MISFOLDING DOMINATES PROTEIN EVOLUTION

Thesis by

David Allan Drummond

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2006

(Defended May 15, 2006)

Advisor: Prof. Frances H. Arnold

ACKNOWLEDGMENTS

Having worked for some stellar managers, I know just how lucky I am to work for Frances Arnold. She makes science fun. Frances has shaped my development as a scientist in ways great and small, some direct (her ability to cut to an issue's heart, and to explain how to make an idea *matter*, continue to amaze me) and some by example (she delivered insightful comments on complex manuscripts within days *at most*, even at the very nadir of her chemotherapy). I am so grateful for her blind investment in me, an older student with no publications, no lab experience, and no knowledge of protein biophysics when I entered her lab in 2003. I've tried to rise to the standard she sets, and imagine that I will keep trying for the rest of my scientific lifetime.

I am indebted to Claus Wilke on innumerable levels. Without Claus's intelligence, gentle suggestions, brutal judgments, and old-fashioned hard labor, most of the results reported here would never have existed. He has been a generous and calm mentor, a joy to work for, and then with, on the considerable corpus we've generated, *P*-value by R-script, in four years. As described in Chapter 4, it was his idea to use principal component regression, and his analysis on yeast, that created the main result reported there. He wrote the first versions (and in my view the most difficult parts) of the lattice protein simulations that became Chapter 6. Claus's influence on me is impossible to overstate; he taught me most of what I know about the gory details of doing science day-to-day, from the difference between thinking I have a result and actually having it, down to the Benjamini-Hochberg false-discovery-rate correction for multiple tests. I envy the students who will get him at 22 instead of 30.

Without Chris Adami, I would still be working a human resources job in Texas. Chris's enthusiasm and aggressive, insightful assaults on the big questions of evolution deeply influenced me. Tactically, I learned more as his TA for CNS 178 than I did from any other

CNS course. Throughout, he guided me with remarkable lightness of touch, but in the end, his influence on this thesis is, as he might say, “completely obvious!”

Discussions with Jesse Bloom have been invaluable: his elegant work on protein mutational tolerance provides the backdrop for Part 1 of this thesis, and his analysis of evolutionary rates inspired Part 2. Michelle Meyer taught me how to think like an experimentalist (my failure to become one cannot be laid at her competent feet!) and has been a wonderful sounding-board for ideas. Conversations with Zhen-Gang Wang built my confidence in theory work, and he provided critical guidance on certain mathematical treatments, particularly for the simple models in Chapters 1 and 3 (errors in conception and execution are solely mine).

Joff Silberg performed the experiments analyzed and modeled in Chapter 3 and helped develop the ideas presented there. Jeff Endelman, and indeed most members of the Arnold Lab, engaged in countless impromptu back-and-forth sessions that greatly improved the ideas here. George Georgiou and Brent Iverson at the University of Texas were generous with their experimental data and insight, and Chapter 2 was the result; Alpan Raval did eye-wateringly thorough work on the failings of partial correlation analyses; and Brian Baer supported my sometimes greedy (Chapter 6!) use of the Alice computing cluster resources at Michigan State University. Alice Sogomonian, Divina Bautista and Mariah Oh helped keep this creaky old vessel in fighting trim. All have my thanks.

The distinguished members of my thesis committee (Frances, Chris, Erik Winfree, Michael Elowitz and Shuki Bruck) provided just the right amount of guidance and were always available when I had questions. By providing me with an NIH Training Grant, the CNS option let me do whatever I pleased for my first three years, including going to work for a chemical engineer unaffiliated with the option.

Most of all, I am grateful to my parents, Dave and Joy Drummond, for their unwavering enthusiasm and support as I spouted quantitative nonsense on vacations during the turn of

the millennium, wrestled with a change of careers, quit a perfectly wonderful job, moved across the country, and subsequently did not get married or produce offspring over the course of four years while grappling with problems of uncertain significance in exchange for poverty wages. I have tried to find the bottom of their love, and I have failed.

ABSTRACT

The diverse array of protein functions depends upon these molecules' reliable ability to fold into the native structures determined by their amino-acid sequences. Because mutations that alter a protein's sequence frequently disrupt its folding, protein evolution explores protein sequence space conservatively, either by point mutations or recombination between related sequences. Attempts to engineer proteins by co-opting the evolutionary algorithm have also largely proceeded by the stepwise accumulation of beneficial mutations. Other strategies for directed evolution have focused on introducing many mutations at once as a way to increase the likelihood of finding improved variants, attempting to balance higher mutational diversity with lower retention of folding. Using simple models, I explore this tradeoff and find that protein misfolding dominates whether increasing mutation levels increase the number of improved variants. I analyze results of a popular mutagenesis protocol, error-prone PCR, for evidence that coupling between mutations might favor higher mutation levels, as claimed by several groups. A comparison of high-mutation-rate mutagenesis to protein recombination between distantly related proteins reveals qualitative differences in protein tolerance for sequence changes introduced by each method. Mutational tolerance may also be reflected in the rate at which proteins accumulate sequence changes over evolutionary time; why proteins evolve at different rates remains a major open question in biology. An analysis of rate determinants suggests that one major variable, linked to how highly expressed the encoding gene is, dominates the rate of yeast protein evolution. To explain this trend, I hypothesize that proteins are selected to fold properly despite mistranslation, a property I call translational robustness, and test it using genomic data. To examine protein evolution at a higher level of detail, a large-scale simulation is constructed in which simulated organisms, with genomes containing genes expressing computationally foldable proteins at different levels, evolve over millions of generations with protein misfolding imposing the only fitness cost. The results suggest that protein misfolding suffices to explain many significant trends in genome evolution

observed across taxa, predict a novel genomic trend which is then identified in yeast, and create insight into the causes of evolutionary rate variation in proteins.

For *Misfolding Dominates Protein Evolution*, by David Allan Drummond.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

PREFACE

Happy families are all alike; every unhappy family is unhappy

in its own way.

Leo Tolstoy*, Anna Karenina*

Functional proteins are all alike; every misfolded protein is misfolded in its own way.

An array of powerful techniques may be swiveled with delight in the direction of a functional protein. There are the countless stereotypical biophysical assays: visualization of circular dichroism and tryptophan fluorescence and NMR spectra, denaturation with heat or chaotropic agents, crystallization, separation by charge and solubility. Biological interrogations may also commence to determine such properties as activities, pathway participation, and subcellular localization. The very existence of huge protein databases with fixed schema attests to the Tolstoyesque alikeness of functional proteins; indeed, as in families of either temperament, protein family members correlate in their behavior down to the very angle of their backbones.

Misfolded proteins are a different story. Or stories—any two misfolded variants of the same protein, to the extent we can (or want to) know anything about them, might differ in every truncated, erroneous, irreversibly modified detail. Even the fates of these fallen soldiers remain uncertain, ranging from aggregation in a gooey blob, to being ground up, or being refolded and sent back into battle. Because all misfolded proteins are different, the diagnosis of misfolding typically signals the end of scientific interest. (The array swivels away.) A database of misfolded proteins seems amusing, or sad, or even gruesome, like a

database of bridge collapses or train derailments. Unhappy families may make great novels, but misfolded proteins make ghastly research subjects.

Such is the view as we pan the camera across basic and applied biology, from biochemistry to biophysics to genetics to protein engineering, and it persists as genomics and evolution enter the frame. The panorama of data is focused through the lens of functionality: catalytic residues, active sites, binding domains, structural motifs, conformational changes, macromolecular complexes, interactome network diagrams. In the genomic era, few annotations are more intriguing than "conserved protein of unknown function."

Yet we continue to grapple unhappily with the unpleasant reality that while functional proteins might be (in some senses) all alike, most protein functions are different. Worse, in the absence of similar sequences with known properties, we cannot reliably predict or engineer the folding or function of a protein. We cannot even reliably predict if or how a single mutation will alter protein fold or function, except to say that the results (like the predictions) probably won't be pretty.

Averaging over all such mutations, though, we might predict two basic outcomes: minimal change, or misfolding-induced loss of function. Such averaging consistently arises in the repeated protein-engineering experiments (by nature or by humans) that generate huge ensembles by the conserved processes of mutation and recombination: the differences dilute out, and the similarities remain. In any genome-wide trend, function dilutes out. In any general directed evolution strategy, function dilutes out. And as it does, misfolding titrates in.

Misfolded proteins are all alike; every functional protein is functional in its own way.

#

In Part 1 of this thesis, I study the influence of mutation-induced misfolding on the average outcomes of many attempts to direct evolution. Chapter 1 introduces several key ideas and tools for studying high-mutation-level directed evolution, and a simple analysis reveals the

powerful effect protein misfolding exerts over optimal mutation rates. Chapter 2 is a detective story in mutagenesis that attempts to explain why the popular method of error-prone PCR mutagenesis, when run at very high mutation rates, seems to produce a startling excess of functional and improved proteins relative to expectations (such as those set in Chapter 1). Chapter 3 compares high-error-rate mutagenesis to protein recombination for the exploration of distant regions of protein sequences space, and provides a simple model for why random mutants lose function at rates up to 16 orders of magnitude higher than chimeric proteins with the same number of amino acid substitutions.

Throughout Part 1, I focus on understanding mutational tolerance in proteins. Intuition suggests that the rate at which proteins evolve in nature should be related to their mutational tolerance. In Part 2, I analyze the natural evolution of proteins.

Chapter 4 analyzes genome-wide data from baker's yeast, a widely used model organism, to determine what variables most strongly predict a protein's evolutionary rate. Surprisingly, the rate of translation appears to be overwhelmingly dominant. Chapter 5 proposes and defends an explanation for this dominance, the hypothesis that proteins are strongly selected to resist mistranslation-induced misfolding, and that this selection for translational robustness slows their evolution. Chapter 6 integrates the ideas from the previous two chapters, but jettisons their retrospective bioinformatic approach. Instead, a massive simulation of the evolution of an entire genome over tens of millions of generations, mutation by mutation, is described and analyzed. This simulation not only reproduces and creates insight into many dominant trends in genome evolution, but makes biological predictions which are then tested.

PART 1

MISFOLDING DOMINATES DIRECTED PROTEIN EVOLUTION

C h a p t e r 1

BALANCING DIVERSITY AND MISFOLDING TO FIND IMPROVED PROTEINS

I begin with a block of marble and chip away the parts that are not statue.

Attributed to Michelangelo Buonarroti

Proteins have evolved to perform an unreasonable number of functions under very reasonable conditions. At room temperature, in water, often with high activity and low toxicity, proteins can be expected to cleave sugar (hexokinase), fix carbon (Rubisco), convert ion gradients into propulsion (flagellar motors), bind oxygen (hemoglobin), cut DNA (restriction enzymes), cut other proteins (proteases), and recognize invaders (antibodies). Yet these sleek, efficient nanomachines turn finicky, balky or useless when aimed at tasks we humans find useful[1], even such related tasks as cutting *other* DNA[2] or recognizing *other* invaders[3].

The diversity of protein functions makes these molecules an equally seductive and daunting engineering target. Given that we cannot yet assign a structure to an amino acid sequence with any reliability, and cannot assign functions to structures without an evolutionary cheat-sheet, how can we hope to engineer these molecules?

One successful answer has been to co-opt nature's engineering algorithm, to direct evolution.[1] By alternating diversity generation, often through random mutation and recombination, with selection for desired properties, we can improve proteins without having to understand the details of the sequence-to-function mapping.

Such a shift may seem positively Faustian: we may obtain engineering results, but only by forfeiting our scientific soul, the imperative to understand *why*. But such a tradeoff is illusory. We have merely shifted problem domains, trading the presently intractable deterministic challenge of designing an improved protein for the (possibly) more tractable probabilistic challenge of designing an ensemble likely to contain such a sequence.

In protein engineering, such ensembles are called libraries[4], and typically they grow, like a small-town branch, from one or a handful of ragged donations, wild-type proteins which in their human usefulness are not bestsellers, but have some promising bits. Most libraries are mutant libraries, in which the variants differ by a few characters. Some are recombination libraries, in which entire folios have been promiscuously swapped around. Whatever the method of generating a library, the goal at the end is to check out of it a better book than we donated—a tall order. Like books, most randomly fiddled-with proteins aren't just bad proteins, they are nonsensical garbage. Rational library design[4-6] seeks general ways to increase our likelihood of finding better proteins, which (in a theme elaborated below and in the following two chapters) often simply involves seeking to minimize the time spent sorting garbage.

A central principle that allows evolutionary library design to be an engineering discipline rather than an anecdotal craft is that, to perform any of their myriad functions, proteins must fold. Sequences encoding folded proteins are exceedingly rare in the space of all possible sequences,[1,7] so the search for folded proteins necessarily guides the search for functional and improved proteins. Most mutations that disrupt function also disrupt folding.[8-11] (Recently, this observation's converse has been examined for one family of enzymes: 94–96% of mutant cytochromes P450 that retained fold also retained at least half the wild-type activity on a target substrate[12].)

Thus, given our probabilistic challenge to design a mutant library likely to contain an improved protein, and knowing little to nothing about how sequence changes affect function, except that: 1) they usually destroy it by disrupting folding; but 2) *some* change is

required to obtain improvement, we must better understand the tradeoff between folding and diversity.

For the rest of this chapter, I develop intuition about the interplay of folding and diversity in a specific class of mutant libraries, develop a simple mathematical treatment of this interplay (elements of which are expanded in the following two chapters), present a protein folding model exercised throughout this thesis, compare model and simulation results for the problem of obtaining mutants with increased stability, and raise questions to be addressed in Chapters 2 and 3.

Modeling improvement, and the Principle of Pessimistic Additivity

Let us assume we can assign a fitness w, a performance rating, to every mutant. We will begin with a wild-type sequence *s* having fitness w_0. The wild type may itself be an engineered mutant; "wild type" and "starting point" will be used interchangeably here. The objective is to isolate an improved mutant having some unspecified number of amino-acid substitutions (mutations) *m* whose fitness exceeds a threshold $w_t > w_0$. I will assume such improvement requires proper folding, where the folding state *f* is encoded by a binary random variable taking values 1 (true) and 0 (false). The probability of improvement in a folded protein having *m* amino-acid mutations generated from a starting sequence *s* I will denote $\Pr(w > w_t | f = 1, m, s)$.

Now suppose we are building a library in which mutants may possess mutation levels drawn from some distribution. For example, the popular method of error-prone PCR generates a library by recursively and sloppily copying DNA in a mixture intially seeded with wild-type sequences, resulting in a distribution of mutations that, as I show using experimental data (Chapter 2), matches a predictable mutant distribution. Alternatively, recombination of several related wild-type sequences (Chapter 3) generates a very different distribution. For our purposes here, I will denote that distribution $\Pr(m|s)$, which omits the

details of mutagenesis but for the moment honors the possibility that the distribution may depend on the wild type's sequence composition.

Some mutations may disrupt protein function, often by destabilizing the native structure enough to cause misfolding.[11] The probability of mutation-induced misfolding depends on the wild type (because more-stable proteins can tolerate a wider array of destabilizing effects)[11] and the number of mutations (because stability changes are roughly additive)[13]. Recognizing this dependence, I will denote the probability of proper folding given m mutations applied to a wildtype sequence $\Pr(f = 1|m, s)$.

These definitions lead to a straightforward formulation of the probability that a library starting from a wild-type sequence s contains an improved mutant:

$$\Pr(w > w_t|s) = \sum_m \Pr(w > w_t|f = 1, m, s) \Pr(f = 1|m, s) \Pr(m|s) \qquad (1.1)$$

where the sum over the number of mutations m runs from zero to the length of the protein, and there is no sum over f because the probability of improvement in a misfolded mutant is zero.

As described above, mutational distributions in libraries have been well-studied, and I devote Chapter 2 to the detailed analysis of one such distribution, so $\Pr(m|s)$ can be considered known in some relevant cases. In a beautifully simple treatment, my colleague Jesse Bloom and others have shown that the probability of proper folding $\Pr(f = 1|m, s)$ can be accurately predicted, for real and simulated proteins, considering only the wild type's thermodynamic stability (free energy of unfolding ΔG) and stability changes ($\Delta\Delta G$) induced by mutations.[11,14] For my purposes here in describing average outcomes as simply as possible, I will use a well-known result, replicated by Bloom *et al.*'s model, that the fraction of folded proteins generated by random mutagenesis declines roughly exponentially on average, [11,15-17]

$$\Pr(f = 1 | m, s) \approx \nu^m \qquad (1.2)$$

where the parameter ν, the *neutrality*, represents the probability that a mutation to a folded protein yields another folded protein. Neutrality describes the average connectivity of the neutral network of folded sequences,[18] hence its name, and it is determined by the protein structure and minimal stability requirement shared by all such sequences.[11,19] With $\Pr(m|s)$ and $\Pr(f = 1|m, s)$ in hand, we are left with only the probability $\Pr(w > w_t | f = 1, m, s)$, the probability of improvement given a folded mutant separated from wildtype by *m* mutations.

Progression past this point requires some knowledge (or, more often, an assumption) about how mutations affect fitness. A common assumption implicitly made in most directed protein evolution experiments is that mutations are roughly additive. Directed evolution is the sequential improvement of protein properties using iterated rounds of mutagenesis and selection, a physical realization of an adaptive walk in protein sequence space.[1] Such adaptive walks have been exhaustively studied elsewhere,[20] and a central result holds that when mutations have strongly coupled (non-additive) effects, the fitness landscape becomes so rugged and decorrelated that most adaptive walks rapidly terminate at sub-optimal local maxima: you take a short walk up a small hill to nowhere. That stepwise directed evolution is so widely used suggests that practitioners are willing to assume mutations are not strongly coupled; that such evolution has produced so many successes suggests they are right to do so.

Additivity confers pleasant mathematical properties which allow the potentially daunting probability of improvement $\Pr(w > w_t | f = 1, m, s)$ to be treated simply. Many mutational effects on highly complex properties are roughly additive,[13] and virtually any property can be treated additively over small enough ranges.

Even given such broad trends, when faced with a particular protein attempting a particular biological task, we may be loath to assume additivity. However, considering strategies for

directing evolution which must encompass the widest range of wild types, a weaker but more palatable (albeit lighthearted) principle might be substituted:

> ***The Principle of Pessimistic Additivity***
>
> *A directed evolution strategy that will not work assuming additive mutational effects will not work at all.*

To the extent we assume anything about multiple mutations, we choose additivity over hopelessness, and even then, we do not expect any mutations to be precisely additive. (In this sense, the assumption of additivity parallels the common statistical assumption of normality, with similar attendant caveats.)

My touchstone question is: "Suppose mutations affected my target property roughly additively. How should I direct evolution?"

I will limit the following analysis in two major ways. First, I will treat cases in which multiple mutations are made simultaneously. Such cases make additivity nontrivial, they have not been treated as thoroughly as the single-mutation-accumulation case, and several approaches and results in directed evolution have emphasized the potential advantages of accessing distant regions of sequence space[5,16,21,22]. Construction and characterization of a library takes time and effort, so one might wonder whether stepwise accumulation of point mutations in an adaptive walk from library to library makes sense. Why walk if you can run? Second, I will examine average trends and properties, because to the extent general approaches to directed evolution are possible, they must work over diverse wild-type proteins and choices of fitness function. Both limits attempt to enhance and harness the key benefit gained by trading off deterministic for stochastic design, namely increased obedience to the law of large numbers.

Directed evolution with additive fitness, given folded mutants

The question, "How should I direct evolution?" may be phrased more tactically: "To increase the probability of finding improved mutants, under what conditions should I increase the mutation level of my library?"

My aim in all that follows is to examine the role of protein misfolding in decisions about mutagenesis. Accordingly, I adopt a rather minimalist and intuitive approach here and relegate detailed quantitative analyses to a specific practical problem in Chapter 2.

Suppose misfolding is not a concern, or we restrict our attention to those mutants which retain fold. Further suppose that libraries in which every sequence has the same number of mutations *m* may be constructed. When should *m* be increased?

Guiding intuitions

Such a question may seem impossible to answer, because the wild-type sequence may have idiosyncratic properties. However, the sequence space for a typical protein, while the size of a multi-universe, has the geography of Mayberry: very little is more than several steps away, and the whole lot may be traversed end-to-end in a few hundred mutational paces[1]. As a result, applying a only handful of mutations to any wild-type sequence without disrupting folding will yield an ensemble of mutants in which, on average, all idiosyncrasies have vanished. Further mutations will tend to generate ensembles with identical properties. A clear example of this behavior is found when looking at the fraction of mutants retaining wild-type fold as a function of mutational distance *m*: after roughly four amino acid changes, the effects of initial wild-type stability have given way to the average properties associated with the protein structure shared by all mutants.[11,14] As a result, an additive model linking mutation-induced stability changes to the probability of folding can be replaced with a simple mean-field model with little accuracy loss.[14]

Such convergence to average properties provides an intuitive guide to the probability of improvement given a folded *m*-mutant of wild-type sequence *s*, $\Pr(w > w_t \,|\, f = 1, m, s)$. Let

the average fitness of a folded protein in sequence space be $\overline{w}$. If the wildtype sequence has $w_0 \approx \overline{w}$, then mutations which preserve folding will also tend to preserve the starting point for future attempts at improvement. If the wildtype has $w_0 < \overline{w}$, then directed evolution is easy: most folded mutants will have higher fitness. If, however, the wild-type sequence has $w_0 > \overline{w}$, then each mutation reduces fitness on average, moving the goalposts farther and farther away.

In a library, the behavior of the mean fitness of a folded protein is rarely crucial, because for any problem requiring real effort, the asymptotic average fitness $\overline{w}$ will be below the desired fitness. The potential value of multiple additive mutations lies in the behavior of the *variance*. For any sum of independent random variables, the variance increases with the number of summands. Thus, if the mean fitness hovers in place ($w_0 \approx \overline{w}$), again restricting our attention to folded proteins, more independent additive mutations are better on average. If the mean fitness tends upward ($w_0 < \overline{w}$), more mutations are even better.

If $w_0 > \overline{w}$, then an additional mutation makes sense only if the increase in variance compensates for the expected fitness reduction. As mutations accumulate and the mean recedes, only those mutational combinations far out on the positive tail have any hope of exceeding the threshold for improvement. A single deleterious mutation can erase all the small positive gains in a stroke, because when $w_0 > \overline{w}$, the average deleterious effect will tend to be larger than the average beneficial effect.

A wild-type protein's history may give us clues about w_0 relative to $\overline{w}$. Properties not under consistent selection will tend to drift[1,23], presumably toward average values, a trend which can explain why proteins are marginally stable[24]. Properties that have been actively traded off for others, such as affinity for unpreferred substrates in enzymes with high specificity, may be pushed below average values, and a property optimized by evolution (natural or directed) will clearly rise above them.

