used for functional and thermostability assays as before.

The L75R substitution conferred naproxen activity on the 5H6 parent (mutant 5H6-13C9). This substitution is in the substrate binding pocket, and destabilized the mutant by 8°C. When we made this substitution in the 21B3 parent, we were only able to recover inactive protein with a CO binding difference spectrum peak at 420 nm (Figure 3.9). A CO binding difference spectrum peak at 420 nm is a characteristic of an improperly folded P450 protein [68, 89], indicating that the L75R substitution is too destabilizing to be tolerated by the 21B3 parent.

The F275S substitution was one of two substitutions found in a 5H6 mutant improved on 3-phenoxytoluene (mutant 5H6-29G8). We made the F275S single mutants of both 5H6 and 21B3. The 5H6-F275S mutant is destabilized relative to the parent, but is active on 3-phenoxytoluene (Figure 3.10). We were unable to recover any folded protein for the 21B3-F275S mutant, suggesting that the substitution is too destabilizing to be tolerated by 21B3.

The F205L substitution was found in conjunction with other substitutions in 21B3 mutant improved on propranolol (21B3-14C10). We made this substitution in both 5H6

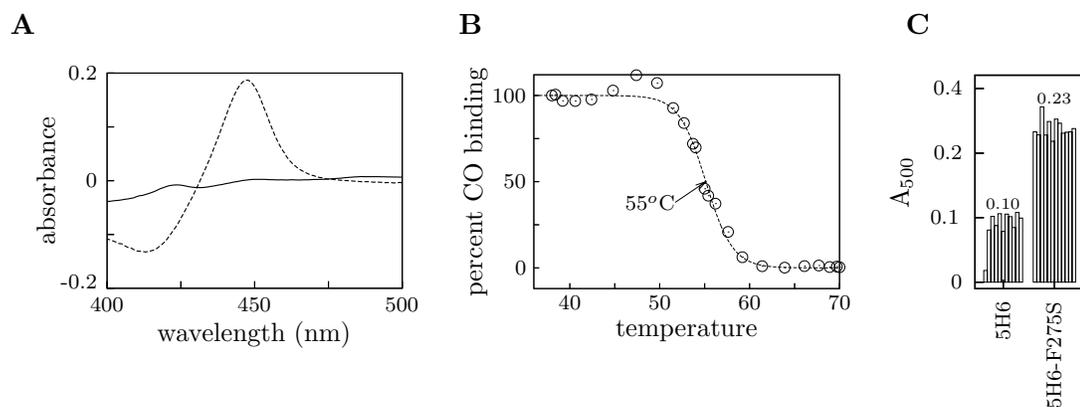


Figure 3.10: The F275S substitution is tolerated only by 5H6. (A) The 21B3-F275S mutant does not yield any folded P450 protein as indicated by the CO binding difference spectrum (solid line). The 21B3 spectrum is shown for comparison (dashed line). (B) The F275S substitution decreases the T_{50} of the 5H6 protein by 7°C to 55°C. (C) The 5H6-F275S mutant is active on 3-phenoxytoluene as measured in the 4-AAP assay.

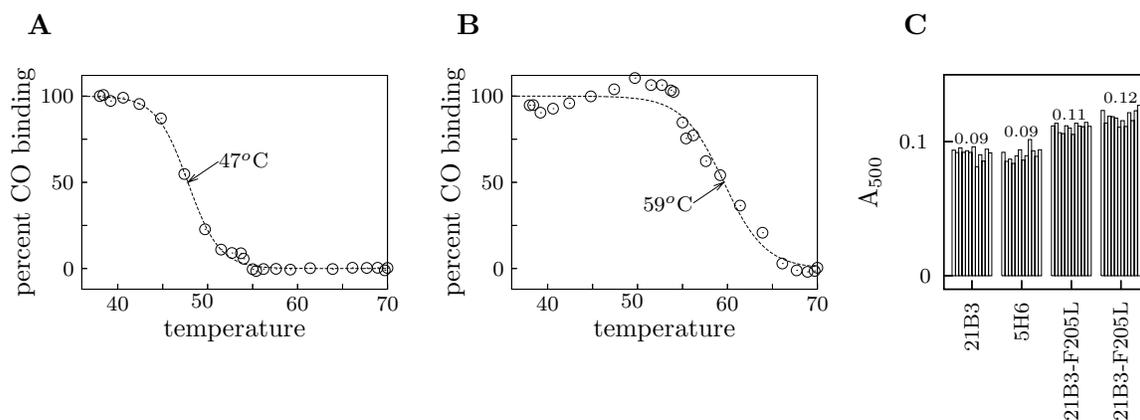


Figure 3.11: The F205L substitution improves the propranolol activity of both 21B3 and 5H6. (A) The 21B3-F205L mutant has the same T_{50} as the 21B3 parent protein. (B) The 5H6-F205L mutant has a T_{50} that is only slightly decreased from that of the 5H6 parent. (C) The F205L substitution slightly increases the propranolol activity of both 21B3 and 5H6 as measured in the 4-AAP assay.

and 21B3. Both mutants folded, and neither was substantially destabilized. Both mutants had slightly improved activity on propranolol, indicating that this beneficial substitution is tolerated by both parents (Figure 3.11).

3.5.8 P450 stability: relationship of T_{50} values to protein stability

In this work, we have argued that more stable variants of a protein should be more tolerant to mutations and therefore more evolvable. Our biophysical arguments and our lattice protein simulations are both based on consideration of the equilibrium thermodynamic stability of the native structure, ΔG_f . It is mutational changes to this true thermodynamic stability ($\Delta\Delta G$ values) that are expected to be largely additive [28], and so ΔG_f is the proper basis for the formulation of our arguments. However, our experiments with the P450 variants do not measure ΔG_f , but instead measure T_{50} , defined as the temperature at which half of the protein irreversibly denatures after at 10 minute incubation. If we were able to measure true melting temperatures (T_m) for reversible thermal denaturation, then we could compellingly argue these T_m values closely correspond with ΔG_f , since both experimental and thermodynamic considerations [90] suggest that changes in T_m are linearly correlated with changes in ΔG_f for proteins with the same structure and high sequence identity (as is the case for all of our mutants). However, our T_{50} values measure the extent of irreversible thermal denaturation, which can be affected by factors other than ΔG_f (such as the kinetics of unfolding, aggregation, or chemical modification of side chains). In this section we discuss why we used T_{50} values, and show that these T_{50} values are highly correlated with the stability to irreversible chemical denaturation.

In order to measure ΔG_f , it is necessary to find a reversible method for denaturing the protein. Thermal denaturation was irreversible for all of our P450 variants (at least in the buffer conditions and at the level of protein purity of our assays), as can be seen from Figure 3.7, which shows that none of the mutants refolded when exposed to sufficiently high temperatures. If chemical denaturation is reversible, ΔG_f scales linearly with the concentration of chemical denaturant, and the slope (m value) is the same for proteins of the same structure and high sequence identity [91]. Therefore, for reversible chemical denaturation, ΔG_f is linearly related to the concentration of denaturant at which half of the protein unfolds. Other researchers have performed chemical denaturation studies on various cytochromes

P450. They have generally found that denaturation by guanidinium chloride is irreversible even at low concentrations [92, 93], but that urea denaturation is sometimes reversible at low urea concentrations before becoming irreversible at higher concentrations [92, 94]. In general, P450 denaturation appears to proceed through multiple intermediates and pathways [92, 93, 94], perhaps explaining why the reversibility of denaturation is so sensitive to the particular conditions and P450 variants used.

We tested the stability of 12 of our P450 BM3 heme domain variants to urea denaturation. We began by obtaining samples of the 12 variants as for the thermal denaturation studies. We then adjusted these samples to roughly 5 μM in the buffer of 100 mM EPPS (pH 8.2). Half of each sample was diluted to 1.8 μM with buffer, while the other half was diluted in the same fashion with 8 M urea in 100 mM EPPS (pH 8.2), resulting in a final urea concentration of 5.1 M. Both samples were incubated overnight at 13°C, and then CO difference spectra were measured. Figure 3.12A shows that all 12 samples were completely or nearly completely denatured by the urea treatment, as indicated by the disappearance of the Soret peak at 450 nm. To see if the denaturation was reversible, we removed the urea from six of the variants by dialysis against a 1000-fold excess of buffer. None of the variants exhibited significant refolding (Figure 3.12A), indicating that urea denaturation was irreversible for our P450 variants and assay conditions.

To measure the stability of our variants to irreversible urea denaturation, we added 75 μl samples of the 5 μM P450 samples to rows of 96-well microtiter plates. We then used the pipetting robot to add and mix 200 μl of solutions of various concentrations of urea in our buffer so that the final urea concentrations were those shown on the x-axes of the plots in Figure 3.12B. The microtiter plates were incubated overnight at 13 °C. We then added 25 μl of 0.1 M sodium hydrosulfite in 1.3 M potassium phosphate (pH 8.0) and read the CO difference spectra. Figure 3.12B shows that the Soret peak diminished with increasing urea concentrations. We fit sigmoidal curves to these denaturation plots, and quantified the stability to irreversible urea denaturation as the concentration of urea at which half of the protein had unfolded ($[\text{urea}]_{50}$). We then compared the $[\text{urea}]_{50}$ values for these 12 variants to the T_{50} values measured in Figure 3.7. Figure 3.12 shows that the $[\text{urea}]_{50}$ and T_{50} values are linearly correlated. Therefore, although we are unable to measure ΔG_f , our measures of stability to irreversible thermal and urea denaturation are nearly the same, supporting

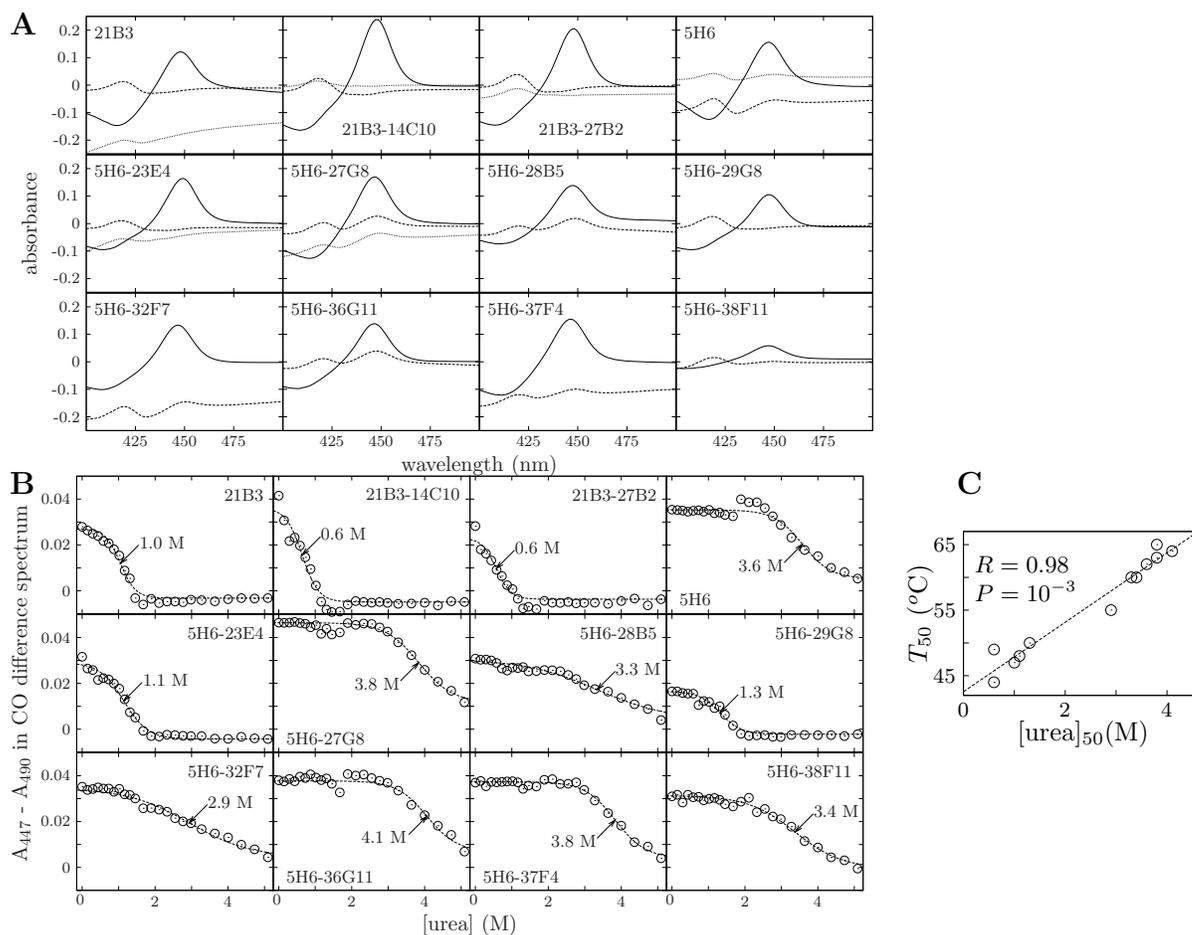


Figure 3.12: The P450 stabilities to irreversible urea denaturation are highly correlated to the T_{50} values. (A) All 12 tested P450 variants exhibited the characteristic Soret peak at 450 nm in buffer (solid lines), but were denatured by 5.1 M urea (dashed lines). The denaturation was irreversible for the 6 variants which we tried to refold by removing the urea by dialysis (dotted lines). (B) Urea denaturation curves, as measured by the Soret peak after treatment with the indicated concentration of urea. Arrows indicate the concentration of urea ($[\text{urea}]_{50}$) at which half of the protein was unfolded. (C) The $[\text{urea}]_{50}$ values are highly correlated with the T_{50} values for the 12 variants, with a Pearson correlation coefficient of $R = 0.98$ and a probability of $P = 10^{-3}$ that a correlation this strong would occur by chance.

the view that the T_{50} values reflect universal aspects of protein stability rather than unique characteristics of the process of irreversible thermal denaturation.

3.6 Acknowledgments

We thank M. Landwehr for bountiful advice and the synthesis of 12-pNCA and C. O. Wilke for helpful advice. J.D.B. is supported by a Howard Hughes Medical Institute pre-doctoral fellowship. This work was supported by NIH Grant R01 GM068664-01. This work was inspired in part by a Santa Fe Institute workshop on Evolutionary Innovations supported by a grant from the Packard Foundation.

Chapter 4

Thermodynamics of Neutral Protein Evolution

A version of this chapter has been published as [95].

4.1 Abstract

Naturally evolving proteins gradually accumulate mutations while continuing to fold to stable structures. This process of neutral evolution is an important mode of genetic change, and forms the basis for the molecular clock. We present a mathematical theory that predicts the number of accumulated mutations, the index of dispersion, and the distribution of stabilities in an evolving protein population from knowledge of the stability effects ($\Delta\Delta G$ values) for single mutations. Our theory quantitatively describes how neutral evolution leads to marginally stable proteins, and provides formulae for calculating how fluctuations in stability can overdisperse the molecular clock. It also shows that the structural influences on the rate of sequence evolution observed in earlier simulations can be calculated using just the single-mutation $\Delta\Delta G$ values. We consider both the case when the product of the population size and mutation rate is small and the case when this product is large, and show that in the latter case the proteins evolve excess mutational robustness that is manifested by extra stability and an increase in the rate of sequence evolution. All our theoretical predictions are confirmed by simulations with lattice proteins. Our work provides a mathematical foundation for understanding how protein biophysics shapes the process of evolution.

4.2 Introduction

Proteins evolve largely through the slow accumulation of amino acid substitutions. Over evolutionary time, this process of sequence divergence creates homologous proteins that differ at the majority of their residues, yet still fold to similar structures that often perform conserved biochemical functions [56]. The maintenance of structure and function during sequence divergence suggests that much of protein evolution is neutral in the sense that observed sequence changes frequently do not alter a protein’s ability to fold and adequately perform the biochemical function necessary to enable its host organism to survive. This comparative evidence for neutrality in protein evolution has been corroborated by experimental studies showing that the mutations separating diverged sequences often have no effect other than modest and additive changes to stability [76], and that a large fraction of random mutations do not detectably alter a protein’s structure or function [20, 24, 23, 18, 7, 49]. In this respect, it seems that protein evolution should be well described by Kimura’s neutral theory of evolution, which holds that most genetic change is due to the stochastic fixation of neutral mutations [96]. One of the key predictions of the neutral theory is that assuming a constant mutation rate, the number of mutations separating two proteins should be proportional to the time since their divergence [96]. Indeed, the observation by Zuckerkandl and Pauling [3] that proteins are “molecular clocks” that accumulate mutations at a roughly constant rate has long been taken as one of the strongest pieces of evidence supporting the neutral theory [97].

However, mutations that are neutral with respect to a protein’s capacity to perform its biological function often affect protein thermodynamics. The biological functions of most proteins depend on their ability to fold to thermodynamically stable native structures [25]. Yet natural proteins are typically only marginally stable, with free energies of folding (ΔG_f) between -5 and -15 kcal/mol [26]. Most random mutations to proteins are destabilizing [77, 24, 79, 98], and their effects on stability (measured as $\Delta\Delta G$, the ΔG_f of the mutant protein minus the ΔG_f of the wildtype protein) are frequently of the same magnitude as a protein’s net stability. The impact of a mutation on a protein’s function can therefore depend on the protein’s stability: a moderately destabilizing mutation that is easily tolerated by a stable parent protein may completely disrupt the folding of a less stable parent. This effect of

protein stability on mutational tolerance has been verified by experiments demonstrating that more stable protein variants are markedly more robust to random mutations [7, 49].

The fact that mutations that are neutral with respect to direct selection for protein function can affect a protein’s tolerance to subsequent mutations is not consistent with the simplest formulation of the neutral theory of evolution, which tends to assume that the fraction of mutations that is neutral remains constant in time. Kimura [99] himself recognized the possibility that the neutrality might change, and Takahata [100] mathematically treated the consequences of a “fluctuating neutral space”. In particular, Takahata showed that fluctuating neutrality could explain the observed overdispersion in the molecular clock [101] (the tendency for the variance in the number of fixed mutations to exceed the expectation for the Poisson process predicted by the neutral theory) long considered troublesome for the neutral theory. However, further progress on this topic was stymied by the lack of a specific model for how or why protein neutrality might fluctuate.

More recently, researchers have preferred to describe neutral evolution using the concept of “neutral networks,” which are networks in the space of possible protein sequences in which each functional protein is linked to all other functional proteins that differ by only a single mutation [102, 103, 104, 105, 9, 106, 107]. A neutrally evolving protein population is then envisioned as moving on the neutral network, and the neutrality of the population may fluctuate if the nodes on the network differ in their connectivities. A general theoretical treatment of evolution on neutral networks by van Nimwegen and coworkers [105] has shown that if the product of the population size and mutation rate is small then members of the population are equally likely to occupy any node, while if this product is large then the population will preferentially occupy highly connected nodes (see also [9, 108, 11]). Simulations with simplified lattice models of proteins have attempted to provide insight into the specific features of protein neutral networks. These simulations have shown that lattice protein neutral networks are centered around highly connected nodes occupied by stable proteins [9, 11, 10, 8], a finding consistent with the experimental observation [7, 49] that stable proteins are more mutationally robust. Lattice protein studies also suggest that protein structures differ in their “designabilities” (defined as the total number of sequences that fold into a structure), and that sequences that fold into more designable structures will neutrally evolve at a faster rate due to the increased size and connectivity of their

neutral networks [104, 12, 43, 34, 10]. Finally, simulations have demonstrated that fluctuations in neutrality as a protein population moves along its neutral network can lead to an overdispersion of the molecular clock [107], as originally suggested by Takahata. However, an extension of these lattice protein simulations of evolution on neutral networks into a quantitative theory has been difficult because protein neutral networks are far too large to be computed for all but the simplest lattice models.

Here we present a mathematical treatment of neutral protein evolution that describes the evolutionary dynamics in terms of the $\Delta\Delta G$ values for single mutations, which are experimentally measurable. Our treatment is based on the experimentally verified [7, 49] connection between protein stability and mutational robustness, as well as a few biophysically supported assumptions about $\Delta\Delta G$ values for random mutations. By linking a protein's tolerance to mutations with stability, we are able to quantitatively describe neutral evolution without a full description of the neutral network. We can then compute the average number of accumulated mutations, the average fraction of neutral mutations, the index of dispersion, and the distribution of stabilities in a neutrally evolving population solely from knowledge of the $\Delta\Delta G$ values for single mutations. In addition, we follow the formalism of van Nimwegen and coworkers [105] to calculate all four of these properties in the limit when the product of the population size and mutation rate is much less than one and in the limit when this product is much greater than one. In demonstrating that these properties are different in these two limits, we show that the rate of fixation of neutral mutations can vary with population size in violation of one of the standard predictions of Kimura's neutral theory [99]. Our work presents a unified view of neutral protein evolution that is grounded in measureable thermodynamic quantities.

4.3 Results

4.3.1 Assumptions and Mathematical Background

In this section we describe the physical view of protein evolution that motivates our work. We begin with the basic observations that evolution selects for protein function, and that most proteins must stably fold in order to function [25], meaning that protein stability is under evolutionary pressure only insofar as it must be sufficient to allow a protein to

fold and function. In taking this view, we ignore those proteins (estimated at 10% of prokaryotic and 30% of eukaryotic proteins) that are intrinsically disordered [109], as well as those rare proteins that are only kinetically stable [110]. Natural selection for function requires a protein to fold with some minimal stability ΔG_f^{\min} , since proteins that lack this minimal stability will be unable to reliably adopt their native structure and perform their biochemical task. A protein's extra stability beyond this minimal threshold is quantified as $\Delta G_f^{\text{extra}} = \Delta G_f - \Delta G_f^{\min}$, meaning that all functional proteins must have $\Delta G_f^{\text{extra}} \leq 0$ (more negative values of ΔG_f indicate increased stability). We further assume that as long as $\Delta G_f^{\text{extra}} \leq 0$, natural selection for protein function is indifferent to the exact amount of extra stability a protein possesses. This assumption is at odds with the persistent speculation that high stability inherently impairs protein function and so is selected against by evolution [72, 70]. But the circular argument most commonly advanced to support this speculation — that the observed marginal stability of natural proteins indicates that higher stability is detrimental to protein function — has now been contradicted both by experiments that have dramatically increased protein stability without sacrificing function [76, 73, 74, 75] and by demonstrations that marginal stability is a simple consequence of the fact that most mutations are destabilizing [41, 83, as well as the current work]. There is a possibility, however, that certain regulatory proteins must be marginally stable to facilitate rapid degradation [111]. To summarize, current biochemical evidence supports our assumption that (with certain well-defined exceptions) the only requirement imposed on protein stability by natural selection for protein function is that stability must meet or surpass some minimal threshold (a protein must have $\Delta G_f^{\text{extra}} \leq 0$).

A mutation to a protein changes its stability by an amount $\Delta\Delta G$, and experimental measurements of $\Delta\Delta G$ values have shown that most mutations are destabilizing (have $\Delta\Delta G > 0$) [24, 77, 79, 98]. A mutation is neutral with respect to selection for stability if $\Delta\Delta G + \Delta G_f^{\text{extra}} \leq 0$ since the mutant protein still satisfies the minimal stability threshold; otherwise the mutant does not stably fold and is culled by natural selection. Of course, mutations can also have specific effects on protein function (such as altering an enzyme's activity), but experiments have shown that such mutations are rare compared to the large number of mutations that affect stability [20, 24, 23, 49]. Mutations can also have effects unrelated to the functioning of the individual protein molecule: they can affect its propensity

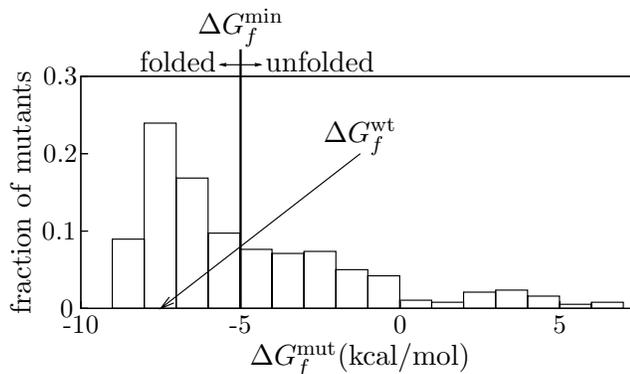


Figure 4.1: A thermodynamic view of protein evolution. A mutant protein stably folds if and only if it possesses some minimal stability, ΔG_f^{\min} (in this case -5 kcal/mol). The stability of the wildtype protein is $\Delta G_f^{\text{wt}} = -7.5$ kcal/mol, meaning that it has $\Delta G_f^{\text{extra}} = -2.5$ kcal/mol of extra stability. The bars show the distribution of $\Delta\Delta G$ values for mutations. Those mutants with $\Delta G_f^{\text{extra}} + \Delta\Delta G \leq 0$ still stably fold, while all other mutants do not fold and so are culled by natural selection. The probability that a mutation will be neutral with respect to stable folding is simply the fraction of the distribution that lies to the left of the threshold. The data in this figure are hypothetical.

to aggregate [5], alter its codon usage [112], change its mRNA stability [113], affect the efficiency or accuracy of translation [112, 114], or change the fraction of mistranslated proteins that fold [85]. These higher-level effects are probably most apparent in the evolution of highly expressed proteins [85, 115]. However, here we ignore such effects and assume that the evolutionary impact of a mutation is mostly determined by its effect on protein stability (an assumption in agreement with a recent bioinformatics analysis [116]). The view we present therefore describes the impact of a mutation solely by its $\Delta\Delta G$ value and the $\Delta G_f^{\text{extra}}$ of the wildtype protein, and is summarized graphically in Fig. 4.1 We have previously used a similar view to successfully describe experimental protein mutagenesis results [7, 49].

To use the view of Fig. 4.1 to construct a useful description of neutral protein evolution, we make one major assumption: that the overall distribution of $\Delta\Delta G$ values for random mutations stays roughly constant as the protein sequence evolves. Actually, this assumption is stronger than is strictly needed for the mathematical theory presented below — the theory can be developed simply by assuming that all proteins with the same ΔG_f have the same distribution of $\Delta\Delta G$ values (in this case the matrix elements W_{ij} introduced below depend on j in addition to the difference $i - j$). However, we make the stronger assumption that the $\Delta\Delta G$ distribution remains constant during sequence evolution, since

we believe that this assumption is consistent with existing evidence. We emphasize that this assumption does not imply that we are arguing that the $\Delta\Delta G$ distribution is identical for every possible protein sequence. Clearly, for any given structure there is a most stable sequence (with all $\Delta\Delta G$ values positive), a least stable sequence (with all $\Delta\Delta G$ values negative), and a vast range of sequences in between. However, most of these sequences fall within a stability range that is never populated by evolution, since simulations [41] and experiments [117, 118] clearly show that the vast majority of protein sequences do not stably fold into any structure (meaning the least stable folded protein is still far more stable than the typical random sequence). Among the subset of sequences that do stably fold, the simple statistical reality that marginally stable sequences are far more abundant than highly stable sequences causes evolution to further confine itself mostly to sequences with stabilities far less than that of the most stable sequence [41, 83, , as well as the current work]. This fact is amply demonstrated by engineering experiments that have greatly increased the stability of natural proteins without sacrificing any of their functional properties [76, 73, 74, 75]. Therefore, although the distribution of $\Delta\Delta G$ values certainly varies widely among all sequences, it is still reasonable to assume that it is relatively constant among those sequences visited by natural evolution. This assumption of a constant $\Delta\Delta G$ distribution among evolved sequences is explicitly supported by simulations [7, 49, 8, 119, , as well as the current work], and is consistent with the observation that the number of neighbors on a protein’s neutral network is approximately determined by its stability [9, 11]. Furthermore, protein mutagenesis experiments indicate that the $\Delta\Delta G$ values for random mutations are usually additive [76, 28], meaning that any given mutation to a protein of length L will alter only $\approx 1/L$ of the other $\Delta\Delta G$ values, leaving the $\Delta\Delta G$ distribution mostly unchanged. Finally, the assumption of a constant $\Delta\Delta G$ distribution has been shown to explain the experimentally observed exponential decline in the fraction of functional proteins with increasing numbers of mutations [7]. However, we acknowledge that at present the assumption of a roughly constant $\Delta\Delta G$ distribution among neutrally evolving proteins can be verified only for lattice proteins — for real proteins the most we can say is that it is consistent with existing experimental evidence.

We begin our mathematical treatment by conceptually dividing the continuous variable of protein stability into small discrete bins of width b . This discretization of stability allows

us to treat mutations as moving a protein from one bin to another — the bins can be made arbitrarily small to eliminate any numerical effects of the binning. The stability of each folded protein in the evolving population (the folded proteins are all those with $\Delta G_f^{\text{extra}} \leq 0$) can be described by specifying its stability bin. Specifically, a protein is in bin i if it has $\Delta G_f^{\text{extra}}$ between $(1-i)b$ and $-ib$, where $i = 1, 2, \dots$. Let W_{ij} be the probability that a random mutation has a $\Delta\Delta G$ value such that it moves a protein's stability from bin j to bin i , where i and j both are in the range $1, 2, \dots$. Then W_{ij} is easily computed as the fraction of $\Delta\Delta G$ values between $b(j-i-1)$ and $b(j-i)$. Since W_{ij} only describes transitions between folded proteins, and since we have assumed that a protein's mutational tolerance is determined by its stability, then the fraction of folded mutants (neutrality) of a protein in bin j is $\nu_j = \sum_i W_{ij}$. Clearly, more stable proteins will have larger values of ν_j .

In the next two sections, we will use the matrix \mathbf{W} with elements W_{ij} to calculate the distribution of stabilities in an evolving protein population of constant size N , the mean number of mutations $\langle m \rangle_T$ after T generations, the corresponding index of dispersion $R_T = \frac{\langle m^2 \rangle_T - \langle m \rangle_T^2}{\langle m \rangle_T}$, and the average fraction of mutations $\langle \nu \rangle$ that do not destabilize the proteins past the minimal stability threshold. We assume that \mathbf{W} is computed from the distribution of $\Delta\Delta G$ values for all random single amino-acid mutations, although in principle it could be for any type of mutation. We also assume that the per-protein-per-generation mutation rate μ is small, so that at each generation a protein undergoes at most one mutation. Our calculations at first follow, and then extend the theoretical treatment by van Nimwegen and coworkers [105] of evolution on a neutral network. In particular, we follow their lead in separately treating the two limiting cases where the product $N\mu$ of the population size and mutation rate is $\ll 1$ and $\gg 1$. We emphasize that all of the equations derived in the next two sections depend only on the mutation rate μ , the number of generations T , and the matrix \mathbf{W} which can be computed from the single-mutant $\Delta\Delta G$ values. The population size N determines the applicable limiting case, but otherwise drops out of all final results.