On the basis of this intuitive treatment, considering the mean and variance of fitnesses of folded mutants, I predict that the mean will predict improvement better when $w_0 < \overline{w}$ (less-fit than average wild type) than when $w_0 > \overline{w}$, and the mean plus the variance will always predict improvement better than the mean.

Quantitative model

A full mathematical treatment is beyond my scope, but a simple model to capture the battle between beneficial and deleterious mutations may be constructed as follows. Suppose that, within the network of folded proteins, there exists a sub-network of improved proteins which are above some threshold of additive fitness. Starting from an unimproved wild-type protein, mutations represent a random walk which occasionally ends on an improved protein. In some cases, an otherwise improved protein might suffer a virtually unrecoverable deleterious mutation, and I will call such proteins "mangled." (The assumption of irreversibility reflects the presumption that the distribution of fitness changes will be strongly skewed toward deleteriousness, such that an average deleterious mutation can only be compensated by several beneficial mutations, leading to an effectively unrecoverable fitness deficit.) Let the transition probability from one unimproved protein to another be *x*, from a folded protein to an improved protein be $1 - x$, from one improved protein to another be *y*, and from an improved to a mangled protein be $1 - y$ (Figure 1.2). Then our system is equivalent to a three-state Markov chain, containing states "improved", "unimproved", and "mangled," with:

$$\text{initial state vector } \mathbf{v} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{ and transition matrix } \mathbf{A} = \begin{bmatrix} y & 1-x & 0 \\ 0 & x & 0 \\ 1-y & 0 & 1 \end{bmatrix}.$$

The probability $\Pr(imp \mid unimp, m)$ of a random walk starting at a folded but unimproved protein and ending at an improved protein (selected by $\mathbf{p} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$) after *m* mutations is:

$$\Pr(imp \mid unimp, m) = \mathbf{p}\mathbf{A}^m\mathbf{v} = \left(\frac{x-1}{x-y}\right)\left(y^m - x^m\right). \qquad (1.3)$$

Equation 1.3 provides an estimate for the form of $\Pr(w > w_t \mid f = 1, m, s)$. The parameters x and y could perhaps be obtained from first principles, but I have been unable to do so. Instead, let us see whether these functional forms can describe the average behavior of a more comprehensive system using computationally folded model proteins.

Lattice proteins as an analytical testbed

To test the above mathematical treatment, one would ideally like a system which allows complex directed evolution experiments to be carried out quickly and cheaply, from mutagenesis through mutant characterization and analysis. Such a system is presently unavailable (real biology remains arduous and time-consuming), but may be approximated *in silico* using lattice proteins, short simulated polymers of all 20 amino acids which fold to a unique, maximally compact lowest-free-energy conformation on a square lattice (Figure 1.1a). These model polymers are valuable theoretical tools because they combine tractability and fidelity: their simplicity allows rapid and exact calculation of their thermodynamic partition function to assess free energy and folding status, and I and my colleagues and others have demonstrated that they reproduce many qualitative nontrivial mutational-tolerance patterns established in real proteins[14,19,24] (cf. Chapters 2, 3 and 6). In the chapters to follow, I will deploy lattice proteins to test various models in directed and natural evolution, so a few words on their properties are warranted.

The 5×5 (25-mer) lattice protein used here can adopt one of 1,081 square conformations not related by rotation or symmetry (*e.g.*, Fig. 1.1a). A sequence conformation's energy equals the sum of its 16 pairwise nearest-neighbor non-bonded interactions, where the energy of each residue-residue interaction is the effective free energy including an implicit solvation term as tabulated by Miyazawa and Jernigan[25] (their Table 3) from a database of real proteins (see *Methods*). The particular energy values have little significance and no attempt will be made to derive conclusions about particular residues, their frequencies, or their

energies in real proteins from lattice-model observations. Similar to real proteins, the stability effects ($\Delta\Delta G$) of mutations are roughly additive (Figure 1.1d).

To simulate the biological requirement for stable folding imposed on most proteins, we apply an arbitrary free-energy cutoff (typically 5 kcal/mol), and define any protein below this stability threshold as misfolded. For 25-mer lattice proteins, roughly one random sequence in 200,000 attains a stability of 5 kcal/mol for any structure, making folded model proteins rare (Figure 1.1b). Accordingly, most mutations are destabilizing (Figure 1.1c).

An alternative noncompact lattice model used in our laboratory[11,12] relaxes the requirement of a maximally compact conformation. Anecdotally, there are two major differences between the noncompact and compact proteins aside from their shape: stable proteins become much rarer (necessitating a higher stability cutoff), and the conformational space grows much more rapidly with chain length such that shorter sequences, typically 18- to 20-mers, must be used for tractability. While I have exercised both models, my reliance on the compact model reflects my preference for longer chain lengths and faster folding times. My colleagues and I have not explored any biologically relevant phenomena uniquely captured by one or the other model. In some cases we have performed the same experiment using both models,[11,14] obtaining qualitatively similar results. In the absence of clear contraindications I will cite results obtained using one model under the assumption that they apply to the other.

Results and Discussion

A typical goal in directed evolution is to find proteins with increased stability. Elevated stability not only can allow high-temperature catalysis[26], but also confers tolerance to destabilizing yet functionally valuable mutations that are otherwise inaccessible[12]. Because lattice proteins naturally model thermodynamic stability, and the stability effects of mutations are roughly additive[13,27] (Fig. 1.1d), stability improvement represents a pertinent and nontrivial test of the model proposed above. In a typical directed evolution experiment,

obtaining *any* improvement is not enough, both because noisy assays may produce false positives and because, in general, our goals are usually not satisfied by a 0.1% increase, so I will impose a nonzero threshold for improvement.

To explore the effect of increasing amino-acid mutation levels on the probability of finding mutants with improved stability, I constructed a simulation as follows (also see *Methods*). A wild-type protein sequence adopting a particular structure above a specified stability threshold ($\Delta G_{\mathrm{wt}} \geq 4$ kcal/mol) was found by hill-climbing. (I will typically use a threshold of 5 kcal/mol, but in this case, the reduced threshold made examination of higher mutation levels tractable.) Fitness was measured by stability. An arbitrary improvement threshold of $w_t - w_0 = 0.5$ kcal/mol of increased stability over the wild type was held constant. For each amino-acid mutation level m=0,1,...,14 (more than 50% of the protein sequence), m-mutants were made at random until either the total number of possible m-mutants had been 1× sampled (expected to cover ~63% of all possible mutants) or 100 improved mutants were found. Each m-mutant was assayed for folding (is $\Delta G \geq 4$ kcal/mol?) and for improvement (is $\Delta G - \Delta G_{\mathrm{wt}} \geq 0.5$ kcal/mol?), allowing the relationships between m, probability of folding, probability of improvement, and probability of improvement given folding to be analyzed. Each mutational level m thus corresponded to library with a mutant "distribution" consisting of a delta function at m. Library size ranged from 100 to more than 10^6 mutants, and the entire series of simulations required folding roughly 10^8 proteins.

Given a set of simulation results, I then fit these three probability curves with the models specified by Eq. 1.2 (probability of folding given m), Eq. 1.3 (probability of improvement given folding and m), and their product (probability of improvement given m), using the three free parameters ν, x and y, fit separately for three starting fitnesses.

Figure 1.3 displays the results. The agreement between model and simulation is reasonable in most cases; the most obvious deviations are linked to the fraction folded, for which I have chosen a simplified model which disregards initial stability values. In the case of

wild-type proteins chosen to be near the average stability of folded mutants (Fig. 1.3b,e)—the case ideally modeled by a mean-field treatment—the agreement is excellent. As expected, the model over-predicts the fraction folded for lower-than-average-stability wild-type sequences, and under-predicts in the opposite case, in both cases biasing the fraction of improved mutants in the same direction. The full model of Bloom *et al*. properly accounts for initial stability differences[11,14]. The main contribution of this work, Eq. 1.3, proves a reasonable approximation for the probability of improvement in a folded sequence (Fig. 1.3c-e).

These results justify the formulation of Eq. 1.1, in the sense that the problem of improvement, at least in this case, can be usefully subdivided into the problem of folding and the problem of improvement given folding. They also suggest that the model can indeed explain the gross average behavior of key terms in Eq. 1.1, although for larger improvement targets, the model deviates more significantly, because more mutations are required to obtain any improvement at all (not shown). However, the mathematical model's reasonable performance does not offer any insight into whether the underlying intuitions used to construct it are correct.

Accordingly, I examined the predictions regarding the fitness mean and variance made above using a simple correlation analysis. Table 1.1 reports squared Spearman rank correlations (r^2) of the mean and variance of folded-mutant fitness (hereafter, folded fitness) with the fraction of improved variants among folded mutants, quantifying the fraction of the latter's variation explained by each statistic. All predictions were confirmed. In addition, as expected, the mean folded fitness gravitated upward with mutational distance for $w_0 < \overline{w}$ (Spearman r = +0.50, $P << 10^{-9}$), hovered virtually unchanged for $w_0 \approx \overline{w}$ ($r = +0.09$, not significant), and declined sharply for $w_0 > \overline{w}$ ($r = -0.95$, $P << 10^{-9}$). These findings suggest the intuitive treatment above has some validity.

Figure 1.3 shows that protein misfolding dominates decisions about directed evolution in this model, in the sense that misfolding virtually determines the answer to the question,

"Under what conditions should I increase the mutation level of my library?" Considering only folded sequences, the answer is, "Almost always," while considering folded and unfolded sequences—the actual problem one faces at the bench—the answer is, "Almost never." Past the first mutation or two, diversity-driven gains in improvement are more than offset by diversity-driven losses in folded proteins.

The dominance of misfolding depends in large part on ν, raising the question of how the neutralities of these lattice proteins compare to those of real proteins. While neutralities have been measured for few proteins, Bloom *et al*. report estimates ranging from 0.38 to 0.55 for proteins of diverse structures, and in Chapter 2, I derive a neutrality of 0.54 from experimental mutagenesis data for the β-lactamase PSE-4. The lattice proteins assayed here have neutralities around 0.6, comparable to real proteins. (It is surprising and fortunate that these compact lattice proteins have similar thermodynamic stabilities [5–10 kcal/mol] and neutralities to real proteins. As we will see in Chapter 6, these biophysical properties are crucial to understanding tradeoffs in the natural evolution of proteins.)

These results bear on the question of optimal diversity in protein engineering. I observe that, consistent with previous results[28,29], there can be such a thing as too much diversity even when folding is preserved. However, excess diversity is practically irrelevant in these simulations, because by the time such effects kick in, half the protein has been mutated and virtually all mutants are misfolded. Instead, the dominant tradeoff is between folding and diversity[4]. Optimal diversity balances the need to explore with the need to survive.

In Chapter 2, results from error-prone PCR mutagenesis lead me to suggest that, contrary to some published claims, optimal mutation rates for protein engineering are protein-dependent (because of differences in ν) and protocol-dependent. The latter follows in large part because different protocols yield different mutant distributions $\Pr(m|s)$ and sample mutants across the spectrum of misfolding and improvement in complex ways. The simulation here, constructed to cut through the fog of sampling, allows us to see that there may be a benefit to increasing mutations, absent folding concerns, even when mutations are

roughly additive. However, they also demonstrate that folding concerns should not be absent.

But what if mutations—particularly the ones which confer improvement—are simply not additive[21]? This question is also taken up in Chapter 2, where my aim is to explain the puzzling observation that high-error-rate random mutagenesis using a popular protocol produces improved proteins more often than low-error-rate mutagenesis, a finding claimed to suggest mutational coupling or non-independence. I find little evidence for this claim. (However, it should be noted that if a protein's function has been optimized by point mutation, but is still improvable, the mutations which confer improvement must logically be coupled in the strong sense of being individually deleterious, but cooperatively beneficial.)

If Chapter 2 describes a case of seeing mutational coupling where none exists, Chapter 3 is a case of finding strong coupling where it had not been fully appreciated or even noted, in the first comparison of the efficiency of protein recombination versus mutation in exploring the space of functional (read: folded) sequences.

These two chapters, adapted with only minor modifications from my first two publications, relate complementary case studies of directed evolution strategies in which many mutations are made at once. Both reveal the profound influence of protein misfolding on the efficient exploration of sequence space.

Methods

Lattice protein folding

Lattice proteins were folded as described[30]; those methods are paraphrased here for convenience. The energy of a conformation *i* is

$$E_i = \sum_{j<k} \gamma(A_j, A_k)\Delta^i_{jk}$$

where $\gamma(A_j, A_k)$ is the contact energy between amino acids A_j at position j and A_k at position k in the sequence, and Δ^i_{jk} is 1 if the two amino acids are in contact in conformation *i* and zero otherwise. The partition function is $Z = \sum_i e^{-E_i/k_BT}$ where *i* runs over all 1,081 conformations and $k_BT = 0.6$ kcal/mol is the Boltzmann constant times the effective temperature *T*. The free energy of folding is defined as

$$\Delta G_f = E_f + k_BT \ln[Z - e^{-E_f/k_BT}]$$

where E_f is the energy of the sequence in its the lowest-energy conformation. Stability, the free energy of unfolding, was then $\Delta G = -\Delta G_f$.

Accelerating interrogation of lattice protein folding

I was able to significantly accelerate evaluation of lattice protein fitnesses by observing that determining whether a protein possesses stability above a threshold is an easier problem than determining its actual stability. The partition function has only positive terms, so by tracking the growth of the sum and the minimum and maximum possible contributions from contact energies, it is possible to determine whether the sum will exceed (or not) an arbitrary threshold without fully evaluating the partition function. Implementing this strategy reduced lattice protein folding times by as much as a factor of 15 at high mutation

levels, where most proteins are highly destabilized, and sped up the overall experiment by roughly an order of magnitude.

Mutagenesis

To generate the results analyzed in this chapter, I began by finding a protein with a target structure and a minimum stability. I chose structure 574 because it is highly designable (more sequences fold to it as their native conformation than any other sequence, my unpublished observation) and is therefore intrinsically mutationally tolerant, allowing access to very high mutation levels with some tractable probability of finding folded sequences. Test runs revealed that, under the conditions I studied (minimum ΔG = 4 kcal/mol for folded sequences), the stability of folded mutants converged around 4.6 kcal/mol, providing an estimate of the average fitness of folded mutants. I then chose three stability ranges for wild-type sequences: ΔG = 4 to 4.6 kcal/mol (below-average fitness), ΔG = 4.6 to 4.65 kcal/mol (average fitness), and ΔG = 5.3 to 5.6 kcal/mol (above-average fitness). Sequences adopting the target conformation with stabilities in these ranges were found by random sequence generation and adaptive walks. Wild-type sequences were obtained by evolving initial sequences for 10,000 generations at a low mutation rate and stability constrained to the ranges given above.

Once a wild-type sequence was found, it was subjected to point mutations as described in the text. Folding (stability greater than or equal to 4 kcal/mol) was assessed for all mutants, and exact stability (hence fitness) was assessed for folded proteins.

Statistical analyis

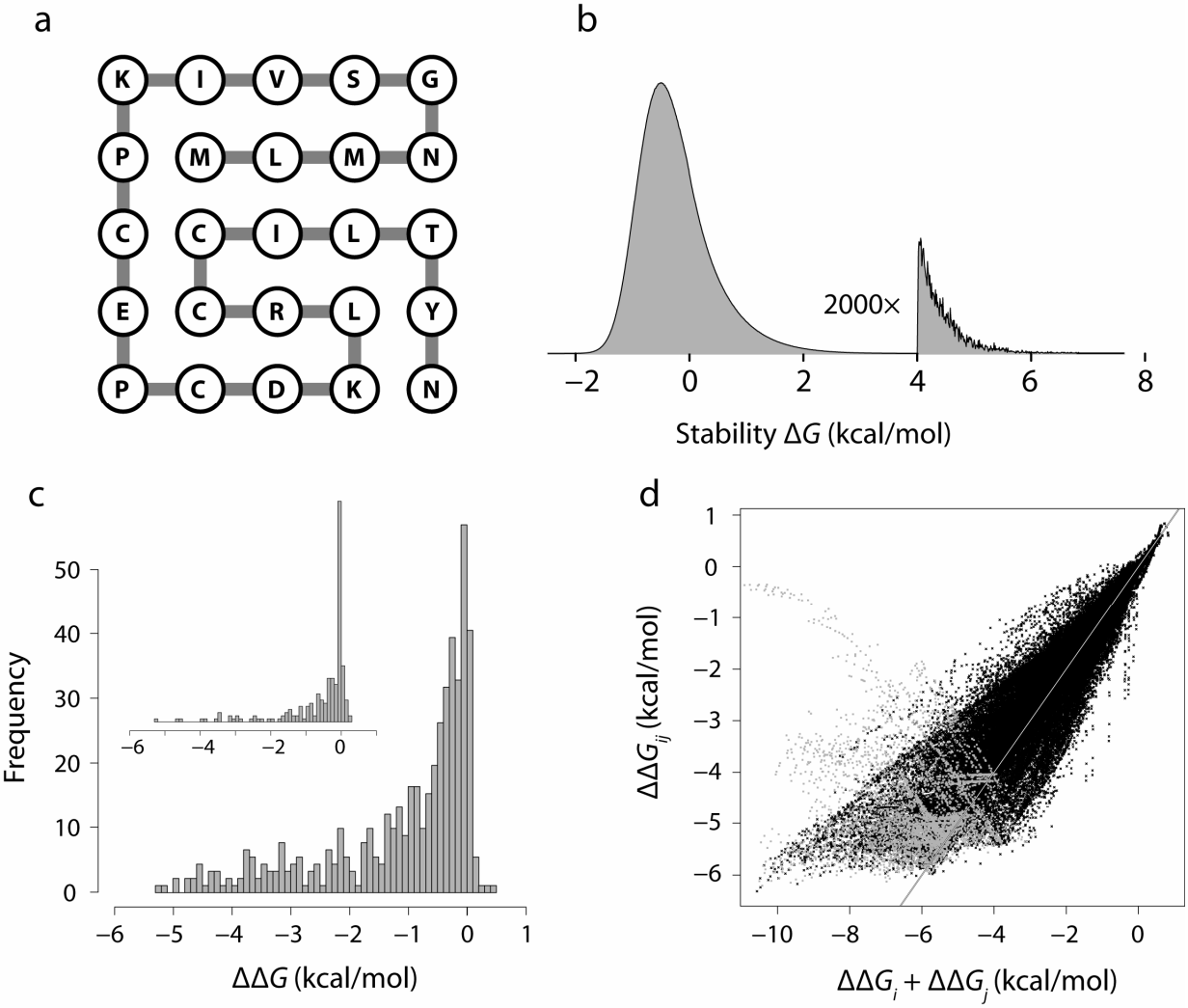
I used R[31] for statistical analyses and plotting.

Table 1.1: Correlation analysis of folded-mutant fitness properties on the likelihood of improved function.

	Fraction of B's variance explained by A (Spearman r^2)*		
Relationship (A & B)	$w_0 < \overline{w}$	$w_0 \approx \overline{w}$	$w_0 > \overline{w}$
Var[folded fitness] & Pr(improved)	0.39	0.73	0.51
Mean+Var[folded fitness] & Pr(improved)	0.41	0.91	0.35
Mean[folded fitness] & Pr(improved)	0.38	0.84	0.22

*All correlations highly significant, $P << 10^{-9}$.

Figure 1.1: Overview of lattice protein properties. **a**, Proteins are simulated as chains of 25 residues, formed using the canonical 20-amino-acid alphabet, which fold into one of 1,081 maximally compact conformations on a square lattice. The native structure is defined to be that conformation adopted with highest stability (free energy of unfolding ΔG). The sequence displayed adopts its native structure with a stability of $\Delta G = -5.04$ kcal/mol. **b**, Stable lattice proteins are rare. 100 million random sequences were folded to obtain stability values, and the distribution of stabilities is shown; above 4 kcal/mol, the distribution is magnified two-thousand-fold to show detail. **c**, Most amino acid substitutions destabilize a lattice protein's native structure. A histogram of stability changes ($\Delta\Delta G$) are shown for all 475 possible single-residue substitutions of the protein shown in **a**. The $\Delta\Delta G$ distribution of all nucleotide-level mutations (inset) for a 75-nucleotide-long gene encoding this protein shows more neutral substitutions ($\Delta\Delta G = 0$) due to the degeneracy and conservative nature of the genetic code. **d**, The stability effects of amino-acid substitutions are roughly additive for lattice proteins. All 108,000 possible double-mutants of one lattice protein were generated, and the stability changes measured as the sum of mutations made separately (x axis) and when both are made together (y axis). The vast majority of double-mutants retain the wild-type native structure despite reduced stability (>95%, black **x**'s), but some mutants adopt a different native structure (gray boxes), especially when both mutations are highly destabilizing. In the model employed throughout this work, "protein misfolding" encompasses destabilization past a threshold and/or altered native structure.


a
K I V S G
P M L M N
C C I L T
E C R L Y
P C D K N
b
2000×
−2 0 2 4 6 8
Stability ΔG (kcal/mol)
c
Frequency
50 40 30 20 10 0
−6 −4 −2 0
−6 −5 −4 −3 −2 −1 0 1
ΔΔG (kcal/mol)
d
ΔΔG_ij (kcal/mol)
1 0 −1 −2 −3 −4 −5 −6
−10 −8 −6 −4 −2 0
ΔΔG_i + ΔΔG_j (kcal/mol)

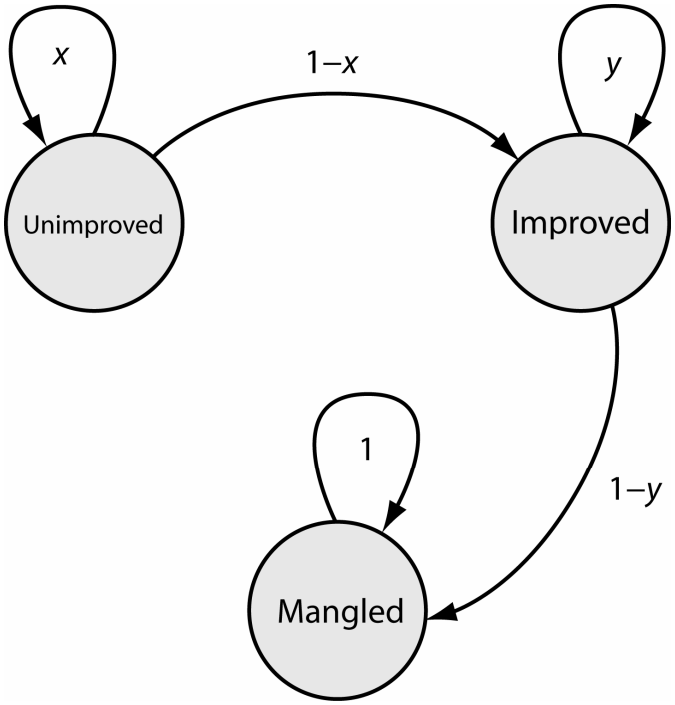


Figure 1.2: A simple Markov-chain model for the probability of protein improvement. The three states and transitition probabilities are shown (see text).