4.3.2 Limit when $N\mu \ll 1$

When $N\mu \ll 1$, the evolving population is usually clonal, since each mutation is either lost or goes to fixation before the next mutation occurs. If a mutation destabilizes a protein in

the population beyond the stability cutoff, then it is immediately culled by natural selection. If a mutation does not destabilize a protein beyond the stability cutoff, it will be lost to genetic drift with probability $\frac{N-1}{N}$ and go to fixation with probability $1/N$ [96]. Since mutations occur rarely ($N\mu \ll 1$), the loss or fixation of the mutant will occur before the next mutant appears in the population. The entire population therefore moves as one entity along its neutral network. The population can thus be described by the column vector $\mathbf{p}(t)$, with element $p_i(t)$ giving the probability that the population is in stability bin i at time t .

If the population is initially in stability bin j , at each generation there is a probability $N\mu W_{ij}$ that a protein experiences a mutation that changes its stability to bin i , and if such a mutation occurs, then there is a probability of $1/N$ that it is eventually fixed in the population. Therefore, at each generation there is a probability μW_{ij} that the population experiences a mutation that eventually causes it to move from stability bin j to bin i . If we define the matrix \mathbf{V} so that the diagonal elements are given by $V_{ii} = \nu_i$ and all other elements are zero, then \mathbf{p} evolves according to

$$\mathbf{p}(t+1) = (\mathbf{I} - \mu\mathbf{V} + \mu\mathbf{W})\mathbf{p}(t) \quad (4.1)$$

where \mathbf{I} is the identity matrix. Note that this equation treats lethal mutations (those that destabilize a protein beyond the cutoff) as immediately being lost to natural selection and so leaving the population in its original stability bin (hence the population accumulates a mutation with probability $\mu\mathbf{V}$ rather than probability μ). Equation 4.1 describes a Markov process with the non-negative, irreducible, and acyclic transition matrix $\mathbf{A} = \mathbf{I} - \mu\mathbf{V} + \mu\mathbf{W}$, and so \mathbf{p} approaches the unique stationary distribution \mathbf{p}_o satisfying

$$0 = (\mathbf{V} - \mathbf{W})\mathbf{p}_o. \quad (4.2)$$

This equation gives the expected distribution of protein stabilities solely in terms of the single-mutant $\Delta\Delta G$ values.

We now calculate the average number of mutations $\langle m \rangle_{T,o}$ that accumulate in an equilibrated population after T generations and the corresponding index of dispersion $R_{T,o}$. We emphasize that $\langle m \rangle_{T,o}$ represents the average number of accumulated mutations during the course of the evolutionary process. When the number of accumulated mutations m is

small compared to the length of the protein sequence L ($m \ll L$), then m is just equal to the number of residues differing from those in the parent protein sequence (the Hamming distance). However, when m becomes substantial relative to L , m becomes larger than the Hamming distance since some sites will undergo multiple mutations [120]. In this case it is necessary to use a substitution model to infer m from the observed Hamming distance. In the treatment that follows, we calculate the expected value of m ; application of these formulae to actual protein sequences requires use of one of the well-established statistical techniques for inferring m from the Hamming distance [120, 121]. We begin the calculation of $\langle m \rangle_{T,o}$ by defining $\mathbf{p}(m, t)$ to be the column vector with element i giving the probability that at time t the population has accumulated m mutations and is in stability bin i . The time evolution of $\mathbf{p}(m, t)$ is given by

$$\mathbf{p}(m, t+1) = (\mathbf{I} - \mu\mathbf{V})\mathbf{p}(m, t) + \mu\mathbf{W}\mathbf{p}(m-1, t). \quad (4.3)$$

The k th moment of the number of mutations at time t is

$$\langle m^k \rangle_t = \mathbf{e} \sum_m m^k \mathbf{p}(m, t), \quad (4.4)$$

where $\mathbf{e} = (1, \dots, 1)$ is the unit row vector. We can write a recursive equation for $\langle m \rangle_t$ in the long-time limit (steady state) by multiplying both sides of Equation 4.3 by m , summing over m , and left multiplying by \mathbf{e} to obtain

$$\begin{aligned} \langle m \rangle_{t+1} &= \mathbf{e}(\mathbf{I} - \mu\mathbf{V}) \sum_m m \mathbf{p}(m, t) + \mu\mathbf{e}\mathbf{W} \sum_m m \mathbf{p}(m-1, t) \\ &= \mathbf{e}\mathbf{A} \sum_m m \mathbf{p}(m, t) + \mu\mathbf{e}\mathbf{W}\mathbf{p}_o \\ &= \langle m \rangle_t + \mu\langle \nu \rangle_o, \end{aligned} \quad (4.5)$$

where we have used the property $\mathbf{e}\mathbf{A} = \mathbf{e}$, noted that in the long-time limit $\sum_m \mathbf{p}(m, t) = \mathbf{p}_o$ and $\sum_m m \mathbf{p}(m-1, t) = \sum_m [(m-1)\mathbf{p}(m-1, t) + \mathbf{p}(m-1, t)] = \sum_m m \mathbf{p}(m, t) + \mathbf{p}_o$, and defined the average neutrality as $\langle \nu \rangle_o = \mathbf{e}\mathbf{W}\mathbf{p}_o = \mathbf{e}\mathbf{V}\mathbf{p}_o$. Summing the recursion yields the

steady-state value for the number of accumulated mutations,

$$\langle m \rangle_{T,o} = T\mu\langle \nu \rangle_o. \quad (4.6)$$

To calculate the index of dispersion $R_{T,o} = \frac{\langle m^2 \rangle_{T,o} - \langle m \rangle_{T,o}^2}{\langle m \rangle_{T,o}}$, we need to find the second moment $\langle m^2 \rangle_{T,o}$. In a fashion analogous to the construction of Equation 4.5, we can write a recursive expression for the long-time limit of $\langle m^2 \rangle_{T,o}$ as

$$\begin{aligned} \langle m^2 \rangle_{t+1} &= \mathbf{e}(\mathbf{I} - \mu\mathbf{V}) \sum_m m^2 \mathbf{p}(m, t) + \mu \mathbf{e}\mathbf{W} \sum_m m^2 \mathbf{p}(m-1, t) \\ &= \mathbf{e}\mathbf{A} \sum_m m^2 \mathbf{p}(m, t) + 2\mu \mathbf{e}\mathbf{W} \sum_m m \mathbf{p}(m, t) + \mu \mathbf{e}\mathbf{W}\mathbf{p}_o \\ &= \langle m^2 \rangle_t + 2\mu \mathbf{e}\mathbf{W} \left[\mathbf{A} \sum_m m \mathbf{p}(m, t-1) + \mu \mathbf{W}\mathbf{p}_o \right] + \mu\langle \nu \rangle_o \\ &= \langle m^2 \rangle_t + 2\mu^2 \mathbf{e}\mathbf{W} \sum_{\tau=0}^{t-1} \mathbf{A}^\tau \mathbf{W}\mathbf{p}_o + \mu\langle \nu \rangle_o \end{aligned} \quad (4.7)$$

where we have used the property (implicit in Equation 4.5) that in the long-time limit, $\sum_m m \mathbf{p}(m, t) = \mathbf{A} \sum_m m \mathbf{p}(m, t-1) + \mu \mathbf{W}\mathbf{p}_o$. Summing the recursion yields the following value for the long-time limit,

$$\begin{aligned} \langle m^2 \rangle_{T,o} &= T\mu\langle \nu \rangle_o + 2\mu^2 \mathbf{e}\mathbf{W} \sum_{t=0}^{T-1} \sum_{\tau=0}^{t-1} \mathbf{A}^\tau \mathbf{W}\mathbf{p}_o \\ &= T\mu\langle \nu \rangle_o + 2\mu^2 \mathbf{e}\mathbf{W} \sum_{t=1}^T (T-t) \mathbf{A}^{t-1} \mathbf{W}\mathbf{p}_o \\ &= T\mu\langle \nu \rangle_o + T(T-1)\mu^2 \langle \nu \rangle_o^2 + 2\mu^2 \mathbf{e}\mathbf{W} \sum_{t=1}^T (T-t) (\mathbf{A}^{t-1} - \mathbf{Q}) \mathbf{W}\mathbf{p}_o, \end{aligned} \quad (4.8)$$

where we have made the substitution $\mathbf{e}\mathbf{W}\mathbf{Q}\mathbf{W}\mathbf{p}_o = \langle \nu \rangle_o^2$ and noted that $\lim_{t \rightarrow \infty} \mathbf{A}^t = \mathbf{Q} = (\mathbf{p}_o, \dots, \mathbf{p}_o)$ since \mathbf{A} is an irreducible, aperiodic, stochastic matrix [122]. This yields a value for the index of dispersion in the long-time limit of

$$R_{T,o} = 1 - \mu\langle \nu \rangle_o + \frac{2\mu}{\langle \nu \rangle_o} \mathbf{e}\mathbf{W} \sum_{t=1}^T \left(1 - \frac{t}{T}\right) (\mathbf{A}^{t-1} - \mathbf{Q}) \mathbf{W}\mathbf{p}_o. \quad (4.9)$$

The above equation is consistent with the generic equation for the index of dispersion

given by Cutler [123, 101], where we now give concrete expressions for the variables ρ and $h(t)$ in Cutler's formula in terms of measurable quantities, namely $\rho = \mu\langle\nu\rangle_o$ and $h(t) = \frac{\mu}{\langle\nu\rangle_o} \mathbf{e} \mathbf{W} \mathbf{A}^{t-1} \mathbf{W} \mathbf{p}_o$.

We can further simplify Equation 4.9 by performing spectral decompositions of \mathbf{A} and \mathbf{Q} . Let $\lambda_1, \dots, \lambda_K$ be the eigenvalues of $\mathbf{V} - \mathbf{W}$, and let $\mathbf{r}_1, \dots, \mathbf{r}_K$ and $\mathbf{l}_1, \dots, \mathbf{l}_K$ be the corresponding right and left eigenvectors, normalized so that $\mathbf{l}_i \mathbf{r}_j = 1$ if $i = j$ and 0 otherwise. These eigenvectors are also eigenvectors of the irreducible, acyclic, stochastic \mathbf{A} , and the corresponding eigenvalues are $1 - \mu\lambda_1, \dots, 1 - \mu\lambda_K$, with Perron-Frobenius theorems guaranteeing that one eigenvalue (chosen here to be $1 - \mu\lambda_1$) is equal to one and all other eigenvalues have absolute values less than one. Then \mathbf{r}_1 and \mathbf{l}_1 are right and left eigenvectors of \mathbf{Q} with eigenvalue 1 (i.e. $\mathbf{r}_1 = \mathbf{p}_o$ and $\mathbf{l}_1 = \mathbf{e}$), and all other eigenvalues of \mathbf{Q} are zero. The spectral decompositions are therefore $\mathbf{Q} = \mathbf{r}_1 \mathbf{l}_1$ and $\mathbf{A} = \mathbf{r}_1 \mathbf{l}_1 + \sum_{i=2}^K (1 - \mu\lambda_i) \mathbf{r}_i \mathbf{l}_i$. Inserting these spectral decompositions into Equation 4.9, we find for the index of dispersion a value of

$$R_{T,o} = 1 - \mu\langle\nu\rangle_o + \frac{2\mu}{\langle\nu\rangle_o} \mathbf{e} \mathbf{W} \sum_{t=1}^T \left(1 - \frac{t}{T}\right) \sum_{i=2}^K (1 - \mu\lambda_i)^{t-1} \mathbf{r}_i \mathbf{l}_i \mathbf{W} \mathbf{p}_o, \quad (4.10)$$

since $\mathbf{A}^t = \mathbf{r}_1 \mathbf{l}_1 + \sum_{i=2}^K (1 - \mu\lambda_i)^t \mathbf{r}_i \mathbf{l}_i$ [122]. In the limit of large T and small μ , the value of $R_{T,o}$ given by the above equation approaches the value

$$\begin{aligned} R_{T,o} &\approx 1 + \frac{2\mu}{\langle\nu\rangle_o} \mathbf{e} \mathbf{W} \sum_{t=1}^T \sum_{i=2}^K (1 - \mu\lambda_i)^{t-1} \mathbf{r}_i \mathbf{l}_i \mathbf{W} \mathbf{p}_o \\ &\approx 1 + \frac{2}{\langle\nu\rangle_o} \mathbf{e} \mathbf{W} \sum_{i=2}^K \lambda_i^{-1} \mathbf{r}_i \mathbf{l}_i \mathbf{W} \mathbf{p}_o, \end{aligned} \quad (4.11)$$

where the $\mu\langle\nu\rangle_o$ term drops out because μ is small and the $\sum_{t=1}^T \frac{t}{T} (1 - \mu\lambda_i)^{t-1}$ term drops out because T is large and $|1 - \mu\lambda_i| < 1$. This equation shows that $R_{T,o}$ approaches a constant value independent of T and μ . Although we could not prove that value of $R_{T,o}$ given by Equation 4.11 is necessarily greater than one (since some of the eigenvalues λ_i could be complex), in all of our simulations we observed $R_{T,o} > 1$, suggesting that when $N\mu \ll 1$, fluctuations in protein stability tend to overdisperse the molecular clock.

4.3.3 Limit when $N\mu \gg 1$

When $N\mu \gg 1$, the population is spread across many nodes of the neutral network rather than converged on a single sequence [105]. In this limit, we treat the evolutionary dynamics of the population deterministically (i.e., we assume an infinite population size), and describe the distribution of stabilities in the population by the column vector $\mathbf{x}(t)$, with element $x_i(t)$ giving the fraction of proteins in the population at time t that have stabilities in bin i . At generation t , the fraction of mutated proteins that continue to fold is $\langle \nu \rangle_t = \mathbf{e}\mathbf{W}\mathbf{x}(t)$. These folded proteins reproduce, and in order to maintain a constant population size, this reproduction must balance the removal of proteins by death, meaning that each folded sequence must produce an average of $\alpha_t = [1 - \mu(1 - \langle \nu \rangle_t)]^{-1}$ offspring. The population therefore evolves according to

$$\mathbf{x}(t+1) = \alpha_t [(1 - \mu)\mathbf{I} + \mu\mathbf{W}]\mathbf{x}(t). \quad (4.12)$$

After the population has evolved for a sufficient period of time, \mathbf{x} approaches an equilibrium distribution of \mathbf{x}_∞ . The corresponding equilibrium neutrality is $\langle \nu \rangle_\infty = \mathbf{e}\mathbf{W}\mathbf{x}_\infty$, and the equilibrium reproduction rate is $\alpha = [1 - \mu(1 - \langle \nu \rangle_\infty)]^{-1}$, so

$$\mathbf{x}_\infty = \alpha [(1 - \mu)\mathbf{I} + \mu\mathbf{W}]\mathbf{x}_\infty. \quad (4.13)$$

This equation can be rewritten to show that \mathbf{x}_∞ is the principal eigenvector of \mathbf{W} ,

$$\langle \nu \rangle_\infty \mathbf{x}_\infty = \mathbf{W}\mathbf{x}_\infty. \quad (4.14)$$

We note that $\langle \nu \rangle_\infty$ approximates the asymptotic neutrality for the decline in the fraction of folded proteins upon random mutagenesis [7, 119].

We now determine the average number of accumulated mutations $\langle m \rangle_{T,\infty}$ and the corresponding index of dispersion $R_{T,\infty}$ by treating the forward evolutionary process. As described in the text immediately prior to Equation 4.3, our calculations describe the actual number of mutations accumulated during the evolutionary process, which may differ from the number of sequence differences relative to the ancestor if a single site undergoes multiple mutations. When $N\mu \gg 1$, it is not *a priori* obvious that the average number of

mutations present in the population is equivalent to number of fixed substitutions along the line of descent. Therefore, in the Appendix, we show that identical results are obtained by tracing a randomly chosen protein backwards in time along its ancestor distribution, proving the treatment we give below is mathematically equivalent to treating the time-reversed process. We define $\mathbf{x}(m, t)$ as the column vector with element i giving the fraction of the population at time t that has accumulated m mutations and is in stability bin i . Once the population has reached the equilibrium distribution of stabilities, the time evolution of $\mathbf{x}(m, t)$ is

$$\mathbf{x}(m, t+1) = \alpha(1-\mu)\mathbf{x}(m, t) + \alpha\mu\mathbf{W}\mathbf{x}(m-1, t). \quad (4.15)$$

The recursion can be solved to obtain

$$\mathbf{x}(m, t) = \alpha^t \sum_{\kappa=0}^t \binom{t}{\kappa} (1-\mu)^{t-\kappa} \mu^\kappa \mathbf{W}^\kappa \mathbf{x}(m-\kappa, 0), \quad (4.16)$$

as can be verified by direct substitution. Since we are assuming the population has equilibrated at time 0 and no mutations have accumulated at that time, $\mathbf{x}(m, 0)$ is \mathbf{x}_∞ for $m = 0$ and 0 otherwise. Furthermore, \mathbf{x}_∞ satisfies Equation 4.14, so multiplying Equation 4.16 by \mathbf{e} yields

$$x(m, t) = \binom{t}{m} \alpha^t (1-\mu)^{t-m} (\mu\langle\nu\rangle_\infty)^m, \quad (4.17)$$

where $x(m, t) = \mathbf{e}\mathbf{x}(m, t)$ gives the fraction of the population that has accumulated m mutations after t generations. The average number of accumulated mutations after T generations is the mean of this binomial distribution,

$$\langle m \rangle_{T, \infty} = \frac{T\mu\langle\nu\rangle_\infty}{1-\mu(1-\langle\nu\rangle_\infty)}. \quad (4.18)$$

Using the well known result for the variance of the binomial distribution, we find that the index of dispersion is

$$R_{T, \infty} = 1 - \frac{\mu\langle\nu\rangle_\infty}{1-\mu(1-\langle\nu\rangle_\infty)}. \quad (4.19)$$

It is important to reiterate that the above equation was derived under the assumption that

there is at most one mutation per sequence per generation. For realistic distributions of mutations (i.e. Poisson), this means that $\mu \ll 1$. In this regime, $R_{T,\infty}$ is close to one.

4.3.4 Lattice Protein Simulations

We tested our theory’s predictions on the evolutionary dynamics of lattice proteins. Lattice proteins are simple protein models that are useful tools for studying protein folding and evolution [34]. Our lattice proteins were chains of 20 amino acids that folded on a two-dimensional lattice. The energy of a lattice protein conformation was equal to the sum of the pairwise interactions between non-bonded amino acids [46]. Each lattice protein has 41,889,578 possible conformations, and by summing over all of these conformations we could exactly determine the partition sum and calculate ΔG_f . We set a minimal stability threshold for the lattice proteins of $\Delta G_f^{\min} = 0$, meaning that we considered all proteins that folded to the target structure with $\Delta G_f \leq 0$ to be folded and functional, while all proteins with $\Delta G_f > 0$ were considered to be nonfunctional. We note that this stability threshold is equivalent to requiring a lattice protein to spend at least half of its time in the target native structure at equilibrium. We began by generating lattice proteins that stably folded to each of the three different structures shown in Figure 4.2. For each of these three proteins, we determined the distribution of $\Delta\Delta G$ values for all 380 single mutations (these distributions are shown in Figure 4.2). These distributions were used to construct the matrix \mathbf{W} and to predict the equilibrium distribution of stabilities, the average number of mutations, and the indices of dispersion for both the $N\mu \ll 1$ and the $N\mu \gg 1$ cases, using the equations presented in the preceding sections.

To test the accuracy of these predictions, we then simulated evolving populations of the lattice proteins with a standard evolutionary algorithm using Wright-Fisher sampling. Briefly, the populations were held at a constant size of either $N = 10$ or $N = 10^5$. At each generation, a new population was created by choosing parents with equal probability from all folded proteins in the previous generation’s population, and copying these parents into the new population with a mutation rate of 5×10^{-4} mutations per residue per generation. Since the proteins have a length of 20 amino acids, this mutation rate corresponds to a per-protein-per-generation mutation rate of $\mu = 10^{-2}$. Therefore, the product $N\mu$ is either 0.1 or 10^3 , corresponding to $N\mu \ll 1$ or $N\mu \gg 1$, respectively. We emphasize that the lattice

protein evolutionary algorithm is the same for both population sizes. When $N = 10$ the population naturally follows dynamics approximating those presented for $N\mu \ll 1$, while when $N = 10^5$ it naturally follows dynamics approximating those presented for $N\mu \gg 1$ (as evidenced by the excellent agreement of the predictions with the simulations). For $N = 10$, we performed 1,000 replicates for each different structure. For $N = 10^5$, computational constraints limited us to 10 replicates for each structure (however the evolutionary dynamics are nearly deterministic in this case, so all replicates yielded similar results). We note that during the simulations we recorded the number of mutations that actually accumulated rather than simply computing the number of differences (Hamming distance) from the original sequence.

Figure 4.2 shows the theoretical predictions and simulation results for each of the three structures. The theoretical predictions are in good agreement with the simulation results. Figure 4.2 clearly shows that when $N\mu \gg 1$, the proteins tend to be more stable than when $N\mu \ll 1$. This extra stability is a biophysical manifestation of the neutrally evolved mutational robustness predicted by van Nimwegen and coworkers [105]. This increase in stability leads to a substantial increase number of accumulated mutations. In accordance with the theoretical predictions, when $N\mu \ll 1$ the index of dispersion is elevated above one by fluctuations in protein stability. Another clear results from the simulations is that proteins of different structure show markedly different distributions of stabilities and rates of sequence evolution due to the differences in their $\Delta\Delta G$ distributions. Overall, the simulations offer strong support for the validity of the theoretical predictions in the preceding sections.

4.4 Discussion

We have presented that a theory that offers quantitative predictions about the distribution of stabilities, the average number of fixed mutations, and the index of dispersion for an evolving protein population in terms of the $\Delta\Delta G$ values for individual mutations. We have demonstrated that these predictions are accurate for simple lattice proteins, and have used existing biophysical evidence to argue that the basic theoretical assumptions should also be accurate for real proteins. In this section, we give qualitative interpretations of

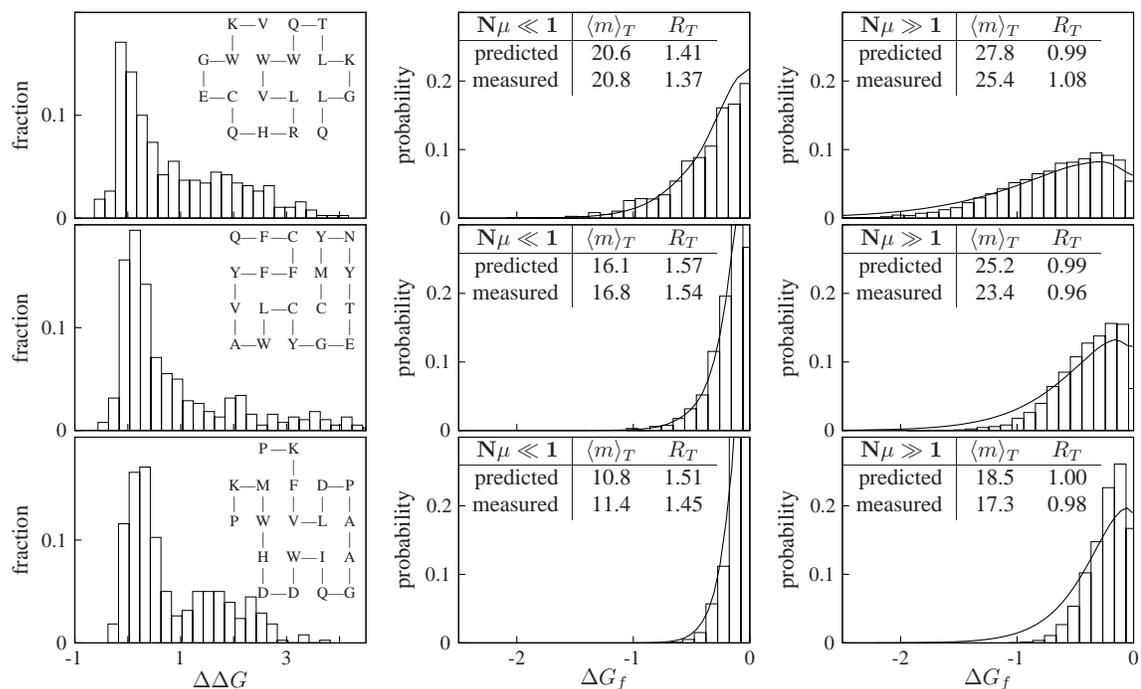


Figure 4.2: The theory gives accurate predictions for the evolution of model lattice proteins. Each row of panels corresponds to a different lattice protein. The graphs at left show the starting protein and the distribution of $\Delta\Delta G$ values for all point mutations. The graphs in the middle and right show the predicted (lines) and measured (boxes) distributions of stabilities among the evolved proteins. The tables embedded in the graphs show the predicted and measured values for the average number of mutations ($\langle m \rangle_T$) and the index of dispersion (R_T) after 5,000 generations of neutral evolution. The center graphs are for a population size of $N = 10$, and the graphs at the right are for $N = 10^5$. In both cases, the per protein per generation mutation rate is $\mu = 0.01$. As predicted, the evolving population with $N\mu \gg 1$ evolved mutational robustness that is manifested by increased protein stability. This additional mutational robustness accelerated the rate of sequence evolution.

the mathematical results and discuss their implications for our understanding of protein evolution.

One major result is to show that the effects of protein structure on the rate of sequence evolution can be quantitatively cast in terms of the $\Delta\Delta G$ values for single mutations. Numerous lattice protein simulations have shown that protein structure can dramatically affect the rate of sequence evolution, since structures that are more “designable” (encoded by more sequences) can evolve their sequences more rapidly (as can be seen in Fig. 4.2 of this work) [104, 9, 106, 11, 12, 34, 10]. Unfortunately, these simulations typically measure structural designability by enumerating a large number of lattice protein sequences, meaning that their findings cannot be extended to real proteins for which such extensive enumeration is impossible. However, recent theoretical work by England and Shakhnovich [43] has made progress in connecting designability to observable structural properties, and a bioinformatics analysis based on this theoretical measure of designability indicates that structure indeed influences the evolutionary rate of real proteins [124]. Our work provides a way to quantitatively relate the structural influences on protein evolution to experimentally measurable $\Delta\Delta G$ values, opening the door to further connecting structural designability and sequence evolution to laboratory stability measurements. Although thousands of $\Delta\Delta G$ values have been measured experimentally [98], at present there are no large sets of measurements for truly random mutations to a single protein. When such sets of measurements become available, it should be possible to use them in conjunction with the theory that we have presented to predict the neutralities of real proteins with different structures.

A second important result is to show that protein evolutionary dynamics can depend on the product of population size and mutation rate, $N\mu$. When $N\mu \gg 1$, the evolving protein population is polymorphic in stability and subject to frequent mutations, so the more stable (and thus more mutationally tolerant) proteins produce more folded offspring. In contrast, when $N\mu \ll 1$, the population is usually monomorphic in stability and so all members of the population are equally likely to produce folded offspring. The general tendency for populations to neutrally evolve mutational robustness when $N\mu \gg 1$ has previously been treated mathematically by van Nimwegen and coworkers [105], and a variety of lattice protein simulations have noted the tendency of evolving protein populations to preferentially occupy highly connected neutral network nodes [9, 108, 11]. Our work shows

that for proteins, in the limiting cases when $N\mu \ll 1$ or $\gg 1$, this process can be rigorously described by considering only protein stability, rather than requiring a full analysis of the neutral network (provided, as we have argued is likely to be the case, that the assumption of a roughly constant $\Delta\Delta G$ distribution holds for real proteins as well as it holds for our lattice proteins). In addition, we prove that the number of accumulated mutations depends on whether $N\mu$ is $\ll 1$ or $\gg 1$. This finding is at odds with the standard prediction [99] of Kimura’s neutral theory that the rate of evolution is independent of population size. The reason for this discrepancy is that the standard neutral theory fails to account for the possibility that increasing the population size so that $N\mu \gg 1$ can systematically increase the fraction of mutations that are neutral.

A third important contribution of our theory is to use the distribution of $\Delta\Delta G$ values for single mutations to predict the distribution of protein stabilities in an evolving population. Several researchers have pointed out that evolved proteins will be marginally stable simply because most mutations are destabilizing [41, 83]; we have described this process quantitatively. In addition, we have shown how the neutral evolution of mutational robustness when $N\mu \gg 1$ will shift the proteins towards higher stabilities (as shown in Fig. 4.2), although this increase in stability is limited by the counterbalancing pressure of predominantly destabilizing mutations. The formulae we provide can in principle be combined with experimentally measured $\Delta\Delta G$ values to predict the expected range of stabilities for evolved proteins.

Our work also weds Takahata’s concept that fluctuating neutral spaces might overdispense the molecular clock [100, 101, 107] to a concrete description of how protein neutrality fluctuates during evolution. When $N\mu \ll 1$, fluctuations in protein stability can cause an overdispersion in the number of accumulated substitutions that can be calculated from the single-mutant $\Delta\Delta G$ distribution. Furthermore, given our assumption of a roughly constant $\Delta\Delta G$ distribution, we show that the index of dispersion will approach a constant value that is independent of time or mutation rate, but will depend on whether $N\mu \ll 1$ or $\gg 1$. Previous simulations have indicated that overdispersion indeed depends on the population size [107, 125] — we have explained this dependence by showing that stability-induced overdispersion does not occur when $N\mu \gg 1$ since the population’s distribution of stabilities equilibrates as it spreads across many sequences. Mathematically, the difference in

the cases $N\mu \gg 1$ and $N\mu \ll 1$ is that, assuming the $\Delta\Delta G$ distribution remains relatively constant, when the population size is sufficiently large, the distribution of protein stabilities no longer fluctuates in a manner that influences the probability of a substitution (Equation 4.3 contains $\mu\mathbf{V}$ in the first term on the right side, while Equation 4.15 does not).

In summary, we have presented a mathematical theory of how thermodynamics shape neutral protein evolution. A major strength of our theory is that it makes quantitative predictions using single-mutant $\Delta\Delta G$ values, which can be experimentally measured. Our work also suggests how neutral and adaptive protein evolution may be coupled through protein thermodynamics. Protein stability represents an important hidden dimension in the evolution of new protein function, since extra stability that is itself neutral can allow a protein to tolerate mutations that confer new or improved functions [49]. Our theory describes the dynamics of protein stability during neutral evolution — adaptive protein evolution is superimposed on these stability dynamics, with proteins most likely to acquire beneficial mutations when they are most stable.