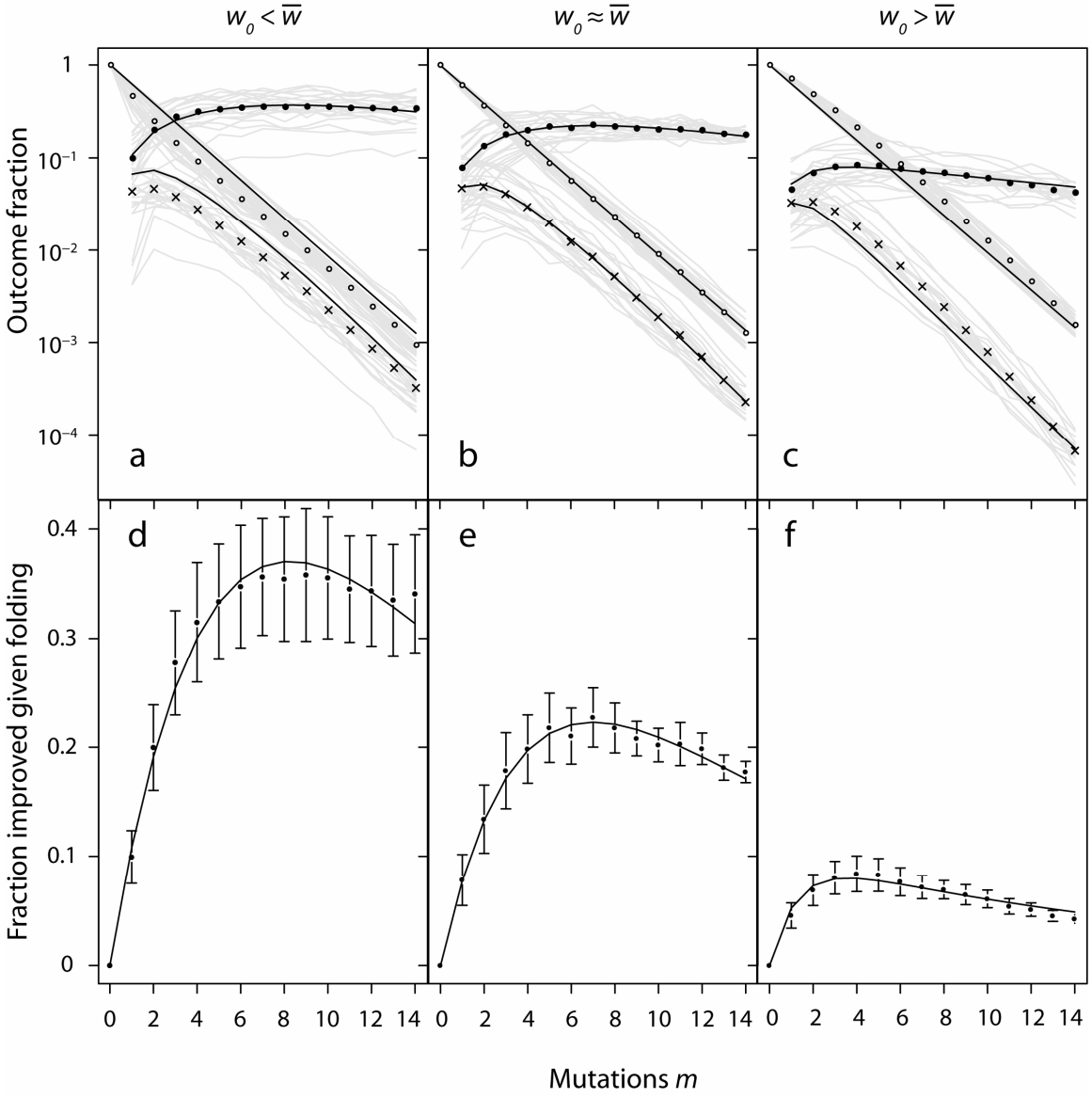


Figure 1.3: Simulation results for the probability of improving lattice protein stability by 0.5 kcal/mol, beginning with wild-type sequences of varying fitnesses (top, see text). **a–c**, The average fractions of mutants that retain fold (○), that show the desired improvement in stability (×), and show improvement conditioning on folding (•) are shown as a function of amino-acid mutational distance m from the wild-type sequence (top). The raw, unaveraged data are also shown (gray lines). Model results (black lines) show reasonable agreement. **c–e**, The fractions of folded proteins showing the desired stability improvement, mean ± s.e.m., are expanded to show the fit of Eq. 1.3.

C h a p t e r 2

COUPLING IN HIGH-ERROR-RATE RANDOM MUTAGENESIS[1]

I don't recall your name but you sure were a sucker for a high inside curve.

Bill Dickey

Summary

The fraction of proteins which retain wildtype function after mutation has long been observed to decline exponentially as the average number of mutations per gene increases. Recently, several groups have used error-prone polymerase chain reactions (PCR) to generate libraries with 15 to 30 mutations per gene on average, and have reported that orders of magnitude more proteins retain function than would be expected from the low-mutation-rate trend. Proteins with improved or novel function were disproportionately isolated from these high-error-rate libraries, leading to claims that high mutation rates unlock regions of sequence space that are enriched in positively coupled mutations. Here, we show experimentally that error-prone PCR produces a broader non-Poisson distribution of mutations consistent with a detailed model of PCR. As error rates increase, this distribution leads directly to the observed excesses in functional clones. We then show that while very low mutation rates result in many functional sequences, only a small number are unique. By contrast, very high mutation rates produce mostly unique sequences, but few retain fold or function. Thus, an optimal mutation rate exists which balances uniqueness

---

[1] Adapted from *Journal of Molecular Biology* **350**, D. Allan Drummond, Brent L. Iverson, George Georgiou, and Frances H. Arnold, "Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins," p. 806-816, copyright (2005), with permission from Elsevier.

and retention of function. Overall, high-error-rate mutagenesis libraries are enriched in improved sequences because they contain more unique, functional clones. The mutational distribution can explain the surprising mutagenesis results; invoking mutational coupling is unnecessary.

Introduction

Laboratory evolution has been used to improve protein properties by mimicking natural evolution's stepwise exploration of sequence space[7], steadily improving protein activity or thermostability through repeated rounds of low-frequency mutation and selection. Because the fraction of proteins retaining function declines exponentially with increasing numbers of amino acid substitutions[11,15-17] (cf. Chapter 1), low mutation rates seek to create mutational diversity without destroying activity[26] so that improved clones can be found.

Recently, several groups reported construction of mutant libraries using high-mutation-rate error-prone polymerase chain reactions (EP-PCR) to probe distant regions of sequence space for an antibody fragment (up to an average $\langle m_{\mathrm{nt}} \rangle = 22.5$ nucleotide mutations per gene)[16,32], hen egg lysozyme (up to $\langle m_{\mathrm{nt}} \rangle = 15.25$)[33], and TEM-1 β-lactamase (up to $\langle m_{\mathrm{nt}} \rangle = 27.2$) [21]. Where both high and low error rates were assessed, the exponential trend in loss of function established for low $\langle m_{\mathrm{nt}} \rangle$ was spectacularly violated at the highest rates, with orders of magnitude more functional clones isolated than would be expected[16,32,33]. Two studies reported improved or novel function more often in these high-mutation-rate libraries[16,21], leading to suggestions that low mutational pressure may not be optimal[16,21] and that hypermutagenesis can, without an exponentially increasing cost in inactivated sequences, explore multiple interacting mutations inaccessible to low-error-rate mutagenesis[21]. These putative interactions could involve synergistic interactions to increase function directly, or combinations in which one or a few mutations increase function at the cost of folding or structural stability, the negative effects of which are suppressed by additional compensatory stabilizing mutations elsewhere in the protein.

The degree to which mutations interact, and mutational effects then deviate from independence, is known as *epistasis*. Independent mutational effects imply an exponential decline in fraction functional with mutational distance, so the above studies' results suggest that mutations interact epistatically on average. Such a finding is of fundamental interest in evolutionary biology[34,35] and is potentially decisive in answering the major open question,

"Why is there sex?"[36] Moreover, the discovery of reservoirs of positively interacting mutations would fundamentally change strategies for *in vitro* enzyme engineering by evolutionary methods[21]. Therefore, a careful analysis of these results is imperative.

Quantitative analysis of high-frequency mutagenesis results often assumes a Poisson distribution of mutations in error-prone PCR, an idea introduced by Shafikhani *et al.*[17]. This group's careful study on *B. lentus* subtilisin found an accurately reproducible exponential decline in fraction functional in all libraries where functional proteins were found, up to $\langle m_{\mathrm{nt}} \rangle$=15, contrary to the upward trend reported later.

To examine the mutational distribution generated by high-error-rate error-prone PCR, we constructed two large libraries of single chain Fv (scFv) antibody mutants. The wildtype scFv antibody fragment derived from the 26-10 monoclonal antibody[37] binds digoxigenin with high affinity, and has been expressed as a fusion to the *E. coli* outer membrane protein Lpp-OmpA', allowing detection of mutants binding fluorescent-dye-conjugated digoxigenin by fluorescence-activated cell sorting (FACS)[16]. Libraries were assayed for mutant retention of wildtype affinity for digoxigenin (briefly, retention of function). These libraries were constructed and assayed exactly as in a previous study[16], making the results of both studies directly comparable. We were able to determine how the mutational statistics relate to PCR experimental parameters and to retention of function.

We show that mutations introduced by error-prone PCR at high error rates do not follow the Poisson distribution, but rather a previously proposed distribution derived from a model of the actual PCR process[38]. We derive the expected fraction of functional mutants based on this more realistic model and show that many reported experimental mutation data follow this model's predictions. We then introduce a simple measure of optimality to evaluate optimal mutation rates for improvement of protein function. Our results show that the trends observed in earlier studies do not constitute evidence for positive epistasis.

Throughout this chapter, we refer to preservation of function rather than of folding, because in typical mutagenesis experiments, only functional assays are performed. See Chapter 1

for a discussion of why, in most cases, loss of folding is the likely culprit for loss of function.

Results

Distribution of mutations generated by error-prone PCR

The probability $\Pr(f)$ that an error-prone PCR-amplified sequence retains function can be obtained as follows (here, and in all that follows, we elide the conditioning on the initial sequence introduced in Chapter 1). Sun[38] modeled error-prone PCR by assuming n thermal cycles during which DNA strands are duplicated with probability λ, the PCR efficiency (assumed constant, realistic for large amounts of starting template[39,40]), resulting in $d=n\lambda$ DNA doublings and an average of $\langle m_{\mathrm{nt}} \rangle$ nucleotide mutations per sequence. The mutational distribution under these assumptions can be written[38], with $x = \frac{\langle m_{\mathrm{nt}} \rangle (1+\lambda)}{n\lambda}$, as

$$\Pr(m_{\mathrm{nt}}) = (1+\lambda)^{-n} \sum_{k=0}^{n} \binom{n}{k} \lambda^k \frac{(kx)^{m_{\mathrm{nt}}} e^{-kx}}{m_{\mathrm{nt}}!}, \qquad (2.1)$$

which has mean $\langle m_{\mathrm{nt}} \rangle$ and variance $\sigma^2_{m_{\mathrm{nt}}} = \langle m_{\mathrm{nt}} \rangle + \frac{\langle m_{\mathrm{nt}} \rangle^2}{n\lambda} = \langle m_{\mathrm{nt}} \rangle \left(1 + \frac{\langle m_{\mathrm{nt}} \rangle}{d} \right)$. At large $\langle m_{\mathrm{nt}} \rangle$, small n or low λ, all of which broaden the variance, deviation from the Poisson assumption that the variance is equal to the mean $\langle m_{\mathrm{nt}} \rangle$ can be profound. We call Equation 2.1 the *PCR distribution*.

Results of mutagenesis

To examine the mutational distribution generated by high-error-rate error-prone PCR, for which the Poisson- and PCR-based models make distinct predictions, two libraries of scFv antibody clones (libraries A and B) were generated using similar mutagenic conditions.

Both libraries were assayed for retention of wildtype-like binding to digoxigenin (retention of function) and 45+ naïve clones from each library were sequenced.

Poisson-distributed mutations will have equal mean and variance, while PCR-distributed mutations will always have a variance larger than the mean. Figure 2.1 shows the distribution of nucleotide mutations observed in library A (46 sequences) and library B (45 sequences); summary statistics are shown in Table 2.1, and mutational spectra are reported in Table 2.2.

While visual inspection of the mutation histograms overlayed with the theoretical distributions cannot distinguish between the two models, the relevant statistics are stark and favor the PCR distribution while rejecting the Poisson distribution. For library A, $\langle m_{\mathrm{nt}} \rangle$ = 15.8 and $\sigma^2_{m_{\mathrm{nt}}}$ = 26.3; for library B, $\langle m_{\mathrm{nt}} \rangle$ = 19.8 and $\sigma^2_{m_{\mathrm{nt}}}$ = 36.1 (Table 2.1). The probability of measuring variances at least this large given an underlying Poisson distribution with the observed mean is $P < 0.005$ for library A and $P < 0.001$ for library B; the joint probability of observing two libraries with variances this high is $P < 10^{-5}$. With a PCR efficiency of $\lambda = 0.6$ (18 doublings), the PCR distribution yields expected variances of 29.6 (library A) and 41.4 (library B), consistent with the observed values.

Using a likelihood ratio test on the mutational samples (see *Methods*), we reject the Poisson distribution in favor of the PCR distribution with two additional degrees of freedom (n and λ) for library A ($\chi^2 = 7.39$, $P < 0.025$) and for library B ($\chi^2 = 8.63$, $P < 0.025$). (Using two additional degrees of freedom is conservative, since n is fixed in each experiment.) Thus, the PCR distribution (Eq. 2.1) better describes the data than the previously assumed Poisson model.

Retention of protein function after mutation

What is the effect of the non-Poisson mutational distribution on the fraction of clones in a library that retains function? We assume the probability an individual protein will retain function after m_{aa} amino acid substitutions declines exponentially according to

$\Pr(f \mid m_{\text{aa}}) = \nu^{m_{\text{aa}}}$, where the neutrality ν can be interpreted as the average fraction of functional one-mutant neighbors on the protein-sequence-space network[34,41] (cf. Chapter 1). This assumption is consistent with experimental results obtained without using PCR[15] and with theoretical considerations[11]. This model assumes no average epistasis.

The probability a nucleotide mutation produces a nonsynonymous change is assumed to be binomial with parameter p_{ns}, corresponding to the assumption that mutations hit distinct codons. This assumption and the value $p_{\text{ns}} = 0.7$ appear realistic[16] (the precise parameter value will vary somewhat based on a gene's codon composition). In the following analysis, nonsynonymous changes include insertions, deletions, mutations to stop codons, and mutations that change the encoded amino acid: $p_{\text{ns}} = p_{\text{ins}} + p_{\text{del}} + p_{\text{stop}} + p_{\text{aa}}$. The first three types of changes are assumed to truncate and inactivate the encoded protein; we assume they constitute a fraction $p_{\text{tr}} = p_{\text{ins}} + p_{\text{del}} + p_{\text{stop}} \approx$ 0.05–0.07 of mutations (*e.g.*, see ref. 19, supporting information) and use the value $p_{\text{tr}} = 0.06$ for our calculations. The probability that a nonsynonymous mutation does not truncate the encoded protein (*i.e.* only changes the encoded amino acid) is $(1 - p_{\text{tr}} / p_{\text{ns}})$. The probability a sequence with m_{nt} nucleotide mutations retains function includes all these effects and is therefore

$$\begin{aligned}
\Pr(f \mid m_{\text{nt}}) &= \sum_{m_{\text{ns}}=0}^{m_{\text{nt}}} \Pr(m_{\text{ns}} \mid m_{\text{nt}}) \Pr(\text{non trunc.} \mid m_{\text{ns}}) \ \Pr(f \mid m_{\text{ns}} \text{ amino acid changes}) \\
&= \sum_{m_{\text{ns}}=0}^{m_{\text{nt}}} \binom{m_{\text{nt}}}{m_{\text{ns}}} p_{\text{ns}}^{m_{\text{ns}}} (1 - p_{\text{ns}})^{m_{\text{nt}} - m_{\text{ns}}} \times (1 - p_{\text{tr}} / p_{\text{ns}})^{m_{\text{ns}}} \times \nu^{m_{\text{ns}}} \qquad (2.2) \\
&= (1 - (1 - \nu(1 - p_{\text{tr}} / p_{\text{ns}}))p_{\text{ns}})^{m_{\text{nt}}}.
\end{aligned}$$

Under the assumption of Poisson-distributed mutations, Shafikhani *et al.*[17] showed that, if a fraction q_i of nucleotide mutations inactivate a protein, the fraction functional declines exponentially as $e^{-\langle m_{\text{nt}} \rangle q_i}$. Because $q_i = (1 - \nu(1 - p_{\text{tr}} / p_{\text{ns}}))p_{\text{ns}}$, we expect

$\Pr(f) = e^{-\langle m_{\mathrm{nt}} \rangle (1-\nu(1-p_{\mathrm{tr}}/p_{\mathrm{ns}}))p_{\mathrm{ns}}}$ in a Poisson-distributed library. This exponential decline became the experimental expectation for subsequent groups, leading to surprise when functional mutants were later found in great excess at high average mutation rates. By combining Equations 2.1 and 2.2 and assuming gene length $L \to \infty$—a mild assumption when $\langle m_{\mathrm{nt}} \rangle << L$—we find the probability a sequence from the library will retain function is

$$\Pr(f) = \sum_{m_{\mathrm{nt}}=0}^{\infty} \Pr(f \mid m_{\mathrm{nt}}) \Pr(m_{\mathrm{nt}}) = \left(\frac{1 + \lambda e^{-\frac{\langle m_{\mathrm{nt}} \rangle (1+\lambda)}{n\lambda} (1-\nu(1-p_{\mathrm{tr}}/p_{\mathrm{ns}}))p_{\mathrm{ns}}}}{1+\lambda} \right)^n . \qquad (2.3)$$

Equation 2.3 makes several predictions. In the limit of many thermal cycles *n*, all else equal, the original expectation $\Pr(f) = e^{-\langle m_{\mathrm{nt}} \rangle (1-\nu(1-p_{\mathrm{tr}}/p_{\mathrm{ns}}))p_{\mathrm{ns}}}$ (above) is recovered. If the number of thermal cycles *n* is proportional to $\langle m_{\mathrm{nt}} \rangle$, following the protocol of Shafikhani *et al.*, then $\Pr(f)$ should be a perfect exponential in $\langle m_{\mathrm{nt}} \rangle$, which is precisely what this group reports. However, if *n* is fixed as in other studies[16,21,33], then $\Pr(f)$ curves upward relative to an exponential decline as $\langle m_{\mathrm{nt}} \rangle$ increases. PCR efficiency λ decreases with increasing $\langle m_{\mathrm{nt}} \rangle$[42], which increases the expected curvature. In other words, there will be more functional sequences than predicted by the exponential decline.

Using the previously reported scFv antibody data[16] for low $\langle m_{\mathrm{nt}} \rangle$, where the Poisson assumption is not unreasonable, and the reported value $q_i = 0.6$, we can estimate $\nu \approx 0.2$ for the antibody binding task. For the subtilisin data[17], we similarly use the reported $q_i = 0.27$ to estimate $\nu \approx 0.65$. With these values for ν, Figure 2.2 compares the predictions of Equation 2.3 to the observed fractions of functional clones at various library mutation levels $\langle m_{\mathrm{nt}} \rangle$ reported by Daugherty *et al.*[16] and in the present work for the scFv antibody fragment (Fig. 2.2a) (see also Table 2.3) and Shafikhani *et al.*[17] for subtilisin (Fig.

2.2b). The agreement is quite good and demonstrates that the excess of functional clones can in fact be consistent with an underlying exponential relationship between number of amino acid substitutions and probability of retained wild-type function. To further test our analytical predictions, we simulated single-round error-prone PCR using template DNA strands encoding a folded "wildtype" lattice protein. The amplified DNA was translated into lattice proteins which were scored as functional if they retained the fold and thermostability of the wildtype. We observed excellent agreement with Equation 2.3 (see Supplemental Material for this chapter).

The reason for deviation from an exponential decline is hinted at in the limit of large average mutation rates, when the exponential part of Equation 2.3 vanishes and $\Pr(f)$ approaches a constant, $\Pr(f) \to (1+\lambda)^{-n}$. For a mutationally fragile protein such as the scFv antibody performing the digoxigenin binding task, this can occur at experimentally accessible mutation rates, as can be seen most clearly in the library originally reported[16] and revisited by Georgiou[32]. As the mutation rate increases, the antibody fragment becomes "quite insensitive to mutational load" and $\Pr(f)$ flattens out at a value of roughly 0.0018 [32]. Most interestingly, this limiting value is a function only of the PCR conditions, and does not depend on the protein at all.

What causes these counterintuitive results? Error-prone PCR at high frequency generates heavily mutated sequences by a process akin to Xeroxing copies of copies: low-fidelity copies give rise to even lower-fidelity copies, yet a copy, once produced, is not replaced, but remains in the final distribution of copies. During the polymerase chain reaction, the first generation of mutants, amplified directly from the wild-type template gene and carrying few mutations, persists in the mix and continues to reproduce copies with few additional mutations throughout subsequent cycles. The protein products of these less-mutated copies retain function at a greatly elevated rate compared to the average sequence, leading to upward bias in the functional fraction.

Why are improved mutants found more often in high-error-rate libraries?

If statistical effects of the mutagenesis protocol can explain the dramatic deviation from exponential in the fraction of functional sequences without recourse to epistasis, why are high-$\langle m_{\mathrm{nt}} \rangle$ libraries enriched in improved clones, despite a smaller number of clones retaining *any* function? To address this question, we now explore another consequence of PCR's broad mutational distribution.