4.5 Materials and Methods

4.5.1 Lattice Protein Simulations

We performed simulations with lattice proteins of $L = 20$ monomers of 20 types corresponding to the natural amino acids. The proteins could occupy any of the 41,889,578 possible compact or non-compact conformations on a two-dimensional lattice. The energy of a conformation \mathcal{C} is the sum of the nonbonded nearest-neighbor interactions, $E(\mathcal{C}) = \sum_{i=1}^L \sum_{j=1}^{i-2} C_{ij}(\mathcal{C}) \times \epsilon(\mathcal{A}_i, \mathcal{A}_j)$, where $C_{ij}(\mathcal{C})$ is one if residues i and j are nearest neighbors in conformation \mathcal{C} and zero otherwise, and $\epsilon(\mathcal{A}_i, \mathcal{A}_j)$ is the interaction energy between residue types \mathcal{A}_i and \mathcal{A}_j , given by Table 5 of [46]. We computed the stability of a conformation \mathcal{C}_t as $\Delta G_f(\mathcal{C}_t) = E(\mathcal{C}_t) + T \ln \{Q(T) - \exp[-E(\mathcal{C}_t)/T]\}$, where $Q(T) = \sum_{\{\mathcal{C}_i\}} \exp[-E(\mathcal{C}_i)/T]$ is the partition sum, made tractable by noting that there are only 910,972 unique contact sets. All simulations were performed at a reduced temperature of $T = 1.0$

We used adaptive walks to find sequences that folded into each of the three arbitrarily chosen conformations shown in Figure 4.2 with $\Delta G_f \leq 0$, and then neutrally evolved these sequences for 10^4 generations with a population size of $N = 100$. Our evolutionary

algorithm was as follows: at each generation we randomly chose a protein that folded to the parental structure with $\Delta G_f \leq 0$ from the population and mutated each residue to some other randomly chosen residue with probability 5×10^{-4} , and continued doing this until we had filled the new population with proteins. At the end of this equilibration evolution, we chose the most abundant sequence in the population as the starting point for further analysis and for the computation of the distribution of $\Delta\Delta G$ values for all 380 point mutations (sequences shown in Figure 4.2). In principle, computing the distribution of $\Delta\Delta G$ values over all sequences in the population rather than just the most abundant one should give a more accurate representation of the true form of this distribution, and indeed we found that doing this slightly increased the accuracy of the predictions shown in Figure 4.2. However, the resulting improvement in accuracy was small, since the approximate constancy of the $\Delta\Delta G$ distribution during neutral evolution (discussed below) means that the distribution computed over a single sequence is representative of that computed over all sequences in the population. Therefore, we chose to compute the $\Delta\Delta G$ distribution over just the most abundant sequence since this choice more closely tracks what would be experimentally feasible with real proteins. (It is experimentally tractable to compute $\Delta\Delta G$ values for a single protein, but would be unmanageable to do so for all proteins in a natural population.)

To collect data for the case when the product $N\mu$ of the population size N and the per protein per generation mutation rate μ is $\ll 1$, we first equilibrated 1,000 replicates by evolving each of them with a population size of $N = 10$ and for 5,000 generations starting with a clonal population of the initial sequence described above. The remainder of the evolutionary algorithm was as described above: the mutation rate stayed at 5×10^{-4} per residue per generation (corresponding to a per protein per generation mutation rate of $\mu = 10^{-2}$), and at each generation all proteins that folded to the target native structure with $\Delta G_f \leq 0$ reproduced with equal probability. We then evolved each of these equilibrated populations for a further 5,000 generations to collect data. We combined the data for all the folded proteins in the final populations of all the replicates to calculate the average number of mutations $\langle m \rangle_T$ after T generations, the corresponding index of dispersion R_T , and the distribution of stabilities shown in Figure 4.2. If we instead simply randomly chose a single folded protein from the final population of each replicate, we obtained results that were

identical within the precision shown in Figure 4.2. We emphasize that $\langle m \rangle_T$ and R_T were computed by keeping track of the actual number of mutations that had occurred during the evolutionary history of each protein, not simply by counting the number of amino acid differences between the ancestral and final sequences (the two quantities may differ if a single site undergoes multiple mutations, as discussed in more detail in later sections).

To generate the data for $N\mu \gg 1$, we used the same procedure but with $N = 10^5$ and only performed 10 replicates. We again computed the statistics shown in Figure 4.2 by combining the data for all of the folded proteins in the final populations of all 10 replicates. Similar results were obtained if we instead computed $\langle m \rangle_T$ and R_T over all of the folded proteins in the final population of a single replicate (average values of $\langle m \rangle_T$ were identical while the R_T values of 1.03, 0.95, and 0.94 were extremely similar to those shown from top to bottom in Figure 4.2). This outcome is expected since the probability distributions for $N\mu \gg 1$ evolve deterministically.

4.5.2 Lattice Protein Predictions

The numerical predictions for the lattice proteins given in Fig. 4.2 were computed by constructing the matrix \mathbf{W} described in the first section of Results with a bin size of $b = 0.005$ and truncating the matrix by assuming that no proteins would have stabilities less than -5.0. For the case when $N\mu \ll 1$, $\langle m \rangle_T$ was calculated using Equation 4.6 and R_T was calculated using Equation 4.11. For $N\mu \gg 1$, $\langle m \rangle_T$ was calculated using Equation 4.18 and R_T was calculated using Equation 4.19.

4.6 Appendix

Here we calculate the properties of the evolving population when $N\mu \gg 1$ by analyzing the time-reversed process to compute the mean and variation in the number of mutations in a single randomly chosen protein over time. We show that the results so obtained are identical to those found in the main text, where we analyzed the forward-time process to compute the mean and variation in the number of mutations across the population of evolving proteins.

When $N\mu \gg 1$, the population is now never converged to a single sequence, so it is not *a priori* obvious that the average number of mutations present in the population is equivalent

to the expected number of fixed substitutions along the line of descent. In fact, in the limit of very large population sizes there may not even be a common line of descent in relevant time frames, since many new mutations will occur before any given mutation goes to fixation. In the main text we calculated the average number of mutations $\langle m \rangle_{T,\infty}$ a sequence in the population has accumulated over the last T generations by treating the forward evolution of the population. Here we trace a randomly chosen protein in the population back in time, and show that the average number of substitutions $\langle s \rangle_T$ that it has accumulated over the last T generations is equal to $\langle m \rangle_{T,\infty}$. We also show that indices of dispersion of $\langle m \rangle_{T,\infty}$ and $\langle s \rangle_T$ have the same value of $R_{T,\infty}$.

To calculate $\langle s \rangle_T$, we first define a vector \mathbf{a} giving the ancestor distribution [126]: element i of $\mathbf{a}(T-t)$ gives the probability that a randomly chosen sequence from the population at time T had a predecessor with stability in bin i at time $T-t$. The transition probabilities of $\mathbf{a}(T-t)$ when the population is in equilibrium are the discrete time analogue of those computed by Hermisson and coworkers [126]. From Equation 4.15 of the main text, it follows that the fraction of sequences in bin i at time $t+1$ that had as their ancestor in the previous generation a sequence in bin j is $\alpha_t [(1-\mu)\delta_{ij} + \mu W_{ij}] x_j(t)$. In order to obtain the probability that a sequence in bin i at time $t+1$ had an ancestor in bin j , we have to divide this fraction by the total number of sequences in bin i at time $t+1$. When the population is at equilibrium, $\alpha_t = \alpha$ and $x_i(t+1) = x_i(t) = x_i$ where x_i is the element from \mathbf{x}_∞ . Hence, the probability that a sequence in bin i had an ancestor in bin j is $\alpha [(1-\mu)\delta_{ij} + \mu H_{ji}]$, where we have defined

$$H_{ji} = W_{ij} x_j / x_i, \quad (4.20)$$

The time evolution of \mathbf{a} is therefore

$$\mathbf{a}(T-t) = \alpha [(1-\mu)\mathbf{I} + \mu\mathbf{H}] \mathbf{a}(T-t+1), \quad (4.21)$$

where the matrix \mathbf{H} is defined by Equation 4.20. Equation 4.21 can be solved to show that the equilibrium value of \mathbf{a} is \mathbf{a}_∞ satisfying

$$\langle \nu \rangle_\infty \mathbf{a}_\infty = \mathbf{H} \mathbf{a}_\infty. \quad (4.22)$$

If we define $\mathbf{a}(s, T - t)$ as the vector with element i giving the probability that a randomly chosen sequence at time T had a predecessor at time $T - t$ in stability bin i and with s substitutions relative to the sequence at time T , then the time evolution for an equilibrated population is

$$\mathbf{a}(s, T - t - 1) = \alpha(1 - \mu)\mathbf{a}(s, T - t) + \alpha\mu\mathbf{H}\mathbf{a}(s - 1, T - t). \quad (4.23)$$

We can solve Equations 4.23 and 4.22 in a manner analogous to the forward process to obtain

$$\mathbf{a}(s, T - t) = \binom{t}{s} \alpha^t (1 - \mu)^{t-s} (\mu\langle\nu\rangle_\infty)^s \mathbf{a}_\infty. \quad (4.24)$$

Again defining $a(s, T - t) = \mathbf{e}\mathbf{a}(s, T - t)$ as the probability of having accumulated s substitutions as one moves back t generations from time T , we obtain the binomial distribution

$$a(s, T - t) = \binom{t}{s} \alpha^t (1 - \mu)^{t-s} (\mu\langle\nu\rangle_\infty)^s. \quad (4.25)$$

Comparison of Equation 4.17 of the main text and Equation 4.25 shows that they are identical. Therefore, all moments computed from the two distributions must be equal. In particular, this proves that $\langle m \rangle_{T,\infty} = \langle s \rangle_T$, and that the corresponding indices of dispersion have the same value of $R_{T,\infty}$ defined by Equation 4.19 of the main text. This shows that when $N\mu \gg 1$, we expect equivalent results regardless of whether we average over the number of mutations in all sequences present in the population, or randomly choose a single sequence and trace back along its ancestor distribution.

4.7 Acknowledgements

We thank Frances H. Arnold for helpful comments and discussion. J.D.B. is supported by a HHMI pre-doctoral fellowship. C.O.W. is supported by the National Institutes of Health grant AI 065960. A.R. is supported by the National Science Foundation grants CCF 0523643 and FIBR 0527023.

Chapter 5

Evolution Favors Mutational Robustness in Sufficiently Large Populations

5.1 Abstract

An important question is whether evolution favors properties such as mutational robustness or evolvability that do not directly benefit any individual, but can influence the course of future evolution. Functionally similar proteins can differ substantially in their robustness to mutations and capacity to evolve new functions, but it has remained unclear whether any of these differences might be due to evolutionary selection for these properties. Here we use laboratory experiments to demonstrate that evolution favors protein mutational robustness if the evolving population is sufficiently large. We neutrally evolve cytochrome P450 proteins under identical selection pressures and mutation rates in populations of different sizes, and show that proteins from the larger and thus more polymorphic population tend towards higher mutational robustness. Proteins from the larger population also evolve greater stability, a biophysical property that is known to enhance both mutational robustness and evolvability. The excess mutational robustness and stability is well described by existing mathematical theories, and can be quantitatively related to the way that the proteins occupy their neutral network. Our work is the first experimental demonstration of the general tendency of evolution to favor mutational robustness and protein stability in highly polymorphic populations. We suggest that this phenomenon may contribute to the mutational robustness and evolvability of viruses and bacteria that exist in large populations.

5.2 Background

Proteins are quite tolerant of mutations, allowing evolution to produce highly diverged sequences that fold to similar structures and perform conserved biochemical functions [3, 56]. However, proteins with nearly identical structures and functions may differ in their robustness to mutation [7, 127, 49], as well as in their capacity to acquire new functions [49]. The fact that mutational robustness and evolvability can vary among the functionally equivalent proteins produced by natural sequence divergence makes these properties important hidden dimensions in evolution — direct selection for protein function is blind to them, yet they can play a crucial role in enabling future evolution. Whether the evolutionary process somehow promotes the acquisition of mutational robustness and evolvability therefore remains a major question [128, 129, 55].

Previous experiments have identified several specific evolutionary conditions that can affect mutational robustness. For example, genetic complementation decreases the mutational robustness of viruses [130], while high mutation rates favor mutational robustness in simulated digital organisms [131]. However, theory [105] makes the much broader — and heretofore experimentally untested — prediction that extra mutational robustness will arise quite generally in sufficiently large populations. This prediction cannot be understood in the standard framework of Kimura’s neutral theory [96], since one of the usual assumptions of the neutral theory is that mutational robustness is constant. (Although Takahata [100] treated the consequences of stochastically fluctuating neutrality on the molecular clock, he did not describe how mutational robustness might change systematically during evolution.) However, changes in mutational robustness can be described by envisioning evolution as occurring on neutral networks, or sets of functionally equivalent proteins that are connected by single mutational steps [102, 132, 103, 9]. In a seminal theoretical analysis of evolution on neutral networks, van Nimwegen and coworkers [105] predicted that the extent of mutational robustness should depend on the degree of population polymorphism. Here we briefly summarize their reasoning, since it motivates our experimental work. We also refer the reader to chapter 16 of [55], which contains an excellent explanation of the densely mathematical work of van Nimwegen and coworkers [105].

If an evolving population is mostly monomorphic, then each mutation is either lost or

goes to fixation before another mutation occurs. The population is therefore usually clustered at a single genotype and rarely experiences mutations, meaning that selection does not distinguish between genotypes of different mutational robustness. All nodes of the neutral network are thus equivalent and will be occupied by the population with equal probability [105]. On the other hand, a highly polymorphic population is always spread across many nodes of the neutral network. When mutations occur, the members of the population at highly connected nodes have a better chance of surviving, causing them to be favored by evolution and increasing the average mutational robustness [105, 9, 133, 108, 11]. Specifically, a highly polymorphic population occupies each node with a probability proportional to its eigenvector centrality [105, 9], a measure of how connected it is to other connected nodes (a variant of eigenvector centrality is used by Google’s PageRank algorithm to rank a webpage’s importance in the network of internet links [134]). Figure 5.1A illustrates how mostly monomorphic and highly polymorphic populations are predicted to occupy a neutral network. For proteins, changes in neutral network occupancy should be manifested by changes in thermodynamic stability [95], with proteins from highly polymorphic populations predicted to be more stable than their counterparts from mostly monomorphic populations (Figure 5.1B). Note that the extent of polymorphism depends on the product of the mutation rate and population size, meaning that protein populations of different sizes are predicted to evolve to different levels of mutational robustness and stability even if they experience the same mutation rate.

5.3 Results and Discussion

5.3.1 Design of neutral evolution experiment

To test whether high population polymorphism drives an increase in mutational robustness and protein stability, we performed laboratory evolution experiments on cytochrome P450 proteins. The basic idea was to neutrally evolve P450s under a constant selection pressure in populations that were either monomorphic or highly polymorphic, and observe whether the proteins evolved to different levels of mutational robustness and stability. The evolution experiments started with a P450 BM3 heme domain that had been engineered to hydroxylate 12-*p*-nitrophenoxydodecanoic acid (12-*p*NCA) [59]. We imposed the selection

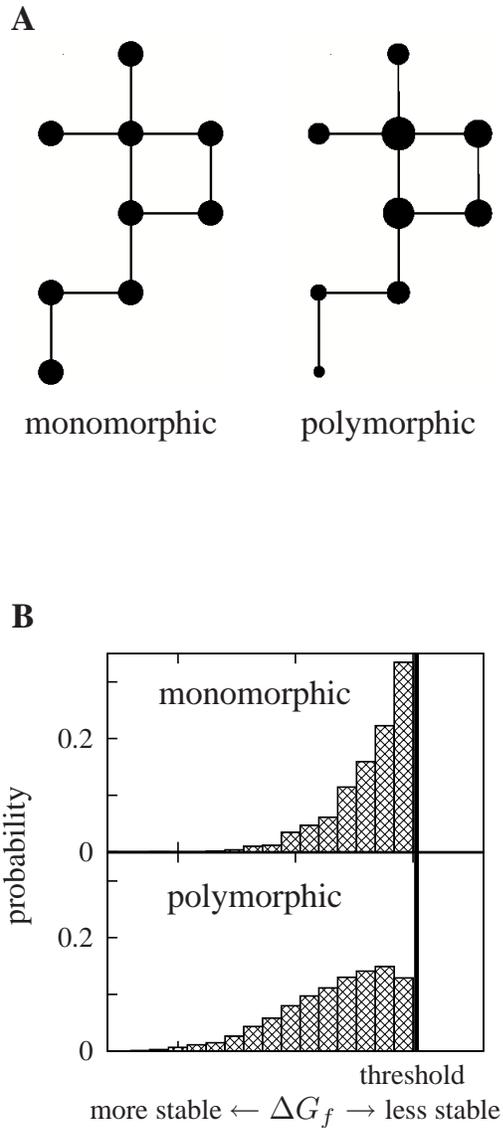


Figure 5.1: Theoretical views of the evolution of protein mutational robustness. **(A)** Theory predicts that a mostly monomorphic population is equally likely to occupy any node of its neutral network, while a highly polymorphic population will prefer more connected nodes [105]. Node sizes are drawn proportional to the occupation probabilities. **(B)** Proteins evolving in a highly polymorphic population are predicted to be more stable than their counterparts in a mostly monomorphic population [95]. The histograms illustrate the distributions of stabilities for the two cases. The increased stability is a biophysical manifestation of excess mutational robustness, since more stable proteins are more mutationally robust [7, 127, 49].

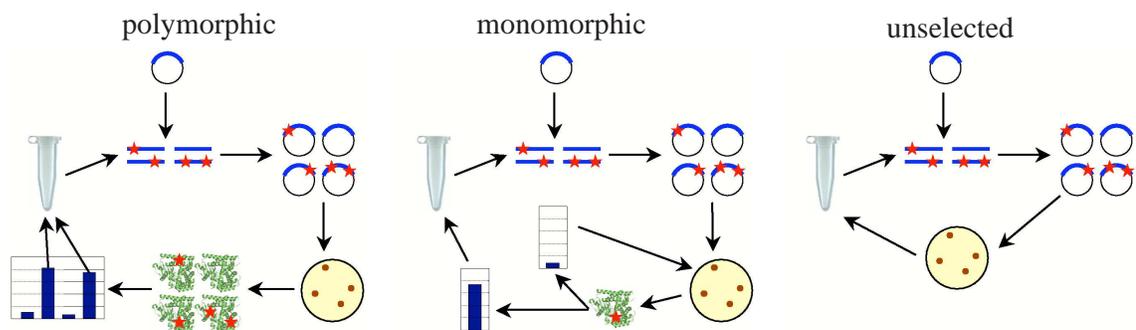


Figure 5.2: Outline of the neutral evolution experimental procedure. For the **polymorphic** population, error-prone PCR was used to generate mutant P450 genes. These genes were ligated into a plasmid and transformed into *E. coli*. Individual mutants (435) were picked, expressed in *E. coli*, and assayed for enzymatic activity. All mutants that met the selection criterion contributed an equal amount of plasmid DNA as template for the next generation of error-prone PCR. The **monomorphic** populations were treated similarly, except only a single mutant was assayed at each generation. If this mutant met the selection criterion then it became the template for the next generation of error-prone PCR; otherwise at the next generation another colony was picked from the same plate. In the **unselected** populations a single mutant was picked and used as the template for the next generation of error-prone PCR.

criterion that *Escherichia coli* cells expressing the P450 had to yield lysate with enough active enzyme to hydroxylate a specified amount of 12-pNCA in 40 minutes. This criterion roughly corresponds to the case in which an enzyme must catalyze a biochemically relevant reaction at some minimal level in order for its host to survive. Note that other properties such as stability and expression level can vary freely, provided that the criterion for total activity is met.

The properties of a neutrally evolving protein eventually “equilibrate,” much as the properties of an isolated physical system under some macroscopic constraint tend towards the values that maximize the system’s internal entropy. For proteins, this usually means that stability, expression, and activity drift towards their lowest tolerable values, since the vast majority of random sequences do not encode stable, well-expressed enzymes (that is, natural selection must work against sequence entropy to maintain a functional protein) [95, 41]. The initial P450 had been engineered for maximal activity [59], meaning that it was not equilibrated to the more mild selection criterion of the experiments. We therefore neutrally evolved this initial P450 for 16 generations, introducing random mutations with error-prone PCR and retaining all mutants that met the selection criterion for total activity

Total nucleotide mutations	67
% synonymous mutations	25
Mutation types (%)	
A →T, T →A	19.4
A →C, T →G	1.5
A →G, T →C	64.2
G →A, C →T	4.5
G →C, C →G	0.0
G →T, C →A	1.5
frameshift	9.0

Table 5.1: Error-prone PCR nucleotide mutation spectrum. Spectrum of nucleotide mutations introduced by the error-prone PCR procedure used in the neutral evolution experiments. The spectrum was determined by sequencing the four final (generation 12) sequences from the unselected population, since in these sequences the mutations accumulate without constraint. As has been previously noted for error-prone PCR with *Taq* polymerase [7, 49, 17], the nucleotide error spectrum is biased towards certain types of mutations.

on 12-pNCA. The procedure used for this equilibration evolution was similar to that for the polymorphic neutral evolution described below. As expected, expression, stability, and activity all dropped during the equilibration evolution. At the end of the equilibration evolution, we chose a single sequence as the parent for the neutral evolution experiments. The gene encoding this parent sequence contained 29 nucleotide mutations and 13 amino acid mutations relative to the initial P450 (a text file with the sequence of this gene is in [135]).

We used this parent gene to begin three parallel sets of neutral evolution experiments, which we named “monomorphic,” “polymorphic,” and “unselected” (Figure 5.2). The monomorphic experiments capture the case where the population moves as a single entity, the polymorphic experiment captures the case where the population spreads across many sequences, and the unselected experiments show how the gene evolves in the absence of selection for protein function. In all experiments, at each generation we used error-prone PCR to introduce an average of 1.4 nucleotide mutations per P450 gene (Table 5.1). The mutant genes were ligated into a plasmid and transformed into *E. coli* [86], and transformants were selected using the plasmid’s antibiotic resistance marker. For the unselected case, we randomly picked one of the mutants, recovered the mutant gene with a plasmid mini-prep, and used this mutant as the template for the next generation of error-prone

PCR. We performed four independent replicates of unselected evolution, evolving each for 12 generations.

For the monomorphic and polymorphic populations, we imposed the selection criterion that the P450s hydroxylate 12-pNCA with at least 75% of the total activity of the original parent gene. We expressed the P450s in *E. coli*, and then assayed the cell lysates for activity in a high-throughput 96-well plate format. The total amount of product produced by 80 μ l of clarified lysate in 40 minutes was compared to the median of four control wells containing the original parent P450 to determine if the mutant met the selection criterion. The only difference between the monomorphic and polymorphic experiments was the size of the evolving populations. In the monomorphic limit, each mutation is either lost or goes to fixation before the next occurs. We enforced this evolutionary dynamic by holding the population size to a single protein sequence, similar to the “blind ant” random walk of [105]. At each generation, we assayed a single mutant. If this mutant met the selection criterion, then it was carried over to the next generation, corresponding to a neutral mutation going to fixation. If the mutant failed the selection criterion, then the population stayed at the previous sequence for the next generation, corresponding to a mutation lost to selection. If all of the mutants assayed had zero or one mutations, then this protocol would correspond exactly to the equations of [105, 95]. However, in order to achieve appreciable sequence evolution on a laboratory time scale, we used a mutation rate that sometimes produced multiple mutations in a generation. We mathematically describe this situation in the Mathematical Appendix; here we simply note that it is possible to think of each generation as introducing a single mutational event rather than a single mutation. We performed 22 independent replicates of monomorphic evolution, evolving each for 25 generations.

In the polymorphic limit, the population spreads across many sequences. To implement this experimentally, we assayed 435 mutants at each generation. The selection criterion was used to classify each mutant as functional or nonfunctional. In neutral evolution, all functional mutants reproduce with equal probability. We therefore pooled equal volumes of stationary-phase cultures of each functional mutant and recovered the pooled genes with a mini-prep. The polymorphic evolution experiment therefore approaches the equations of [105, 95], again with the exception that a sequence may undergo multiple mutations at a

single generation. We give the equations describing this situation in the Mathematical Appendix. Since the population evolves deterministically in the polymorphic limit [105, 95], a single replicate was performed. Because mutations accumulate more rapidly in the polymorphic experiments than the monomorphic ones, we evolved the polymorphic population for 15 generations rather than 25.

5.3.2 Mutations and mutational robustness

Figure 5.3 shows how mutations accumulated during the course of the neutral evolution experiments (the full data can be found in text file format in [135]). Since the unselected protein populations evolve without constraint, mutations accumulate at the same rate at which they are introduced by error-prone PCR, 1.4 nucleotide mutations per generation. Because selection eliminates mutations that disrupt P450 activity, mutations accumulate more slowly in the monomorphic and polymorphic populations. Mutations accumulate more rapidly in the polymorphic population than in the monomorphic populations. This difference in rates is predicted by the equations in the Mathematical Appendix to be a consequence of the fact that the polymorphic population is more mutationally robust, and so can tolerate more of the possible mutations.

To test directly whether the polymorphic population evolves higher average mutational robustness, we measured the fraction of 435 random mutants that met the selection criterion. Figure 5.4 shows that the polymorphic population neutrally evolved to a markedly higher mutational robustness than the monomorphic populations, with $50 \pm 2\%$ of the final polymorphic population mutants continuing to function versus $39 \pm 2\%$ for the final monomorphic populations (Chi-square P -value of 10^{-3} that these values are significantly different). The only difference between the two types of populations was their size, so evolution has clearly favored mutational robustness in the larger and thus more polymorphic population. This finding represents the first experimental support for the prediction that highly polymorphic populations evolve excess mutational robustness [105].

Theory predicts that the excess mutational robustness of a highly polymorphic protein population comes from increased protein stability [95]. Because the P450 variants unfold irreversibly, an equilibrium thermodynamic stability ΔG_f cannot be measured. We therefore determined stability to irreversible thermal and chemical denaturation, two highly corre-

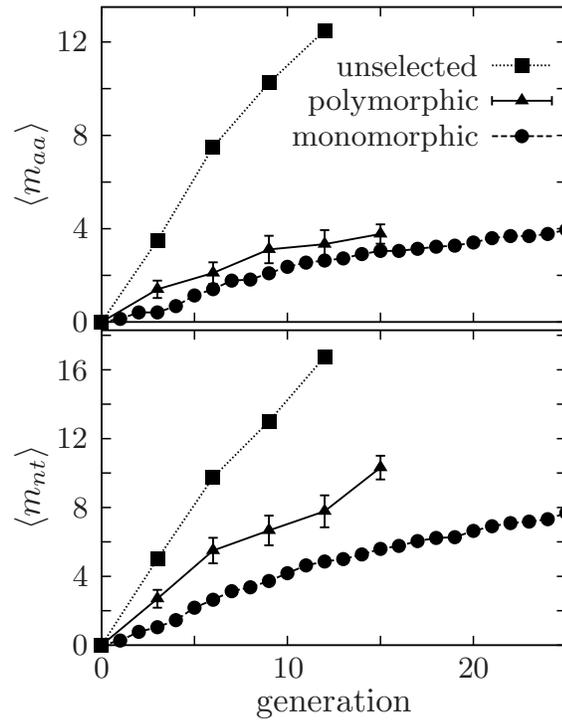


Figure 5.3: Accumulation of nucleotide ($\langle m_{nt} \rangle$) and nonsynonymous ($\langle m_{aa} \rangle$) mutations in the experimentally evolved P450 populations. For the unselected and monomorphic populations, numbers are the average over all replicates at the indicated generation; for the polymorphic population they are from a random sample, with sampling standard error shown.

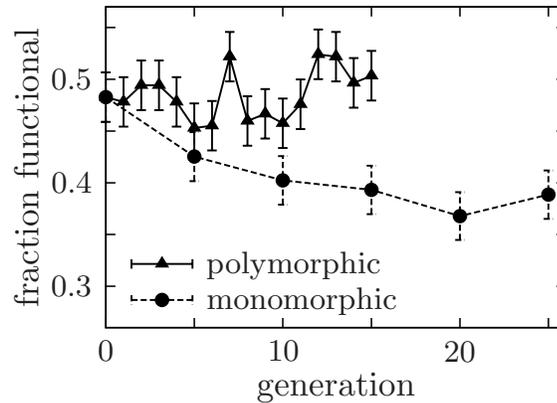


Figure 5.4: The polymorphic population neutrally evolved a higher average mutational robustness than the monomorphic populations. The fraction functional was determined by assaying 435 mutants (average of 1.5 nucleotide mutations per gene). Error bars show binomial standard error. For the monomorphic population, numbers are the average over all replicates.

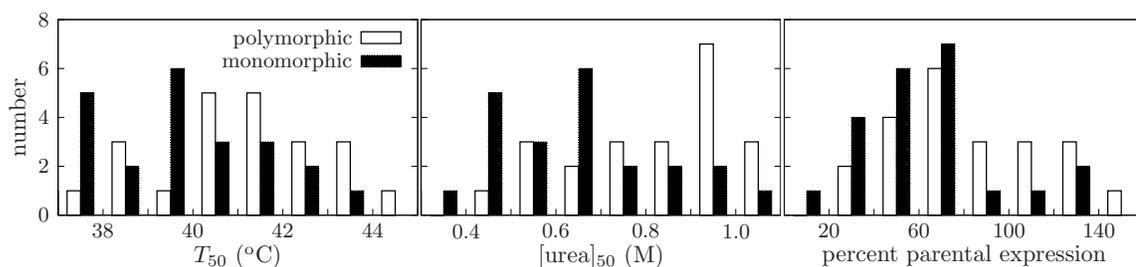


Figure 5.5: The more mutationally robust proteins are more stable. The P450s from the polymorphic population neutrally evolved higher stability and expression levels than their counterparts from the monomorphic populations. The histograms show the distributions for the final protein from all monomorphic replicates and for the same number of randomly chosen proteins from the final polymorphic population. The plots show (left to right) the temperature at which half the protein irreversibly denatured after 10 minutes (T_{50}), the urea concentration at which half the protein denatured after 4 hours ($[\text{urea}]_{50}$), and the expression level relative to that of the original parental P450. The means are significantly different, with unequal variance t-test P -values of 0.02, 0.005, and 0.04, respectively.

lated measures of P450 stability that have previously been shown to contribute to mutational robustness [49] (Figures 5.6, 5.7, and 5.8). Figure 5.5 shows that proteins from the polymorphic population were in fact more stable than their counterparts from the monomorphic population. We also observed that proteins in the polymorphic population tended to accumulate to higher levels in *E. coli* (Figure 5.5). Elevated expression could be a byproduct of increased stability, or it could independently increase mutational robustness by allowing the proteins to better tolerate mutations that decrease codon adaptation or reduce folding efficiency. It is possible that additional unrecognized biophysical factors also contributed to the excess mutational robustness of the polymorphic population, but no such factors were immediately obvious.

5.3.3 Interpretation in terms of the P450 neutral network

The higher mutational robustness of the polymorphic population is due to the fact that it occupies the P450 gene neutral network differently than the monomorphic populations. Measurements from the evolution experiments can therefore be used to infer basic properties of the underlying neutral network of P450 genes, as originally noted by van Nimwegen and coworkers [105]. In the Mathematical Appendix, we derive approximations for the normalized principal eigenvalue $\langle \nu \rangle_{\infty}$ and the normalized average connectivity $\langle \nu \rangle_o$ of the

neutral network, where in both cases the normalization is obtained by dividing by the network coordination number. We obtain $\langle \nu \rangle_\infty = 0.51$ and $\langle \nu \rangle_o = 0.35$ for the P450 gene neutral network. Our ability to consistently estimate these two parameters from four different experimental measurements supports the idea that the theory that we elaborate in the Mathematical Appendix appropriately describes the experiments. The difference between $\langle \nu \rangle_\infty$ and $\langle \nu \rangle_o$ is a measure of the extent to which some P450 neutral network nodes have more connections than others. We note that $\langle \nu \rangle_\infty$ is approximately equal to the exponential decline parameter for the asymptotic decline in the fraction of functional mutants with increasing numbers of random nucleotide mutations [7, 17, 18] (see Mathematical Appendix). Previous studies looking at this exponential decline have reported $\langle \nu \rangle_\infty = 0.7$ for subtilisin [17], $\langle \nu \rangle_\infty = 0.7$ for 3-methyladenine DNA glycosylase [18], and $\langle \nu \rangle_\infty = 0.7 - 0.8$ for TEM1 β -lactamase [7]. These comparisons suggest that P450 has a sparser neutral network (smaller $\langle \nu \rangle_\infty$) than these other proteins. We suspect, however, that these earlier studies (one of which is our own) overestimate $\langle \nu \rangle_\infty$ due to insufficient equilibration of the starting sequence. We believe that the approach of the current work is more accurate for determining $\langle \nu \rangle_\infty$ because the measurements are made after many mutations have equilibrated the initial sequence. This approach could be used in future experiments to compare the neutral network connectivities of proteins from different families.

5.4 Conclusions

We have demonstrated that neutral evolution favors more mutationally robust proteins when the evolving population is highly polymorphic. Strikingly, the excess mutational robustness is due only to population polymorphism, and so will arise in any population of sufficiently large size. Our work is the first experimental demonstration of this phenomenon, which is predicted to occur quite generally in neutrally evolving proteins and nucleic acids [105]. Furthermore, we were able to identify one of the biophysical factors underlying the increase in mutational robustness by showing that proteins from the highly polymorphic population are more stable. We recognize that evolution in a biological context will be more complex. In our experiments, fitness was the P450's ability to be expressed in active form by bacteria grown to saturation in an environment with plentiful nutrients. Biological fitness, however,

depends on numerous additional and subtle effects such as the metabolic costs of synthesis or the burdens imposed by misfolded molecules. Some mutations that are neutral in the experiments may therefore have deleterious effects in a biological setting [72]. The experiments nonetheless capture the overriding constraint that proteins retain their biochemical functions. Our success in quantitatively explaining the results supports the notion that important aspects of protein evolution can be described simply in terms of mutational effects on stability [95, 72].

An obvious question is whether evolution in nature favors mutational robustness by the process we have demonstrated. Whether natural populations will neutrally evolve mutational robustness depends on whether they are sufficiently polymorphic, which will be the case if the product of their effective population size N and per protein per generation mutation rate μ is much greater than one [105, 96]. Accurately estimating $N\mu$, which is closely related to the widely used parameter θ in population genetics, for natural populations is difficult [136, 137] (note that since mutational robustness is a protein-wide property, the relevant mutation rate is per protein, which is $\approx 10^2$ to 10^3 larger than the per codon mutation rate). For humans and other multicellular organisms, $N\mu$ is probably too small [138] for their proteins to neutrally evolve mutational robustness. But estimates [138, 139] place $N\mu \approx 10$ to 100 for typical-length proteins in bacteria, and it is probably much higher for many viruses [140, 141]. It is therefore likely that many viral and some bacterial proteins have neutrally evolved extra mutational robustness.

The neutral evolution of protein mutational robustness may also contribute to adaptive evolution. Experiments have shown that extra stability increases a protein's evolvability by allowing it to tolerate a wider range of functionally beneficial but destabilizing mutations [49]. A similar phenomenon seems to occur in natural evolution, where functionally neutral but stabilizing mutations can play a key role in adaptive evolution by counterbalancing the destabilizing effects of other functionally beneficial mutations [36]. Viruses and perhaps bacteria may thus benefit from large population sizes and high mutation rates that drive an increase in the mutational robustness and stability of their proteins, which in turn enhances the capacity of these proteins to rapidly change their sequences and evolve new functions.

5.5 Methods

5.5.1 Equilibration evolution of the P450 protein

We began with a 21B3 P450 peroxygenase that had been engineered for highly efficient hydroxylation of 12-pNCA [59]. This P450 was not well equilibrated to the constant selection criterion that we planned to impose, since it had substantially higher total activity. We therefore neutrally evolved it for 16 generations in order to create P450s that were better equilibrated to the selection criterion. We evolved two parallel populations, which we named R1 and R2. The procedure was exactly identical to that described below for the polymorphic evolution with the following exceptions:

- Starting sequence: the starting sequence for the equilibration evolution was the 21B3 sequence.
- Population size: each of the two equilibration evolution populations had a size of 174 sequences rather than the 435 used for the polymorphic evolution.
- Selection criterion: the sequences were required to have at least 75% of the total activity of the 21B3 P450.
- Mutation rate: the mutation rate for the equilibration evolution was much higher than for the polymorphic evolution. The error-prone PCR protocol used 200 μM manganese chloride (MnCl_2), rather than the 25 μM used for the polymorphic evolution. We estimate that this error-prone PCR protocol introduced ≈ 4 nucleotide mutations per P450 gene at each generation during the equilibration evolution.

We performed 16 generations of equilibration evolution, and then randomly selected 23 functional mutants from each of the R1 and R2 populations (text files with sequences can be found in [135]). We picked one of these mutants, R1-11, for use as the parent for the neutral evolution experiments.

5.5.2 Detailed protocol for evolution experiments

We began with the R1-11 P450 BM3 heme domain variant (sequence given in text format in [135]) cloned into the pCWori [86] plasmid with a 5' *BamHI* and 3' *EcoRI* site as described

in [49]. The cloning primers were *pCWori_for* (5'-GAAACAGGATCCATCGATGCTTAGGAGGTCAT-3') and *pCWori_rev_clone* (5'-GCTCATGTTTGACAGCTTATCATCG-3'). We used error-prone PCR to generate mutants, taking great care to make the error-prone PCR protocol repeatable by using a relatively small number of thermal cycles. This was both to control the mutation rate by ensuring that the reaction did not saturate the reagents (which would cause the number of doublings to become sensitive to the initial template concentration), and to avoid the PCR-based recombination events which can occur during with the last few thermal cycles of PCR reactions [142, 143]. The PCR reactions were 100 μ l in volume, and contained \approx 13 ng of plasmid template (corresponding to \approx 3 ng of template gene), 7 mM magnesium chloride $MgCl_2$, 1 \times Applied BioSystems PCR Buffer II without $MgCl_2$, 25 μ M $MnCl_2$, 0.5 μ M *pCWori_for* primer, 0.5 μ M *pCWori_rev* primer, 200 μ M of dATP and dGTP, 500 μ M of dTTP and dCTP, and 5 units of Applied Biosystems AmpliTaq polymerase. The reactions were run on the BLOCK setting of a MJ Research PCR machine with a program of 95°C for 2 minutes, then 15 cycles of (95°C for 30 seconds, 57°C for 30 seconds, 72°C for 90 seconds), and then cooling to 4°C. This protocol yielded roughly 1-1.5 μ g of product gene (as quantified by gel electrophoresis versus a known standard), for a PCR efficiency of \approx 0.5. Sequencing the unselected populations at the end of the experiment indicated that this protocol introduced an average of 1.4 ± 0.2 nucleotide mutations, with the nucleotide error-spectrum shown in Table 5.1. Because the number of PCR doublings is large compared the average mutation rate, the distribution of mutations among sequences should be well-described by the Poisson distribution [48, 64].

The mutant genes from the error-prone PCR were purified over a ZymoResearch DNA clean and concentrator column, and digested at 37°C with *EcoR1* and *BamH1*. The digested genes were then purified from an agarose gel with ZymoResearch DNA gel extraction columns, and ligated into pCWori plasmid that had been digested with *BamH1* and *EcoR1* and dephosphorylated. The ligations were transformed into electro-competent catalase-free *Escherichia coli* [86] (the catalase is removed because it breaks down the hydrogen peroxide utilized by the P450 peroxygenase), plated on Luria Broth (LB) plates containing 100 μ g/ml of ampicillin to select for the plasmid's antibiotic resistance marker, and grown at 37°C. Transformation of a control ligation reaction without any digested gene yielded at least 100-fold fewer colonies, indicating that the rate of plasmid self-ligation was less than

T	$\langle m_{nt} \rangle_U$	$\langle m_{aa} \rangle_U$	$\langle m_{nt} \rangle_P$	$\langle m_{aa} \rangle_P$	$\langle m_{nt} \rangle_M$	$\langle m_{aa} \rangle_M$	$\langle \mathcal{F} \rangle_P$	$\langle \mathcal{F} \rangle_M$
0	0	0	0	0	0	0	0.48 (210 / 435)	0.48 (210 / 435)
1	NA	NA	NA	NA	0.1 (3 / 22)	0.3 (6 / 22)	0.48 (208 / 435)	NA
2	NA	NA	NA	NA	0.4 (9 / 22)	0.8 (17 / 22)	0.49 (215 / 435)	NA
3	5.0 (20 / 4)	3.5 (14 / 4)	2.7 (27 / 10)	1.4 (14 / 10)	1.0 (23 / 22)	0.4 (9 / 22)	0.49 (215 / 435)	NA
4	NA	NA	NA	NA	1.5 (32 / 22)	0.7 (15 / 22)	0.48 (208 / 435)	NA
5	NA	NA	NA	NA	2.2 (48 / 22)	1.1 (25 / 22)	0.45 (197 / 435)	0.43 (185 / 435)
6	9.8 (39 / 4)	7.5 (30 / 4)	5.5 (55 / 10)	2.1 (21 / 10)	2.6 (58 / 22)	1.4 (31 / 22)	0.46 (198 / 435)	NA
7	NA	NA	NA	NA	3.1 (69 / 22)	1.8 (39 / 22)	0.52 (227 / 435)	NA
8	NA	NA	NA	NA	3.4 (74 / 22)	1.8 (40 / 22)	0.46 (200 / 435)	NA
9	13.0 (52 / 4)	10.3 (41 / 4)	6.7 (61 / 9)	3.1 (28 / 9)	3.7 (82 / 22)	2.1 (46 / 22)	0.47 (203 / 435)	NA
10	NA	NA	NA	NA	4.2 (92 / 22)	2.4 (52 / 22)	0.46 (199 / 435)	0.40 (175 / 435)
11	NA	NA	NA	NA	4.6 (102 / 22)	2.5 (56 / 22)	0.48 (207 / 435)	NA
12	16.8 (67 / 4)	12.5 (50 / 4)	7.8 (70 / 9)	3.3 (30 / 9)	4.9 (107 / 22)	2.6 (58 / 22)	0.52 (228 / 435)	NA
13	NA	NA	NA	NA	5.0 (110 / 22)	2.7 (60 / 22)	0.52 (227 / 435)	NA
14	NA	NA	NA	NA	5.3 (116 / 22)	2.9 (64 / 22)	0.50 (216 / 435)	NA
15	NA	NA	10.3 (227 / 22)	3.8 (83 / 22)	5.6 (123 / 22)	3.0 (67 / 22)	0.50 (219 / 435)	0.39 (171 / 435)
16	NA	NA	NA	NA	5.8 (127 / 22)	3.0 (67 / 22)	NA	NA
17	NA	NA	NA	NA	6.0 (133 / 22)	3.1 (69 / 22)	NA	NA
18	NA	NA	NA	NA	6.3 (137 / 22)	3.2 (71 / 22)	NA	NA
19	NA	NA	NA	NA	6.3 (138 / 22)	3.3 (72 / 22)	NA	NA
20	NA	NA	NA	NA	6.6 (145 / 22)	3.4 (75 / 22)	NA	0.37 (160 / 435)
21	NA	NA	NA	NA	6.9 (152 / 22)	3.6 (79 / 22)	NA	NA
22	NA	NA	NA	NA	7.1 (156 / 22)	3.7 (81 / 22)	NA	NA
23	NA	NA	NA	NA	7.2 (158 / 22)	3.7 (81 / 22)	NA	NA
24	NA	NA	NA	NA	7.3 (161 / 22)	3.8 (83 / 22)	NA	NA
25	NA	NA	NA	NA	7.7 (169 / 22)	4.0 (87 / 22)	NA	0.39 (169 / 435)

Table 5.2: Neutral evolution robustness and mutation data. Each row is for a different generation, T . Entries of NA indicate that no measurement was made. The $\langle m_{nt} \rangle$ and $\langle m_{aa} \rangle$ are the average number of nucleotide mutations and nonsynonymous mutations, respectively. Numbers in parentheses are total counts over the total samples. Subscripts indicate the population type: U for unselected, P for polymorphic, and M for monomorphic. For the unselected and monomorphic populations, numbers represent averages of all replicates. For the polymorphic population, numbers are for a random sample of functional mutants. $\langle \mathcal{F} \rangle_P$ and $\langle \mathcal{F} \rangle_M$ are the fraction of functional mutants out of 435 assayed.

one percent.

Individual mutant colonies from the plates were picked into 96-well 2 ml deep-well plates containing 400 μl of LB supplemented with 100 $\mu\text{g}/\text{ml}$ ampicillin. Each plate contained four parental control wells with cells carrying the parent R1-11 gene, four null control wells with cells carrying the pCWori plasmid without a P450 gene, and a non-inoculated well to check for contamination. For the polymorphic population, we picked five such plates with all 87 other wells containing different mutants for a total population size of $5 \times 87 = 435$ mutants. For the 22 monomorphic populations (we began with 24 populations but two had to be discarded due to contamination), we picked a single colony for growth and screening. For the unselected populations we picked a single colony for growth without screening. The LB deep-well plates were grown for 16-20 hours at 30°C, 210 revolutions per minute (rpm), and 80% relative humidity in a Kuhner humidified shaker. To express the P450 mutants, we prepared 2 ml deep well plates containing 400 μl per well of terrific broth (TB) supplemented with 200 μM isopropyl β -D-thiogalactoside (IPTG), 100 $\mu\text{g}/\text{ml}$ ampicillin, and 500 μM of δ -aminolevulinic acid. We used a pipetting robot inoculated these TB plates with 100 μl from the LB plates. We stored the LB deep-well plates at 4°C, and grew the TB deep-well plates in the humidified shaker at 30°C, 210 rpm, and 80% relative humidity for 22-24 hours. After this growth, the cells were harvested by centrifuging the TB deep-well plates at 4000 \times g for 5 minutes and discarding the liquid. The cell pellets were flash-frozen in liquid nitrogen to aid in cell lysis.

To lyse the cells for the assays, we resuspended the cell pellets in 300 μl of 100 mM [4-(2-hydroxyethyl)-1-piperazinepropanesulfonic acid] (EPPS) (pH 8.2) with 0.5 mg/ml lysozyme and 4 units/ml of deoxyribonuclease by pipetting 40 times with the pipetting robot. We then incubated the plates at 37°C for 30 minutes, again resuspended with the pipetting robot, and put back at 37°C for an addition 30 minutes. We then pelleted the cell debris by centrifugation at 6000 \times g for 5 minutes at 4°C. The pipetting robot was used to dispense 80 μl of the clarified lysate into 96-well microtiter plates (Rainin). We prepared a 6 \times stock of 1.5 mM 12-pNCA in 36% dimethyl sulfoxide (DMSO) and the EPPS buffer (the 12-pNCA was stored in the DMSO solution and combined with the buffer immediately before use). We used a multichannel pipette to add 20 μl of this substrate stock to each well of the microtiter plate. We briefly mixed the plates with “shake” setting of a 96-well

plate spectrophotometer, and read an absorbance baseline at 398 nm. We then immediately added 20 μl of a freshly prepared solution of 24 mM hydrogen peroxide in the EPPS buffer to initiate the reaction, and mixed again. The final reaction conditions were therefore the EPPS buffer with 6% DMSO, 4 mM hydrogen peroxide, and 250 μM 12-pNCA. After 40 minutes we quantified the amount of enzymatic product by the increase in absorbance at 398 nm. This absorbance increase is due to the 4-nitrophenolate molecule released after the P450 hydroxylates the twelfth carbon of the 12-pNCA molecule [59]. To score the mutants as functional or nonfunctional, we compared their gain in absorbance minus the median null control reading to that of the median parental control reading minus the median null control reading. All mutants that had at least 75% of the parental gain were scored as functional, all other mutants were scored as nonfunctional.

We used the information from these assays to select the parents for the next generation. For the unselected population we did not require the mutants to be functional, so the selected mutant was used to start a 4 ml culture of LB with 100 $\mu\text{g}/\text{ml}$ ampicillin, and the plasmid DNA was harvested with a mini-prep. This plasmid DNA was used as the template for the next round of error-prone PCR. Therefore, after the first generation the four unselected replicates diverged into four separate error-prone PCR reactions. These unselected replicates were evolved for a total of twelve generations, and were sequenced at every third generation.

For the polymorphic population, all mutants that were functional contributed an equal amount of plasmid DNA as template for the next generation. In order to do this, we collected 50 μl of the culture from the LB deep-well plate for each mutant that was scored as functional. All of these LB aliquots were pooled, and then the plasmid DNA was collected with a mini-prep. The pool of plasmid DNA was used as template for the next generation's error-prone PCR reactions. We performed 15 generations of evolution for this polymorphic population. Note that at each generation we are assaying 435 mutants as part of the evolutionary procedure, so this provides information on mutational robustness. At every third generation, we also selected a random sample of functional mutants for sequencing. After 15 generations, we randomly selected 22 mutants for stability measurements and sequencing analysis. The random selections were made from all functional mutants with the Python computer language random number generator.

For the monomorphic populations, at each generation we assayed just a single mutant. If that mutant was nonfunctional, then at that generation the population stayed at its original sequence. In that case, for the next generation we simply picked a new mutant from the previous generation's plate of transformed mutants. If the mutant we screened was functional, then that mutant represented the new population. We therefore grew a 4 ml LB culture with 100 $\mu\text{g/ml}$ of ampicillin, and collected the plasmid DNA with a miniprep. That plasmid DNA was then used as the template for the next generation's error-prone PCR reaction. We thus had 22 (actually 24 before 2 were contaminated) independent monomorphic populations that were being evolved in parallel. Each was evolved for 25 generations, and at the end of these 25 generations we measured the stability of the final sequence of each population. Each time an assayed mutant was functional, we sequenced the new P450 gene. We also measured the average mutational robustness of the monomorphic populations at every fifth generation. To do this, we did a pooled mini-prep of equal volumes of LB cultures of all 22 replicates to obtain an equal mix of plasmid DNA. We then performed error-prone PCR on this mix, and assayed 435 mutants to measure the fraction functional. Full neutral evolution data are given in text file format in [135].

5.5.3 Test for recombination during error-prone PCR

During the polymorphic population evolution, we performed error-prone PCR on a mix of different plasmids. It is common for PCR on mixed templates to lead to recombination events during the reaction [142, 143]. We attempted to reduce this recombination by using a small number of thermal cycles. However, in order to test for recombination, we analyzed the sequences of the final 22 selected members of the polymorphic population. There are a variety of statistical tests to detect recombination in a set of sequences. A comparison of these tests by Posada [144] found that the Max-Chi² method developed by John Maynard Smith [145] performs well. A publicly available implementation of this method [146] is at <http://www.lifesci.sussex.ac.uk/CSE/test/maxchi.php>. We used this implementation to analyze the 22 final polymorphic sequences, and the resulting P -value was 0.29 after 100 random permutations, indicating that there is not significant recombination.

5.5.4 Measurement of P450 stabilities

We measured the stabilities to both irreversible thermal and irreversible urea denaturation of the final (generation 25) member of each monomorphic population, as well as of the 22 randomly selected members of the polymorphic population. As discussed in the Supplementary Information of [49], cytochrome P450 BM3 heme domains (and indeed most P450s) denature irreversibly, forcing us to use resistance to irreversible denaturation to quantify protein stability. The first stability measure is the T_{50} , defined as the temperature at which half of the protein is denatured after a 10 minute incubation. The second stability measure is the $[\text{urea}]_{50}$, defined as the urea concentration at which half of the protein denatures after a 4 hour room-temperature incubation. Each set of measurements (those of T_{50} and $[\text{urea}]_{50}$) was performed on all of the mutants in the same day, and each mutant was treated identically. Therefore, it is possible to make accurate comparisons of the relative values of the measurements within the data set. However, the absolute values of the T_{50} and $[\text{urea}]_{50}$ values may be less accurate. Therefore, care should be taken in comparing the absolute value of these measurements to those of other studies (such as [49]).

Both the T_{50} and $[\text{urea}]_{50}$ measurements were performed in clarified cell lysate. The protein was expressed using catalase-free *E. coli* [86] containing the encoding gene on the IPTG inducible pCWori [86] plasmid. We used freshly streaked cells to inoculate 2 ml cultures of LB supplemented with 100 $\mu\text{g}/\text{ml}$ of ampicillin, and grew these starter cultures overnight with shaking at 37°C. We then used 0.5 ml from these starter cultures to inoculate 1 L flasks containing 200 ml of TB supplemented with 100 $\mu\text{g}/\text{ml}$ of ampicillin. The TB cultures were grown at 30°C and 210 rpm until they reached an optical density at 600 nm of ≈ 0.9 , at which point IPTG and δ -aminolevulinic acid were added to a final concentration of 0.5 mM each. The cultures were grown for an additional 19 hours, then the cells were harvested by pelleting 50 ml aliquots at 5,500 g and 4°C for 10 min, and stored at -20°C. To obtain clarified lysate, each pellet was resuspended in 8 ml of 100 mM EPPS (pH 8.2) and lysed by sonication, while being kept on ice. The cell debris was pelleted by centrifugation at 8,000 g and 4°C for 10 minutes, and the clarified lysate was decanted and kept on ice.

For the T_{50} measurements, 125 μl of clarified lysate from a single mutant was added to all 12 wells in a row of a 96-well hard-shell thin-wall microplate (MJ Research). The plate was heated for 10 minutes using the gradient method of an Eppendorf Mastercycler gradient

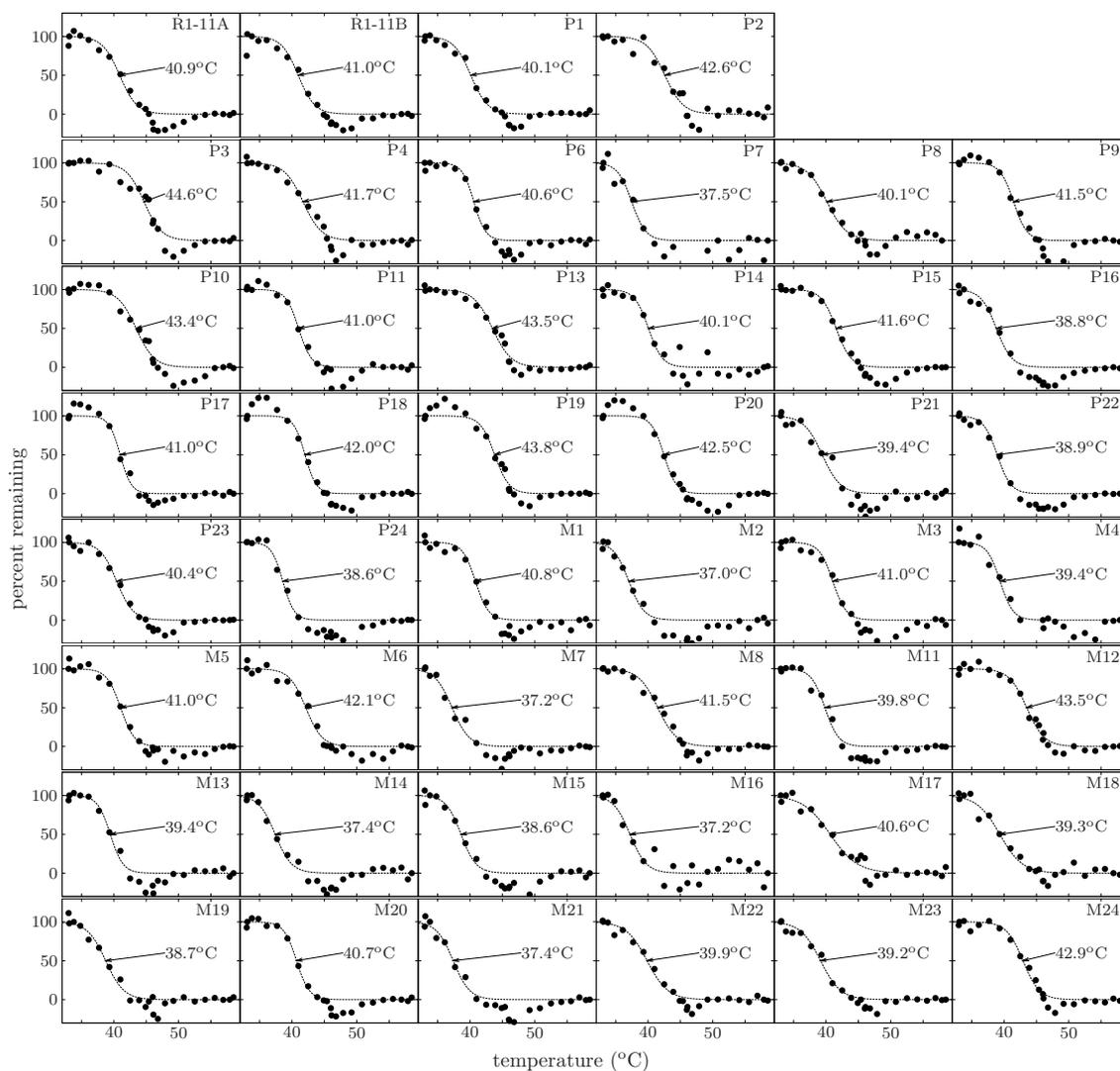


Figure 5.6: Thermostability measurements. The stability to irreversible thermal denaturation was quantified as T_{50} , the temperature at which half of the protein irreversibly denatured after a 10 minute incubation. Curves show the percent of the protein remaining after incubation at the indicated temperatures. The T_{50} was determined by fitting the data to a sigmoidal curve of the form $f(t) = 100 / (1 + e^{a(T-T_{50})})$ where T is the temperature and $f(T)$ is the percent of protein remaining at temperature T .

PCR machine, with the gradient set at either 33°C-45°C or 46°C-58°C (each mutant was exposed to both of these gradients), the machine on the BLOCK setting, and the heated lid set to 75°C with the lid WAIT option. The plate was then cooled to 4°C, removed from the PCR machine, and centrifuged at 5,500 g and 4°C for 5 minutes to pellet any debris. A pipetting robot was used to dispense 80 μ l of the supernatant into a 96-well microtiter plate (Rainin), and the amount of remaining properly folded P450 was quantified from the carbon monoxide difference spectrum as described below. The T_{50} values were determined by fitting sigmoidal curves the percent of remaining protein as shown in Figure 5.6. Our ability to accurately compare T_{50} values within the data set requires that each well in a given column of the gradient PCR machine be at the same temperature. We used a thermocouple to measure the temperature of the wells with the machine lid open, and confirmed that the wells were within a few tenths of a degree of the same temperature. Further evidence for the consistency of our T_{50} values comes from the fact that two independent measurements of the T_{50} for our R1-11 parent yielded values that differed by only 0.1°C. However, the absolute values of the measured temperatures are less accurate. Thermocouple measurements indicated that, with the machine lid open, the wells were \approx 1°C cooler than the indicated temperature. We were unable to ascertain the temperatures with the heated lid closed, but based on comparisons water bath measurements, the temperatures with the lid closed slightly exceeded the indicated temperatures.

For the $[\text{urea}]_{50}$ measurements, 125 μ l of the clarified lysate from a single mutant was added to all 12 wells in a row of a 96-well microtiter plate. A pipetting robot was then used to add and mix 125 μ l of a 2X solution of urea in 100 mM EPPS (pH 8.2) so that each subsequent column had a higher concentration of urea, and so that the final urea concentrations were those shown in Figure 5.7. The plates were left on the bench at room temperature for 4 hours, and the amount of remaining properly folded P450 was quantified from the carbon monoxide difference spectrum as described below. The $[\text{urea}]_{50}$ values were determined by fitting sigmoidal curves to the percent of remaining protein. Evidence for the consistency of the $[\text{urea}]_{50}$ measurements comes from the fact that two independent measurements of the $[\text{urea}]_{50}$ for our R1-11 parent yielded values that differed by only 0.01 M. In addition, the $[\text{urea}]_{50}$ and T_{50} values are highly correlated (Figure 5.8), indicating that they provide consistent measures of stability.