The effective size of a library is not the number of mutants screened, the number usually reported, but rather the number of *unique* mutants screened. In a library of 10^6 transformants of the scFv antibody gene (726 bp, 242 aa) with an average of one mutation per sequence, most of the 2,178 possible 1-mutants will occur on the order of 100 times, reducing the effective library size by roughly two orders of magnitude. Most mutagenesis is concerned with protein sequences, where additional losses occur. Truncations due to frameshift mutations or mutations to stop codons eliminate a significant fraction of sequences. With one nucleotide mutation per codon, an average of 5.7 amino acid substitutions (out of a maximum of 19) are accessible due to the conservatism of the genetic code, for a total of 242 × 5.7 = 1,379 accessible amino acid sequences with one substitution. (We ignore the effects of synonymous mutations.) One million transformants thus yield just over one thousand unique protein sequences, about a 1,000-fold reduction in the effective library size.

We estimate the number of unique sequences in an error-prone PCR library in the following way. We derive the distribution of nonsynonymous substitutions $\Pr(m_{\mathrm{ns}})$ after error-prone PCR, estimate the number of non-truncated amino acid sequences $N_{m_{\mathrm{ns}}}$ with each m_{ns} in a library of a given size, compute the expected number of unique sequences $U_{m_{\mathrm{ns}}}$ at each m_{ns} by accounting for recurrence among the $N_{m_{\mathrm{ns}}}$ sequences, and then find the expected number of unique sequences U by summing the $U_{m_{\mathrm{ns}}}$.

With PCR conditions denoted as before and an average number of nucleotide mutations per sequence $\langle m_{\mathrm{nt}} \rangle$, what is the distribution of the number of nonsynonymous substitutions per sequence $\Pr(m_{\mathrm{ns}})$? We assume, as before, that each nucleotide mutation causes a nonsynonymous change with probability p_{ns}, so we obtain

$$\begin{aligned} \Pr(m_{\mathrm{ns}}) &= \sum_{m_{\mathrm{nt}}=m_{\mathrm{ns}}}^{L} \Pr(m_{\mathrm{nt}}) \binom{m_{\mathrm{nt}}}{m_{\mathrm{ns}}} p_{\mathrm{ns}}^{m_{\mathrm{ns}}} (1-p_{\mathrm{ns}})^{m_{\mathrm{nt}}-m_{\mathrm{ns}}} \\ &= (1+\lambda)^{-n} \sum_{k=0}^{n} \binom{n}{k} \lambda^{k} \frac{(ky)^{m_{\mathrm{ns}}} e^{-ky}}{m_{\mathrm{ns}}!} \end{aligned} \tag{2.4}$$

with $y = \dfrac{\langle m_{\mathrm{nt}} \rangle p_{\mathrm{ns}} (1+\lambda)}{n\lambda}$. That is, the distribution of nonsynonymous substitutions $\Pr(m_{\mathrm{ns}})$ is equivalent, in form, to the distribution of nucleotide mutations $\Pr(m_{\mathrm{nt}})$, but with an average of $\langle m_{\mathrm{ns}} \rangle = \langle m_{\mathrm{nt}} \rangle p_{\mathrm{ns}}$ substitutions. For simplicity, we will drop the subscript for nonsynonymous substitutions and use *m*.

Of the sequences with *m* nonsynonymous substitutions, some will also be truncated by frameshifts or stop codons. Because we treat all truncations as nonsynonymous changes, the fraction of non-truncated sequences with *m* substutions is $\Pr(\text{non-truncated}|m) = (1 - p_{\mathrm{tr}} / p_{\mathrm{ns}})^{m}$. Given an error-prone PCR library of *N* transformants, $N_m = N \Pr(m) \Pr(\text{non-truncated}|m)$ on average are non-truncated proteins with *m* amino acid substitutions.

Of these proteins with *m* substitutions, how many unique sequences exist? Only one unique sequence has *m* = 0. For any *m* there are on average $M_m = \binom{L/3}{m} 5.7^{m}$ total unique proteins with at most one mutation per codon, where *L* is the length of the gene in nucleotides.

Given N_m samples, how many of these M_m unique proteins can we expect to find? This is the classic "coupon collector problem"[43] and directly addresses the question of mutant recurrence, since any sample either yields a new, unique protein or one that has been sampled before. The expected number of unique sequences produced by equiprobably sampling M_m sequences N_m times is

$$U_m = M_m - M_m(1-1/M_m)^{N_m} \approx M_m(1-e^{-N_m/M_m}). \quad (2.5)$$

For example, to sample 99% of the M_m = 1,379 accessible 1-mutants of scFv requires 4.6-fold oversampling (N_m = 6,350 samples) on average. Taking 1,379 samples, $N_m = M_m$, on average yields only 872 unique proteins, or 63% of the total. In practice, for proteins of a few hundred amino acids and libraries of a few million transformants, recurrence need only be considered for small values of *m* ($m < 3$), because sequence space becomes large enough to make recurrence extremely unlikely at higher *m* values so that $U_m \approx N_m$. The total number of unique sequences in a library is simply the sum over all unique sequences with a specific number of substitutions:

$$U = \sum_{m=0}^{L/3} U_m \ . \quad (2.6)$$

Figure 2.3a shows the fraction of unique sequences *U/N* obtained from simulations (see Methods) in which the scFv gene was mutated according to PCR statistics with the observed frequencies (Table 2.2, with 3% frameshift rate) or unbiased frequencies (all mutations equally weighted, with 3% frameshift rate). The prediction from Equation 2.6 is also plotted and agrees well. Increasing the mutation rate increases the number of unique sequences because fewer are lost to recurrence. Note that, even at the highest mutation rates, the fraction of unique sequences does not approach 1.0, because sequences truncated by frameshifts and stop codons are not considered unique and accumulate at increasing levels as the mutation rate is increased.

Of greater interest is the expected number of unique sequences in the library that are expected to retain at least wildtype function, because these sequences are a superset of potentially improved sequences. We can estimate the number of unique, functional sequences as

$$U_f = \sum_{m=0}^{L/3} U_m \nu^m \,. \tag{2.7}$$

Figure 2.3b shows the fraction of unique, functional sequences U_f/N obtained from the same simulations as in Fig. 2.3a, with Eq. 2.7 plotted for comparison. Biases in mutation frequencies decrease the fraction of unique sequences, but preserve the overall form. Results using unbiased frequencies are predicted accurately by our theoretical treatment.

Clearly, low-error-rate libraries suffer from dramatic mutant recurrence, an effect avoided at high error rates. Improved proteins are found often in high-error-rate libraries *because these libraries contain more unique functional sequences*.

Optimal random mutagenesis

A typical and important goal in protein engineering is to improve an existing protein function, for example by increasing catalytic rate, thermostability, binding affinity, or specificity. While rational engineering has made significant strides, high-throughput screening of large mutant libraries for improved clones is both a dominant strategy to achieve this goal and an area of active research[32].

Given a choice of protein scaffold, a library of fixed size, and no reliable basis for rational engineering, a simple measure of library optimality is the number of unique functional sequences it contains. Figure 2.3b shows that, given this measure, an optimal mutation rate exists which balances diversity (uniqueness is lost if $\langle m_{\mathrm{nt}} \rangle$ is too low) with retained function (functional sequences are rare if $\langle m_{\mathrm{nt}} \rangle$ is too high). Mutational biases do not significantly affect the optimal mutation rate.

The optimum depends on the number of transformants sampled, the PCR protocol used, and the wildtype protein being mutated, among other parameters. Figure 2.4a compares predicted optimal mutation rates under identical PCR conditions for the scFv antibody ($\nu \approx 0.2$), depending on whether a thousand or a million clones are screened. The difference, 1.3 average nucleotide substitutions, corresponds to one amino acid substitution on average. Figure 2.4b compares predicted optimal mutation rates under identical conditions and with the same wildtype protein, but using 30 thermal cycles (as in the present work) in one case and 2 cycles (as in ref. [21]) in the other. A difference of one nucleotide mutation results. Optimal rates also depend on protein mutational tolerance as reflected by ν : the more tolerant the protein, the higher the optimal mutation rate (not shown).

Table 2.3 lists estimates for U_f given the scFv library experimental conditions reported here and previously[16]. Despite the over 200-fold lower observed percentage of functional transformants isolated from the highest-$\langle m_{\mathrm{nt}} \rangle$ library relative to the lowest, and the 14-fold fewer functional sequences observed, only 60% fewer unique functional sequences are expected in the highest-$\langle m_{\mathrm{nt}} \rangle$ library. Given the experimental parameters of the highest-$\langle m_{\mathrm{nt}} \rangle$ library and altering only the mutation rate, the rate $\langle m_{\mathrm{nt}} \rangle$ = 11.0 is predicted to produce more unique functional sequences (>10,000) than any of the reported libraries. The *optimal* mutation rate given the highest-$\langle m_{\mathrm{nt}} \rangle$ experimental parameters is predicted to be roughly $\langle m_{\mathrm{nt}} \rangle$ = 3.0, which is predicted to yield >34,000 unique, functional sequences. These results do not account for gains in probability of improvement treated in Chapter 1, but such gains are expected to be small relative to the cost of loss of folding and function.

Discussion

Laboratory evolution by random mutagenesis remains the most effective known strategy for improving enzyme properties given a choice of scaffold and no reliable basis for rational engineering. The possibility that distant regions of sequence space harbor excesses

of functional and, for at least some enzymatic tasks, improved proteins has been advanced several times, with significant experimental evidence to bolster the claims. We have shown that a more accurate model of error-prone PCR than previously used, due to Sun[38], is required to adequately describe the mutational distribution resulting from high-error-rate error-prone PCR. This model, in turn, provides straightforward explanations for the previously observed experimental findings: 1) the excess functional proteins observed at high $\langle m_{\mathrm{nt}} \rangle$ is predictable using our Equation 2.3, is due to low-mutation sequences generated early in the reaction, and is consistent with an exponential decrease in retention of function with amino acid substitution level; and 2) loss of functional sequences at high mutation rates can be balanced by diversity in the form of more unique sequences, improving sampling of sequence space and leading to a higher probability that improved mutants will be found if they exist. We have demonstrated the often-overlooked importance of accounting for recurrence of mutants when estimating how much of sequence space a library covers, extending previous work on modeling effects of mutational bias[44]. With our simple definition of library optimality as maximizing the number of unique, functional proteins, these two observations lead to an optimal mutation rate for error-prone PCR which can be estimated using our analytical results. However, optimal mutation rates are both protocol- and protein-dependent. Optimal rates derived for error-prone PCR using one set of conditions do not necessarily hold for another set (Fig. 2.4), and are highly unlikely to hold for saturation mutagenesis or site-directed mutagenesis, for which uniqueness is rarely a problem and the distribution of mutation levels in a typical library is tight and easily controllable.

We have explained several disparate mutagenesis results using only a single parameter unrelated to experimental protocols: ν, the average probability of retaining wildtype function after a random amino acid substitution[11]. It follows that these experiments can be used to measure ν using the analytical tools we have introduced here, with an important caveat. Because multiple mutations per codon, rarely found in error-prone PCR even at high mutation rates (though not always[45]), are necessary to experimentally measure ν, such experiments cannot directly measure this parameter but can provide a credible upper bound

due to the conservative nature of the genetic code. While ν relates simply to the "structural plasticity" $q_i = (1-\nu(1-p_{\mathrm{tr}}/p_{\mathrm{ns}}))p_{\mathrm{ns}}$ proposed by Shafikhani *et al.* [17], our results show that the emergence of a perfect exponential decline in their experiments likely depended both on a fundamental property of proteins and the particular experimental protocol employed. We also distinguish between genetic mutations which produce truncated protein products, essentially all of which lack function, and those which produce full-length proteins whose structural properties determine whether mutations are tolerated. We believe ν more accurately captures the idea of *structural* plasticity.

Because optimal mutation rates depend on ν, we can suggest measures which influence ν and which therefore may be used to manipulate the optimal mutation rate. All else being equal, proteins with higher thermodynamic stability (free energy of unfolding) have a higher ν [11] and tolerate more destabilizing substitutions, suggesting that more stable variants of a protein represent more promising departure points for mutagenesis. If longer proteins are more tolerant of substitutions, as seems plausible, then longer genes will tend to have higher optimal mutation rates. Codon usage may influence ν indirectly, through protein expression; in cases where high protein expression is required for the relevant function, replacement of rare codons with common synonyms may allow higher mutation rates. When a protein's crystal structure is available, ν can be estimated computationally[11]. We also note that the exponential decline in fraction functional holds when many mutations are introduced, as in the present work, but may not always hold for small numbers of mutations[11] (e.g., see Fig. 1.3 in the previous chapter).

A protein's intrinsic functional tolerance to substitutions is only one of many ways in which genetic mutations may affect the fraction of active clones in a library. Biologically relevant or screenable activity may depend on the action of many molecules in an organism, so mutations which hinder expression (e.g. through introduction of non-preferred codons, or in rarer cases by altering mRNA secondary structure) may decrease the fraction of clones scored as active. Disruption of signal sequences may result in improper targeting to cellular locations such as the periplasm or cell membrane. Mutations may destabilize the

protein, hindering its folding or exposing it to proteolysis or irreversible misfolding without actually destroying the function of the natively folded molecule. The dominant effect of most random mutagenesis is changes in the primary sequence of a target protein, most of which disrupt native function, and our simple treatment appears to work well under these circumstances.

Our results also illuminate potentially serious methodological flaws in previous studies. For example, the accuracy in measuring average library mutation rate by nucleotide sequencing depends on the variance of the mutational distribution, which at high mutation rates is far broader than that of the Poisson distribution previously assumed. The expected standard error of measurement on a library with $\langle m_{\mathrm{nt}} \rangle$ average mutations assessed by sequencing N_{seq} clones is $\sigma_m / \sqrt{N_{seq}} = \sqrt{\langle m_{\mathrm{nt}} \rangle \left(1 + \langle m_{\mathrm{nt}} \rangle / n\lambda\right) / N_{seq}}$. Zaccolo and Gherardi[21], for example, report four libraries averaging $\langle m_{\mathrm{nt}} \rangle$ = 8.2, 19.7, 21.3 and 27.2 mutations per coding region of a 1,088 base-pair gene constructed using 2, 5, 10 and 20 thermal cycles with $\langle m_{\mathrm{nt}} \rangle$ measured by sequencing at least 2,500 base pairs, effectively N_{seq} = 2.5. Even if the true value of $\langle m_{\mathrm{nt}} \rangle$ is as measured and perfect PCR efficiency assumed, these measurements have an expected 1σ standard error of 4.3, 6.5, 5.4 and 5.3 mutations per gene, respectively, calling into question the actual levels of hypermutagenesis achieved in these experiments.

The analysis presented here has important consequences for understanding the natural and directed evolution of proteins. Importantly, we have provided a thorough analysis of an apparent manifestation of mutational epistasis.

Two issues are often confused: whether mutations interact epistatically on average in *individual* folded sequences, and whether mutations interact epistatically on average in a library or *ensemble* that contains both folded and unfolded sequences. Ensemble epistasis is the only measure of interest in studies of the evolutionary persistence of sexual

recombination[36] and of primary interest in deciding which regions of sequence space should be targeted for efficient directed evolution.

If ensemble epistasis existed, as implied by earlier interpretations of the less-than-exponential decline in retention of function with mutational distance discussed in the present work, then individual epistasis would also be found on average. Importantly, the reverse is not true. Though folded or improved proteins may display cooperative effects (mutations which are better together than individually), many polypeptides in a random library may also carry mutations that are more deleterious together than apart. However, the latter are unlikely to be found by investigators, because such mutants are disproportionately likely to fail to fold, and little if any attention is given to the vast numbers of unfolded proteins in mutant libraries. Confusion arising from the asymmetry between types of epistasis—ensemble epistasis implies individual epistasis, but individual epistasis does not imply ensemble epistasis—may have inspired prior claims that high mutation rates can be used to access reservoirs of cooperative mutations while only a "small proportion" of clones will be lost to disruptive mutations[21].

As a result of our analysis, several data sets probing high mutation rates can now be seen, despite appearances to the contrary, to provide no evidence for ensemble epistasis—of particular biological interest given the recent discoveries of multiple native error-prone polymerases in bacteria and higher organisms[46]. Meanwhile, recent work providing a explanation for why the fraction of mutant proteins retaining function will decline exponentially[11] suggests that ensemble epistasis is unlikely. We cannot rule out the existence of epistasis; our analysis merely points out one way in which a mutation process can produce results which give the *appearance* of epistasis when there is none.

Exploration of distant regions of sequence space by random mutation alone appears highly inefficient, reinforcing the role of other search processes such as homologous recombination in creating sequence diversity[47,48], a subject treated in the next chapter. High-mutation-rate error-prone PCR, however, can be used to overcome the "uniqueness sink" that occurs at low mutation rates when using selection or high-throughput screening

to assay large numbers of clones. Finally, optimal mutation rates cannot be decoupled from the physical process of mutation, making them dependent on the particular organism or protocol under consideration. There can be no "optimal mutational load for protein engineering," as has previously been suggested[45], without specification of the engineering methodology.

Methods

Library construction, sequencing and functional assay

Two libraries, A and B, were constructed from error prone PCR reactions as described.[42] Identical mutagenesis conditions were used for both libraries but produced different mutation levels in each library. In particular, 2.50 mM $MgCl_2$, 0.5 mM $MnCl_2$, 0.35 mM dATP, 0.40 mM dCTP, 0.20 dGTP, and 1.35 mM dCTP were used along with Taq DNA Polymerase. The PCR reaction was continued for 30 cycles rather than 16 as in the reference. All other parameters, and subsequent ligation, transformation and FACS functional analysis procedures were performed as previously described.[16]

Statistical characterization of mutational distributions

To characterize the sequencing results and relate them to two theoretical distributions (the Poisson distribution, $\Pr(m;\langle m_{\mathrm{nt}}\rangle) = \frac{\langle m_{\mathrm{nt}}\rangle^m e^{-\langle m_{\mathrm{nt}}\rangle}}{m!}$, and the PCR distribution, Eq. 2.1), we used the likelihood ratio test, which compares the probabilities of observing a particular mutational sample under competing distributions. A mutational sample obtained by sequencing consists of N sequences $i = 1...N$ having m_i mutations. Given a theoretical mutational distribution $\Pr(m)$ which gives the probability of randomly choosing a sequence having m mutations, the likelihood of a sample is $L = \prod_{i=1}^{N} \Pr(m_i)$. The likelihood ratio test evaluates the statistic $LR = 2[\ln(L_{Poisson} / L_{PCR})]$ which has approximately a χ^2 distribution[49]. Significance values (P values) can be computed from the likelihood ratio statistic, the χ^2 distribution, and a number of degrees of freedom, which in this case is 2, corresponding to the two additional parameters in the PCR distribution, the number of thermal cycles n, and the replication efficiency λ.

Simulation

To simulate the error-prone PCR process, two approaches were taken. First, we exhaustively simulated the error-prone PCR process using genes encoding simplified model proteins (compact lattice model, 25 residues consisting of any of 20 amino acids) which were then folded and assayed for retention of wildtype structure. Details and results of this simulation are presented below in Supplemental Material.

We found that a vastly simpler simulation produced nearly identical results (see Figure 2.S2 in Supplemental Material, below) and we used this simulation to generate Fig. 2.3. The scFv gene was mutated $N = 50{,}000$ times at each $\langle m_{\mathrm{nt}} \rangle$ according to the observed mutation frequencies (Table 2.2, Library A) and the PCR distribution, Eq. 2.1, with parameters as indicated in the figure legend. Each mutated gene was translated into a protein sequence according to the universal genetic code. Truncated proteins, either from stop codons or frameshifts, were discarded. Whether a full-length sequence was functional or not was estimated by counting the number of amino acid substitutions relative to wildtype and designating the protein functional with probability $\Pr(f \mid m_{\mathrm{aa}}) = \nu^{m_{\mathrm{aa}}}$. All full-length protein sequences were inserted in a set which retained only unique sequences. Numbers and fractions of unique, functional and jointly unique and functional sequences were then tabulated.

Supplemental Material

To test our analytical results, we carried out simulations of error-prone PCR. Because we wished to accurately model the effect of mutations on proteins, yet do so in a tractable way, we used lattice proteins for our *in silico* work (cf. Chapter 1).

Supplemental methods

To model mutagenesis results, we used the lattice protein model described in Chapter 1. Each simulation run begins with an arbitrarily chosen target conformation and a minimum stability (free energy of unfolding) of 5.0 kcal/mol. Proteins are defined as functional if they fold to this conformation with free energy at or above this value.

Our analytical work describes the effects of mutation on genes of several hundred base pairs, the biologically relevant regime, but not on the 75bp genes encoding these lattice proteins due to the breakdown of the Poisson assumption. Thus, we extended the protein model in a simple way: genes are 750 base pairs long and encode ten independently folding 25-residue "domains," initially identical in the wildtype, which must each fold to a target structure with the required free energy in order for the overall protein to retain fold.

Error-prone PCR was simulated as follows. Beginning with a set of 2000 identical template genes in the mix, sequences are duplicated with a probability equal to the PCR efficiency λ and a per-site mutation rate $x = \frac{\langle m_{\mathrm{nt}} \rangle (1+\lambda)}{n\lambda}$. This process is repeated for n cycles. A sample of N = 20,000 sequences is then taken of the resulting mix, translated according to the universal genetic code, and assayed for function according to the folding assay described above. The mutation rate was determined by sequencing these N sequences; excellent agreement was found between the predicted rate $\langle m_{\mathrm{nt}} \rangle$ and the actual rate, as well as with the standard error and that expected (see main text, Discussion; data not shown). The probability of truncation, p_{tr}, was set to 0.045; in this simulation, frameshifts do not occur, though stop codons do arise at a low frequency. The fraction of

nonsynonymous mutations p_{ns} was also determined from these sequences, and generally was in the range 0.7 to 0.8. The observed average value for each gene was used when evaluating Equation 2.3. The number of unique genes, unique proteins, functional proteins, and unique and functional proteins was tabulated for each sample.

Because PCR is an exponential-growth process, simulation is notoriously difficult. We implemented an efficient simulation allowing us to obtain libraries at high mutation rates of $>10^6$ sequences on a modest desktop PC with a 2.8GHz Intel Pentium IV processor and 500MB of RAM. Performance is significantly better at low mutation rates due to the nature of the optimization (storing only mutational changes rather than entire sequences).