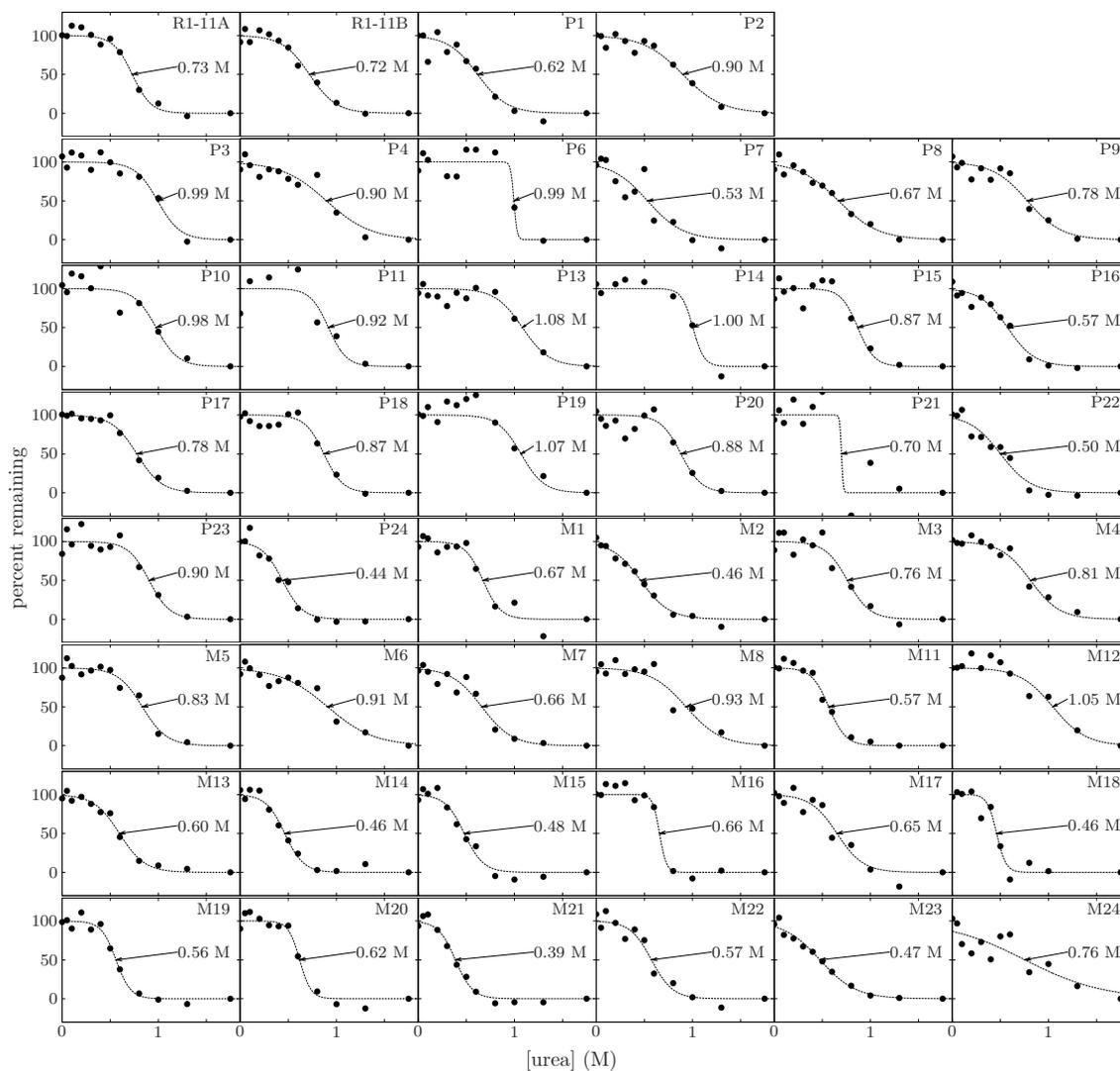


Figure 5.7: Urea stability measurements. The stability to irreversible urea denaturation was quantified as the $[\text{urea}]_{50}$, the urea concentration at which half of the protein irreversibly denatured after a 4 hour incubation at room temperature. Curves show the percent of the protein remaining after incubation at the indicated urea concentrations. The $[\text{urea}]_{50}$ was determined by fitting the data to a sigmoidal curve of the form $f(u) = 100 / \left(1 + e^{a(u - [\text{urea}]_{50})} \right)$ where u is the urea concentration and $f(u)$ is the percent of protein remaining at urea concentration u .

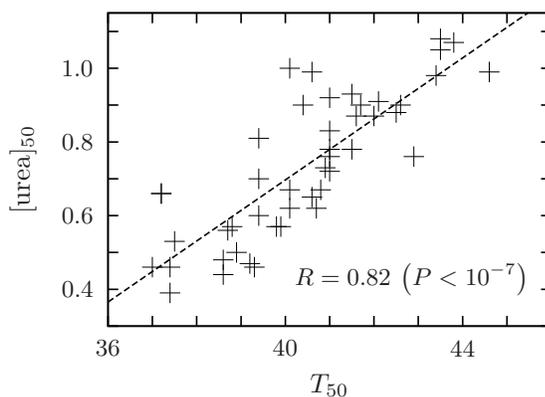


Figure 5.8: Correlation of thermal and urea stabilities. The T_{50} and $[\text{urea}]_{50}$ values are highly correlated. The plot shows the Pearson correlation coefficient and associated P -value.

For both the T_{50} and $[\text{urea}]_{50}$ measurements, the folded P450 was quantified from the carbon monoxide difference spectrum [66]. The microtiter plates containing the P450 samples were first used to read blank spectra at 450 and 490 nm using a Tecan Safire 2 plate reader. The plates were then incubated for 10 minutes in an airtight oven with carbon monoxide. The plates were removed from the oven and 10 μl of 0.1 M sodium hydrosulfite in 1.3 M potassium phosphate (pH 8.0) was immediately added to each well. After 5-10 minutes, spectra were again read at 450 and 490 nm. The amount of P450 is proportional to the increase in the signal at 450 nm after this procedure minus the change in the signal at 490 nm.

5.6 Mathematical Appendix

5.6.1 Mathematical background

The first purpose of this appendix is to provide mathematical equations that describe the experiments. The second is to show how four measurements from the experiments can be used to calculate two quantities that describe the topology of the underlying protein neutral network. We will derive two equations for both quantities, each in terms of a different measurement. The fact that the four equations will be seen to yield consistent results provides evidence for the accuracy of the following calculations.

Our calculations are based on a view of neutral protein evolution as a process constrained

by a stability threshold, a view that we originally introduced to explain experimental protein mutagenesis results [7]. The calculations closely parallel our earlier work [95], which is in turn based on a general theoretical treatment of evolution on neutral networks by van Nimwegen and coworkers [105]. These calculations will probably be most thoroughly understood by first reading those works. The primary difference between the current calculations and [95] is that previously we assumed that the per generation per protein mutation rate μ was $\ll 1$, so that at each generation a protein was either unmutated (with probability $1 - \mu$) or experienced a single mutation (with probability μ). In contrast, here we allow the mutation rate to be arbitrarily large, so that a protein may experience multiple mutations in a single generation (in this sense the calculations resemble the generalization by Wilke [133] of [105]). Specifically, let f_m be the probability that a protein experiences m mutations in a single generation. Here we derive results for arbitrary f_m , and then approximations relevant to the form of f_m in the experiments. In the limiting case of small mutation rate (where $f_0 = 1 - \mu$, $f_1 = \mu$, and $f_m = 0$ for $m > 1$), the calculations here reduce to those in [95]. Proteins evolving in nature typically experience very low mutation rates, so [95] probably offers the best description of natural protein evolution. The calculations presented here are designed to specifically treat the evolutionary dynamics of the experiments.

A protein's thermodynamic stability is described by its free energy of folding, ΔG_f , with more negative values indicating more stable proteins. As described in several previous papers [7, 95, 49], we assume that selection requires a protein to fold with some minimal stability ΔG_f^{\min} , so that a protein adequately folds if and only if $\Delta G_f \leq \Delta G_f^{\min}$. The amount of extra stability a protein possesses relative to the stability threshold is given by $\Delta G_f^{\text{extra}} = \Delta G_f - \Delta G_f^{\min}$; note that all folded proteins will have $\Delta G_f^{\text{extra}} \leq 0$. We further assume that as long as $\Delta G_f^{\text{extra}} \leq 0$, selection is indifferent to the exact amount of extra stability that a protein possesses (see [95] for a discussion of the limitations of this assumption). We conceptually divide the continuous variable of protein stability into small discrete bins of width b . Specifically, a protein is in bin i if it has $\Delta G_f^{\text{extra}}$ between $(1 - i)b$ and $-ib$, where $i = 1, 2, \dots$. Mutating a protein changes its stability by an amount $\Delta \Delta G$ (defined as the stability of the mutant protein minus the stability of the initial protein), and so may move it to a new stability bin. In [95], we defined a matrix \mathbf{W} with elements W_{ij} giving the transition probabilities that a single mutation changes a protein's stability

from bin j to bin i . We noted that \mathbf{W} could be computed from the distribution of $\Delta\Delta G$ values for all single mutations, and argued that \mathbf{W} remains fairly constant during neutral evolution since the distribution of $\Delta\Delta G$ values remains relatively unchanged. However, we emphasize that (as discussed in detail in [95]) the constancy of the $\Delta\Delta G$ distribution remains an assumption, albeit one that has now been shown to be quite accurate for lattice proteins [7, 95, 119] and provide a consistent theoretical explanation for a growing body of experimental results (the current work as well as [7]).

Since we are allowing for larger mutation rates, and we must consider the possibility that a protein’s stability might change due to multiple mutations at a single generation. Therefore, we make a more general definition of $W_{ij,m}$ as the probability that m random mutations to a protein in stability bin j change its stability to bin i , and let \mathbf{W}_m be the matrix with elements $W_{ij,m}$. Note that \mathbf{W}_m only describes mutations that cause transitions from one folded protein to another, since the stability bins $i = 1, 2, \dots$ all correspond to folded proteins. As before [95], we assume that \mathbf{W}_m is roughly constant during evolution, meaning that the distribution of $\Delta\Delta G$ values for multiple mutations is roughly constant during neutral evolution. Note that if $m = 1$, then \mathbf{W}_m is just the matrix \mathbf{W} that can be computed from the distribution of single-mutant $\Delta\Delta G$ values [95]. We will now use the matrices \mathbf{W}_m to calculate the following characteristics of a population that has evolved to equilibrium: the distribution of stabilities, the average number of mutations $\langle m \rangle_T$ accumulated after T generations, and the average fraction $\langle \mathcal{F} \rangle$ of stably folded proteins in the population. We then introduce a few approximations (that should be quite accurate for the experimental work in this paper) that greatly simplify these calculations. Finally, we relate the calculations to properties of the underlying protein neutral network.

As described generally by van Nimwegen and coworkers [105], the evolutionary dynamics depend on whether the evolving population tends to be monomorphic or highly polymorphic. When the per sequence per generation mutation rate μ is $\ll 1$, whether the population is mostly monomorphic or highly polymorphic is determined by the product of the population size N and μ : when $N\mu \ll 1$ the population is mostly monomorphic, and when $N\mu \gg 1$ the population is highly polymorphic. However, with multiple mutations per generation, $N\mu$ is no longer an appropriate parameter to distinguish between mono- and polymorphism, since if the population size is sufficiently small the population can still be monomorphic

even if there are multiple mutations per generation. Specifically, in one set of experiments we constrained the population to be monomorphic (by maintaining a population size of one), but still allowed the single protein in this population to experience more than one mutation at a generation. So we instead denote the populations as either monomorphic or polymorphic. We indicate quantities calculated for the monomorphic population by the subscript M (i.e. $\langle \mathcal{F} \rangle_M$) and those calculated for the polymorphic population by the subscript P (i.e. $\langle \mathcal{F} \rangle_P$).

5.6.2 Monomorphic limit

In the limit of a completely monomorphic population, all of the proteins are in a single stability bin. Let $p_i(t)$ be the probability that the population is in stability bin i at time t , and let $\mathbf{p}(t)$ be the column vector with elements $p_i(t)$. At each generation there is a probability f_0 that there is no mutation that becomes fixed in the population, a probability of $\sum_{m=1}^{\infty} f_m \sum_j W_{ij,m} p_j$ that the population experiences a mutational event (which could be a single mutation or several simultaneous mutations) that moves it into bin i , and a probability $\sum_{m=1}^{\infty} f_m p_i \sum_j W_{ji,m}$ that the population is in bin i and experiences one or more mutations that move it to another bin of stably folded proteins. Define $\nu_{i,m} = \sum_j W_{ji,m}$ to be the fraction of m -mutants of a protein in bin i that still fold, and let \mathbf{V}_m be the matrix with diagonal elements given by $V_{ii,m} = \nu_{i,m}$ and all other elements zero. The time evolution of \mathbf{p} is

$$\mathbf{p}(t+1) = \left[\mathbf{I} + \sum_{m=1}^{\infty} f_m (\mathbf{W}_m - \mathbf{V}_m) \right] \mathbf{p}(t) \quad (5.1)$$

where \mathbf{I} is the identity matrix. Note that mutations that destabilize a protein beyond the stability threshold are immediately lost to natural selection, and so leave the population in its original stability bin. This describes the experiments for the monomorphic populations, where we retain the parental sequence if the single mutant we generate is nonfunctional. Equation 5.1 corresponds to Equation (1) of [95], and the blind ant random walk described by van Nimwegen and coworkers [105].

Equation 5.1 describes a Markov process with a non-negative, irreducible, and acyclic transition matrix, and so \mathbf{p} approaches a unique stationary distribution (equilibrium value)

of \mathbf{p}_M given by the eigenvector equation

$$\mathbf{p}_M = \left[\mathbf{I} + \sum_{m=1}^{\infty} f_m (\mathbf{W}_m - \mathbf{V}_m) \right] \mathbf{p}_M. \quad (5.2)$$

Once \mathbf{p} has reached equilibrium, the average fraction of proteins that still stably fold at each generation is

$$\langle \mathcal{F} \rangle_M = \mathbf{e} \left(f_0 \mathbf{I} + \sum_{m=1}^{\infty} f_m \mathbf{W}_m \right) \mathbf{p}_M \quad (5.3)$$

where $\mathbf{e} = (1, \dots, 1)$ is the unit row vector.

To calculate $\langle m \rangle_{T,M}$, the average number of mutations accumulated after T generations once the population has equilibrated, we note that at each generation there is a probability of $f_m p_j \sum_i W_{ij,m}$ that a randomly chosen protein is in bin j , experiences m mutations, and still stably folds. The average number of mutations accumulated in a single generation is simply the average of m weighted over this probability. So summing over all values of m and j , we see that

$$\langle m \rangle_{T,M} = T \mathbf{e} \sum_{m=0}^{\infty} m f_m \mathbf{W}_m \mathbf{p}_M. \quad (5.4)$$

This equation corresponds to Equation (6) of [95], which was derived using an embedded Markov process formalism. Here we have foregone this formalism for the more intuitive argument presented above, since we do not attempt to calculate higher moments of the number of mutations.

5.6.3 Polymorphic limit

In the limit when the population is highly polymorphic, at each generation there are sequences in many different stability bins. In this case, we describe the distribution of stabilities by the column vector $\mathbf{x}(t)$, with element $x_i(t)$ giving the fraction of proteins in stability bin i at time t . At generation t , the fraction of mutants that continue to fold is

$$\langle \mathcal{F} \rangle_t = \mathbf{e} \left(f_0 \mathbf{I} + \sum_{m=1}^{\infty} f_m \mathbf{W}_m \right) \mathbf{x}(t). \quad (5.5)$$

Therefore, in order to maintain a constant population size, each remaining protein must produce an average of $\alpha_t = \langle \mathcal{F} \rangle_t^{-1}$ offspring. The population therefore evolves according

to

$$\mathbf{x}(t+1) = \alpha_t \left(f_0 \mathbf{I} + \sum_{m=1}^{\infty} f_m \mathbf{W}_m \right) \mathbf{x}(t). \quad (5.6)$$

After the population evolves for a sufficiently long period of time, \mathbf{x} will approach an equilibrium value of \mathbf{x}_P . At this equilibrium, the average fraction of mutants that fold at each generation is

$$\langle \mathcal{F} \rangle_P = \mathbf{e} \left(f_0 \mathbf{I} + \sum_{m=1}^{\infty} f_m \mathbf{W}_m \right) \mathbf{x}_P, \quad (5.7)$$

and the equilibrium reproduction rate is $\alpha = \langle \mathcal{F} \rangle_P^{-1}$. Therefore,

$$\mathbf{x}_P = \alpha \left(f_0 \mathbf{I} + \sum_{m=1}^{\infty} f_m \mathbf{W}_m \right) \mathbf{x}_P. \quad (5.8)$$

Equations 5.7 and 5.8 can be combined to show that \mathbf{x}_P and $\langle \mathcal{F} \rangle_P$ can be calculated from the eigenvector equation

$$(\langle \mathcal{F} \rangle_P - f_0) \mathbf{x}_P = \sum_{m=1}^{\infty} f_m \mathbf{W}_m \mathbf{x}_P, \quad (5.9)$$

with $(\langle \mathcal{F} \rangle_P - f_0)$ the principal eigenvalue of the nonnegative and irreducible matrix $\sum_{m=1}^{\infty} f_m \mathbf{W}_m$. Equation 5.9 corresponds to Equation (14) of [95], Equation (6) of the work by van Nimwegen and coworkers [105], and Equation (13) of the work by Wilke [133].

We now calculate $\langle m \rangle_{T,P}$, the average number of mutations accumulated in T generations after the population has equilibrated. At equilibrium, there is a probability of $f_m x_j \sum_i W_{ij,m}$ that a protein is in bin j , experiences m mutations, and still stably folds. Subsequently, all of these folded proteins produce an average of α offspring. The average number of mutations accumulated in a single generation is simply the average of m weighted over this probability, and then multiplied by the average reproduction rate. So summing over all values of m and j , we obtain

$$\langle m \rangle_{T,P} = \alpha T \mathbf{e} \sum_{m=0}^{\infty} m f_m \mathbf{W}_m \mathbf{x}_P = \frac{T}{\langle \mathcal{F} \rangle_P} \mathbf{e} \sum_{m=0}^{\infty} m f_m \mathbf{W}_m \mathbf{x}_P. \quad (5.10)$$

This equation is the counterpart of Equation (18) of [95], where we have again foregone the embedded Markov process formalism for a more intuitive derivation.

5.6.4 Approximations for polymorphic limit

We can dramatically simplify the results from the previous sections with several reasonable approximations. The first approximation is that the $\Delta\Delta G$ values for random mutations are roughly additive, and is supported by a number of experimental studies of the thermodynamic effects of mutations [28, 29, 76]. We have previously shown that this approximation can be used to accurately describe experimental protein mutagenesis data with a simple stability threshold model [7]. Under this approximation, the distribution of net $\Delta\Delta G$ values for multiple mutations can be computed from the distribution of $\Delta\Delta G$ values for single mutations by performing convolutions of the single-mutation $\Delta\Delta G$ distribution [7], meaning that \mathbf{W}_m for arbitrary m can be computed solely from the distribution of $\Delta\Delta G$ values for single mutations. However, to simplify the equations from previous sections, we need to express \mathbf{W}_m for arbitrary m only in terms of \mathbf{W} (recall that $\mathbf{W} = \mathbf{W}_1$). Since \mathbf{W} only contains information about stability transitions from folded proteins to other folded proteins, if we make the second approximation that a protein that is destabilized beyond the minimal stability threshold by one mutation is not re-stabilized to a folded protein by a subsequent mutation, then $\mathbf{W}_m = \mathbf{W}^m$. This approximation that unfolded proteins are not re-stabilized should be quite accurate since stabilizing mutations tend to be relatively rare and small in magnitude [77, 78, 79, 27] (this is the underlying idea behind the Markov chain approximation that was shown to be highly accurate for lattice proteins [119]). To summarize, if $\Delta\Delta G$ values are roughly additive and stabilizing mutations are rare, we have the approximation

$$\mathbf{W}_m \approx \mathbf{W}^m. \quad (5.11)$$

Simplifying the equations of the previous sections also requires assigning a specific functional form to f_m , the probability that a sequence undergoes m mutations. Here we assume that mutations are Poisson distributed among sequences, so that

$$f_m = \frac{e^{-\mu} \mu^m}{m!} \quad (5.12)$$

where $\mu = \sum_{m=0}^{\infty} m f_m$ is the average number of mutations per protein per generation. When the mutations are introduced by error-prone PCR, the Poisson distribution is an excel-

lent approximation to the true theoretical distribution of mutations created by error-prone PCR [48, 64] provided that μ is much less than the number of PCR doublings, as is the case in all of the experiments in the current work.

We now use the approximations of Equations 5.11 and 5.12 to simplify the results given above for the highly polymorphic limit. We begin by using these approximations to rewrite Equation 5.9 as

$$(\langle \mathcal{F} \rangle_P - e^{-\mu}) \mathbf{x}_P = e^{-\mu} \sum_{m=1}^{\infty} \frac{\mu^m}{m!} \mathbf{W}^m \mathbf{x}_P. \quad (5.13)$$

This equation makes clear that \mathbf{x}_P is the principal eigenvector of the matrix $\sum_{m=1}^{\infty} \frac{\mu^m}{m!} \mathbf{W}^m$, therefore \mathbf{x}_P must also be the principal eigenvector of \mathbf{W} . Now in our earlier work [95], we defined the principal eigenvector of \mathbf{W} as \mathbf{x}_∞ , called the corresponding eigenvalue $\langle \nu \rangle_\infty$, and showed that this eigenvalue is shown the average fraction of single mutations that are neutral in a population that is evolving with $N\mu \gg 1$ and $\mu \ll 1$. Therefore, with the approximation of Equation 5.11, \mathbf{x}_P and \mathbf{x}_∞ are equal, and are both defined by the same eigenvector equation,

$$\langle \nu \rangle_\infty \mathbf{x}_P = \mathbf{W} \mathbf{x}_P = \mathbf{W} \mathbf{x}_\infty = \langle \nu \rangle_\infty \mathbf{x}_\infty. \quad (5.14)$$

Combining Equations 5.13 and 5.14 we have,

$$\begin{aligned} \langle \mathcal{F} \rangle_P \mathbf{x}_P &= e^{-\mu} \sum_{m=0}^{\infty} \frac{(\mu \langle \nu \rangle_\infty)^m}{m!} \mathbf{x}_P \\ &= e^{-\mu(1 - \langle \nu \rangle_\infty)} \mathbf{x}_P \end{aligned} \quad (5.15)$$

Equation 5.15 can be solved to yield

$$\langle \nu \rangle_\infty = 1 + \frac{\ln \langle \mathcal{F} \rangle_P}{\mu}. \quad (5.16)$$

Similarly, we can simplify Equation 5.10,

$$\begin{aligned}
\langle m \rangle_{T,P} &= \frac{T}{\langle \mathcal{F} \rangle_P} \mathbf{e} \sum_{m=1}^{\infty} m f_m \mathbf{W}_m \mathbf{x}_P \\
&= T e^{\mu(1-\langle \nu \rangle_{\infty})} \sum_{m=1}^{\infty} m e^{-\mu} \frac{\mu^m}{m!} \mathbf{e} \mathbf{W}^m \mathbf{x}_P \\
&= T e^{-\mu \langle \nu \rangle_{\infty}} \sum_{m=1}^{\infty} m \frac{(\mu \langle \nu \rangle_{\infty})^m}{m!} \\
&= T \mu \langle \nu \rangle_{\infty} e^{-\mu \langle \nu \rangle_{\infty}} \sum_{m=0}^{\infty} \frac{(\mu \langle \nu \rangle_{\infty})^m}{m!} \\
&= T \mu \langle \nu \rangle_{\infty}.
\end{aligned} \tag{5.17}$$

Solving this equation for $\langle \nu \rangle_{\infty}$ yields

$$\langle \nu \rangle_{\infty} = \frac{\langle m \rangle_{T,P}}{T \mu}. \tag{5.18}$$

5.6.5 Approximations for monomorphic limit

We now simplify the equations for the monomorphic limit. This requires several further approximations. We begin by approximating that the stability probability distribution \mathbf{p}_M given by Equation 5.2 by the distribution \mathbf{p}_0 defined in [95] as satisfying

$$0 = (\mathbf{W} - \mathbf{V}) \mathbf{p}_0. \tag{5.19}$$

The basic rationale behind approximating \mathbf{p}_M with \mathbf{p}_0 is that Equation 5.2 can be viewed as a perturbation to Equation 5.19 [147]. Essentially, \mathbf{p}_0 is an eigenvector of the matrix $\mathbf{W} - \mathbf{V}$ while \mathbf{p}_M is the corresponding eigenvector of the matrix $\mathbf{W} - \mathbf{V} + \sum_{m=2}^{\infty} \frac{\mu^{m-1}}{m!} (\mathbf{W}^m - \mathbf{V}_m)$.

The latter matrix can be viewed as a perturbation to the first, since the sum $\sum_{m=2}^{\infty} \frac{\mu^{m-1}}{m!} (\mathbf{W}^m - \mathbf{V}_m)$ is small. This smallness is due to the fact that \mathbf{W}^m tends to zero with large m , causing \mathbf{V}_m to tend towards the identity matrix. In addition, the $\mu^m/m!$ terms tend to zero with large m . Therefore, the terms in the summation are all simply either a perturbation to $\mathbf{W} - \mathbf{V}$ or involve subtracting terms that are fractions of the identity matrix. The perturbations lead to bounded changes in the eigenvectors [147], while the identity matrix terms do not change the eigenvectors. Below we give a more rigorous justification of the assumption that

\mathbf{p}_M is approximately equal to \mathbf{p}_o .

We need one additional approximation to make further progress. Both Equations 5.3 and 5.4 contain terms of the form $\mathbf{W}_m \mathbf{p}_o$, and even if we use Equation 5.11 to rewrite these terms as $\mathbf{W}^m \mathbf{p}_o$, there are no further clear simplifications. However, any probability vector that is multiplied repeatedly by \mathbf{W} and normalized will eventually converge to $\mathbf{x}_\infty = \mathbf{x}_P$ (since this is the principal eigenvector of \mathbf{W}). We make the approximation that this convergence is sufficiently rapid to be essentially complete after a single multiplication. This approximation is supported by both protein mutagenesis studies [7, 17, 18] that indicate that proteins rapidly converge to an exponential decline in the fraction folded (indicating the stability distribution has equilibrated, as discussed below, and by lattice protein studies showing the same [7, 119]. Therefore, we make the approximation that $\mathbf{e} \mathbf{W}^m \mathbf{p}_o = \langle \nu \rangle_o \mathbf{e} \mathbf{W}^{m-1} \mathbf{x}_\infty = \langle \nu \rangle_o \langle \nu \rangle_\infty^{m-1}$ where $\langle \nu \rangle_o = \mathbf{e} \mathbf{W} \mathbf{p}_o$ has the same definition as in [95], where it was defined as the average fraction of functional single mutants of a population evolving with $\mu \ll 1$ and $N\mu \ll 1$.

We use these approximations to simplify Equation 5.3 as

$$\begin{aligned}
\langle \mathcal{F} \rangle_M &= \mathbf{e} \left(f_0 \mathbf{I} + \sum_{m=1}^{\infty} f_m \mathbf{W}_m \right) \mathbf{p}_M \\
&= e^{-\mu} \left[1 + \sum_{m=1}^{\infty} \frac{\mu^m}{m!} \mathbf{e} \mathbf{W}^m \mathbf{p}_o \right] \\
&= e^{-\mu} \left[1 + \mu \langle \nu \rangle_o \sum_{m=1}^{\infty} \frac{(\mu \langle \nu \rangle_\infty)^{m-1}}{m!} \right] \\
&= e^{-\mu} \left[1 + \frac{\langle \nu \rangle_o}{\langle \nu \rangle_\infty} \left(-1 + \sum_{m=0}^{\infty} \frac{(\mu \langle \nu \rangle_\infty)^m}{m!} \right) \right] \\
&= e^{-\mu} \left[1 + \frac{\langle \nu \rangle_o}{\langle \nu \rangle_\infty} \left(e^{\mu \langle \nu \rangle_\infty} - 1 \right) \right]. \tag{5.20}
\end{aligned}$$

Solving this equation for $\langle \nu \rangle_o$, we find

$$\langle \nu \rangle_o = \frac{\langle \nu \rangle_\infty (\langle \mathcal{F} \rangle_M e^\mu - 1)}{e^{\mu \langle \nu \rangle_\infty} - 1}. \tag{5.21}$$

We now use the approximations to simplify Equation 5.4 as

$$\begin{aligned}
\langle m \rangle_{T,M} &= T e \sum_{m=0}^{\infty} m f_m \mathbf{W}_m \mathbf{P}_M \\
&= T e^{-\mu} \sum_{m=1}^{\infty} m \frac{\mu^m}{m!} e \mathbf{W}^m \mathbf{P}_o \\
&= T e^{-\mu} \langle \nu \rangle_o \sum_{m=1}^{\infty} m \frac{\mu^m}{m!} \langle \nu \rangle_{\infty}^{m-1} \\
&= \mu T e^{-\mu} \langle \nu \rangle_o \sum_{m=0}^{\infty} \frac{(\mu \langle \nu \rangle_{\infty})^m}{m!} \\
&= \mu T \langle \nu \rangle_o e^{\mu(\langle \nu \rangle_{\infty}-1)}. \tag{5.22}
\end{aligned}$$

Solving this equation for $\langle \nu \rangle_o$ yields

$$\langle \nu \rangle_o = \frac{\langle m \rangle_{T,M} e^{\mu(1-\langle \nu \rangle_{\infty})}}{\mu T}. \tag{5.23}$$

To recap, we now have equations to calculate $\langle \nu \rangle_{\infty}$ and $\langle \nu \rangle_o$ from experimentally measurable quantities. Equations 5.16 and 5.18 allow us to calculate $\langle \nu \rangle_{\infty}$ from $\langle \mathcal{F} \rangle_P$ and $\langle m \rangle_{T,P}$, respectively. Given this calculated value of $\langle \nu \rangle_{\infty}$, Equations 5.21 and 5.23 then allow us to calculate $\langle \nu \rangle_o$ from $\langle \mathcal{F} \rangle_M$ and $\langle m \rangle_{T,M}$, respectively. The fact that we have two equations each for $\langle \nu \rangle_{\infty}$ and $\langle \nu \rangle_o$ allows us to assess the self-consistency of the approach.

5.6.6 Interpretation in terms of neutral networks

Throughout the preceding calculations, we have referred to $\langle \nu \rangle_{\infty}$ and $\langle \nu \rangle_o$ as we defined them in [95]: namely, as the average neutrality of protein populations evolving with $\mu \ll 1$ and $N\mu$ either $\gg 1$ or $\ll 1$, respectively. However, van Nimwegen and coworkers [105] have shown that they can also be interpreted in terms of the underlying neutral network. In the experiments we make mutations at the nucleotide (rather than amino acid) level, so each point in our sequence space corresponds to a different gene. Every gene that yields an amount of protein sufficient to hydroxylate the twelfth carbon of 12-*p*-nitrophenoxydodecanoic acid with at least 75% of the total activity conferred by the original R1-11 parent gene represents a node on this neutral network. We note that in the experiments (and also usually

in natural evolution), the edges on the neutral network are not all completely equivalent or fully undirected, since some mutations are more likely to occur than others (for example, error-prone PCR with *Taq* polymerase is more likely to cause an A→G mutation than an A→C mutation). In the analysis that follows, we ignore this complication and assume all neutral network edges are equivalent.

In an extremely insightful analysis, van Nimwegen and coworkers [105] have shown that important characteristics of a neutral network can be inferred from evolutionary quantities. Specifically, they have shown that if a population is evolving with $\mu \ll 1$ and $N\mu \gg 1$, then the average neutrality (which we have denoted by $\langle \nu \rangle_\infty$) is equal to the principal eigenvalue of the adjacency matrix of the neutral network, normalized by the network coordination number (number of possible connections per node). In addition, they pointed out that a population evolving with $\mu \ll 1$ and $N\mu \ll 1$ moves like a blind ant random walk, meaning that the average neutrality (which we have denoted by $\langle \nu \rangle_o$) is equal to the average connectivity of a neutral network node divided by the network coordination number. In our P450 experiments, we have measured the values needed to estimate $\langle \nu \rangle_\infty$ and $\langle \nu \rangle_o$ using Equations 5.16, 5.18, 5.21, and 5.23. Using the final values listed in Table 5.2, $\langle \mathcal{F} \rangle_P = 0.50$ and $\langle \mathcal{F} \rangle_M = 0.39$. Taking the final nucleotide mutation values from Table 5.2, $\langle m \rangle_{T,P}/T = 0.69$ and $\langle m \rangle_{T,M}/T = 0.31$. The average mutation rate, computed from the unselected population, is $\mu = 1.40$. So using Equation 5.16, $\langle \nu \rangle_\infty = 0.53$, while using Equation 5.18, $\langle \nu \rangle_\infty = 0.49$. The consistency of these two values supports the idea that the calculations above accurately describe the evolutionary process. Taking the average value of these two measurements as $\langle \nu \rangle_\infty = 0.51$, we can then use Equations 5.21 and 5.23 to calculate $\langle \nu \rangle_o$. We calculate values of 0.28 and 0.43, respectively. These estimates differ by more than those for $\langle \nu \rangle_\infty$, perhaps because additional approximations have gone into the derivation of the relevant equations (in addition, we have made no attempt to carry out the rather complex propagation of the sampling errors of Table 5.2). However, the values are still in a similar range. Taking the average of these two values, we estimate that $\langle \nu \rangle_o = 0.35$. So overall, we predict that each functional P450 gene should have an average fraction of 0.35 of its sequence nearest neighbors also encoding a functional gene, for an average of about 1,500 neighbor genes. We predict that the principal eigenvalue of the neutral network adjacency matrix is $0.51 \times 3L$. The fact that principal eigenvalue

exceeds the average connectivity indicates that the neutral network is not a regular graph, but instead has some nodes more highly connected than others.

The value for $\langle \nu \rangle_\infty$ calculated above can also be related to measurements from protein mutagenesis experiments. A number of studies [7, 17, 18] have observed that the probability that a protein remains functional after m mutations falls off exponentially with the number of mutations. In fact, the decline is not always exponential for the first few mutations if the starting protein has especially high or low stability [7] or activity [148], but will still converge to this exponential form after a few mutations [7, 119, 149]. The stability threshold model can be used to relate this decline to $\langle \nu \rangle_\infty$, as is done indirectly in the Markov chain approximation of [119]. Here we make that connection explicit. The initial protein has a stability that falls into some stability bin i . Therefore, its stability can be described by the column vector \mathbf{y}_0 , which has element i equal to one and all other elements equal to zero. Now imagine constructing all single mutants of this protein. The fraction of these single mutants that still fold is just $\mathbf{e}\mathbf{W}\mathbf{y}_0$, and the distribution of stabilities among the single mutants is $\mathbf{y}_1 = \mathbf{W}\mathbf{y}_0$ (note that the elements of \mathbf{y}_1 no longer sum to one). Similarly, after m mutations, the fraction of mutants that still fold is $\mathbf{e}\mathbf{W}_m\mathbf{y}_0$, and the distribution of stabilities among the m -mutants is $\mathbf{y}_m = \mathbf{W}_m\mathbf{y}_0$. With the approximation of Equation 5.11, $\mathbf{y}_m = \mathbf{W}^m\mathbf{y}_0$. This makes it clear that \mathbf{y}_m will converge to a vector proportional to \mathbf{x}_∞ , the principal eigenvector of \mathbf{W} . Once this convergence is complete, each new mutation simply reduces the fraction of mutants that fold by a factor of $\langle \nu \rangle_\infty$, the principal eigenvalue of \mathbf{W} (and the spectral radius of the neutral network normalized by the coordination number). Therefore, what we have called $\langle \nu \rangle_\infty$ in the present work and [95] is equal to what is called x in [18], q in [17], and $\langle \nu \rangle$ in [7]. The major difficulty that is faced in extracting $\langle \nu \rangle_\infty$ by the method of those three studies [7, 17, 18] is that it is not possible to directly assay mutants with m mutations, but instead only possible to assay a set of mutants with a distribution of m . All three studies use different (and valid) methods to account for this distribution, but this accounting is still difficult because most of the functional mutants come from the low m end of the distribution. This makes it hard to get accurate values for the fraction functional after large numbers of mutations, since most of the functional mutants in the set come from sequences with few mutations. For this reason, we believe the current method of measuring $\langle \nu \rangle_\infty$ is more accurate. A second caution about comparing values of $\langle \nu \rangle_\infty$ from different

studies is that its value depends on the nucleotide error-spectrum of the experiment, since different mutagenesis methods create different distributions of nucleotide and amino acid mutation types.

We also briefly mention how we arrived at an estimate of $\langle \nu \rangle_\infty$ for 3-methyladenine DNA glycosylase from the data of [18]. This paper reports that a fraction $x = 0.34$ of amino acid mutations inactivate the protein. We would like to determine the fraction $\langle \nu \rangle_\infty$ of nucleotide mutations that do not inactivate the protein. Roughly 75% of random mutations to a gene will be synonymous. Therefore, m amino acid mutations should cause about $4m/3$ nucleotide mutations. The study of [18] measures that after m mutations, a fraction $(1 - x)^m$ of the mutants are functional. That means that $\langle \nu \rangle_\infty^{4m/3}$ fraction should be functional. Equating these expressions yields $\langle \nu \rangle_\infty = \exp\left(\frac{3}{4} \log(1 - x)\right)$. So using $x = 0.34$, we arrive at $\langle \nu \rangle_\infty = 0.73$.

5.6.7 Detailed justification for approximating \mathbf{p}_M by \mathbf{p}_o

Here we provide a detailed justification for the approximation that \mathbf{p}_M is about equal to \mathbf{p}_o . In the monomorphic limit, the time evolution of \mathbf{p} is given by Equation 5.1, and the stationary distribution \mathbf{p}_M is given by Equation 5.2. We assume the approximations of Equations 5.11 and 5.12 and show that we can approximate \mathbf{p}_M by \mathbf{p}_o , where \mathbf{p}_o is given by Equation 5.19. To justify this approximation, we insert \mathbf{p}_o into the right hand side of Equation 5.1 and ask to what extent \mathbf{p}_o is left unaltered by the dynamics. If \mathbf{p}_o is found to be stationary to good approximation then, by uniqueness of the stationary distribution of an ergodic process, \mathbf{p}_o would be a good approximation to \mathbf{p}_M .

We therefore suppose that at some time t the distribution is given by \mathbf{p}_o and compute, using Equation 5.1, the change in \mathbf{p}_o after one generation. The new distribution at time $t + 1$ is given by

$$\mathbf{p}(t + 1) = \left[\mathbf{I} + \sum_{m=1}^{\infty} f_m (\mathbf{W}^m - \mathbf{V}_m) \right] \mathbf{p}_o. \quad (5.24)$$

Using $(\mathbf{V} - \mathbf{W}) \mathbf{p}_o = 0$, and taking components of the above equation, we obtain

$$p_i(t + 1) = p_{0i} + \sum_{m=2}^{\infty} f_m [(\mathbf{W}^m - \mathbf{V}_m) \mathbf{p}_o]_i. \quad (5.25)$$

Thus \mathbf{p}_o would be an approximately stationary distribution of the dynamics if $|\sum_{m=2}^{\infty} f_m [(\mathbf{W}^m - \mathbf{V}_m) \mathbf{p}_o]_i| \ll p_{0i}$. We now proceed to show that this will be the case in most situations of interest by deriving upper and lower bounds on the second term of the right hand side of Equation 5.25.

Consider first the term $(\mathbf{W}^m \mathbf{p}_o)_i$, which can be written as

$$\begin{aligned} (\mathbf{W}^m \mathbf{p}_o)_i &= \sum_{k_1, \dots, k_m} W_{ik_1} W_{k_1 k_2} \cdots W_{k_{m-1} k_m} p_{0k_m} \\ &= \sum_{k_1, \dots, k_{m-1}} W_{ik_1} W_{k_1 k_2} \cdots W_{k_{m-2} k_{m-1}} \nu_{k_{m-1}} p_{0k_{m-1}}, \end{aligned} \quad (5.26)$$

where we have used $\mathbf{W} \mathbf{p}_o = \mathbf{V} \mathbf{p}_o$ in the second equality. We now note that $\nu_k \leq \nu_{\max}$ for all k , where ν_{\max} is the maximum neutrality, maximized over all bins. This leads to the successive inequalities

$$\begin{aligned} (\mathbf{W}^m \mathbf{p}_o)_i &\leq \nu_{\max} \sum_{k_1, \dots, k_{m-1}} W_{ik_1} W_{k_1 k_2} \cdots W_{k_{m-2} k_{m-1}} p_{0k_{m-1}} \\ &= \nu_{\max} \sum_{k_1, \dots, k_{m-2}} W_{ik_1} W_{k_1 k_2} \cdots W_{k_{m-3} k_{m-2}} \nu_{k_{m-2}} p_{0k_{m-2}} \\ &\leq \nu_{\max}^2 \sum_{k_1, \dots, k_{m-2}} W_{ik_1} W_{k_1 k_2} \cdots W_{k_{m-3} k_{m-2}} p_{0k_{m-2}} \\ &\leq \nu_{\max}^{m-1} \sum_{k_1} W_{ik_1} p_{0k_1}, \end{aligned} \quad (5.27)$$

yielding the upper bound

$$(\mathbf{W}^m \mathbf{p}_o)_i \leq \nu_{\max}^{m-1} \nu_i p_{0i}. \quad (5.28)$$

In an identical manner, we obtain the lower bound

$$(\mathbf{W}^m \mathbf{p}_o)_i \geq \nu_{\min}^{m-1} \nu_i p_{0i}, \quad (5.29)$$

where ν_{\min} is the smallest neutrality, minimized over all bins. Note that both inequalities above become exact equalities when all bins have the same neutrality ν , which could be interpreted as either ν_{\min} or ν_{\max} .

Having obtained inequality constraints on $(\mathbf{W}^m \mathbf{p}_o)_i$, we now consider the term $(\mathbf{V}_m \mathbf{p}_o)_i$,

which can be written as

$$\begin{aligned}
(\mathbf{V}_m \mathbf{p}_0)_i &= p_{0i} \nu_{i,m} \\
&= p_{0i} \sum_j (\mathbf{W}^m)_{ji} \\
&= p_{0i} \sum_{j, k_1, \dots, k_{m-1}} W_{jk_1} W_{k_1 k_2} \cdots W_{k_{m-1} i} \\
&= p_{0i} \sum_{k_1, \dots, k_{m-1}} \nu_{k_1} W_{k_1 k_2} \cdots W_{k_{m-1} i} \\
&\leq p_{0i} \nu_{\max} \sum_{k_1, \dots, k_{m-1}} W_{k_1 k_2} \cdots W_{k_{m-1} i} \\
&\leq p_{0i} \nu_{\max}^{m-1} \sum_{k_{m-1}} W_{k_{m-1} i}, \tag{5.30}
\end{aligned}$$

which yields an identical upper bound to that on $(\mathbf{W}^m \mathbf{p}_0)_i$, namely,

$$(\mathbf{V}^m \mathbf{p}_0)_i \leq \nu_{\max}^{m-1} \nu_i p_{0i}, \tag{5.31}$$

and similarly

$$(\mathbf{V}^m \mathbf{p}_0)_i \geq \nu_{\min}^{m-1} \nu_i p_{0i}. \tag{5.32}$$

It should again be noted that both the above inequalities become exact equalities when all bins have a common neutrality ν .

We are now in a position to estimate bounds on the magnitude of the second term of Equation 5.25. Using the four inequalities of Equations 5.28, 5.29, 5.31, and 5.32 above, we have

$$-(\nu_{\max}^{m-1} - \nu_{\min}^{m-1}) \nu_i p_{0i} \leq [(\mathbf{W}^m - \mathbf{V}_m) \mathbf{p}_0]_i \leq (\nu_{\max}^{m-1} - \nu_{\min}^{m-1}) \nu_i p_{0i}, \tag{5.33}$$

or equivalently,

$$|[(\mathbf{W}^m - \mathbf{V}_m) \mathbf{p}_0]_i| \leq (\nu_{\max}^{m-1} - \nu_{\min}^{m-1}) \nu_i p_{0i}, \tag{5.34}$$

where the inequality above becomes an exact equality when all bins have the same neutrality. However, in this limit, the right hand side of the above equation vanishes, and therefore the second term of Equation 5.25 is identically zero in this case, giving the result that \mathbf{p}_M

is exactly equal to \mathbf{p}_o when all bins have the same neutrality, even if μ is arbitrarily large.

We now carry out the sum over m to obtain an upper bound on the second term of Equation 5.25 in the more general and realistic case of unequal neutrality bins. Using Equation 5.34 and the specific Poisson form of f_m , we obtain an upper bound on the fractional change in p_{0i} in one generation:

$$\begin{aligned} \left| \frac{p_i(t+1) - p_{0i}}{p_{0i}} \right| &\leq \nu_i e^{-\mu} \sum_{m=2}^{\infty} \frac{\mu^m}{m!} (\nu_{\max}^{m-1} - \nu_{\min}^{m-1}) \\ &= \nu_i e^{-\mu} \left[\frac{e^{\mu\nu_{\max}} - 1}{\nu_{\max}} - \frac{e^{\mu\nu_{\min}} - 1}{\nu_{\min}} \right]. \end{aligned} \quad (5.35)$$

The above bound vanishes for small μ , is an increasing function of $\nu_{\max} - \nu_{\min}$, and is typically much smaller than 1. An extreme estimate of the size of the fractional change can be made when $\nu_{\max} = 1$ and $\nu_{\min} = 0$. In this case, using $\mu = 1.4$ (the value in our experiments), the above inequality simplifies to

$$\left| \frac{p_i(t+1) - p_{0i}}{p_{0i}} \right| \leq \nu_i (1 - e^{-\mu} - \mu e^{-\mu}) \simeq 0.41\nu_i. \quad (5.36)$$

Noting that $\nu_i < 1$, the fractional change in p_{0i} is therefore reasonably controlled even in the most extreme case. For realistic situations, the fractional change in p_{0i} is expected to be much lower, thus justifying the use of \mathbf{p}_o as the stationary distribution of the dynamics of Equation 5.1.

5.7 Acknowledgments

We thank Claus O Wilke for helpful advice and comments. JDB is supported by a HHMI predoctoral fellowship. ZL and DC were supported by Summer Undergraduate Research Fellowships from the California Institute of Technology. AR is supported by NSF grants CCF 0523643 and FIBR 0527023.

Chapter 6

Neutral Genetic Drift Can Aid Functional Protein Evolution

6.1 Abstract

Many of the mutations accumulated by naturally evolving proteins are neutral in the sense that they do not significantly alter a protein's ability to perform its primary biological function. However, new protein functions evolve when selection begins to favor other, "promiscuous" functions that are incidental to a protein's original biological role. If mutations that are neutral with respect to a protein's primary biological function cause substantial changes in promiscuous functions, these mutations could enable future functional evolution. Here we investigate this possibility experimentally by examining how cytochrome P450 enzymes that have evolved neutrally with respect to activity on a single substrate have changed in their abilities to catalyze reactions on five other substrates. We find that the enzymes have sometimes changed as much as four-fold in the promiscuous activities. The changes in promiscuous activities tend to increase with the number of mutations, and can be largely rationalized in terms of the chemical structures of the substrates. The activities on chemically similar substrates tend to change in a coordinated fashion, potentially providing a route for systematically predicting the change in one function based on the measurement of several others. Our work suggests that initially neutral genetic drift can lead to substantial changes in protein functions that are not currently under selection, in effect poisoning the proteins to more readily undergo functional evolution should selection "ask new questions" in the future.

6.2 Background

Nature employs proteins for a vast range of tasks, and their capacity to evolve to perform diverse functions is one of the marvels of biology. Recently, it has become possible to reconstruct convincing scenarios for how new protein functions evolve. One of the most important conclusions of this work is that the initial steps may occur even before the new functions come under selection [150, 54, 151, 152, 153, 154]. The reason is that in addition to their primary biological functions, most proteins are at least modestly effective at performing a range of other “promiscuous” functions [150, 54, 155, 156, 157, 158]. In laboratory experiments, selection can rapidly increase these promiscuous functions, often without much immediate cost to a protein’s original function [54]. In a particularly compelling set of experiments, Tawfik and coworkers have shown that selection for promiscuous activity likely explains the origin and evolution of a bacterial enzyme that hydrolyzes a synthetic compound only recently introduced into the environment [54, 159, 160]. Mounting evidence therefore supports the idea that new protein functions evolve when selection favors mutations that increase an existing weak promiscuous function.

But for as long as 50 years, since Linus Pauling and Emile Zuckerkandl published their seminal analysis of molecular change in proteins [3], it has been clear that just a small fraction of the mutations that accumulate in naturally evolving proteins are driven by selection for a new function. Instead, most of the mutations responsible for natural sequence divergence do not change a protein’s primary biological function, but rather are due to either neutral genetic drift [96] or pressure for a subtle recalibration of protein properties unrelated to the acquisition of an entirely new function [161]. However, even though most mutations accumulate under the constraint that they not interfere with a protein’s primary function, they could still substantially alter other, promiscuous functions. Such alterations could then aid in the subsequent evolution of new functions.

Here we have experimentally investigated this possibility using a set of enzymes that have undergone genetic drift that is neutral with respect to a well-defined laboratory selection criterion for enzymatic activity on a single substrate [135]. We have examined how these enzymes have changed in their promiscuous activities on five other substrates. As described below, we find that the enzymes have often undergone substantial changes in

their promiscuous activities, suggesting that neutral genetic drift could play an important role in enabling future functional evolution.

6.3 Results and Discussion

6.3.1 A set of neutrally evolved cytochrome P450 enzymes

We focused our analysis on cytochrome P450 proteins. P450s are excellent examples of enzymes that can evolve to catalyze new reactions, since they are involved in a wide range of important functions such as drug metabolism and steroid biosynthesis [60, 61]. We worked with P450 BM3, a cytosolic bacterial enzyme that catalyzes the subterminal hydroxylation of medium- and long-chain fatty acids [62]. We have previously described a set of P450 BM3 heme domain variants that were created by laboratory neutral evolution from a common parent sequence [135]. Here we briefly recap the procedure used to create these P450s in order to explain their origin and why they can properly be viewed as the product of neutral genetic drift.

The essential difference between neutral genetic drift and adaptive evolution is that in the former case mutations that have no substantial effect on fitness spread stochastically in a population, while in the latter case mutations spread because they are beneficial and so favored by selection. Of course, it may be difficult to discern whether a specific mutation in a natural population has spread neutrally or due to favorable selection. But in the laboratory it is possible to define an arbitrary selection criterion to ensure that all mutations spread due to neutral genetic drift. Specifically, we imposed the requirement that the P450s had to hydroxylate the substrate 12-*p*-nitrophenoxydodecanoic acid (12-pNCA) with an activity exceeding a specific threshold [135]. All mutant P450s were therefore straightforwardly classified as either functional (if they exceeded the threshold) or nonfunctional (if they did not). While this selection criterion is obviously a simplification of natural evolution, we believe that for the current purpose it is a reasonable abstraction of the evolutionary requirement that an enzyme's primary activity exceed some critical level in order to allow its host organism to robustly survive and reproduce. To implement laboratory neutral evolution using this selection criterion, we began with a single parent P450 BM3 heme domain variant (called R1-11) and used error-prone PCR to create random mutants of

this parent [135]. Mutants that failed to yield sufficient active protein to hydroxylate at least 75% of the 12-pNCA of the R1-11 parent when expressed in *Escherichia coli* were immediately eliminated, while all other mutants were carried over to the next generation with equal probability. Any mutations that spread among the offspring sequences were therefore by definition due to neutral genetic drift, since there was no opportunity for any functional mutant to be favored over any other. We emphasize that the fact that the mutations spread due to neutral genetic drift does not mean that they have no effect on the protein's properties. Indeed, one of the growing realizations about protein evolution is that mutations that spread by neutral genetic drift may still have an impact on future evolution [72, 95]. One mechanism for this impact is that neutral genetic drift can change a protein's stability and so alter its tolerance to future mutations [7, 127, 49]. As will be demonstrated below, another mechanism is that neutral genetic drift can alter a protein's promiscuous functions.

As described previously [135], the end result of the neutral evolution was 44 different P450 variants, each of which satisfied the selection criterion for activity on 12-pNCA (these are the combined final sequences from the monomorphic and polymorphic populations in [135]). For the current study, we analyzed the promiscuous activities of 34 of these neutrally evolved P450 variants. The sequence diversity of these P450s is shown in the phylogenetic tree of Figure 6.1; they have accumulated an average of four nonsynonymous mutations each.

6.3.2 Activities of the neutrally evolved P450 enzymes

All of the P450 variants had evolved under selection solely for their ability to hydroxylate 12-pNCA. We examined their promiscuous hydroxylation activities on the five other substrates shown at the top of Figure 6.2. Two of these promiscuous substrates, propranolol and 2-amino-5-chlorobenzoxazole (also known as zoxazolamine), are drugs that are metabolized by human P450s [162, 163]. The other three promiscuous substrates, 11-phenoxyundecanoic acid, 2-phenoxyethanol, and 1,2-methylenedioxybenzene, are organic compounds of increasing structural dissimilarity to 12-pNCA. The parent P450 possessed at least some hydroxylation activity on all of these substrates (throughout the remainder of this work, "activity" refers to total substrate turnovers per enzyme).

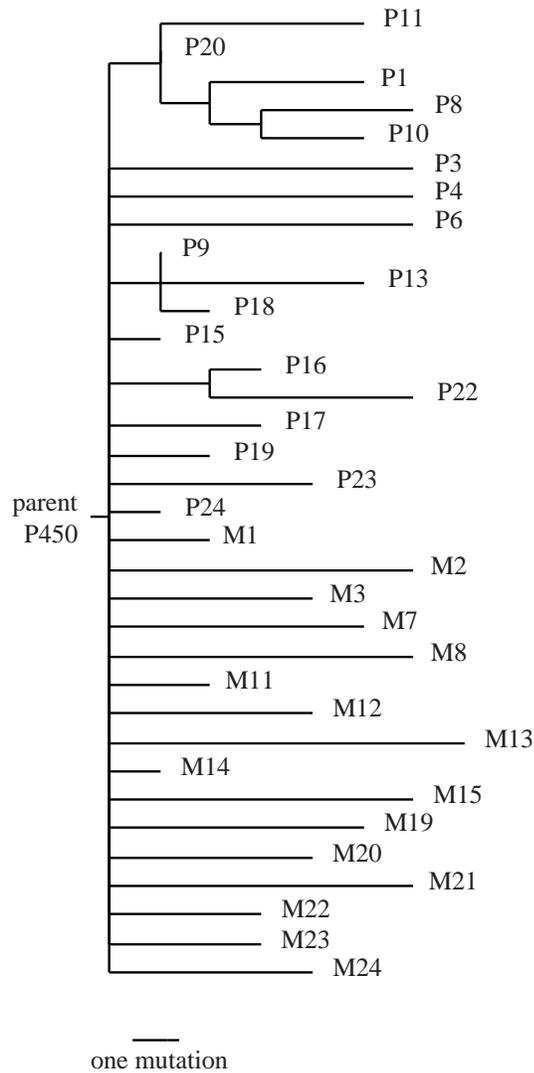


Figure 6.1: Phylogenetic tree of the neutrally evolved P450s. The tree shows the relationship among the 34 neutrally evolved P450 variants examined in this study. All of the P450s neutrally evolved from the same R1-11 parent P450. The horizontal lengths of the branches are proportional to the number of nonsynonymous mutations, as indicated by the scale bar. The vertical arrangement of the branches is arbitrary.

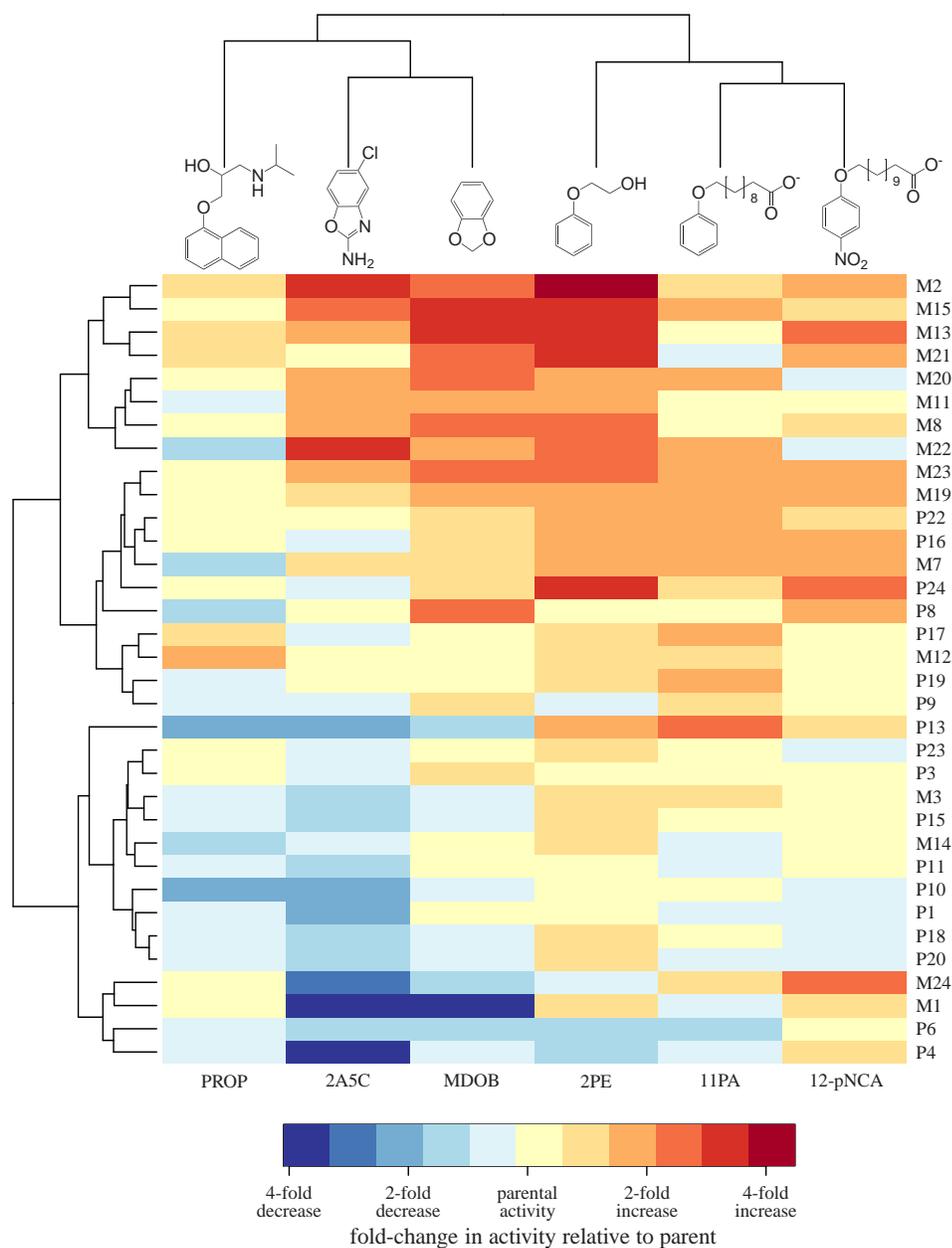


Figure 6.2: Activities of the neutrally evolved P450s on the six substrates. The heat map shows the fold change in activity of all 34 neutrally evolved P450 variant on all six substrates. Each row shows the data for a different P450 variant, while each column shows the activity on a different substrate. The fold change in activity is the ratio of the variant's activity to that of the neutral evolution parent. Both the substrates and the P450 variants are hierarchically clustered according to the activity profiles, as shown by the dendrograms at top and left. Substrate abbreviations: PROP - propranolol, 2A5C - 2-amino-5-chlorobenzoxazole, MDOB - 1,2-methylenedioxybenzene, 2PE - 2-phenoxyethanol, 11PA - 11-phenoxyundecanoic acid. The standard errors for the changes in activity displayed in the heat map tend to be much smaller than the changes themselves; these errors are shown explicitly in Figure 6.3.

We measured the activities of all 34 neutrally evolved P450s on the five promiscuous substrates as well as 12-pNCA. Figure 6.2 shows the fold change in activity of each of the variants relative to the parent P450 on all six substrates, and Figure 6.3 shows the same data with standard errors. As is apparent from these figures, many of the neutrally evolved P450s have undergone changes in their activities that substantially exceeded the standard errors of the measurements. Even on 12-pNCA, some of the variants have undergone modest increases or very mild decreases in activity. The modest increases in 12-pNCA activity were unsurprising, since the parent P450 only hydroxylates 12-pNCA with about a quarter of the activity reported for a P450 engineered for maximal 12-pNCA activity [59]. Likewise, the mild decreases in 12-pNCA activity were due to the fact that during neutral evolution the P450s were only required to maintain this activity above a minimal threshold (75% of the total 12-pNCA conversion of the parent protein when expressed in *E. coli* [135]). The changes in the promiscuous activities were often much larger than those on 12-pNCA. For example, several of the neutrally evolved variants have undergone nearly four-fold increases in activity on one or more of 2-phenoxyethanol, 2-amino-5-chlorobenzoxazole, and 1,2-methylenedioxybenzene. Other variants have experienced equally large decreases in one or more of the promiscuous activities.