Supplemental results

Using the protein model described in *Methods*, we found four genes encoding proteins with a wide range of *v* values, from 0.13 to 0.8. We amplified these genes by simulated error-prone PCR per above. We also performed a mutagenesis run in which all mutations are introduced at once, the conditions under which a Poisson distribution of mutations should arise corresponding to the assumption made originally by Shafikhani *et al.*[17] discussed in the main text. Figure 2.S1 shows the results of these simulations. The observed close agreement is typical and repeatable.

Figure 2.S2 shows the results of lattice-protein simulations compared to the simplified simulations described in the main text and with our theoretical results. The agreement is excellent and shows that essentially identical results can be obtained without a full simulation of the PCR process, as stated in the main text.

Table 2.1: scFv antibody mutational results and corresponding predictions for PCR and Poisson-distributed mutations.

Library	# seq'd	$\langle m_{\mathrm{nt}} \rangle$	$\sigma^2_{m_{\mathrm{nt}}}$ ($P(\sigma^2_{m_{\mathrm{nt}}})$ if Poisson)	PCR $\sigma^2_{m_{\mathrm{nt}}}$ [a]	Poisson $\sigma^2_{m_{\mathrm{nt}}}$
A	46	15.8 ± 0.8	26.3 ($P < 0.005$)	29.6	15.8
B	45	19.8 ± 0.9	36.1 ($P < 0.001$)	41.4	19.8

[a] Assumed efficiency $\lambda = 0.6$ (18 DNA doublings).

Table 2.2: Mutational spectra for libraries.[a]

	Library A (33,396 bp sequenced)		Library B (32,670 bp sequenced)	
Type	Number	Fraction	Number	Fraction
A→T, T→A	172	0.24	106	0.12
A→C, T→G	7	0.01	7	0.01
A→G, T→C	336	0.46	202	0.23
G→A, C→T	188	0.26	529	0.60
G→C, C→G	11	0.02	28	0.03
G→T, C→A	11	0.02	17	0.02
Total mutations	725		889	
Nonsynonymous	501	0.69	634	0.71
Termination	19	0.03	44	0.05

[a] In each gene, 726 nucleotides were sequenced. Sequences containing frameshift events were discarded, but occurred at a very low level (<5%).

Table 2.3: Comparison of retention of wildtype digoxigenin binding for scFv antibody libraries with analytical predictions.

$\langle m_{\mathrm{nt}} \rangle$	N	Observed functional	Observed % funct.	Predicted % funct. [a] (Poisson)	Predicted % funct. [a] (Eq. 2.3)	Predicted U_f
1.7	3×10^5	1.4×10^5	40.0	36.1	38.8	2,473
3.8	1×10^6	6.7×10^4	6.7	10.2	12.9	8,811
15.8 [b]	–	–	0.12	0.0076	0.095	–
19.8 [b]	–	–	0.041	0.00069	0.029	–
22.5	6×10^6	1×10^4	0.17	0.00014	0.15	1,463

[a] Assumed scFv $\nu = 0.2$ (see text), efficiency $\lambda = 0.6$ for all but highest-$\langle m_{\mathrm{nt}} \rangle$ library, for which we estimate efficiency $\lambda = 0.3$.

[b] Only fractions functional were recorded for these libraries.

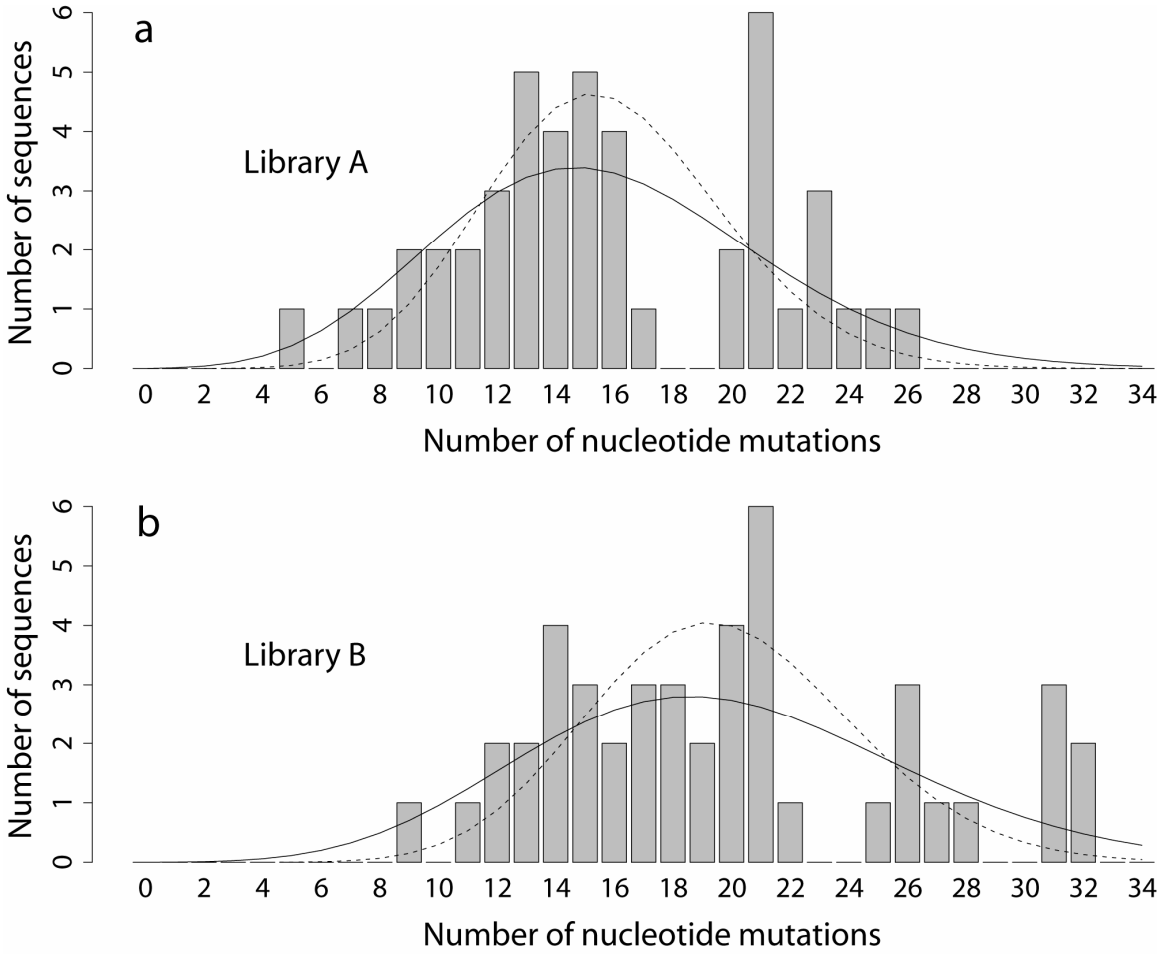


Figure 2.1: Mutational distributions for two high-error-rate scFv antibody libraries compared with Poisson and PCR distributions. **a**, Library A, 46 sequences. **b**, Library B, 45 sequences. The corresponding PCR distributions with the same means (see Table 2.1) (solid line, n = 30 cycles and efficiency λ = 0.6) and Poisson distribution (dashed line) are shown for comparison. For these histograms, the Poisson distribution may be rejected in favor of the PCR distribution (see text).

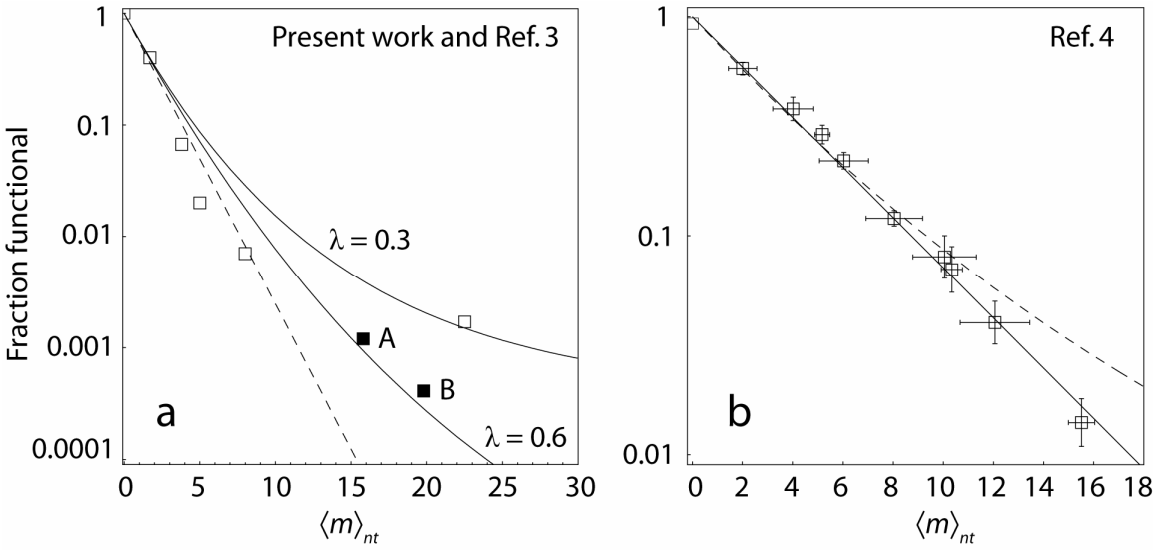


Figure 2.2: Equation 2.3 explains previously reported experimental results. **a**, Comparison to scFv antibody data from Daugherty *et al.*[16] (□) and present work (■); for conditions, see the footnotes to Table 2.3. Dashed line is the original fit reported[16], $e^{-\langle m_{\mathrm{nt}}\rangle q_i}$ with $q_i = 0.6$. Solid lines show Eq. 2.3 for the two libraries reported here (bottom) and for the highest-$\langle m_{\mathrm{nt}}\rangle$ library conditions reported previously[16] (top). Changes in line curvature are due entirely to changes in PCR efficiency λ. **b**, Comparison to high-$\langle m_{\mathrm{nt}}\rangle$ subtilisin data from Shafikhani *et al.*[17] (open squares with standard error bars), which were produced by a multi-round protocol. Conditions (all per-round): $d = n\lambda = 10$ DNA doublings, n=13 thermal cycles, $\langle m_{\mathrm{nt}}\rangle$ = 2.01 or 5.17 nucleotide mutations per gene. The fractions functional predicted by Eq. 2.3 for a multi-round protocol (solid line) and a single-round protocol (dotted line) show that the theory properly predicts the observed exponential decline in fraction functional.

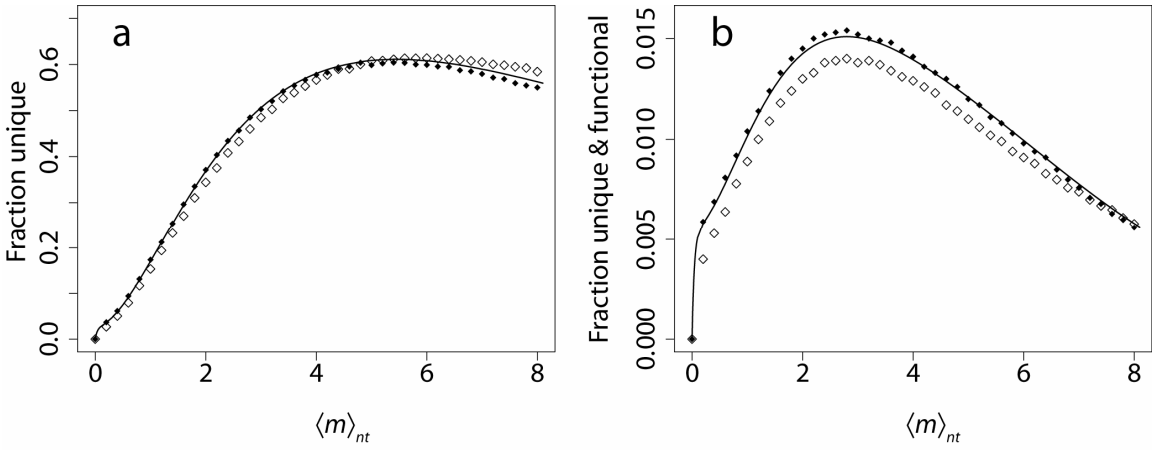


Figure 2.3: Error-prone PCR error rates strongly influence the fraction of unique and functional sequences. **a**, Fraction of unique sequences in a simulated library of $N = 50{,}000$ scFv clones ($\nu = 0.2$) using the observed mutational spectrum (◇) or an unbiased spectrum (◆). Line is Eq. 2.6 (divided by N) evaluated with n = 30 thermal cycles, efficiency $\lambda = 0.6$, p_{ns} = 0.76 and p_{tr} = 0.07. **b**, Fraction of unique and functional sequences in the same library. Line is Eq. 2.7 (divided by N) evaluated using the same parameters. An optimal mutation rate exists which balances uniqueness with retention of function. Mutational biases lower the fraction of unique and functional sequences, but do not significantly alter the optimal mutation rate.

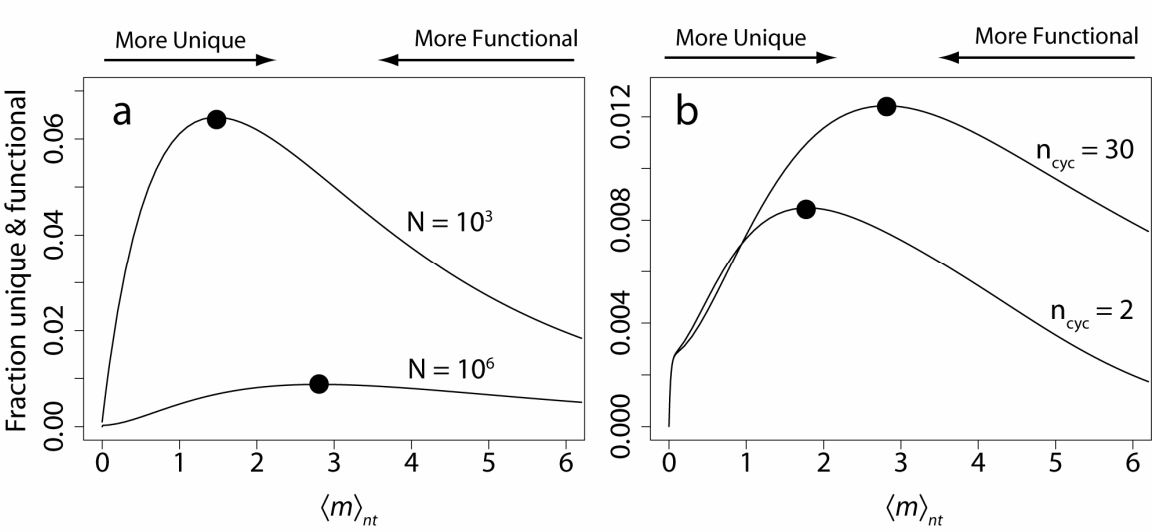


Figure 2.4: The requirement for uniqueness reduces effective library size and leads to library- and protocol-dependent optimal library mutation rates. **a**, Optimal mutation rate (●) depends on library size. Predicted fractions of unique functional sequences given by Eq. 2.7 for the same protocol (n = 30 thermal cycles with efficiency $\lambda = 0.6$, $p_{\mathrm{ns}} = 0.76$ and $p_{\mathrm{tr}} = 0.07$) and protein (scFv-like, $\nu = 0.2$) are shown at each average mutation rate $\langle m_{\mathrm{nt}} \rangle$ if 10^3 transformants (top, $\langle m_{\mathrm{nt}} \rangle_{\mathrm{opt}} = 1.5$) or 10^6 transformants (bottom, $\langle m_{\mathrm{nt}} \rangle_{\mathrm{opt}} = 2.8$) are screened. **b**, Optimal mutation rate (●) depends on PCR protocol. Predicted fractions of unique functional sequences given by Eq. 7 are shown for the same protein (scFv-like, $\nu = 0.2$) and library size (10^5 transformants) using n = 30 thermal cycles (top, $\langle m_{\mathrm{nt}} \rangle_{\mathrm{opt}} = 2.8$) or n = 2 thermal cycles (bottom, $\langle m_{\mathrm{nt}} \rangle_{\mathrm{opt}} = 1.8$). In all cases, recurrence leads to profound loss of uniqueness at low $\langle m_{\mathrm{nt}} \rangle$, and the optimal $\langle m_{\mathrm{nt}} \rangle$ balances uniqueness and retention of function.

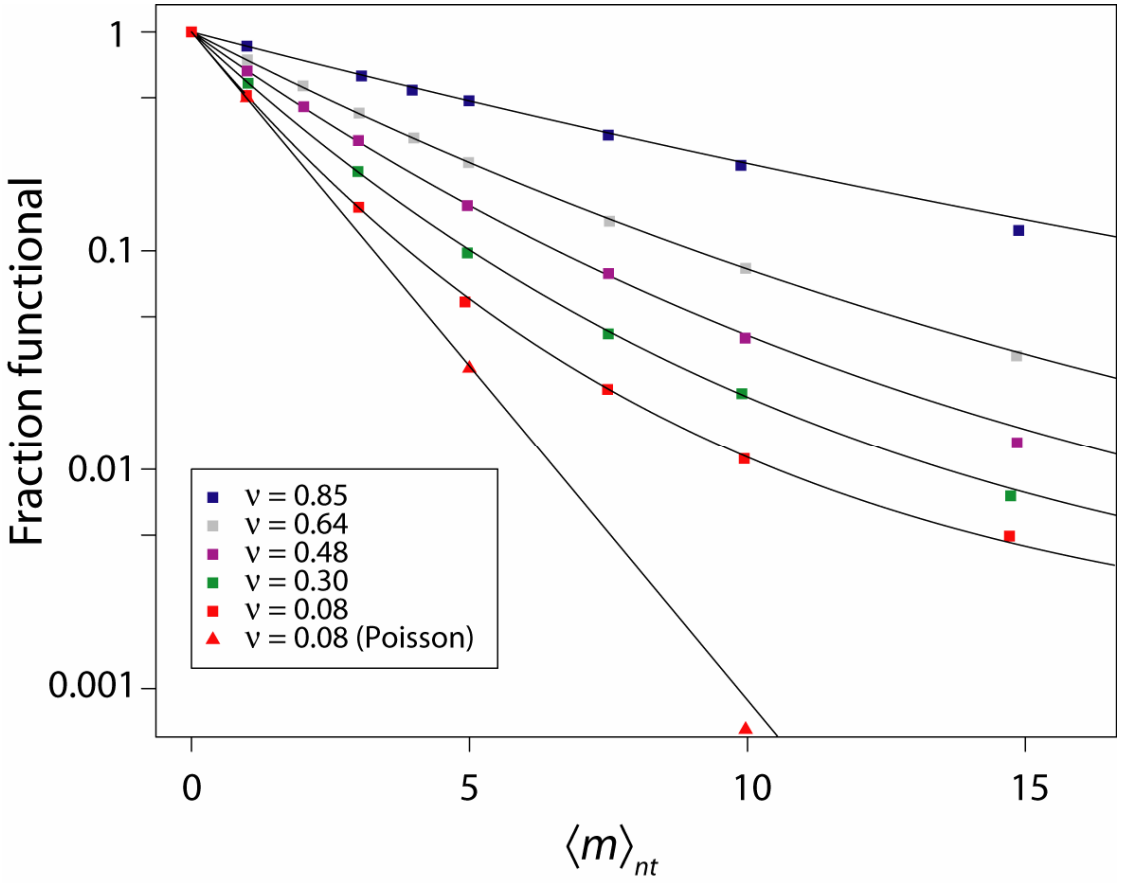


Figure 2.S1: Comparison of Equation 2.3 to simulation results. Five proteins having domain structures with differing v were assayed after error-prone PCR at n = 16 cycles, efficiency λ = 0.5. The lowest-v structure was also subjected to single-round mutagenesis with Poisson-distributed mutations. The fraction of functional proteins is plotted (points) along with predictions using Equation 2.3 and, for the Poisson-distributed library, the equation $\Pr(f) = e^{-\langle m_{\mathrm{nt}} \rangle (1-\nu(1-p_{\mathrm{tr}}/p_{\mathrm{ns}}))p_{\mathrm{ns}}}$ (see main text).

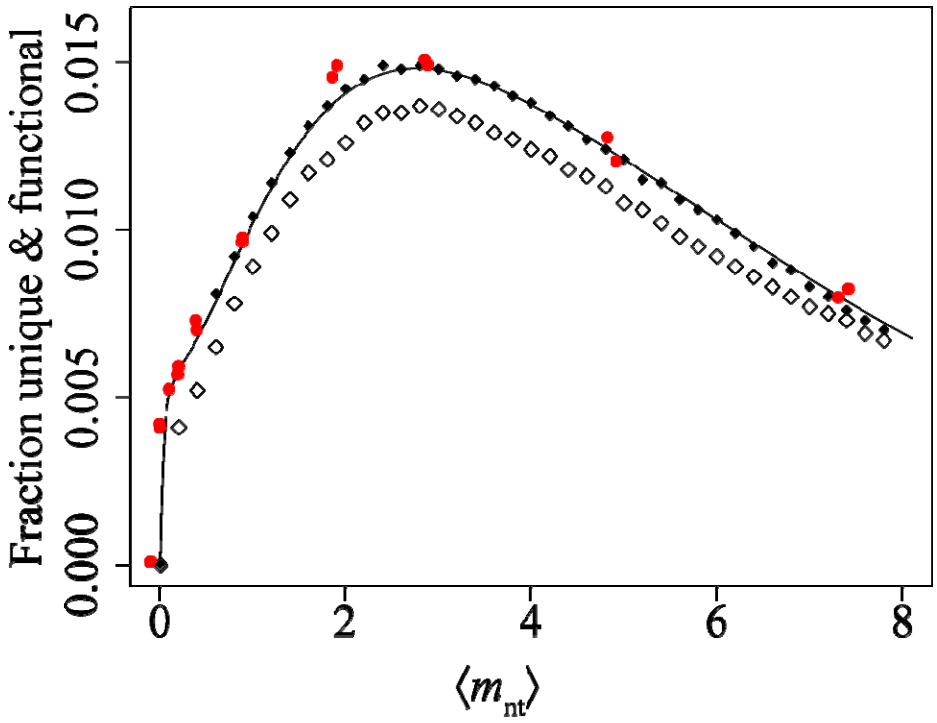


Figure 2.S2: Simulation results match predictions for number of unique, functional proteins. Simulation results (red points) are compared to predictions (filled circles, no mutation biases; open circles, with biases as in main text). Error-prone PCR conditions: $n = 14$ cycles, efficiency $\lambda = 0.71$, $\nu = 0.2$, $p_{ns} = 0.76$ and $p_{tr} = 0.07$.