6.3.3 Broad patterns of change in activity can be rationalized in terms of substrate properties

The data in Figures 6.2 and 6.3 clearly indicate that some of the P450s have undergone substantial changes in their activities. In an effort to understand the nature of these changes, we sought to determine whether there were any clear patterns in the activities. In Figure 6.2, the substrates have been hierarchically clustered so that each successive cluster contains substrates on which the P450s have increasingly similar activities (the clustering is illustrated by the tree-like dendrogram at the top of the figure, with similar substrates in adjacent columns). The clustering of the substrates is readily rationalized in terms of their chemical structures. For example, 2-amino-5-chlorobenzoxazole and 1,2-methylenedioxybenzene cluster, meaning that P450s with high activity on one of these substrates also tend to have high activity on the other. Presumably, they cluster because the similarity of their structures (both are fusions of six and five membered rings) means that they have similar modes

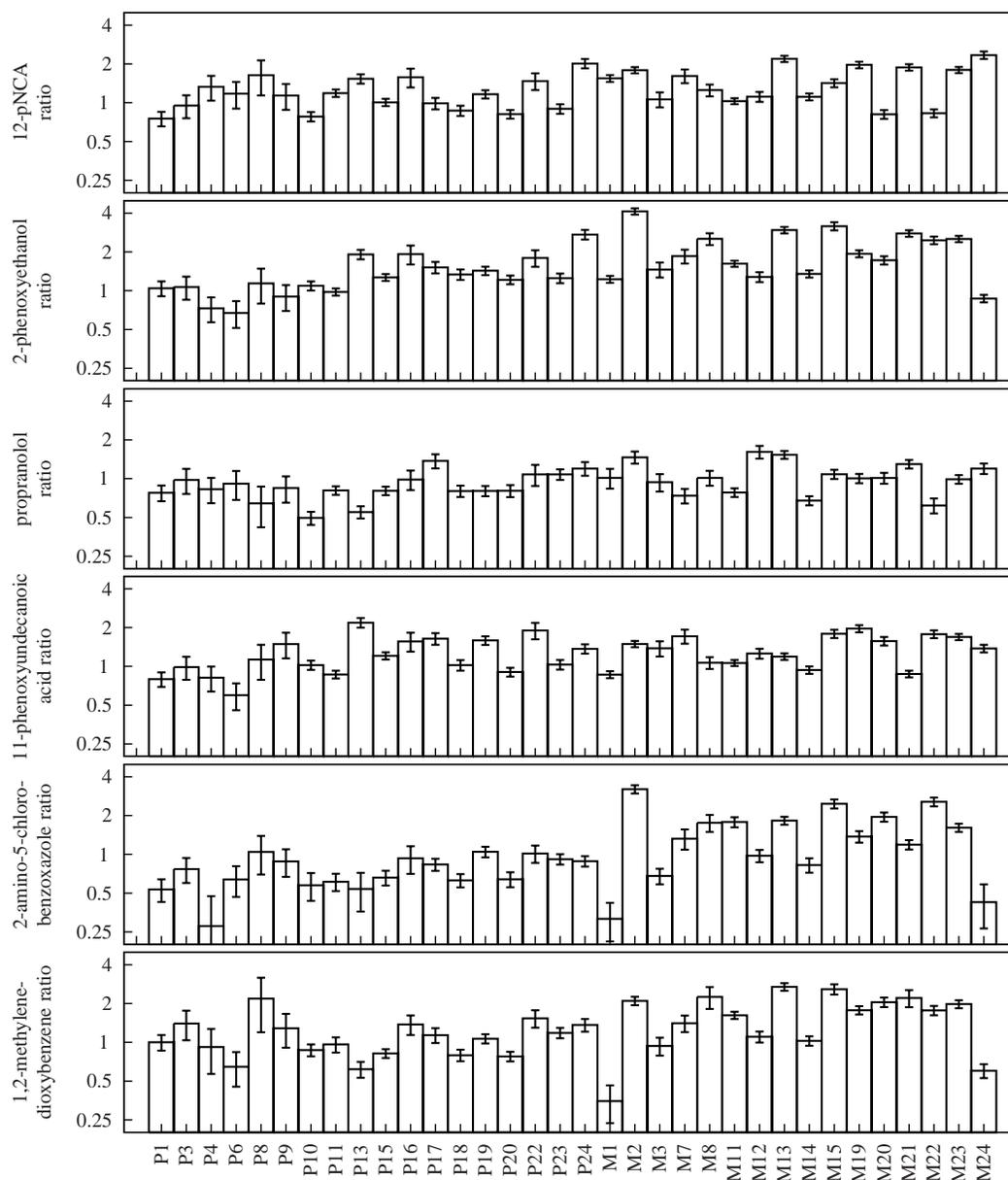


Figure 6.3: Fold changes in P450 activities with standard errors. The bar graphs show the fold change in activity of all 34 neutrally evolved P450 variants on all six substrates. This is the same data as in Figure 6.2, except these graphs also give error bars showing the standard errors in two separate measurements of the activities. In most cases the standard errors are much smaller than the activity changes themselves.

	12-pNCA	2PE	PROP	11PA	2A5C	MDOB
PC1 (explains 62% of variance)	0.25	0.61	0.05	0.29	0.47	0.51
PC2 (explains 20% of variance)	-0.48	-0.39	0.14	-0.27	0.70	0.19

Table 6.1: Principal component analysis of activity profiles. The first two principal components explain 82% of the variance in P450 activity profiles. The table shows the composition of these two components and the variance explained by each. The first component contains positive contributions from all substrates and can be thought of as representing a general high catalytic ability. The second component can be thought of as representing discrimination between fused ring substrates (PROP, 2A5C, and MDOB) and phenolic ether substrates (12-pNCA, 11PA, 2PE). Substrate abbreviations are as defined in the legend to Figure 6.2. The analysis was performed on the logarithms of the fold changes in activity.

of docking in the substrate binding pocket. Likewise, 12-pNCA and 11-phenoxyundecanoic acid are phenoxy-carboxylic acids of similar chain length, and are in the same cluster. To a lesser extent, 2-phenoxyethanol resembles 12-pNCA and 11-phenoxyundecanoic acid in its phenolic ether structure, and it falls into a higher level cluster with these two substrates. Propranolol shares a fused ring structure with 2-amino-5-chlorobenzoxazole and 1,2-methylenedioxybenzene, and these three substrates share a common higher level cluster. Overall, the hierarchical clustering indicates that substrates that appear similar to the human eye are also “seen” this way by the P450s, since the P450s tend to increase or decrease their activities on these substrates in a coordinated fashion.

Figure 6.2 also shows the P450 variants arranged in hierarchical clusters. A visual inspection immediately indicates that there is an overall association among all of the activities. Some of the P450 variants (redder rows) tend to show improved activity on most substrates, while others (bluer rows) tend to show decreased activity on most substrates. Taken together with the clustering of the similar substrates, this overall association suggests that there are two main trends in the activity changes. First, the P450s appear to have undergone general changes in their catalytic abilities that are manifested by broad increases or decreases in activity on all substrates. Second, the P450s appear to have experienced shifts in specificity to favor either the fused ring or the phenolic ether substrates.

To test whether these two apparent trends in activity changes are supported by a quantitative examination of the data, we performed principal component analysis. Principal component analysis is a well-established mathematical technique for finding the dominant components of variation in a data set, essentially by diagonalizing the covariance matrix.

As suggested by the foregoing visual inspection, principal component analysis revealed that two components explained most of the changes in P450 activity (Table 6.1). The first component contained positive contributions from all six substrates, and so represents a general improvement in catalytic ability. The second component contained positive contributions from the fused ring substrates and negative contributions from the phenolic ether substrates, and so represents an increased preference for the former class of substrates over the latter. Together, these two components explain 82% of the variance in activities among the 34 P450 variants. The remaining 18% of the variance is explained by the four remaining components, which represent more subtle shifts in activity that are less easily rationalized with intuitive chemical arguments.

6.3.4 Overall distributions of change in the activities

The preceding sections have demonstrated that neutral genetic drift can lead to substantial changes in P450 activities, and that many of these changes can be understood as resulting from either fairly general increases/decreases in catalytic ability or shifts in preference for different broad classes of substrate structures. In this section, we examine whether there are any pervasive trends in the distributions of activity changes — for example, did most of the promiscuous activities tend to increase or decrease? If a property is not under any evolutionary constraint, then during neutral genetic drift its values might be expected to be distributed in a roughly Gaussian fashion, as the neutrally evolving proteins freely sample from the presumably normal underlying distribution. On the other hand, if a property is constrained by selection to remain above a certain threshold, then during neutral genetic drift its values should display a truncated distribution since selection culls proteins with values that fall below the threshold (such a distribution has been predicted for protein stability by simulations [41] and theory [95]).

Figure 6.4 shows the distribution of changes in activity for each of the six substrates. The distribution for 12-pNCA appears to be truncated on the left, as expected since the P450s neutrally evolved under a requirement to maintain the ability to hydroxylate 12-pNCA. Some of the P450s have undergone a mild decrease in 12-pNCA activity, reflective of the fact that the neutral evolution selection criterion provided a small amount of latitude by allowing the total amount of hydroxylated 12-pNCA to drop to 75% of the parental

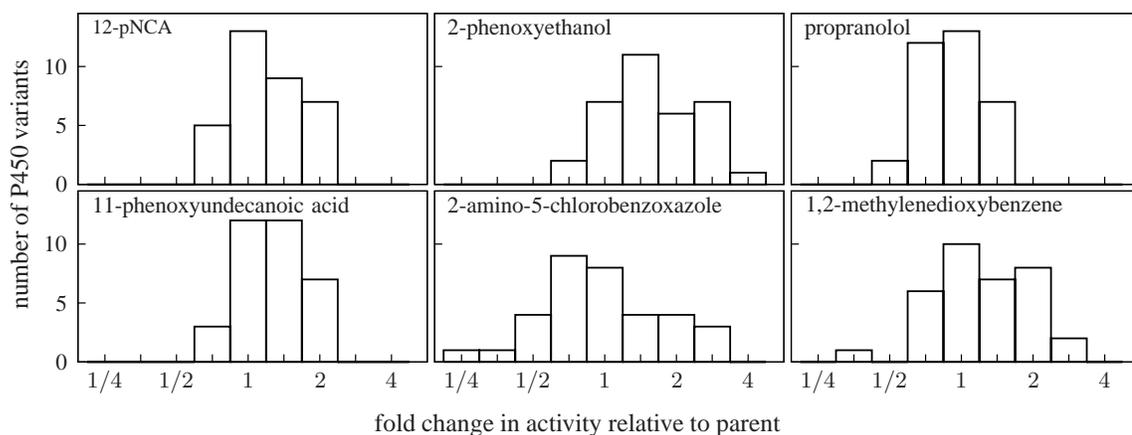


Figure 6.4: Distributions of activity changes on each of the six substrates. The histograms show the distributions of fold changes in activity for all 34 neutrally evolved P450 variants on each of the six substrates, with a value of one indicating that the activity is the same as the neutral evolution parent.

value [135]. A number of P450s have neutrally evolved 12-pNCA activity that modestly exceeds that of the parent — again unsurprising, because the parental 12-pNCA activity falls well below the maximal value achievable for this type of protein [59]. The distribution for 11-phenoxyundecanoic acid resembles that for 12-pNCA, probably because activities on these two chemically similar substrates are highly linked, as discussed in the previous section.

The other four promiscuous activities are less linked to 12-pNCA activity, and their distributions are much more symmetric. The symmetric shapes of these distributions suggest that neutral genetic drift has sampled from a roughly Gaussian distribution for these four promiscuous activities. For three of the substrates (propranolol, 2-amino-5-chlorobenzoxazole, and 1,2-methylenedioxybenzene), the distributions of activities are approximately centered around the parental activity. This centering indicates that the promiscuous activities of the parent on these three substrates are typical of what would be expected of a neutrally evolved P450. The distribution for 2-phenoxyethanol, on the other hand, is shifted towards activities higher than that of the parent. This shift indicates that the parent is less active on 2-phenoxyethanol than a typical neutrally evolved P450.

If the activity distributions of Figure 6.4 truly reflect what would be expected after a very long period of neutral genetic drift (i.e., if they are “equilibrium” distributions), then each variant represents a random sample from the underlying distribution of activities among

12-pNCA	2PE	PROP	11PA	2A5C	MDOB	ALL
0.32 (0.06)	0.27 (0.11)	0.06 (0.71)	0.12 (0.48)	0.20 (0.24)	0.36 (0.04)	0.22(10^{-3})

Table 6.2: Correlations between changes in activity and number of mutations. The extent of change in activity is positively correlated with the number of nonsynonymous mutations the P450 has undergone relative to the neutral evolution parent. Each column shows the Pearson correlation between the number of nonsynonymous mutations and the absolute value of the logarithm of the fold change in activity for a different substrate, computed over all 34 P450 variants. The final column (ALL) is the correlation among the 6×34 pooled data points for all six substrates. The P -values are shown in parentheses; none of the correlations for the individual substrates are significant at a 1% level (due to the small number of data points), but the overall correlation for all substrates is highly significant. Substrate abbreviations are as defined in the legend to Figure 6.2.

all P450s that can neutrally evolve under this selection criterion. In this case, there should be no correlation between the extent of change in activity and the number of accumulated mutations, since the P450s should have lost all “memory” of the parent’s activity. On the other hand, if there has not been enough neutral genetic drift to completely eliminate residual memory of the parent’s activity, then variants with fewer mutations should more closely resemble the parent’s activity profile. To test whether the activity distributions of the P450 variants had equilibrated, we computed the correlation between the magnitude of each variant’s change in activity and the number of nonsynonymous mutations it possessed relative to the parent. Table 6.2 shows that the magnitude of activity change is positively correlated with the number of mutations for all six substrates. Although the correlations for the individual substrates are mostly not statistically significant due to the small number of samples, the overall correlation for all six substrates is highly significant ($P = 10^{-3}$). Therefore, the P450 activities are still in the process of diverging from the parental values by neutral genetic drift. If the variants were to undergo further neutral genetic drift, we would expect to see even larger changes in their promiscuous activities.

We also examined whether P450 variants with mutations near the substrate binding pocket were more likely to have undergone large changes in their activities. Five of the P450 variants had a mutation to a residue that was within 5 Å of the surrogate substrate in the P450 BM3 crystal structure [69]: variant M2 had A74V, M8 had A330V, M13 had M354I, M15 had A74P, and M24 had I263V [135]. Two of these mutated residues are of clear importance, since mutating residue 74 has previously been shown to shift substrate specificity [164, 165, 162] and residue 263 plays a role in the substrate-induced conforma-

tional shift [166]. We compared the activity changes for the five variants with mutations near the binding pocket to those for the 29 variants without any such mutations, computing the magnitude of activity change as the absolute value of the logarithm (base two) of the fold change in activity averaged over all six substrates. The average magnitude of activity change for the five variants with mutations near the active site was 0.88, while the average for the other 29 variants was 0.47. These averages are significantly different, with an unequal variance T-test P -value of 10^{-2} . Therefore, variants with mutations near the substrate binding pocket are especially likely to have altered activities, although many variants without mutations near the pocket also underwent substantial activity changes.

6.4 Conclusions

We have shown that neutral genetic drift can lead to changes of as much as four-fold in the promiscuous activities of P450 proteins. The ubiquity of these changes is striking — even though many of the neutrally evolved P450s had only a handful of mutations, most of them had experienced at least some change in their promiscuous activities. P450s may be especially prone to this type of change, since their catalytic mechanism involves large substrate-induced conformational shifts [167] that can be modulated by mutations distant from the active site [168, 169, 164]. In addition, P450s have a tendency to eventually undergo irreversible inactivation that can be promoted by reduced coupling between substrate binding and conformational shifts, as well as by other poorly understood determinants of catalytic stability [62, 170, 171]. There are therefore ample opportunities for mutations that spread by neutral genetic drift to cause subtle alterations in a P450's promiscuous activities. But we believe that neutral genetic drift is also likely to cause substantial changes in the promiscuous activities of enzymes with other catalytic mechanisms. In support of this idea, a recent study by Tawfik and coworkers [172] indicates that mutations with little effect on the native lactonase activity of serum paraoxonase can alter this enzyme's promiscuous activities. Taken together, this study and our work suggest that neutral genetic drift allows for changes in promiscuous protein functions. These changes could in turn have important implications for future functional evolution. For example, one can easily imagine a scenario in which neutral genetic drift enhances a promiscuous protein function, and then a

subsequent gene duplication allows natural selection to transform one of the genes into the template for a protein with a full-fledged new functional role [150, 54, 151, 152, 153, 154].

One of the most attractive aspects of our study is the degree to which the changes in P450 activities during neutral genetic drift could be understood in terms of the chemical structures of the substrates. Neutral genetic drift did not simply cause unpredictable shifts in activities. Instead, most of the variation was explained by two eminently intuitive components: an overall increase or decrease in catalytic ability, and a preference for either fused ring or phenolic ether substrates. We have suggested that neutral genetic drift under a fixed selection criterion can be viewed as sampling underlying “equilibrium” distributions of activities. The distributions for different activities are linked, since we have shown that P450s with good activity on one substrate will frequently also be highly active on chemically similar substrates (similar linkages have been observed in P450s created by recombination [173]). So while it may be impossible to know exactly how any specific mutation will affect a given activity, measuring a handful of activities allows one to make relatively accurate predictions about other closely linked activities. The prerequisite for making such predictions is an understanding of the linkages among activities in the set of sequences explored by neutral genetic drift (the neutral network). We have made the first steps in elucidating these linkages for P450s that have neutrally evolved under one specific selection regime. The linkages are very similar to those that would have been made by an organic chemist grouping the substrates on the basis of their chemical structures. Knowledge of these linkages is of use in understanding the origins of enzyme specificity [158, 174] — if an enzyme displays high activity on one substrate but low activity on another, then either these two activities are negatively linked during neutral genetic drift or selection has explicitly disfavored one of them.

Our work also has implications for the general relationship between neutral genetic drift and adaptive evolution. A number of studies focused on RNA [103, 175, 176] or computational systems [177, 178] have suggested that genetic drift might aid in adaptive evolution. Our study and that of Tawfik and coworkers [172] support this notion for the evolution of new protein functions. However, the way that drift in promiscuous functions promotes adaptive evolution is slightly different than the paradigm proposed for RNA [103, 175, 176] and computational systems [177, 178]. In those systems, neutral genetic drift

is envisioned as allowing a sequence to move along its neutral network until it reaches a position where it can jump to a new higher-fitness and non-overlapping neutral network. In contrast, promiscuous protein functions change even as a protein drifts along a single neutral network. The adaptive benefits of this drift come when new selective pressures suddenly favor a previously irrelevant promiscuous function, in effect creating a new neutral network that overlaps with parts of the old one.

Overall, experiments have now demonstrated two clear mechanisms by which neutral genetic drift can aid in the evolution of protein functions. In the first mechanism, neutral genetic drift fixes a mutation that increases a protein's stability [76, 72, 95], thereby improving the protein's tolerance for subsequent mutations [7, 127, 49], some of which may confer new or improved functions [49]. In the second mechanism, which was the focus of this work and the recent study by Tawfik and coworkers [172], neutral genetic drift enhances a promiscuous protein function. This enhancement poises the protein to undergo adaptive evolution should a change in selection pressures make the promiscuous function beneficial at some point in the future.

6.5 Methods

6.5.1 Determination of P450 activities

We attempted to determine the activities of all 44 neutrally evolved P450 variants described in [135] (22 from the final monomorphic populations and 22 from the final polymorphic population). Ten of these variants expressed relatively poorly in the procedure used here (as described in more detail below), and so were eliminated from further analysis since their low expression led to large errors in the activity measurements. That left activity data for the 34 neutrally evolved P450 variants listed in Figures 6.2 and 6.3, as well as for the R1-11 neutral evolution parent. The activities for each of these P450 variants were measured on all six substrates (12-pNCA, 2-phenoxyethanol, propranolol, 11-phenoxyundecanoic acid, 2-amino-5-chlorobenzoxazole, and 1,2-methylenedioxybenzene). In all cases, the activities represent the total amount of product produced after two hours, and so are in units of total turnovers per enzyme. P450 BM3 enzymes typically catalyze only a finite number of reaction cycles before becoming irreversibly inactivated, and we believe that all reactions

were essentially complete after two hours, so these activities should represent the total turnovers of the enzymes during their catalytic lifetimes.

To obtain P450 protein for the activity measurements, we expressed the protein using catalase-free *Escherichia coli* [86] containing the encoding gene on the isopropyl β -D-thiogalactoside (IPTG) inducible pCWori [86] plasmid (the catalase is removed since it breaks down the hydrogen peroxide used by the P450). The sequences of the P450 variants are detailed in [135]. We used freshly streaked cells to inoculate 2 ml cultures of Luria Broth (LB) supplemented with 100 $\mu\text{g}/\text{ml}$ of ampicillin, and grew these starter cultures overnight with shaking at 37°C. We then used 0.5 ml from these starter cultures to inoculate 1 L flasks containing 200 ml of terrific broth (TB) supplemented with 100 $\mu\text{g}/\text{ml}$ of ampicillin. The TB cultures were grown at 30°C and 210 rpm until they reached an optical density at 600 nm of ≈ 0.9 , at which point IPTG and δ -aminolevulinic acid were added to a final concentration of 0.5 mM each. The cultures were grown for an additional 19 hours, then the cells were harvested by pelleting 50 ml aliquots at 5,500 g and 4°C for 10 min, and stored at -20°C. To obtain clarified lysate, each pellet was resuspended in 8 ml of 100 mM [4-(2-hydroxyethyl)-1-piperazinepropanesulfonic acid] (EPPS), pH 8.2 and lysed by sonication, while being kept on ice. The cell debris was pelleted by centrifugation at 8,000 g and 4°C for 10 minutes, and the clarified lysate was decanted and kept on ice.

To perform the assays, various dilutions of the clarified lysate were used to construct a standard curve. For each sample, we prepared dilutions of the clarified lysate in the 100 mM EPPS (pH 8.2) buffer to create samples for the standard curves. The dilutions were 100% clarified lysate (undiluted), 67% lysate, 40% lysate, 25% lysate, 17% lysate, 10% lysate, 6.7% lysate, and 4.0% lysate. Similar dilutions were also prepared of the clarified lysate of *E. coli* cells carrying a null pCWori plasmid in order to assess the background readings from lysate without any P450. A pipetting robot was then used to dispense 80 μl of this series of clarified lysate dilutions into 96-well microtiter plates. Duplicate microtiter plates were then assayed for P450 concentration and total enzymatic activity on each of the six substrates. The R1-11 parent was assayed four times rather than in duplicate. To minimize variation, all of these assays were performed in parallel, with the same stock solutions, and on the same day.

The P450 concentration was determined using the carbon monoxide (CO) difference

spectrum assay [66]. Immediately before use, we prepared a 5× stock solution of 50 mM sodium hydrosulfite in 1.3 M potassium phosphate, pH 8.0. A multichannel pipette was used to add 20 μl of this stock solution to each well of the microtiter plates (which contained 80 μl of a dilution of clarified lysate), so that the final sodium hydrosulfite concentration was 10 mM in each well. The plates were briefly mixed and the absorbances were read at 450 and 490 nm. The plates were then incubated in a CO binding oven [66] for 10 minutes to bind CO to the iron. The absorbance was then again read at 450 and 490 nm. The amount of P450 is proportional to the increase in the magnitude of the absorbance at 450 nm minus the absorbance at 490 nm. At each dilution along the standard curve, the reading for the null control (lysate dilutions without P450) was subtracted from the reading for each P450 variant to control for clarified lysate background. Ten P450 variants had standard curve slopes less than or equal to 0.020, indicating a low P450 concentration. These were the ten P450 variants that we discarded from further analysis, since the low P450 concentration decreased the accuracy of the measurements.

To determine the activity on 12-pNCA, we monitored the formation of the yellow 4-nitrophenolate compound that is released upon hydroxylation of the twelfth carbon in the 12-pNCA molecule [179, 59]. Immediately before use, we prepared a 6× stock solution of 12-pNCA by adding 3.6 parts of 4.17 mM 12-pNCA in DMSO to 6.4 parts 100 mM EPPS, pH 8.2. A multichannel pipette was used to add 20 μl of this stock solution to each well of the microtiter plates (which contained 80 μl of a dilution of clarified lysate). The plates were briefly mixed, and the absorbance was read at 398 nm. To initiate the reactions, we then prepared a 6× stock solution of 24 mM hydrogen peroxide in 100 mM EPPS, pH 8.2, and immediately added 20 μl of this solution to each well of the microtiter plate and mixed. The final assay conditions were therefore 6% DMSO, 250 μM 12-pNCA, and 4 mM hydrogen peroxide. The reactions were incubated on the benchtop for two hours, and the total amount of enzymatic product was quantified by the gain in absorbance at 398 nm. At each dilution along the standard curve, the corresponding null control lysate dilution was subtracted from the reading to control for lysate background.

The activities on 2-phenoxyethanol, propranolol, 11-phenoxyundecanoic acid, 2-amino-5-chlorobenzoxazole, and 1,2-methylenedioxybenzene were determined using the 4-aminoantiprene (4-AAP) assay [67, 87], which detects the formation of phenolic compounds. For each of

these five substrates, immediately before use we prepared a 6× substrate stock solution. These stock solutions were 6% DMSO and 6% acetone in 100 mM EPPS, pH 8.2, with an amount of substrate added so that the substrate concentrations in the stock solutions were: 150 mM for 2-phenoxyethanol, 30 mM for propranolol, 5 mM for 11-phenoxyundecanoic acid, 12 mM for 2-amino-5-chlorobenzoxazole, and 120 mM for 1,2-methylenedioxybenzene. The stock solutions were prepared by first dissolving the substrate in the DMSO and acetone, and then adding the EPPS buffer. In some cases, the stock solution became cloudy upon addition of the buffer, but there was no immediate precipitation, so we could still pipette the stock solution. A multichannel pipette was used to add 20 μl of the appropriate substrate stock solution to each well of the microtiter plates (which contained 80 μl of a dilution of clarified lysate). To initiate the reactions, we then added 20 μl of the freshly prepared 6× hydrogen peroxide stock solution (24 mM hydrogen peroxide in 100 mM EPPS, pH 8.2) and mixed. We incubated the plates on the benchtop for two hours. To detect the formation of phenolic products, a pipetting robot was used to add and mix 120 μl of quench buffer (4 M urea in 100 mM sodium hydroxide) to each well. We then used the robot to add and mix 36 μl per well of 0.6% (w/v) of 4-aminoantipyrene in distilled water, and immediately read the absorbance at 500 nm. To catalyze formation of the red compound produced by coupling a phenolic compound to 4-aminoantipyrene [67, 87], we then used the pipetting robot to add and mix 36 μl per well of 0.6% (w/v) of potassium peroxodisulfate in distilled water. The plates were incubated on the benchtop for 30 minutes, and the amount of product was quantified by the gain in absorbance at 500 nm. At each dilution along the standard curve, the corresponding null control lysate dilution was subtracted from the reading to control for lysate background.

In order to extract enzymatic activities from the standard curves, we fit lines to the data points. For some of the substrates (most notably 12-pNCA and 2-phenoxyethanol), many of the P450 variants were sufficiently active to either saturate the substrate or exceed the linear range of absorbance readings. Therefore, we examined each standard curve by eye to determine which points remained in the linear range. Lines were then fit to the points in the linear range. These slopes are averaged for a best estimate of the slope, and the standard error computed over these two measurements is also reported.

To compare the activities (total substrate turnovers per enzyme) among the different

P450 variants, it is first necessary to normalize to the enzyme concentration. To do this, we took the ratio of the slope for each substrate divided by the slope of the CO difference spectrum, propagating the errors. These normalized slopes are proportional to the activity on each substrate. These normalized slopes allow for accurate comparisons among the P450 variants, and were used in the analyses in this paper. To convert these normalized slopes into total substrate turnovers per enzyme, it is necessary to multiply them by the ratio of extinction coefficients. The extinction coefficient for the CO difference spectrum reading (the absorbance at 450 nm minus that at 490 nm) is $91 \text{ mM}^{-1}\text{cm}^{-1}$ [66], and we calculated the extinction coefficient at 398 nm for the 4-nitrophenolate group in our buffer to be $12,000 \text{ M}^{-1}\text{cm}^{-1}$. Therefore, for 12-pNCA, the total number of substrate turnovers per P450 enzyme is 7.58 times the ratio of the 12-pNCA standard curve slope to the CO difference spectrum slope. This indicates that our parent protein had about 250 12-pNCA turnovers per enzyme, compared to the 1,000 reported for a variant engineered for maximal 12-pNCA activity [59]. For the other substrates assayed with the 4-AAP assay, the extinction coefficient at 500 nm for the 4-AAP/phenol complex has been reported to be 4,800 [87]. However, we believe that this extinction coefficient could be of dubious accuracy for our data. Depending on the exact type of phenolic compound created by P450 hydroxylation, the extinction coefficient for the 4-AAP/phenol complex may vary. Assuming the extinction coefficient of 4,800 $\text{M}^{-1}\text{cm}^{-1}$ is accurate, then the total number of substrate turnovers per P450 enzyme is 19.0 times the ratio of the substrate standard curve slope to the CO difference spectrum slope. Using this coefficient, the parent P450 had roughly 1,000 turnovers on 2-phenoxyethanol, 30 turnovers on propranolol, 400 turnovers on 11-phenoxyundecanoic acid, 50 turnovers on 2-amino-5-chlorobenzoxazole, and 80 turnovers on 1,2-methylenedioxybenzene. The high activities on 2-phenoxyethanol and 11-phenoxyundecanoic acid are presumably due to the fact that lack of polar substituents on the aromatic ring allows these compounds to enter the hydrophobic P450 BM3 binding pocket [69] more easily than 12-pNCA. However, we emphasize that the exact numerical values for the turnovers for these five substrates are questionable. Definitive determination of the extinction coefficients would require analytical analysis of the enzymatic products for each P450 variant on each substrate, which is beyond the scope of this study.

6.5.2 Analysis of activity data

To analyze and display the activity data, we computed the fold change in activity of each variant relative to the R1-11 parent P450. The fold change is simply the variant activity divided by the parent activity on each substrate, with the standard errors propagated to give an error on the fold change. In Figures 6.2 and 6.3, these fold changes are displayed on a logarithmic scale so that each unit corresponds to a two-fold increase or decrease in activity. In Figure 6.2, the substrates and the P450 variants have both been clustered, as shown by dendrograms on the side of the heat map. The clustering was performed using the standard hierarchical clustering function of the R statistical package. This is complete linkage hierarchical clustering, with the distances computed as the Euclidian distance between the logarithms of the fold changes in activity. The standard errors on the fold changes in activity are not incorporated into Figure 6.2 or any of the related analysis. However, these standard errors are shown in Figure 6.3; it is apparent from this figure that the errors tend to be much less than the fold changes in activity themselves.

In Figure 6.4, the histogram bins are logarithmically spaced so that each bin contains a $2^{0.5}$ -fold range of activities. For example, the histogram bin centered at one contains all variants with between $2^{-0.25} = 0.84$ and $2^{0.25} = 1.19$ fold the parental activity, while the bin centered at 1.5 contains all variants with between $2^{0.25} = 1.19$ and $2^{0.75} = 1.68$ fold the parental activity.

The principal component analysis shown in Table 6.1 was performed using the R statistical package, with inputs being the logarithms of the fold changes in activity. Since these log fold changes in activity contained no arbitrary units (they were already normalized to the parent), the data was neither scaled nor zeroed before performing the analysis. Table 6.1 shows the composition and the percent of variance explained (the eigenvalue for that component divided by the sum of all eigenvalues) for the first two components. The remaining four components were relatively unimportant, explaining 7%, 5%, 4%, and 2% of the total variance.