C h a p t e r 3

ON THE CONSERVATIVE NATURE OF INTRAGENIC RECOMBINATION[2]

All armies prefer high ground to low.

Sun Tzu

Summary

Intragenic recombination rapidly creates protein sequence diversity compared with random mutation, but little is known about the relative effects of recombination and mutation on protein function. Here, we compare recombination of the distantly related β-lactamases PSE-4 and TEM-1 to mutation of PSE-4. We show that among β-lactamase variants containing the same number of amino acid substitutions, variants created by recombination retain function with a significantly higher probability than those generated by random mutagenesis. We present a simple model which accurately captures the differing effects of mutation and recombination, in both real and simulated proteins, with only four parameters: the amino acid sequence distance between parents, the number of substitutions, and the average probabilities that random substitutions and substitutions generated by recombination will preserve function. Our results expose a fundamental functional enrichment in regions of protein sequence space accessible via recombination, and provide a framework for evaluating whether the relative rates of mutation and recombination observed in nature reflect the underlying imbalance in their effects on protein function.

Introduction

A major goal in understanding the molecular basis of evolution is to quantitatively describe how effectively mutation and recombination traverse protein sequence space to create new functional proteins[7]. Protein sequence distance, measured by counting the number of amino acid substitutions *m* separating two sequences, is a fundamental metric of evolutionary rate and relationships[50], diversity of structure and function[51], and a key variable in protein engineering[16,28], while mutation and recombination are its biochemical cause. Genetic studies[52,53] and algorithmic inferences from biological sequence data[54-56] have revealed that recombination can occur preferentially within coding sequences, at times with a higher frequency than mutation[57,58]. When sequences encoding divergent but related proteins recombine, large distances may be traveled in sequence space relative to random mutation[59-62] without disturbing function and/or structure. However, a complete understanding of the underlying relative efficiency of mutation and recombination in accessing nearby or distant regions of sequence space cannot be gained from genomic sequences, because these become available only after natural selection has acted.

Laboratory[1] and *in silico*[63] evolution experiments, in contrast, can be used to quantitatively differentiate the effects of mutation or recombination on protein structure and function. By screening or selecting libraries of proteins for retention of parental function and determining the sequences of both functional and nonfunctional proteins, one can determine how the retention of function or structure depends on *m*, the sequence distance. This type of analysis has been used to determine the effects of random mutation on the function of subtilisin[17], DNA polymerase and HIV reverse transcriptase[15], an antibody fragment[16], lysozyme[33], a DNA repair enzyme[64], and a β-lactamase and lattice proteins[11], revealed a consistent exponential decline in the proportion of variants retaining function with increasing distance from wildtype. As discussed in the previous two chapters, this

---

[2] Adapted from *Proceedings of the National Academy of Sciences, USA,* **102**(15), D. Allan Drummond, Jonathan J. Silberg, Michelle M. Meyer, Claus O. Wilke, and Frances H. Arnold, "On the conservative nature of intragenic recombination," p. 5380-5385, copyright (2005).

exponential dependence occurs because a random amino acid substitution preserves protein function with some average probability[17,64], referred to as mutational tolerance or *neutrality* ν. Multiple independent substitutions lead to an exponential decline in the probability of retaining protein function, *i.e.*, $\Pr(f \mid m) = \nu^m$.

Effects of recombination on protein function have not been similarly characterized, although anecdotal and qualitative studies abound. Structurally-related polypeptides have been swapped among homologous single-domain proteins to create functional chimeras with substitution levels much higher than in random mutation experiments[5,47,65-69]. The more conservative nature of recombination is likely to arise at least in part because the individual amino acid substitutions created by recombination, having proved compatible with a similar structure, are less likely to be incompatible in the homolog structure than substitutions created by mutation. Whether differences in residue-structure compatibility alone are sufficient to explain the conservative nature of recombination relative to mutation has remained unclear.

Here, we attempt to answer the following related questions: what is the relationship between retention of function and the number of amino acid substitutions *m* introduced by homologous recombination; how does this relationship compare to random mutation; and how is it influenced by neutrality and homolog sequence identity? To set the stage, we derive a simple model comparing retention of protein function after *m* amino acid substitutions generated by either random mutation or recombination. We show that under the simple assumption that protein function depends on compatibility of residues with the protein backbone and with each other, recombination benefits from fundamental advantages over mutation. To test our model's predictions, we measured the effects of random mutation and recombination on the function of β-lactamases. Detailed tests using *in silico* evolution of lattice proteins confirm the generality of the model predictions and demonstrate that recombinational tolerance depends on the neutrality of the parental structures.

Results

A model comparing mutation and recombination. We want to answer the question, "What is the probability a protein will retain fold after *m* amino acid substitutions, generated either by mutation or by recombination?" We analyze retention of fold rather than attempt to explicitly model function for two reasons. First, the definition of function depends strongly on the particular assay or selective environment used (*e.g.*, the precise concentration of antibiotic), while fold does not, and thus is more tractable. Second, function requires that the protein be folded, so results for conservation of fold create an upper bound on functional conservation.

For mutation, probability of retaining fold declines exponentially with the number of substitutions,

$$\Pr(f \mid m)_{\text{mutation}} = \nu^m, \tag{3.1}$$

where ν is the neutrality and the exponential relationship results from the approximate independence of random substitutions[11].

For recombination, the exponential relationship cannot hold. Consider recombination of two protein sequences which fold into the same structure. A chimera is formed, in essence, by taking *m* residues from one protein and placing them at the corresponding positions in the other protein. Two proteins differing at *D* amino acids can produce chimeras with at most *D*−1 substitutions, and $\Pr(f \mid 0) = \Pr(f \mid D) = 1$. Moreover, for parental proteins with similar properties, the probability of retaining fold will be *symmetrical*, $\Pr(f \mid m) = \Pr(f \mid D - m)$, since the choice of which homolog is at $m = 0$ and $m = D$ is arbitrary.

Let us assume that chimeras fold if all their residues are compatible with the native structure (*e.g.*, have a hydrophobicity consistent with the structure's hydropathic pattern) and compatible with all other residues (*e.g.*, not in steric clash). As in Chapter 1, we suppose that each incompatibility on average reduces the stability, in some cases enough to disrupt folding. For proteins which share a structure, all residues must be compatible with that structure, so only pairwise interactions enter into $\Pr(f \mid m)$.

Each of the m substitutions in a chimera come from one parental protein and are therefore compatible with each other. The only possible incompatibilities result from interactions between the m substitutions and the $(D-m)$ remaining residues which are not identical between the homologs (all but D residues are the same). The number of possible pairwise incompatibilities resulting from these interactions is $m(D-m)$.

If each interaction has an independent probability q of not disrupting folding, then a chimera with m substitutions (and thus $m(D-m)$ possible incompatibilities) will have a probability $\Pr(f \mid m) = q^{m(D-m)}$ of retaining fold. (If only local interactions in the folded structure can create incompatibilities, larger proteins will have a higher apparent q than smaller proteins; we do not attempt to distinguish these effects in this analysis.) Notably, this simple expression satisfies the symmetry and end-point considerations introduced above. Because we wish to directly compare mutation and recombination, we write the probability as

$$\Pr(f \mid m)_{\text{recombination}} = \rho^{\frac{m(D-m)}{D-1}} \tag{3.2}$$

so that $\Pr(f \mid 1)_{\text{recombination}} = \rho$ and $\Pr(f \mid 1)_{\text{mutation}} = \nu$.

We have now formulated $P_f(m)$ in terms of two unknown parameters, which allow us to compare mutation and recombination in a simple way: ν (the *neutrality*) represents the average probability that a random residue substitution will preserve fold, and ρ (the *recombinational tolerance*) measures the average probability that a substitution coming

from a homolog via recombination will preserve fold. $\nu < \rho$ indicates that substitutions created by recombination are more conservative than random substitutions, and $\nu > \rho$ the opposite. See Box 3.1 for a more rigorous derivation of Eqs. (3.1) and (3.2).

Lactamase evolution supports model predictions. Our model predicts that substitutions created by recombination should have distinct effects on protein function from those created randomly. The logarithm of the fraction of functional chimeras is predicted to have a parabolic shape with the vertex center at the maximal substitution level. We also expect that $\nu < \rho$ when recombining structurally related proteins, since recombination incorporates substitutions that have been pre-selected for compatibility with the structures being recombined.

To investigate these qualitative predictions, we took advantage of a previously reported library of lactamase chimeras in which the related PSE-4 and TEM-1 β-lactamases (43% amino acid identity and 0.98 Å backbone RMS deviation) were divided into 14 fragments, which were then synthesized as oligonucleotides and combinatorially ligated, to produce a maximum of 2^{14} (= 16,384) unique chimeric sequences[5]. This construction protocol allowed us precise knowledge of the maximum number of chimeric sequences at each substitution level m, where m = 0 for PSE-4 and m = 150 for TEM-1. The structural conservation of these chimeras was assessed by selecting the library for variants that enabled *E. coli* growth on an ampicillin concentration that is approximately two orders of magnitude lower than the minimal inhibitory concentrations for cells expressing TEM-1 and PSE-4 [5].

A total of 30 functional chimeras were identified upon sequencing the lactamase genes obtained from the functional selection. Of the 136 substitution levels sampled by the library, 27 contained at least one functional chimera. We calculated the fraction of chimeras that retained β-lactamase activity over all substitution levels by partitioning all possible chimeras in our library into ten bins and dividing the number of functional chimeras by the number of total chimeras in each bin. These data represent a lower bound

on the fraction of functional chimeras. Figure 3.1a shows that the minimum fraction of chimeras retaining function does not decrease exponentially, as it does for random amino acid substitution[15-17,33]. Rather, the logarithm of the minimum fraction of functional chimeras has a parabolic shape with its vertex found near the substitution level farthest from both parents (m = 75), as predicted by Eq. 3.2. A fit of Eq. 3.2 to the recombination data yielded $\rho = 0.79 \pm 0.02$ ($P << 0.0001$) (asymptotic standard error), indicating that at least 79% of the substitutions generated by recombination preserve function. We believe that this minimum ρ is not larger than what would be found on average in other PSE-4 and TEM-1 chimeric libraries.

To determine the effects of mutation on lactamase function, we mutated the PSE-4 gene using error-prone PCR and analyzed the fractions functional in the resulting libraries. (Mutagenesis was performed by Dr. Joff Silberg.) Four libraries were created, and 9-10 unselected variants from each library were sequenced and used to calculate the average nucleotide mutation level in each library, $\langle m_{\mathrm{nt}} \rangle$. Figure 3.1b shows that, as observed with other proteins [15-17,33], increasing mutations cause an exponential decrease in PSE-4 function. A fit of Eq. 3.1 to our experimental data revealed that the neutrality for random single amino acid substitutions is $\nu = 0.54 \pm 0.03$ ($P < 0.0001$) (asymptotic standard error). Thus, the individual amino acid substitutions created by error-prone PCR are tolerated 54% of the time, versus at least 79% for substitutions created by recombination. We plotted ν^m for random mutation along with the recombination data in Figure 3.1a to compare the effects on function of multiple substitutions created by mutation and recombination. Extrapolation of random mutation effects to the highest substitution level accessible by recombination (m = 75) suggests that recombination is at least *sixteen orders of magnitude* more effective than random mutation at creating the most highly substituted chimeras.

The effects of parental sequence and structure on ρ. We would like to know to what extent the value of ρ depends on the sequence identity of parents recombined and on parental structure. To approach this question, we evaluated the effects of mutation and

recombination on lattice proteins, simple simulated polymers that that have been used to rapidly assess the general features of protein sequence space[41,63,70].

In initial experiments, libraries of chimeras were created by recombining structurally-related proteins exhibiting a range of sequence identities (20% to 80%), and the fraction of all functional mutants (see Methods) that differed by one to five substitutions from the parents was calculated. Figure 3.2a and 3.2b show the results from recombination experiments using distinct protein structures exhibiting high and low neutrality, respectively. For both structures, the results mirrored those from the lactamase experiments. Recombination produced proteins with parent-like structures at a rate that is orders of magnitude higher than random substitution of the same structure. The logarithm of the fraction of folded chimeras at each m is parabolic as predicted by our model, regardless of parental sequence identity or the neutrality of the proteins recombined.

Comparable mutation and recombination data were collected for ten distinct structures. The four trials for each structure correspond to the results from mutating and recombining four pairs of structural homologs with sequence identity of 20%, 40%, 60% and 80%. Figure 3.3 shows that recombination was more conservative than random substitution ($\nu < \rho$) for all structures examined, and that ρ correlates strongly with ν, as anticipated (see Box 3.1). We fit our model to the 50-run average for each trial independently and found that fits to each data set were highly significant for both ρ and ν ($P < 0.0001$ in all cases). While ν varied several-fold, ρ varied less (Figure 3.3). The standard deviation in both ν and ρ across differing choices of homolog sequence identity was less than 15% of the average values, suggesting that neutrality and recombinational tolerance are determined primarily by protein structure. The values of ρ anti-correlated with sequence distance D, with high significance but low variation (mean $R^2 = 0.75$, mean slope -0.002).

Discussion

We have directly demonstrated that recombination of structurally related proteins preserves function with a higher probability than does random mutation. A simple model captures the interplay of amino acid substitutions (m), parental sequence divergence (D), neutrality (ν) and recombinational tolerance (ρ) to a high degree of accuracy: retention of function declines exponentially as ν^m after random mutation, but curves symmetrically and log-parabolically as $\rho^{\frac{m(D-m)}{D-1}}$ after recombination. For a pair of β-lactamases, we find that recombination is significantly more conservative than mutation ($\nu < \rho$), as predicted. Notably, this is true even though mutations were generated by error-prone PCR, which creates less-deleterious changes than truly random substitution would, due to the conservative nature of the genetic code.

Computational work using lattice proteins both reinforces our experimental findings and allows us to explore consequences of the model that point out potentially general phenomena and suggest future experiments. For these simulated proteins, we find that mutationally tolerant proteins are likely to be recombinationally tolerant as well (Figure 3.3). The neutrality ν reflects the connectivity of function or fold networks in sequence space and has been studied as a key determinant of mutational tolerance in proteins[11,41] and RNA sequences[18,34]; our results demonstrate its importance for recombination through the correlation of recombinational tolerance ρ with neutrality. We find that the proportion of functional sequences after homologous recombination is a simple function of sequence identity and the recombinational tolerance ρ for homologs sharing 80% to as little as 20% of their primary sequence, in support of the idea that, at least for these simulated proteins, recombinational tolerance is a property of the structure.

The negative correlation between recombinational tolerance and parental sequence divergence may be explained by considering the line of descent. As two proteins diverge from a common ancestor, they accumulate substitutions at different sites. Substitutions along these lines of descent, not the total number of substitutions separating the homologs,

define the potential pairwise incompatibilities considered in our model. Thus, our model under-counts substitutions and incompatibilities for highly diverged homologs, decreasing the estimate of recombinational tolerance relative to less-diverged homologs.

Specific physical observations motivate our model. Our assumptions that protein folding can be modeled by considering single (residue-backbone) and pairwise (residue-residue) interactions and that residue-backbone incompatibility is more deleterious than residue-residue incompatibility are inspired in part by a plausible source of such interactions and incompatibilities: the hydrophobic and mixing energies[71] contributing to the free energy of folding. The hydrophobic force—a residue-backbone contribution—is a dominant force in protein folding[71]. Our finding that retention of function after homologous recombination can be modeled by consideration of pairwise interactions alone is consistent with the findings that proteins sharing more than 40% sequence identity are likely to have a shared structure[72], and that model proteins undergoing homologous recombination are overwhelmingly likely to retain the parental structure[73], thereby conserving pairwise spatial relationships.

Our finding that $\nu < \rho$ is consistent with the idea that substitutions generated by recombination have been pre-tested for structural compatibility[47]. The preservation of hydrophobic-polar (HP) patterning via recombination of similarly patterned sequences (TEM-1 and PSE-4 have 76% HP identity) is one likely source of this pre-testing[73]. Conserved residue charge and side-chain volume may also improve the odds that recombination preserves fold and/or function[66].

The *qualitative* difference between the effects of substitutions generated by random mutation and homologous recombination also has an intuitive basis: while random substitutions move variant proteins away from all functional sequences on average, substitutions from homologs always move chimeras toward at least one functional sequence. Figure 3.4 illustrates this fundamental difference schematically by compressing sequence space into a landscape with the average probability of retaining parental function represented by height. While random mutants fall down exponentially sloped hills,

chimeras traverse a ridge connecting the two parental sequences. Pure mutants and chimeras occupy the axes, and mutated chimeras fill the landscape. Under the assumption that the two parents and their chimeras have the same structure, mutation of these chimeras must produce the same exponential slope on average as the schematic suggests.

Various methods have been described that attempt to anticipate the effects of recombination on protein structure and function using sequence and structural information. Among sequence-based measures, number of crossovers[47] and crossover position[66] have been shown to affect the likelihood that recombination will preserve protein function. Our results suggest that, on average, the number of substitutions which result from a set of crossovers is the more important underlying variable. The choice of a particular structure-based measure used to anticipate chimera folding—the number of broken residue-residue contacts (SCHEMA disruption)[4,5,67]—is supported by the present work because these residue-residue interactions are predicted to be the dominant contributors to retention of chimera fold. For mutation, residue-backbone interactions dominate, and our work suggests that strategies to reduce these conflicts (*e.g.*, by preserving side-chain volume and avoiding prolines) should play a correspondingly larger role.

Our simple analytical model integrates the effects of a variety of other design parameters of interest in protein engineering (mutational tolerance, substitution level, and parental sequence divergence), providing a basis for optimizing the design of a recombination library and some general rules for obtaining libraries with a higher fraction of folded sequences[5]. When sequence diversity (folded sequences with high values of m) is a goal, choosing parents with the minimum divergence necessary to achieve that goal will maximize the yield of functional proteins, all else being equal. We recently showed that mutational tolerance depends on thermodynamic stability[11], suggesting that another way to increase the efficiency of recombination for a particular structure is to choose parents with high stability. Many important questions, *e.g.*, regarding recombination effectiveness at or between domain boundaries[65], must go beyond our average metric, but our findings create a null-model baseline for evaluating recombination strategies. Our model is limited to studying retention of function or fold using homologs of similar structure. Furthermore, we

have neglected the effects of mutations on expression, *e.g.*, through changes in mRNA half life or secondary structure, because TEM-1 and PSE-4 are low-expression proteins for which effects on expression are unlikely to be significant relative to the inactivating effects of amino acid substitutions. The effect of mutations on expression determinants remains an important open question.

One question raised by our observations is whether relative rates of intragenic mutation and recombination reflect the underlying imbalance in their effects on protein function. This can be partly answered. In both natural and laboratory evolution, recombination allows creation of broad sequence diversity with relatively low cost in loss of function compared to mutation. Pathogens under immune surveillance wage combinatorial warfare with their hosts, recombining homologous surface proteins to create folded proteins with diverse epitopes to escape immune responses[48,59]. In the laboratory, gene shuffling[47] and site-directed recombination[67] have proven useful in evolving new enzyme functions by generating diversity while preserving overall fold. By contrast, random mutation allows access to only narrow regions of sequence space because of its tendency to induce misfolding, though it can be used to search exhaustively for local optima inaccessible by recombination. Our results may explain why recombination is so strongly favored when diversity is the goal: intragenic recombination efficiently creates protein sequence diversity while conserving structure via preservation of interactions[65], symmetry, and conservatively chosen substitutions. Conservation of fold allows exploration of function.

Methods

Materials

E. coli XL1-Blue was from Stratagene (La Jolla, CA). Enzymes for DNA manipulations were obtained from New England Biolabs (Beverly, MA) or Roche Biochemicals (Indianapolis, IN). Synthetic oligonucleotides were obtained from Invitrogen (Carlsbad, CA). DNA purification kits were from Zymo Research (Orange, CA) and Qiagen (Valencia, CA), and other reagents were from Sigma (St. Louis, MO).

Functional conservation and recombination

In a previous study, we recombined PSE-4 and TEM-1 to create a well-defined library of chimeras[5], and selected for those that allowed *E. coli* XL1-Blue to grow on 20 μg/mL ampicillin. Approximately 100 colonies were observed, and sequencing fifty of these clones identified 23 unique functional chimeras. Sequencing of the remaining clones revealed an additional seven sequences for a total of 30 unique functional chimeras (see Table 3.S1). While no point mutations were found in the newly characterized chimeras, one of those previously identified as functional has two adjacent amino acid substitutions[5]. Sequencing of unselected chimeras showed that nine of 13 (69%) contained frameshifts introduced during oligonucleotide synthesis. To calculate the fraction of functional chimeras at each amino acid substitution level *m*, we divided the number of functional chimeras by the number of possible chimeras at each *m*. At many substitution levels, no functional chimeras were found despite large sample sizes. To determine the average effects of recombining PSE-4 and TEM-1 over all possible substitution levels, we partitioned all chimeras into bins of substitution levels containing at least one functional chimera. The number of unique synthesized chimeras in each bin sets an upper bound on the denominator of the fraction of functional chimeras; due to the possibilities that frameshifts inactivated some chimeras and that certain fragments were over-represented due to biases in library construction [5], it is unlikely this upper bound was reached. The calculated fraction of functional chimeras therefore represents a lower bound on the true fraction functional at each *m*.

Creation and functional analysis of random mutants

PCR under mutagenic conditions was used to create libraries of PSE-4 variants with a range of amino acid substitutions. An initial library was created by amplifying 1 ng of the PSE-4 gene (100 µL, total volume) in the presence of 0.5 mM $MnCl_2$, 0.2 mM dATP and dGTP, 1.0 mM dCTP and dTTP, 7 mM $MgCl_2$, 50 pmol of each primer (with restriction sites for cloning), and 5 U AmpliTaq polymerase. The temperature cycling scheme was 95°C for 5 min. followed by 13 cycles of 95°C for 30s, 50°C for 30s, and 72°C for 30s. PCR products (~0.9 kb) were purified using a 1% agarose gel and a Zymoclean gel purification kit. Libraries with increasing levels of mutation were generated by sequentially mutating 1 ng of product from each previous reaction. Each round of PCR resulted in ~0.5 µg of a 0.9 kb amplified fragment, corresponding to nine doublings. This procedure is expected to produce an exponential decline in the fraction of functional variants at increasing library mutation levels, simplifying analysis[74].

The gene products from each library were digested with *Hind*III and *Sac*I, purified using a Zymo DNA Clean and Concentrator Kit, and ligated into pMon-1A2 as in a previous study[5]. *E. coli* XL1-Blue were transformed with plasmids containing each library as recommended by the manufacturer and plated on three or more non-selective (10 µg/mL kanamycin) and selective (20 µg/mL ampicillin and 10 µg/mL kanamycin) plates. The fraction of functional variants in each library $\Pr(f \mid \langle m_{nt} \rangle)$ was determined by dividing the average number of colonies on selective medium by the average number on non-selective medium; all fractions reported are ± standard error (S.E.). The fraction of functional clones in the control populations created by cloning the PSE-4 gene into pMon-1A2 was 1.05±0.06.

To determine the average mutation level $\langle m_{\mathrm{nt}} \rangle$ for each library, 6,000 to 8,000 base pairs of unselected clones were sequenced. Error-prone PCR by the multi-round method used here produces a known distribution of nucleotide mutations in the resulting gene library and is expected to produce an exponential decline in the fraction functional with increasing average library nucleotide mutation level $\langle m_{\mathrm{nt}} \rangle$. To calculate ν, we must first take into

account the fraction p_{ns} of nonsynonymous nucleotide mutations, the probability of truncated/frameshifted and therefore inactive gene products p_{tr} due to deletions and stop codons, and the physical process of DNA amplification by error-prone PCR with n_{cyc} thermal cycles per round and PCR efficiency λ[74]. The resulting experimentally observed fractions functional can be fitted with a model incorporating all these factors to obtain a value for ν, given by Equation 2.3 in Chapter 2.

Lattice protein simulations

We used the lattice protein model described in Chapter 1. Each simulation run began with an arbitrarily chosen wildtype conformation and a minimum stability of 5.0 kcal/mol. An initial DNA sequence, 75 nucleotides long and encoding a functional lattice protein, was found by an adaptive walk, equilibrated for one million generations, and used to seed two populations of 500 DNA sequences. In each generation, sequences coding for functional lattice proteins were randomly chosen to reproduce with a nucleotide mutation rate of 0.0002/site until the new population contained 500 sequences. Evolution continued until the two populations had diverged by *D* amino acid substitutions. From these populations, two homologous DNA sequences were chosen, and the encoded lattice proteins designated the parental homologs. The DNA sequences were no longer considered. Site-directed amino-acid recombination between these parental homologs was carried out at seven randomly chosen protein crossover points (equivalent to gene-level recombination constrained to codon boundaries) to make 512 chimeras. The number of chimeras retaining function that differed from a given parent at *m* residues was tabulated. Random amino acid substitutions were made to each parental sequence; all 475 1-mutants and 10,000 each of 2-mutants, 3-mutants, and so on were generated, evaluated for function, and tabulated. The fraction functional at each level of substitution is the number of functional lattice proteins divided by the number generated. This process was repeated 50 times with the same initial DNA sequence to obtain means and variances.

Error analysis and fitting procedures are described in Supplemental Material below.

Box 3.1: A model comparing mutation and recombination.

Here, we more rigorously derive Equations 3.1 and 3.2 from the main text, which quantify the probability with which mutants or chimeras with *m* substitutions retain function. Consider recombining two homologous parental proteins having *L* amino acid residues differing at *D* sites and a conserved structure (fold). We make three simplifying assumptions: 1) the fraction of recombined proteins that retain function is an unbiased subset of those retaining fold; 2) the probability of retaining fold is determined by the independent probabilities that each residue is compatible with the parental structure and with all other residues; and 3) residues found in parental sequences are compatible with the structure and each other, while all other amino acids have an unknown average probability of incompatibility.

Under these assumptions, the probability that a protein containing residues $r_1 ... r_L$ retains the parental fold can be written as

$$\Pr(f \mid r) = \prod_{i}^{L} \Pr(r_i \text{ compatible}) \prod_{j<k}^{L} \Pr(r_j, r_k \text{ compatible}).$$

While this probability cannot be practically computed for a particular protein due to the intricate details of the molecular interactions determining compatibility, we may estimate it *on average* over a large number of mutants or chimeras by examining the quantity $\Pr(f \mid m) = \langle \Pr(f \mid r) \rangle$, the average fraction of proteins with *m* substitutions that retain fold. Assumption 2 asserts independence, so

$$\Pr(f \mid m) = \langle \Pr(f \mid r) \rangle = \prod_{i}^{L} \langle \Pr(r_i \text{ compatible}) \rangle \prod_{j<k}^{L} \langle \Pr(r_j, r_k \text{ compatible}) \rangle,$$

and according to Assumption 3, these average probabilities can be written in terms of an average residue-residue incompatibility p_{rr} and a residue-backbone incompatibility p_{rb},

$$\left\langle \Pr(r_i \text{ compatible}) \right\rangle = \begin{cases} 1 & \text{if } r_i \text{ is in a parental structure,} \\ p_{rb} < 1 & \text{otherwise;} \end{cases}$$

$$\left\langle \Pr(r_j, r_k \text{ compatible}) \right\rangle = \begin{cases} 1 & \text{if } r_j \text{ and } r_k \text{ are in a parental structure,} \\ p_{rr} < 1 & \text{otherwise.} \end{cases}$$

Our final assumption therefore reduces determination of the probability of retaining fold to counting the number of possible residue-backbone and residue-residue incompatibilities resulting from m substitutions. In the case of random mutation, m substitutions create m possible residue-backbone incompatibilities and $m(L-(m+1)/2)$ residue-residue incompatibilities. Recombination, by contrast, does not create any residue-backbone incompatibilities, because residues from both parents have proven compatible with the conserved structure, but alters a possible $m(D-m)$ residue-residue compatibilities. As a result, we have

$$\Pr(f \mid m)_{\text{mutation}} = p_{rb}^{m} p_{rr}^{m(L-(m+1)/2)} \approx (p_{rb} p_{rr}^{L})^{m} \equiv \nu^{m} \qquad (3.S1)$$

$$\Pr(f \mid m)_{\text{recombination}} = p_{rr}^{m(D-m)} \equiv \rho^{\frac{m(D-m)}{D-1}}. \qquad (3.S2)$$

The definitions introduce the parameters ν and ρ to enable a direct comparison: the fraction of functional variants with a single substitution (m = 1) is ν for mutation and ρ for recombination. The approximation in Eq. 3.S1 follows if $m << L$, which is generally true for random mutagenesis, and if p_{rr} is, on average, less than p_{rb}. We have now formulated $\Pr(f \mid m)$ in terms of two unknown parameters, which allow us to compare mutation and recombination in a simple way: ν (the *neutrality*) represents the average probability that a random residue substitution will preserve fold, and ρ (the *recombinational tolerance*) measures the average probability that a substitution coming from a homolog via recombination will preserve fold. $\nu < \rho$ indicates that substitutions created by recombination are more conservative than random substitutions, and $\nu > \rho$ the opposite. In

all cases, we expect $\nu < \rho$ because, as the intermediate expressions in Eqs. 3.S1 and 3.S2 show, $\Pr(f \mid m)_{\text{recombination}}$ is strictly greater than $\Pr(f \mid m)_{\text{mutation}}$. Moreover, Eqs. 3.S1 and 3.S2 indicate that ν and ρ should correlate through their mutual dependence on p_{rr}. As would be expected in this model, $\Pr(f \mid m)_{\text{recombination}}$ is symmetric, such that it makes no difference which parent *m* is measured from.

Supplemental Material

Error analysis and fitting procedure

Best-fit parameters and fit statistics were obtained using Mathematica's NonlinearRegress function with data weighted by inverse standard error on the dependent variable. Lactamase mutation data were fit to Equation 2.3 and recombination data to Equation 3.2. For lactamase mutation data, standard error on the fraction functional was calculated using results from replicates, and standard error on the assessment of library average nucleotide mutation level $\langle m_{\mathrm{nt}} \rangle$ was calculated as described in Chapter 2. Standard errors for the lactamase recombination data were approximated under the assumption that each bin's fraction functional was generated by a binomial process with proportion equal to the minimum fraction functional. Lattice protein mutation data were fit to Equation 3.1 and recombination data to Equation 3.2. We examined four values of D for each of ten lattice protein structures, and fits were performed independently on each of the four resulting 100-run sets of data. Standard errors were calculated over each 100-run set.

Identified functional chimeras of TEM-1 and PSE-4

Table 3.S1 lists the modular composition of functional chimeras isolated from the recombination library discussed in the main text. The polypeptide modules inherited from either PSE-4 (P) or TEM-1 (T) correspond to TEM-1 residues 1-39 (A), 40-57 (B), 58-67 (C), 68-84 (D), 85-102 (E), 103-115 (F), 116-131 (G), 132-146 (H), 147-163 (I), 164-204 (J), 205-222 (K), 223-249 (L), 250-264 (M), 265-286 (N) and structurally related residues in PSE-4 identified using a structure-based alignment with Swiss-PDB Viewer[75]. Substitution level (m) is the minimum number of mutations required to convert a chimera into PSE-4, excluding residues comprising the periplasmic secretory signal sequences.

Calculation of neutrality ν from error-prone PCR library data

The fraction of functional clones in a mutant library generated by error-prone PCR can be modeled using experimental parameters and knowledge of protein neutrality[74]. Multi-

round error-prone PCR (see *Methods* and Chapter 2) ensures that $\langle m_{nt} \rangle$ is proportional to n_{cyc}, which in turn means that $\Pr(f \mid \langle m_{nt} \rangle)$ (Equation 2.3) will decline exponentially with a slope related to ν, consistent with our data. In general, the observed $\Pr(f \mid \langle m_{nt} \rangle)$ slope will be significantly higher than ν^m or even predictions which assume a Poisson distribution of mutations in the library, because error-prone PCR generates a mutation distribution of particularly high variance as described in Chapter 2. The excess of sequences with fewer than average mutations inflate the fraction functional relative to the Poisson-based (smaller variance) expectation.

We calculated p_{ns} and p_{tr} from the sequencing data shown in Table 3.S2. p_{ns} is the fraction of all mutations excluding deletions that were nonsynonymous = 0.677; p_{tr} is the fraction of all mutations that produced a deletion or a stop codon = 0.059. Our error-prone PCR protocol used 13 thermal cycles per round (n_{cyc} = number of rounds × 13), produced DNA 9 doublings per round for an efficiency λ = 9/13 = 0.69, and yielded the observed fractions functional at four values of $\langle m_{nt} \rangle$ shown in Table 3.S2.

To obtain a best-fit value for ν in a simple way, we made an auxiliary assumption that the number of thermal cycles n_{cyc} was proportional to the observed library average nucleotide mutation level $\langle m_{nt} \rangle$, $n_{cyc} = 13 \langle m_{nt} \rangle / 8.37$, where 8.37 is the average number of nucleotide mutations introduced per round. Substituting this expression for n_{cyc} into Eq. 2.3 allowed us to express $\Pr(f \mid \langle m_{nt} \rangle)$ as a function only of $\langle m_{nt} \rangle$ and ν (the remaining values are constants). Using Mathematica's NonlinearRegress function on the five pairs of data for $\Pr(f \mid \langle m_{nt} \rangle)$ (Table 3.S2 and ($\langle m_{nt} \rangle$=0, $\Pr(f \mid \langle m_{nt} \rangle)$=1.05 ± 0.06) reported in the main text) with values weighted by the inverse standard error on $\Pr(f \mid \langle m_{nt} \rangle)$ for each point, we obtained a best-fit value of ν = 0.54 ± 0.03 ($P < 0.0001$) (error is asymptotic standard error). To check that this result did not depend strongly on our auxiliary assumption, we then evaluated Eq. 2.3 for $\Pr(f \mid \langle m_{nt} \rangle)$ using the actual number of thermal cycles at each

round. The resulting data shown in Table 3.S2 does not differ meaningfully from the predicted exponential line, and falls within a standard error of all but one datum.

Table 3.S1. Functional PSE-4/TEM-1 chimeras.

Chimera	A	B	C	D	E	F	G	H	I	J	K	L	M	N	*m*
1	P	P	P	P	P	P	P	P	T	P	P	P	P	P	7
2	P	P	P	P	P	T	P	P	P	P	P	P	P	P	7
3	P	P	P	P	P	P	P	T	P	P	P	P	P	P	7
4	P	P	P	P	P	P	P	P	P	P	P	T	P	P	11
5	P	P	P	P	P	T	P	P	T	P	P	P	P	P	14
6	P	P	P	P	T	P	P	P	P	P	P	P	P	P	14
7	P	P	P	P	P	T	P	P	P	P	T	P	P	P	16
8	P	P	P	P	T	T	P	P	P	P	P	P	P	P	21
9	P	P	P	P	T	P	P	P	T	P	P	P	P	P	21
10	P	P	P	P	P	P	P	P	P	T	P	P	P	P	22
11	P	P	P	P	P	T	P	P	T	P	T	P	P	P	23
12	P	P	P	P	T	T	P	P	T	P	P	P	P	P	28
13	P	P	P	P	T	T	P	P	P	P	T	P	P	P	30
14	P	P	T	P	T	T	P	P	T	P	P	P	P	P	35
15	P	P	P	P	P	T	P	P	T	P	P	T	T	P	36
16	P	P	T	P	P	T	T	T	P	P	P	T	P	P	40
17	P	P	P	P	T	T	P	P	P	P	P	T	T	P	43
18	P	P	P	T	T	T	T	T	T	P	P	P	P	P	53
19	P	P	P	T	T	T	T	T	P	P	P	T	P	P	58
20	P	T	P	P	T	T	T	T	P	P	P	T	P	P	60
21	P	P	P	T	T	T	T	P	P	T	P	T	T	P	67
22	P	P	P	T	T	T	T	T	P	T	P	T	P	P	71
23	P	P	P	T	P	T	T	T	T	T	P	T	P	P	73
24	P	T	T	T	T	P	T	P	P	T	P	T	T	P	94
25	P	P	P	T	T	T	T	T	T	T	P	T	T	P	96
26	T	T	P	T	T	T	T	T	T	P	P	T	T	T	111
27	T	T	P	T	T	T	T	T	P	T	P	T	T	T	126
28	T	T	P	T	T	T	T	T	T	T	P	T	T	T	133
29	T	T	T	T	P	T	T	T	T	T	T	T	T	T	135
30	T	T	T	T	T	T	T	T	P	T	T	T	T	T	142

Table 3.S2. Characteristics of PSE-4 mutant libraries.

	Library A	Library B	Library C	Library D
nucleotides sequenced [a]	7879	6824	6344	7656
synonymous substitutions	18	38	52	84
nonsynonymous subst.	41	57	114	191
nucleotide deletions	3	4	6	5
nonsynonymous subst. producing stop codons	2	4	4	8
library average nucleotide subst./gene ($\langle m_{nt} \rangle$) [b]	7.20 ± 1.23	13.27 ± 1.76	24.81 ± 2.62	33.46 ± 2.78
fraction of clones surviving selection ($\Pr(f \mid \langle m_{nt} \rangle)$)	0.13 ±0.015	0.0142 ±0.0032	0.00158 ±0.0004	0.00007 ±0.00007
Eq. 1 with $\nu = 0.54$ and auxiliary assumption	0.112	0.0170	0.00063	0.000049
Eq. 1 with $\nu = 0.54$, no auxiliary assumption	0.118	0.0196	0.00064	0.000049

[a] Nine to ten clones were partially sequenced from each library.

[b] Library average nucleotide mutations per gene $\langle m_{nt} \rangle$ equals the sum of synonymous mutations, nonsynonymous mutations, and deletions divided by the number of gene equivalents sequenced (base pairs sequenced / 915). Errors are expected standard errors following[74].

Table 3.S3: Values of neutrality ν and recombinational tolerance ρ for lattice protein structures.

Structure # [a]	$\langle \nu \rangle \pm \sigma_\nu$	$\langle \rho \rangle \pm \sigma_\rho$
415 [b]	0.104 ± 0.013	0.699 ± 0.053
414 [c]	0.128 ± 0.002	0.749 ± 0.056
820	0.196 ± 0.019	0.754 ± 0.057
873	0.275 ± 0.020	0.830 ± 0.038
19	0.280 ± 0.030	0.805 ± 0.048
350	0.314 ± 0.016	0.858 ± 0.032
55	0.380 ± 0.025	0.850 ± 0.027
200	0.385 ± 0.016	0.849 ± 0.015
300	0.426 ± 0.004	0.882 ± 0.028
1080	0.480 ± 0.012	0.891 ± 0.022

[a] See Table 3.S4 for pictures of each structure.

[b] Only two values of D (5 and 10) were evaluated.

[c] Only three values of D (5, 10 and 15) were evaluated.

Table 3.S4. Lattice protein structures used in this study.

ID	Structure	ID	Structure
19	01--02--03--04--05 | 20--21--22--23 06 | | | 19 14--13 24 07 | | | | | 18 15 12 25 08 | | | | 17--16 11--10--09	414	15--16 01--02--03 | | | 14 17 22--23 04 | | | | | 13 18 21 24 05 | | | | | 12 19--20 25 06 | | 11--10--09--08--07
55	01--02--03--04--05 | 14--13--12--11 06 | | | 15--16--17 10 07 | | | 20--19--18 09--08 | 21--22--23--24--25	415	15--16 01--02--03 | | | 14 17 20--21 04 | | | | | 13 18--19 22 05 | | | 12 25--24--23 06 | | 11--10--09--08--07
200	01--02--03 16--17 | | | 08--07 04 15 18 | | | | | 09 06--05 14 19 | | | 10--11--12--13 20 | 25--24--23--22--21	820	07--06--05--04--03 | | 08 21--20 01--02 | | | 09 22 19--18--17 | | | 10 23--24--25 16 | | 11--12--13--14--15
300	01--02 21--22 25 | | | | 04--03 20 23--24 | | 05 18--19 14--13 | | | | 06 17--16--15 12 | | 07--08--09--10--11	873	25 06--05--04--03 | | | 24 07--08 01--02 | | 23--22 09 12--13 | | | | 20--21 10--11 14 | | 19--18--17--16--15
350	01--02 11--12--13 | | | 04--03 10--09 14 | | | 05--06--07--08 15 | 22--21--20--19 16 | | | 23--24--25 18--17	1080	19--20 23--24--25 | | | 18 21--22 03--04 | | | 17--16 01--02 05 | | 14--15 10--09 06 | | | | 13--12--11 08--07

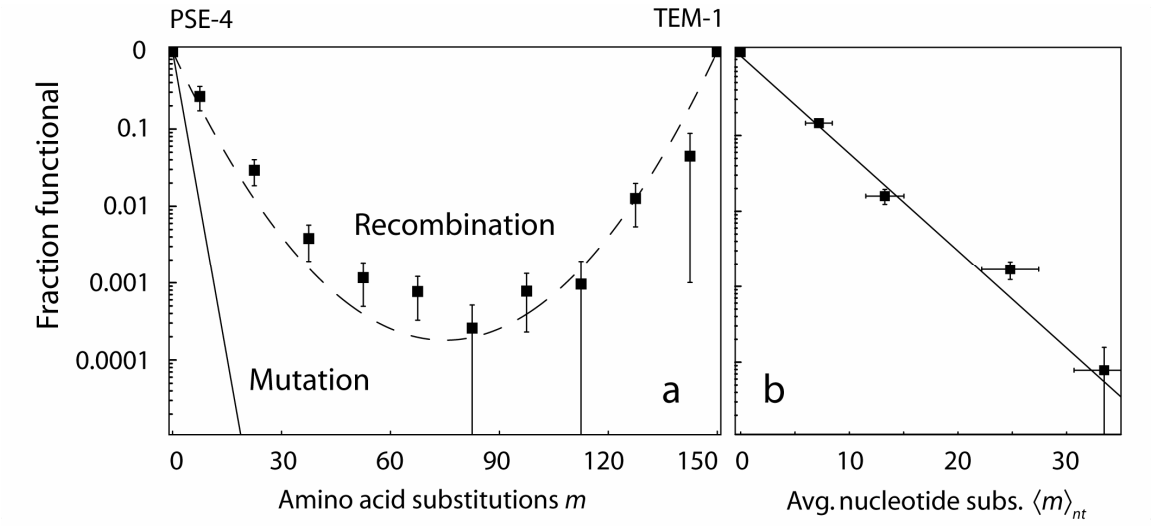


Figure 3.1: Effects of recombination and mutation on lactamase function. **a**, Recombination results in a higher fraction of functional lactamase variants than mutation. The (minimum) fractions of functional chimeras (■) in each bin of substitution levels m are shown relative to PSE-4 ($m = 0$) and TEM-1 ($m = 150$) (see *Methods*). Eq. 3.2 using the best-fit value $\rho = 0.79 \pm 0.02$ (dashed line) agrees well with these data. Mutation produces a lower fraction of functional variants (Eq. 2.3 with a best-fit value of ν, solid line; see caption for **b** and Supplemental Material for Chapter 3) than recombination at all values of m. **b**, Error-prone PCR mutagenesis of PSE-4 results in exponentially declining retention of lactamase function with increasing substitutions. The fractions of functional PSE-4 random mutants in each of four libraries and a no-mutation control (■) are plotted against each library's average nucleotide mutation level $\langle m_{\mathrm{nt}} \rangle$ ± standard error. The exponential best-fit of the random mutation data to Eq. 2.3 yields $\nu = 0.54 \pm 0.03$ (solid line).