6.5.3 Phylogenetic tree

The phylogenetic tree shown in Figure 6.1 is based on the number of nonsynonymous mutations the P450 variants have relative to the R-11 neutral evolution parent, as reported

in [135]. Each of the P450s that evolved in a monomorphic population (prefix of M) are known to have diverged independently, and so are drawn on their own branch regardless of any sequence identity to other variants. The exact phylogenetic relationship of the P450s that evolved in the polymorphic population (prefix of P) is not known, so the portion of the tree for these mutants was reconstructed by maximum parsimony. The tree is based only on the nonsynonymous mutations, and all mutations weighted equally. Full nucleotide and amino acid sequences of the P450s can be found in [135].

6.6 Acknowledgements

We thank Andrew Sawayama and Sabine Bastian for helpful comments. JDB was supported by a Howard Hughes Medical Institute predoctoral fellowship. ZL was supported by a summer undergraduate research fellowship from the California Institute of Technology.

Bibliography

- [1] Pauling L, Itano HA, Singer SJ, Wells IC: **Sickle cell anemia, a molecular disease.** *Science* 1949, **110**:543–548.
- [2] Ingram VM: **Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin.** *Nature* 1957, **180**:326–328.
- [3] Zuckerkandl E, Pauling L: **Evolutionary divergence and convergence in proteins.** In *Evolving genes and proteins*, New York, NY: Academic Press 1965:97–166.
- [4] Godoy-Ruiz R, Ariza F, Rodriguez-Larrea D, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM: **Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments.** *J. Mol. Biol.* 2006, **362**:966–978.
- [5] Chiti F, Taddei N, Bucciantini M, White P, Ramponi G, Dobson CM: **Mutational analysis of the propensity for amyloid formation by a globular protein.** *EMBO J.* 2000, **19**:1441–1449.
- [6] Park C, Marqusee S: **Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding.** *Nature Methods* 2005, **2**:207–212.
- [7] Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH: **Thermodynamic prediction of protein neutrality.** *Proc. Natl. Acad. Sci. USA* 2005, **102**:606–611.
- [8] Broglia RA, Tiana G, Roman HE, Vigezzi E, Shakhnovich E: **Stability of designed proteins against mutations.** *Phys. Rev. Lett.* 1999, **82**:4727–4730.

- [9] Bornberg-Bauer E, Chan HS: **Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space.** *Proc. Natl. Acad. Sci. USA* 1999, **96**:10689–10694.
- [10] Wingreen NS, Li H, Tang C: **Designability and thermal stability of protein structures.** *Polymer* 2004, **45**:699–705.
- [11] Xia Y, Levitt M: **Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution.** *Proteins* 2004, **55**:107–114.
- [12] Li H, Helling R, Tang C, Wingreen N: **Emergence of preferred structures in a simple model of protein folding.** *Science* 1996, **273**:666–669.
- [13] Shakhnovich EI, Gutin AM: **Influence of point mutations on protein structure: probability of a neutral mutation.** *J. Theor. Biol.* 1991, **149**:537–546.
- [14] Govindarajan S, Goldstein RA: **Why are some protein structures so common?** *Proc. Natl. Acad. Sci. USA* 1996, **93**:3341–3345.
- [15] Rennell D, Bouvier SE, Hardy LW, Poteete AR: **Systematic mutation of bacteriophage T4 lysozyme.** *J. Mol. Biol.* 1991, **222**:67–87.
- [16] Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH: **Genetic studies of the *lac* repressor reveals essential and non-essential residues, as well ‘spacers’ which do not require a specific sequence.** *J. Mol. Biol.* 1994, **240**:421–433.
- [17] Shafikhani S, Siegel RA, Ferrari E, Schellenberger V: **Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization.** *BioTechniques* 1997, **23**:304–310.
- [18] Guo HH, Choe J, Loeb LA: **Protein tolerance to random amino acid change.** *Proc. Natl. Acad. Sci. USA* 2004, **101**:9205–9210.
- [19] Daugherty PS, Chen G, Iverson BL, Georgiou G: **Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies.** *Proc. Natl. Acad. Sci. USA* 1999, **97**:2029–2034.

- [20] Shortle D, Lin B: **Genetic analysis of staphylococcal nuclease: identification of three intragenic “global” suppressors of nuclease-minus mutations.** *Genetics* 1985, **110**:539–555.
- [21] Poteete AR, Rennell D, Bouvier SE, Hardy LW: **Alteration of T4 lysozyme structure by second-site reversion of deleterious mutations.** *Protein Sci.* 1997, **6**:2418–2425.
- [22] Wilke CO, Adami C: **Evolution of mutational robustness.** *Mutat. Res.* 2003, **522**:3–11.
- [23] Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, III CAH: **Complete mutagenesis of the HIV-1 protease.** *Nature* 1989, **340**:397–400.
- [24] Pakula AA, Young VB, Sauer RT: **Bacteriophage λ *cro* mutations: effects on activity and intracellular degradation.** *Proc. Natl. Acad. Sci. USA* 1986, **83**:8829–8833.
- [25] Anfinsen CB: **Principles that govern the folding of protein chains.** *Science* 1973, **181**:223–230.
- [26] Fersht AR: *Structure and Mechanism in Protein Science.* New York: W. H. Freeman and Company 1999.
- [27] Bava KA, Gromiha MM, Uedaira H, Kitajimi K, Sarai A: **ProTherm, version 4.0: thermodynamic database for proteins and mutants.** *Nucleic Acids Res.* 2004, **32**:D120–D121.
- [28] Wells JA: **Additivity of mutational effects in proteins.** *Biochemistry* 1990, **29**:8509–8517.
- [29] Zhang XJ, Baase WA, Shoichet BK, Wilson KP, Matthews BW: **Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive.** *Protein Eng.* 1995, **8**:1017–1022.
- [30] Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS: **Principles of protein folding – a perspective from simple exact models.** *Protein Sci.* 1995, **4**:561–602.

- [31] Hinds DA, Levitt M: **Exploring conformational space with a simple lattice model for protein–structure.** *J. Mol. Biol.* 1994, **243**:668–682.
- [32] Shakhnovich EI, Gutin AM: **Engineering of stable and fast–folding sequences of model proteins.** *Proc. Natl. Acad. Sci. USA* 1993, **90**:7195–7199.
- [33] Socci ND, Onuchic JN, Wolynes PG: **Protein folding mechanisms and the multidimensional folding funnel.** *Proteins* 1998, **32**:136–158.
- [34] Chan HS, Bornberg-Bauer E: **Perspectives on protein evolution from simple exact models.** *Applied Bioinformatics* 2002, **1**:121–144.
- [35] Xia Y, Levitt M: **Simulating protein evolution in sequence and structure space.** *Curr. Opin. Struct. Biol.* 2004, **14**:202–207.
- [36] Wang X, Minasov G, Shoichet BK: **Evolution of an antibiotic resistance enzyme constrained by stability and activity trade–offs.** *J. Mol. Biol.* 2002, **320**:85–95.
- [37] Gilis D, Rooman M: **Predicting protein stability changes upon mutation using database–derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence.** *J. Mol. Biol.* 1997, **272**:276–290.
- [38] Gilis D, Rooman M: **PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins.** *Protein Eng.* 2000, **13**:849–856.
- [39] Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.** *J. Mol. Biol.* 2002, **320**:369–387.
- [40] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH – a hierarchic classification of protein domain structures.** *Structure* 1997, **5**:1093–1108.
- [41] Taverna DM, Goldstein RA: **Why are proteins marginally stable?** *Proteins* 2002, **46**:105–109.

- [42] Wolynes PG: **Symmetry and the energy landscapes of biomolecules.** *Proc. Natl. Acad. Sci. USA* 1996, **93**:14249–14255.
- [43] England JL, Shakhnovich EI: **Structural determinant of protein designability.** *Phys. Rev. Lett.* 2003, **90**:218101.
- [44] Arnold FH: **Design by directed evolution.** *Acc. Chem. Res.* 1998, **31**:125–131.
- [45] Bloom JD, Wilke CO, Arnold FH, Adami C: **Stability and the evolvability of function in a model protein.** *Biophysical J.* 2004, **86**:1–7.
- [46] Miyazawa S, Jernigan RL: **Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation.** *Macromolecules* 1985, **18**:534–552.
- [47] Meyer MM, Silberg JJ, Voigt CA, Endelman JB, Mayo SL, Wang ZG, Arnold FH: **Library analysis of SCHEMA-guided protein recombination.** *Protein Sci.* 2003, **12**:1686–1693.
- [48] Sun F: **The polymerase chain reaction and branching processes.** *J. Comput. Biol.* 1995, **2**:63–86.
- [49] Bloom JD, Labthavikul ST, Otey CR, Arnold FH: **Protein stability promotes evolvability.** *Proc. Natl. Acad. Sci. USA* 2006, **103**:5869–5874.
- [50] Kauffman SA: *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford: Oxford University Press 1993.
- [51] Kirschner M, Gerhart J: **Evolvability.** *Proc. Natl. Acad. Sci. USA* 1998, **95**:8420–8427.
- [52] Bloom JD, Meyer MM, Meinhold P, Otey CR, MacMillan D, Arnold FH: **Evolving strategies for enzyme engineering.** *Curr. Opin. Struct. Biol.* 2005, **15**:447–452.
- [53] Kicinger R, Arciszewski T, Jong KD: **Evolutionary computation and structural design: a survey of the state-of-the-art.** *Computers & Structures* 2005, **83**:1943–1978.

- [54] Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS: **The ‘evolvability’ of promiscuous protein functions.** *Nat. Genetics* 2005, **37**:73–76.
- [55] Wagner A: *Robustness and Evolvability in Living Systems*. Princeton, New Jersey: Princeton University Press 2005.
- [56] Lesk AM, Chothia C: **How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins.** *J. Mol. Biol.* 1980, **136**:225–270.
- [57] Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J.* 1986, **5**:823–826.
- [58] Nikolova PV, Wong KB, DeDecker B, Henckel J, Fersht AR: **Mechanism of rescue of common p53 cancer mutations by second-site suppressor mutations.** *EMBO J.* 2000, **19**:370–378.
- [59] Cirino PC, Arnold FH: **A self-sufficient peroxide-driven hydroxylation biocatalyst.** *Angew. Chem. Int. Ed.* 2003, **42**:3299–3301.
- [60] de Montellano PRO: *Cytochrome P450: Structure, Mechanism, and Biochemistry*. New York: Plenum Press 1995.
- [61] Lewis DFV: *Cytochromes P450: Structure and Function*. London: Taylor and Francis 2001.
- [62] Munro AW, Leys DG, McLean KJ, Marshall KR, Ost TWB, Daff S, Miles CS, Chapman SK, Lysek DA, Moser CC, Page CC, Dutton PL: **P450 BM3: the very model of a modern flavocytochrome.** *Trends Biochem. Sci.* 2002, **27**:250–257.
- [63] Salazar O, Cirino PC, Arnold FH: **Thermostabilization of a cytochrome P450 peroxygenase.** *ChemBioChem* 2003, **4**:891–893.
- [64] Drummond DA, Iverson BL, Georgiou F, Arnold FH: **Why high-error-rate random mutagenesis libraries are highly enriched in functional and improved proteins.** *J. Mol. Biol.* 2005, **350**:806–816.

- [65] Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C*, Cambridge University Press. 2 edition 2002 chap. 14, :620–628.
- [66] Otey CR: **High-throughput carbon monoxide binding assay for cytochromes P450**. In *Directed enzyme evolution: screening and selection methods, Volume 230 of Methods in Molecular Biology*. Edited by Arnold FH, Georgiou G, Humana press 2003:137–139.
- [67] Otey CR, Joern JM: **High-throughput screen for aromatic hydroxylation**. In *Directed enzyme evolution: screening and selection methods, Volume 230 of Methods in Molecular Biology*. Edited by Arnold FH, Georgiou G, Humana press 2003:141–148.
- [68] Martinis SA, Blanke SR, Hager LP, Sligar SG: **Probing the heme iron coordination structure of pressure-induced cytochrome P420cam**. *Biochemistry* 1996, **35**:14530–14536.
- [69] Haines DC, Tomchick DR, Machius M, Peterson JA: **Pivotal role of water in the mechanism of P450BM-3**. *Biochemistry* 2001, **40**:13456–13465.
- [70] Somero GN: **Proteins and temperature**. *Annu. Rev. Physiol.* 1995, **57**:43–68.
- [71] Fields PA: **Review: protein function at thermal extremes: balancing stability and flexibility**. *Comp. Biochem. Physiol. A* 2001, **129**:417–431.
- [72] DePristo MA, Weinreich DM, Hartl DL: **Missense meanderings in sequence space: a biophysical view of protein evolution**. *Nat. Rev. Genetics* 2005, **6**:678–687.
- [73] Giver L, Gershenson A, Freskgard PO, Arnold FH: **Directed evolution of a thermostable esterase**. *Proc. Natl. Acad. Sci. USA* 1998, **95**:12809–12813.
- [74] van den Burg B, Vriend G, Veltman OR, Venema G, Eusink VGH: **Engineering an enzyme to resist boiling**. *Proc. Natl. Acad. Sci. USA* 1998, **95**:2056–2060.
- [75] Zhao H, Arnold FH: **Directed evolution converts subtilisin E into a functional equivalent of thermitase**. *Protein Eng.* 1999, **12**:47–53.

- [76] Serrano L, Day AG, Fersht AR: **Step-wise mutation of barnase to binase: a procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability.** *J. Mol. Biol.* 1993, **233**:305–312.
- [77] Godoy-Ruiz R, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM: **Relation between protein stability, evolution and structure as probed by carboxylic acid mutations.** *J. Mol. Biol.* 2004, **336**:313–318.
- [78] Pakula AA, Sauer RT: **Genetic analysis of protein stability and function.** *Annu. Rev. Genet.* 1989, **23**:289–310.
- [79] Matthews BW: **Structural and genetic analysis of protein stability.** *Annu. Rev. Biochem.* 1993, **62**:139–160.
- [80] Meiering EM, Serrano L, Fersht AR: **Effect of active site residues in barnase on activity and stability.** *J. Mol. Biol.* 1992, **225**:585–589.
- [81] Shoichet BK, Baase WA, Kuroki R, Matthews BW: **A relationship between protein stability and protein function.** *Proc. Natl. Acad. Sci. USA* 1995, **92**:452–456.
- [82] Beadle BM, Shoichet BK: **Structural bases of stability-function tradeoffs in enzymes.** *J. Mol. Biol.* 2002, **321**:285–296.
- [83] Arnold FH, Wintrode PL, Miyazaki K, Gershenson A: **How enzymes adapt: lessons from directed evolution.** *Trends Biochem. Sci.* 2001, **26**:100–107.
- [84] Earl DJ, Deem MW: **Evolvability is a selectable trait.** *Proc. Natl. Acad. Sci. USA* 2004, **101**:11531–11536.
- [85] Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proc. Natl. Acad. Sci. USA* 2005, **102**:14338–14343.
- [86] Barnes HJ, Arlotto MP, Waterman MR: **Expression and enzymatic activity of recombinant cytochrome P450 17 α -hydroxylase in *Escherichia coli*.** *Proc. Natl. Acad. Sci. USA* 1991, **88**:5597–5601.

- [87] Otey CR, Silberg JJ, Voigt CA, Endelman JB, Bandara G, Arnold FH: **Functional evolution and structural conservation in chimeric cytochromes P450: calibrating a structure-guided approach.** *Chem. Biol.* 2004, **11**:309–318.
- [88] Higuchi R, Krummel B, Saiki RK: **A general method of *in vitro* preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions.** *Nucleic Acids Res.* 1988, **16**:7351–7367.
- [89] Wells AV, Li P, Champion PM: **Resonance raman investigations of *Escherichia coli*-expressed *Pseudomonas putida* cytochrome P450 and P420.** *Biochemistry* 1992, **31**:4384–4393.
- [90] Rees DC, Robertson AD: **Some thermodynamic implications for the thermostability of proteins.** *Protein Sci.* 2001, **10**:1187–1194.
- [91] Myers JK, Pace CN, Scholtz JM: **Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding.** *Protein Sci.* 1995, **4**:2138–2148.
- [92] Yu XC, Shen S, Strobel HW: **Denaturation of cytochrome P450 2B1 by guanidine hydrochloride and urea: evidence for a metastable intermediate state of the active site.** *Biochemistry* 1995, **34**:5511–5517.
- [93] Munro AW, Lindsay JG, Coggins JR, Kelly SM, Price NC: **Analysis of the structural stability of the multidomain enzyme flavocytochrome P-450 BM3.** *Biochimica et Biophysica Acta* 1996, **1296**:127–137.
- [94] Murugan R, Mazumdar S: **Role of substrate on the conformational stability of the heme active site of cytochrome P450cam: effect of temperature and low concentrations of denaturants.** *J. Biol. Inorg. Chem.* 2004, **9**:477–488.
- [95] Bloom JD, Raval A, Wilke CO: **Thermodynamics of neutral protein evolution.** *Genetics* 2007, **175**:255–266.
- [96] Kimura M: *The Neutral Theory of Molecular Evolution.* Cambridge, U.K.: Cambridge Univ. Press 1983.

- [97] Ohta T, Kimura M: **On the constancy of the evolutionary rate of cistrons.** *J. Mol. Evol.* 1971, **1**:18–25.
- [98] Kumar MD, Bava KA, Gromiha MM, Parabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.** *Nucleic Acids Res* 2006, **34**:D204–D206.
- [99] Kimura M: **Molecular evolutionary clock and the neutral theory.** *J. Mol. Evol.* 1987, **26**:24–33.
- [100] Takahata N: **On the overdispersed molecular clock.** *Genetics* 1987, **116**:169–179.
- [101] Cutler DJ: **Understanding the overdispersed molecular clock.** *Genetics* 2000, **154**:1403–1417.
- [102] Smith JM: **Natural selection and the concept of a protein space.** *Nature* 1970, **225**:563–564.
- [103] Huynen MA, Stadler PF, Fontana W: **Smoothness within ruggedness: the role of neutrality in adaptation.** *Proc. Natl. Acad. Sci. USA* 1996, **93**:397–401.
- [104] Govindarajan S, Goldstein RA: **The foldability landscape of model proteins.** *Biopolymers* 1997, **42**:427–438.
- [105] van Nimwegen E, Crutchfield JP, Huynen M: **Neutral evolution of mutational robustness.** *Proc. Natl. Acad. Sci. USA* 1999, **96**:9716–9720.
- [106] Tiana G, Broglia RA, Shakhnovich EI: **Hiking in the energy landscape in sequence space: a bumpy road to good folders.** *Proteins* 2000, **39**:244–251.
- [107] Bastolla U, Porto M, Roman HE, Vendruscolo M: **Lack of self-averaging in neutral evolution of proteins.** *Phys. Rev. Lett.* 2002, **89**:208101.
- [108] Taverna DM, Goldstein RA: **Why are proteins so robust to site mutations?** *J. Mol. Biol.* 2002, **315**:479–484.
- [109] Uversky VN, Oldfield CJ, Dunker AK: **Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling.** *J. Mol. Recognit.* 2005, **18**:343–384.

- [110] Jaswal SS, Sohl JL, Davis JH, Agard DA: **Energetic landscape of α -lytic protease optimizes longevity through kinetic stability.** *Nature* 2002, **415**:343–347.
- [111] Huntzicker EG, Estay IS, Zhen H, Lokteva LA, Jackson PK, Oro AE: **Dual degradation signals control Gli protein stability and tumor formation.** *Genes and Development* 2006, **20**:276–281.
- [112] Akashi H: **Translational selection and yeast proteome evolution.** *Genetics* 2003, **164**:1291–1303.
- [113] Chamary JV, Hurst LD: **Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals.** *Genome Biology* 2005, **6**:R75.
- [114] Rocha EPC, Danchin A: **An analysis of determinants of amino acids substitution rates in bacterial proteins.** *Mol. Biol. Evol.* 2004, **21**:108–116.
- [115] Pal C, Papp B, Hurst LD: **Highly expressed genes in yeast evolve slowly.** *Genetics* 2001, **158**:927–931.
- [116] Sanchez IE, Tejero J, Gomez-Moreno C, Medina M, Serrano L: **Point mutations in protein globular domains: contributions from function, stability, and misfolding.** *J. Mol. Biol.* 2006, **363**:422–432.
- [117] Keefe AD, Szostak JW: **Functional proteins from a random-sequence library.** *Nature* 2001, **410**:715–718.
- [118] Davidson AR, Lumb KJ, Sauer RT: **Cooperatively folded proteins in random sequence libraries.** *Nat. Struct. Biol.* 1995, **2**:856–864.
- [119] Wilke CO, Bloom JD, Drummond DA, Raval A: **Predicting the tolerance of proteins to random amino acid substitution.** *Biophysical J.* 2005, **89**:3714–3720.
- [120] Jukes TH, Cantor CR: *Mammalian protein metabolism, III*, New York, NY: Academic Press 1969 chap. Evolution of protein molecules, :21–132.
- [121] Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol. Biol. Evol.* 1994, **11**:725–736.

- [122] Ewens WJ, Grant GR: *Statistical methods in bioinformatics*. New York, NY: Springer, 2 edition 2005.
- [123] Cutler DJ: **The index of dispersion of molecular evolution: slow fluctuations**. *J. Theor. Population Biol.* 2000, **57**:177–186.
- [124] Bloom JD, Drummond DA, Arnold FH, Wilke CO: **Structural determinants of the rate of protein evolution in yeast**. *Mol. Biol. Evol.* 2006, **23**:1751–1761.
- [125] Wilke CO: **Molecular clock in neutral protein evolution**. *BMC Genetics* 2004, **5**:25.
- [126] Hermisson J, Redner O, Wagner H, Baake E: **Mutation-selection balance: ancestry, load, and maximum principle**. *Theor. Pop. Biol.* 2002, **62**:9–46.
- [127] Besenmatter W, Kast P, Hilvert D: **Relative tolerance of mesostable and thermostable protein homologs to extensive mutation**. *Proteins* 2007, **66**:500–506.
- [128] Lenski RE, Barrick JE, Ofria C: **Balancing robustness and evolvability**. *PLoS Biology* 2006, **4**:e428.
- [129] Sniegowski PD, Murphy HA: **Evolvability**. *Current Biology* 2006, **16**:R831–R834.
- [130] Montville R, Froissart R, Remold SK, Tenailon O, Turner PE: **Evolution of mutational robustness in an RNA virus**. *PLoS Biol.* 2005, **3**:e381.
- [131] Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C: **Evolution of digital organisms at high mutation rates leads to survival of the flattest**. *Nature* 2001, **412**:331–333.
- [132] Lipman DJ, Wilbur WJ: **Modeling neutrality and selective evolution of protein folding**. *Proc. R. Soc. London Ser. B* 1991, **245**:7–11.
- [133] Wilke CO: **Adaptive evolution on neutral networks**. *Bull. Math. Biol.* 2001, **63**:715–730.
- [134] Brin S, Page L: **The anatomy of a large-scale hypertextual Web search engine**. *Computer Networks and ISDN Systems* 1998, **30**(1–7):107–117.

- [135] Bloom JD, Lu Z, Chen D, Raval A, Venturelli OS, Arnold FH: **Neutral evolution favors mutational robustness in sufficiently large populations.** *Submitted* 2007. [Currently available on ArXiv preprint server at www.arxiv.org as arXiv:0704.1885v1].
- [136] Plotkin JB, Dushoff J, Desai MM, Fraser HB: **Codon usage and selection in proteins.** *J. Mol. Evol.* 2006, **63**:635–653.
- [137] Berg OG: **Selection intensity for codon bias and the effective population size of *Escherichia coli*.** *Genetics* 1996, **142**:1379–1382.
- [138] Lynch M, Conery JS: **The origins of genome complexity.** *Science* 2003, **302**:1401–1404.
- [139] Hartl DL, Moriyama EN, Sawyer SA: **Selection intensity for codon bias.** *Genetics* 1994, **138**:227–234.
- [140] Moya A, Holmes EC, Gonzalez-Candelas F: **The population genetics and evolutionary epidemiology of RNA viruses.** *Nature Reviews Microbiology* 2004, **2**:279–288.
- [141] Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R: **Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population.** *Nature* 2006, **439**:344–348.
- [142] Meyerhans A, Vartanian JP, Wain-Hobson S: **DNA recombination during PCR.** *Nucleic Acids Res.* 1990, **18**:1687–1691.
- [143] Kanagawa T: **Bias and artifacts in multitemplate polymerase chain reactions (PCR).** *Journal of Bioscience and Bioengineering* 2003, **96**:317–323.
- [144] Posada D: **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** *Mol. Biol. Evol.* 2002, **19**:708–717.
- [145] Smith JM: **Analyzing the mosaic structure of genes.** *J. Mol. Evol.* 1992, **34**:126–129.

- [146] Piganeau G, Gardner M, Eyre-Walker A: **A broad survey of recombination in animal mitochondria.** *Mol. Biol. Evol.* 2004, **21**:2319–2325.
- [147] Franklin JN: *Matrix Theory.* Englewood Cliffs, N.J.: Prentice-Hall, Inc. 1968.
- [148] Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS: **Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein.** *Nature* 2006, **444**:929–932.
- [149] Bloom JD, Arnold FH, Wilke CO: **Breaking proteins with mutations: threads and thresholds in evolution.** *Molecular Systems Biology* 2007, **3**:76.
- [150] O'Brien PJ, Herschlag D: **Catalytic promiscuity and the evolution of new enzymatic activities.** *Chemistry and Biology* 1999, **6**:R91–R105.
- [151] Kondrashov FA: **In search of the limits of evolution.** *Nat. Genetics* 2005, **37**:9–10.
- [152] Chothia C, Gough J, Vogel C, Teichmann SA: **Evolution of the protein repertoire.** *Science* 2003, **300**:1701–1703.
- [153] Copley SD, Novak WRP, Babbitt PC: **Divergence of function in the thioredoxin fold suprafamily: evidence for evolution of peroxiredoxins from a thioredoxin-like ancestor.** *Biochemistry* 2004, **43**:13981–13995.
- [154] Bridgham JT, Carroll SM, Thornton JW: **Evolution of hormone-receptor complexity by molecular exploitation.** *Science* 2006, **312**:97–101.
- [155] Copley SD: **Enzymes with extra talents: moonlighting functions and catalytic promiscuity.** *Curr. Opin. Chem. Biol.* 2003, **7**:265–272.
- [156] Khersonsky O, Roodveldt C, Tawfik DS: **Enzyme promiscuity: evolutionary and mechanistic aspects.** *Curr. Opin. Chem. Biol.* 2006, **10**:498–508.
- [157] O'Brien PJ: **Catalytic promiscuity and the divergent evolution of DNA repair enzymes.** *Chem. Rev.* 2006, **106**:720–752.
- [158] O'Loughlin TL, Greene DN, Matsumura I: **Diversification and specialization of HIV protease function during in vitro evolution.** *Mol. Biol. Evol.* 2007, **23**:764–772.

- [159] Roodveldt C, Tawfik DS: **Shared promiscuous activities and evolutionary features in various members of the amidohydrolase superfamily.** *Biochemistry* 2005, **44**:12728–12736.
- [160] Afriat L, Roodveldt C, Manco G, Tawfik DS: **The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase.** *Biochemistry* 2006, **45**:13677–13686.
- [161] Blundell TL, Wood SP: **Is the evolution of insulin Darwinian or due to selectively neutral mutations?** *Nature* 1975, **257**:197–203.
- [162] Otey CR, Bandara G, Lalonde J, Takahashi K, Arnold FH: **Preparation of human metabolites of propranolol using laboratory-evolved bacterial cytochromes P450.** *Biotechnol. and Bioeng.* 2005, **93**:494–499.
- [163] Lasker JM, Huang MT, Conney AH: **In vitro activation of zoxazolamine metabolism by flavone.** *Science* 1982, **216**:1419–1421.
- [164] Li QS, Ogawa J, Schmid RD, Shimizu S: **Engineering cytochrome P450 BM-3 for oxidation of polycyclic aromatic hydrocarbons.** *Appl. Environ. Microbiol.* 2001, **67**:5735–5739.
- [165] Li QS, Schwaneberg U, Fischer M, Schmitt J, Pleiss J, Lutz-Wahl S, Schmid RD: **Rational evolution of a medium chain-specific cytochrome P-450 BM-3 variant.** *Biochim. Biophys. Acta* 2001, **1545**:114–121.
- [166] Pylypenko O, Schlichting I: **Structural aspects of ligand binding to and electron transfer in bacterial and fungal P450s.** *Annu. Rev. Biochem.* 2004, **73**:991–1018.
- [167] Modi S, Sutcliffe MJ, Primrose WU, Lian LY, Roberts GCK: **The catalytic mechanism of cytochrome P450 BM3 involves a 6 angstrom movement of the bound substrate on reduction.** *Nat. Struct. Biol.* 1996, **3**:414–417.
- [168] Glieder A, Farinas ET, Arnold FH: **Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase.** *Nat. Biotech.* 2002, **20**:1135–1139.

- [169] Meinhold P, Peters MW, Chen MMY, Takahashi K, Arnold FH: **Direct conversion of ethan to ethanol by engineered cytochrome P450 BM3.** *ChemBioChem* 2005, **6**:1765–1768.
- [170] Loida PJ, Sligar SG: **Molecular recognition in cytochrome P450 - mechanism for the control of uncoupling reactions.** *Biochemistry* 1993, **32**:11530–11538.
- [171] Bernhardt R: **Cytochromes P450 as versatile biocatalysts.** *J. Biotechnol.* 2006, **124**:128–145.
- [172] Amitai G, Gupta RD, Tawfik DS: **Latent evolutionary potentials under the neutral mutation drift of an enzyme.** *HFSP Journal* 2007, **in press**.
- [173] Landwehr M, Carbone M, Otey CR, Li Y, Arnold FH: **Diversification of catalytic function in a synthetic family of chimeric cytochrome P450s.** *Chem. Biol.* 2007, **14**:269–278.
- [174] Varadarajan N, Gam J, Olsen MJ, Georgiou G, Iverson BL: **Engineering of protease variants exhibiting high catalytic activity and exquisite substrate selectivity.** *Proc. Natl. Acad. Sci. USA* 2005, **102**:6855–6860.
- [175] Huynen MA: **Exploring phenotype space through neutral evolution.** *J. Mol. Evol.* 1996, **43**:165–169.
- [176] Fontana W, Schuster P: **Continuity in evolution: on the nature of transitions.** *Science* 1998, **280**:1451–1455.
- [177] van Nimwegen E, Crutchfield JP, Mitchell M: **Finite populations induce metastability in evolutionary search.** *Physics Letters A* 1997, **229**:144–150.
- [178] van Nimwegen E, Crutchfield JP: **Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths?** *Bull. Math. Biol.* 2000, **62**:799–848.
- [179] Schwaneberg U, Schmidt-Dannert C, Schmitt J, Schmid RD: **A continuous spectrophotometric assay for P450 BM-3, a fatty acid hydroxylating enzyme, and its mutant F87A.** *Analytical Biochemistry* 1999, **269**:359–366.