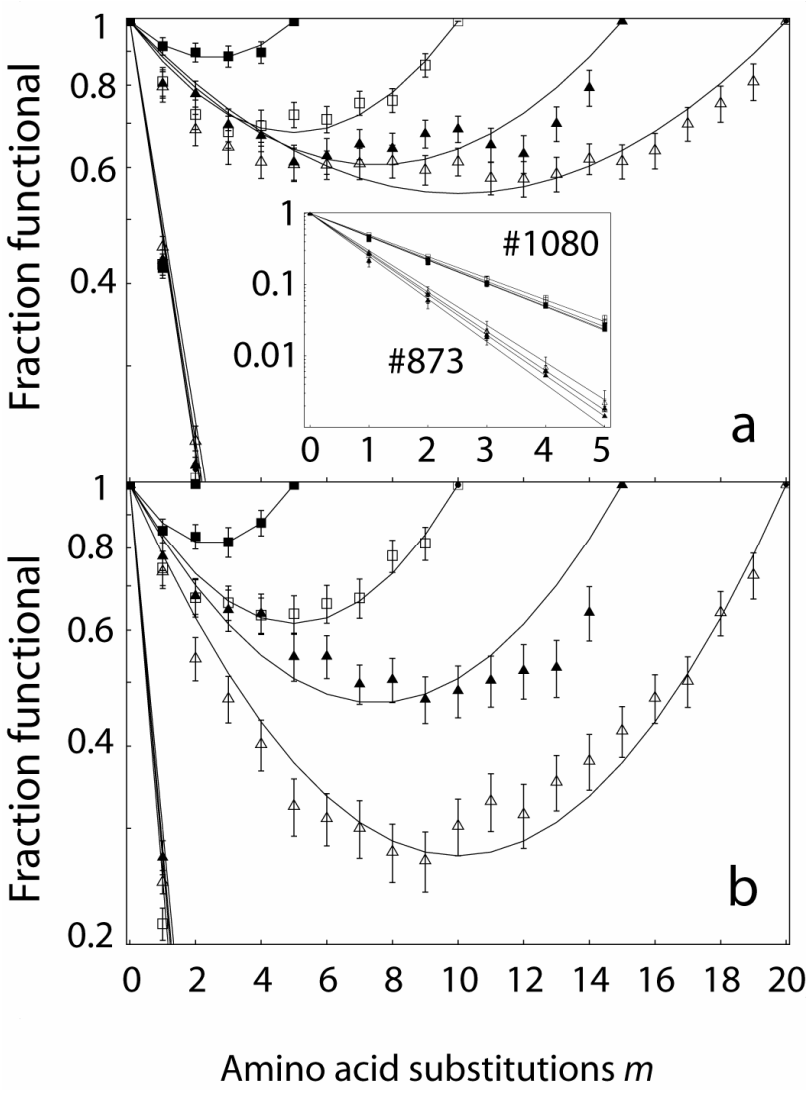


Figure 3.2: Lattice protein results mirror experimental findings. Shown are average fractions of functional chimeras over 50 replicates using parents sharing 20-80% sequence identity (D = 20, 15, 10, or 5) for a high-v structure, #1080 (**a**) and a low-v structure, #873 (**b**) (see Supplemental Material for Chapter 3). Independent fits for ρ and v are plotted. **Inset**: Mutation data for each structure, collected from homologs used to construct **a** and **b**. Curves show four independent best fits to Equations 3.1 and 3.2 (see *Methods*); error bars are $\pm$ 1 S.E.

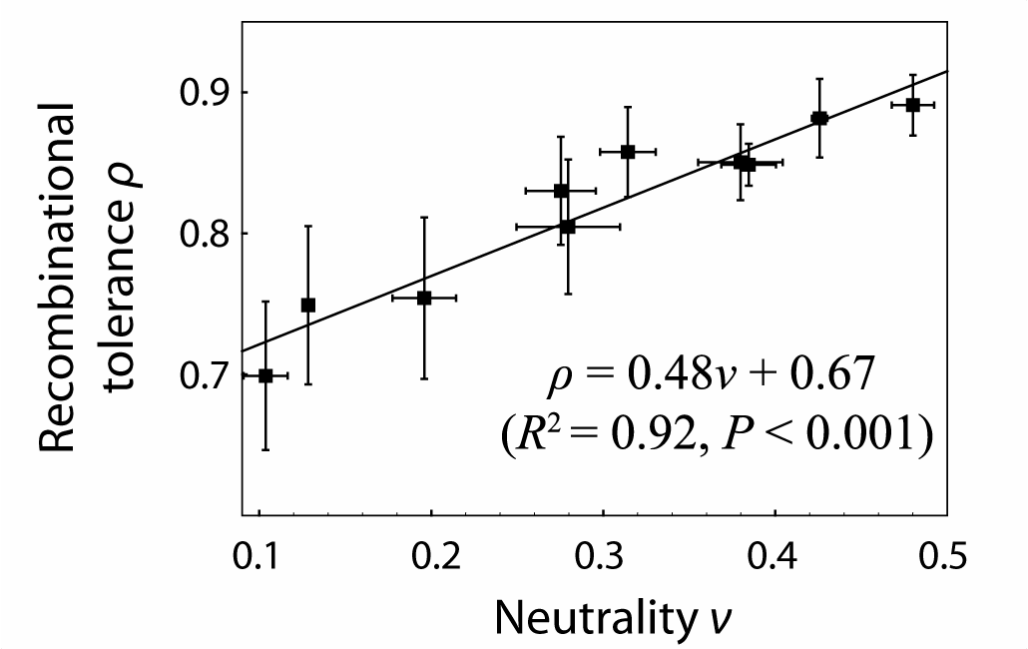


Figure 3.3: Neutrality ν is correlated with recombinational tolerance ρ for lattice proteins. Results are from 10 different structures. Error bars show s.d. of averages of ν and ρ taken at four values of sequence identity (20, 40, 60 and 80%, as in Fig. 3.2). [For the two lowest-neutrality structures, error bars reflect two and three sequence identities, respectively, because no highly diverged homologs were found.]

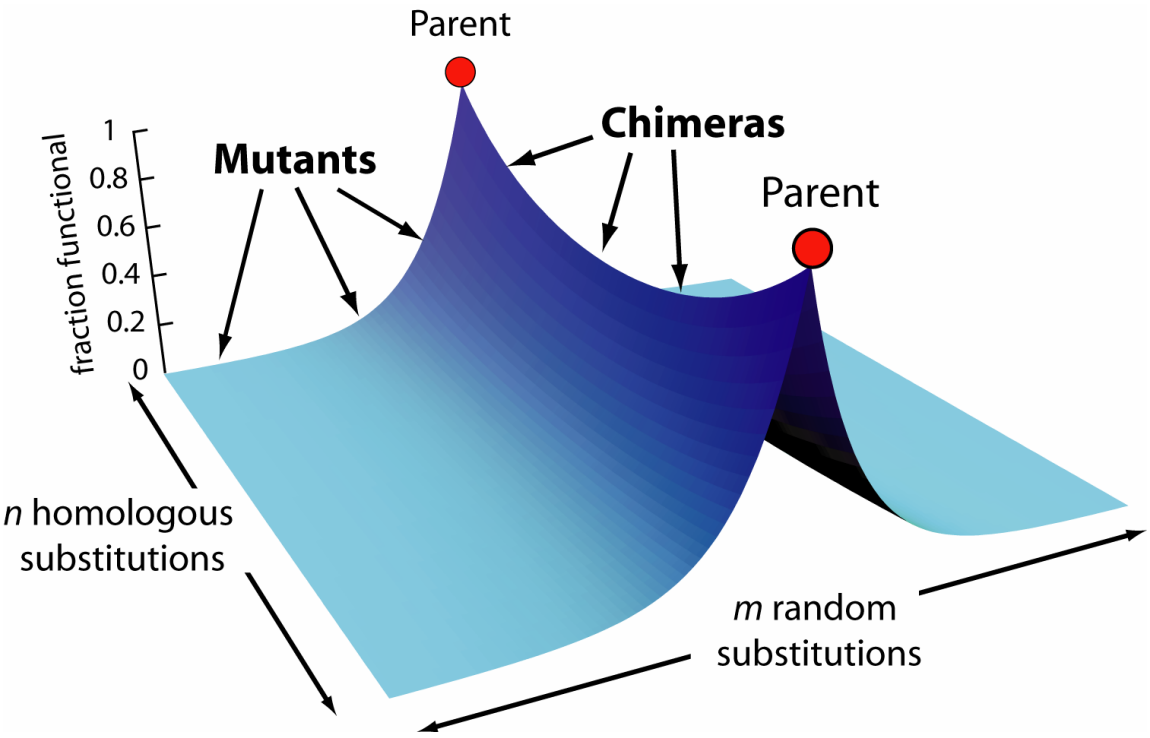


Figure 3.4: Chimeras occupy a functionally enriched ridge in sequence space. Surface height, the product of Equations 3.1 and 3.2, represents the probability of retaining parental fold (and therefore function) given independent random and homologous substitutions. Mutants lie along the near and far edges (slope determined by ν), chimeras lie on the ridge (slope determined by ρ), and mutated chimeras lie on the hillsides.

PART 2

MISFOLDING DOMINATES NATURAL PROTEIN EVOLUTION

C h a p t e r 4

A SINGLE DOMINANT CONSTRAINT ON PROTEIN EVOLUTION[3]

Explanations should not be multiplied beyond necessity.

attributed to William of Ockham

Summary

Proteins evolve at different rates, and while these rates are used ubiquitously in molecular evolutionary biology, why rates differ between proteins has remained unclear. An explosion of genome-wide data sets has produced the surprising discovery that a gene's expression level strongly predicts the evolutionary rate of the protein it encodes. Simultaneously, many other correlates of evolutionary rate have been found, but because each of these may co-vary with expression level, controlling for expression's influence is necessary to establish an independent effect of any quantity on evolutionary rate. We show that typical methods used to statistically control for expression produce spurious results given co-varying noisy data, the rule in genomic analyses, calling into question the conclusions of several influential analyses of the causes of evolutionary rate. Using a technique that does not suffer from these problems, we carry out a comprehensive analysis

---

[3] Portions of this chapter adapted from D. Allan Drummond, Alpan Raval, and Claus O. Wilke, "A single determinant dominates the rate of yeast protein evolution," *Molecular Biology and Evolution* (2006) **23**(2):327–337, reprinted by permission of Oxford University Press.

of seven variables designed to uncover the major independent correlates of evolutionary rate in the model eukaryote *Saccharomyces cerevisiae*. Strikingly, our analysis suggests that, at least among these variables, there is only one major independent correlate, and all others are either relatively minor or entirely spurious. We argue that this dominant determinant represents the translation frequency of a gene, raising the question of how and why translation physically influences protein evolution (treated in Chapters 5 and 6).

Introduction

The rate at which proteins accumulate changes over evolutionary time is the hallmark measurement of the molecular age of evolutionary biology. Protein evolutionary rates, usually measured by the number of nonsynonymous (amino-acid-altering) nucleotide substitutions per site separating related genes in divergent lineages, are now routinely used to detect the tempo and mode of natural selection[76], identify gene relatives[77] and the molecular signatures of disease[78], create phylogenetic trees[79], and infer the time of evolutionary events from molecular evidence (most famously, the divergence time of humans from other primates[80]). The major controversy among molecular evolutionists in the latter half of the twentieth century, the selectionist-neutralist debate[81], revolves around our understanding of what determines the evolutionary rates of genes and their encoded proteins.

That debate continues, and what determines a protein's evolutionary rate remains the subject of active speculation and ongoing research[82-84]. It has long been noted that functionally important portions of protein sequences evolve slowly, and the view that functional importance governs differences in evolutionary rates[85] held sway for decades. However, the advent of the genomic era has rendered this view untenable[78]. Recent studies examining rates of evolution across entire genomes have uncovered significant correlates, often argued to be causes, for evolutionary rate among many disparate variables: proteins have been reported to evolve slower if they interact with more protein partners (have higher "degree")[86], play a more central role in interaction networks (higher "centrality")[87], have shorter sequence length[88], or if their encoding genes have a higher codon adaptation index (CAI)[83,89], or yield a larger fitness effect upon gene knockout (lower "dispensability")[90-92]. Perhaps surprisingly, the strongest known predictor of a protein's evolutionary rate is its encoding gene's expression level measured in mRNA molecules per cell[78,82], an effect

which spans the tree of life: highly expressed proteins evolve slowly, from bacteria[83,93], yeast[82,93-95] and algae[96], to worm[95], cress[97], fly[98], mouse[99] and human[95,99].

Here, we first demonstrate that the analytical techniques widely used to establish independent roles for many effects—partial correlation and multivariate regression—generate highly significant but entirely spurious effects given noisy data such as those available for evolutionary analyses. Then, using a technique which does not suffer from these problems, we carry out a comprehensive analysis designed to uncover the major independent correlates of evolutionary rate in the model eukaryote *Saccharomyces cerevisiae*. We determine the number of such correlates, their strength, and their relationship to the biological variables used in previous studies. Finally, we ask what these correlates reveal about the biological constraints on protein sequence evolution.

Results

Correlation and partial correlation analysis

We used the yeast *Saccharomyces cerevisiae* to examine the determinants of evolutionary rate because it has been the subject of many previous analyses[82,86,89,90,100] and has an enormous amount of available genomic, proteomic, and functional data. We first examined the raw correlation of six previously assessed biological variables (expression, CAI, length, dispensability, degree, and centrality) with protein evolutionary rate, as measured by the number of nonsynonymous (amino-acid-altering) nucleotide substitutions per nonsynonymous site in the underlying gene, dN. A seventh variable, the number of protein molecules per cell ("abundance"), was also considered, and later analyses also consider dS, the number of synonymous substitutions per synonymous site. Table 4.1 shows that all variables except centrality correlated significantly with evolutionary rate, as previously reported. (The original analysis by Drummond *et al.*[93] used parametric [Pearson] correlations; the present analysis uses nonparametric [Spearman rank] correlations, and arrives at similar conclusions.)

Expression level strongly correlates with evolutionary rate, and higher-expressed genes have higher CAIs[101], are less dispensable[102], more abundant[103], and more likely to be found in protein-protein interaction experiments[104] than lower-expressed genes. No inverse relationships have been posited by which these variables alter expression level. Thus, it is imperative to establish whether these variables play a role independent of expression level. Following previous analyses[89,105,106], we computed the (nonparametric) partial correlation of our seven variables with evolutionary rate, controlling for expression level. Table 4.1 shows that CAI, dispensability and degree all showed reduced but highly significant partial correlations consistent with previous studies[89,107], as did abundance.

Partial correlations and noisy data

What can we conclude from highly significant partial correlations? Yeast expression-level measurements from multiple groups, even two using the same commercial oligonucleotide array, correlated with coefficients of only 0.39 to 0.68 [108], demonstrating that expression level measurements either are inaccurate and/or simply reflect the variability of gene expression across growth conditions and strains. We refer to all such variability as noise, regardless of its source. Noisy data are the rule in genome-wide molecular studies, leading us to explore what effect noise has on partial correlation analyses. As a concrete example, CAI is so tightly bound to expression level that a recent analysis used CAI as its preferred expression-level measurement[89]. Might CAI's significant partial correlation only reflect our inability to control for the true (*i.e.*, evolutionarily relevant) underlying expression level? More generally, we can ask: what is the expected partial correlation of two variables, controlling for a third, when *i*) the two variables relate only through dependence on the third "master" variable, and *ii*) all measurements contain noise?

Given these conditions, Drummond *et al.* reported explicit formulas for the expected partial correlation, its statistical significance, and its behavior under various limiting cases[93]. The expected partial correlation is, in general, larger than zero, because the full correlation reflects the true underlying master variable's influence, while partial correlations can only remove the portion of this influence that is visible through a noisy measurement (Box 4.1). Surprisingly, if measurements of an underlying causal variable (*e.g.*, expression level) are noisy, highly significant partial correlations of virtually any strength between the dependent predictors can be obtained[93].

As a case in point, dispensability's role has been vigorously debated[89,105,107] with correlation and partial correlations acting as key analytical tools. Given a model in which expression level X and noise completely determine dispensability D and evolutionary rate K (see Box 4.1), what is the observed partial correlation $r_{DK|X'}$ if we fit variables to

approximately match the observed correlations between *X'*, *D* and *K*? As a concrete example, previous reports show that, using parametric Pearson's correlations, $r_{X'K} \approx -0.6$ [82,89], $r_{DK} \approx 0.25$ [89], $r_{DX'} \approx 0.2$ [105], and $r_{DK|X'} \approx 0.24$ [89]. We can obtain roughly the reported full correlations and $r_{DK|X'} \approx 0.23 \pm 0.02$, $P << 10^{-9}$ with 3,000 observations if the true expression level *X* is normally distributed with mean 0.5 and standard deviation 0.25 and the observable predictors *X'*, *D*, and *K* are equal to *X* plus zero-mean normally distributed noise with standard deviations of 0.3, 0.7 and 0.1, respectively. This highly significant partial correlation is entirely spurious: in this model, expression level and random noise completely determine dispensability. Thus, the observed statistical relationship between dispensability and evolutionary rate, established by correlation and partial correlation, would arise *even if no actual relationship existed* except mutual dependence on noisily measured expression level.

Drummond *et al.* show that multivariate regression analysis fails in virtually the same way[93]: collinear predictors confound the technique, which implicitly assumes statistical independence among its input variables (this analysis was originally done by Claus Wilke; the present analysis is a nonparametric version of that analysis). The variance inflation factor (VIF) may be used to quantify the degree of predictor collinearity, and Table 4.1 reports VIF's for our data. These VIF's indicate some collinearity but are not high enough to raise significant concerns. However, for our toy model (Box 4.1) in which the two predictors reflect the same underlying variable plus noise, the VIF's are only 1.21 in both cases, yet the analysis demonstrates that multivariate regression and partial correlation break down anyway. Collinearity and noise work together to undermine these techniques.

Principal component regression analysis

An alternative approach is to first identify independent sources of variation in the data, and then determine the contribution of each biological predictor to each source. The technique

of principal component regression offers a standard way to carry out such an analysis. (The idea to use principal component regression, and the two main analyses on yeast reported here, are due to Claus Wilke.)

In principal component regression[109], multiple linear predictors (*e.g.*, expression level, dispensability, etc.) are scaled to zero mean and unit variance, inserted in a matrix, and rotated such that the new coordinate axes point in the directions of greatest predictor variation. The new axes define variables, called principal components, which are linear combinations of the original predictors. Subsequent linear regression of the response (*e.g.*, the nonsynonymous rate dN or synonymous rate dS) on the rotated predictor data yields several pieces of information per principal component: the proportion of the response's variance, R^2, explained by the component, the significance of this R^2, and the fractional contribution of each original predictor to the component. Because all principal components are orthogonal and independent, the total proportion of response variance explained by the data is the sum of the component R^2's. Principal component regression thus circumvents the debilitating problems of partial correlation and multivariate regression analyses (Box 4.1) while yielding results which are, in some ways, easier to interpret.

Drummond *et al*. [93] carried out principal component regression on the seven predictors analyzed above. Because the determination of principal components involves only the predictors and not the response (*i.e.*, dN or dS), there is only one set of components and contributions from biological predictors. The regression analysis generates response-specific results, in particular, the proportions of variance in dN and dS, which each component explains. Figures 4.1a and 4.2a show the results of principal component regression of dN and dS using the seven predictors of expression, CAI, abundance, length, dispensability, degree and centrality. (Here, we report results of a nonparametric analysis, showing that the results differ little from the parametric analysis of Drummond *et al*.)

Strikingly, for the rate of protein evolution, dN, one principal component explained 41% of the variance with high significance, while all other components explained less than 2% (Fig. 4.1a). The single dominant component was mostly (>75%) determined by roughly equal contributions from three predictors: expression level, abundance, and CAI.

While the causes of dN's variation have remained unclear, the rate of synonymous-site evolution dS is constrained by translational selection. Selection for preferred codons, which correspond to abundant tRNAs and are translated faster and more accurately[101,110], makes many synonymous changes unfavorable and thus reduces dS[111]. Figure 4.2 shows that the dS results mirror those using dN: the first component again dominates the rate of evolution (32% of dS variation).

The size of the seven-component data set (568 genes) was severely limited by the requirement for genes having measures for all seven predictors. In particular, we used high-quality interactions measurements[112] for degree and betweenness-centrality; eliminating these measurements, which apparently contribute negligible amounts to evolutionary rate, more than triples the data set size to 1,939 genes. We performed the same analysis on this expanded set and obtained similar results (Figures 4.1b and 4.2b).

It is common practice to interpret dS as the rate of selectively neutral divergence, and the ratio dN/dS as the deviation of protein evolutionary rate from neutral, putatively allowing detection of purifying selection or adaptive evolution. We analyzed dN/dS and found trends that were similar to those observed in dN and dS alone (not shown). The dominant principal component explained only half the variation in dN/dS compared to dN or dS, but the reason seems obvious in light of our results: dN and dS appear to reflect the same underlying selective force, so dividing one by the other removes much of the shared influence. (We will return to this issue in greater detail in Chapter 6.) In yeast, as in many other organisms, dS does not reflect neutral divergence but rather divergence constrained by translational selection for preferred codons, as previous authors have noted[111].

Evolutionary rates reflect the accumulation of differences between orthologous sequences over long times, and noise (both actual and inherent in various estimation methodologies) likely varies with phylogentic distance. To assess the importance of phylogeny on our results, we carried out principal component regression on dN and dS values calculated using two relatives of *S. cerevisiae*, *S. paradoxus* and *Kluyveromyces waltii*, which diverged roughly 5 and 100 million years ago, respectively[94].

For *S. paradoxus*, we obtained almost identical results for dN as for the Hirsh *et al*. data. However, dS showed a much weaker, though still dominant, first component that explained 15% of the dS variance including interaction data and 6% without these data, five-fold more than any other variable. We traced the weaker dS signal to differences in gene filtering (Hirsh *et al.*'s smaller data set omits sequences whose gene-level phylogeny did not match the species-level pattern, and sequences containing introns and potential frameshifts) and in codon frequency estimates. Controlling for gene filtering, the nine-free-parameter codon frequency model used by Hirsh *et al.* produced a larger signal than the sixty-free-parameter model used by Drummond *et al*. (data not shown), indicating that analyses of dS may be sensitive to estimation methodologies.

For the distant relative *K. waltii*, we again obtained nearly identical results for dN. For the 2,412 genes without (and 752 genes with) interaction data, one principal component determined by CAI, abundance and expression explained 41% of the variance in dN, while all other components explained < 2%. For dS, no dominant component emerged, and the best component (mostly expression and CAI) explained 1.7% of the variance. The lack of any predictive signal for dS is not surprising, since the dS values relative to *K. waltii* average more than 14 substitutions per synonymous site, far beyond the range of reliable estimation. These high dS values may result from a combination of the large amount of time separating the species, changes in synonymous pressures, and difficulties in ortholog identification and alignment. The robust dN results lend weight to the first two

explanations. We expect that as even more distant relatives are analyzed, the dN results will be attenuated by noise, alignment degradation, and phenotypic changes that must, in some cases, be linked to changes in relative gene expression levels.

To assess whether the trends we identified for yeast extend to other species, we examined evolutionary rates in 2,605 *Escherichia coli* genes relative to *Salmonella typhimurium*. Lacking global protein abundance, interaction and dispensability data for *E. coli*, we used length, two measures of expression level, reflecting growth in minimal M9 and rich LB media, and two measures of codon optimization, CAI and the frequency of optimal codons F_{op} [113], as predictors. Again, a dominant component emerged which explained 36% of the dN variance (16-fold more than any other) and 25% of the dS variance (38-fold more than any other). Since most of the included predictors are translation-oriented in some way, our results offer no conclusion as to the possible influence of other predictors in *E. coli*. However, the remarkable similarity to the yeast results, including the large portion of variance explained, suggests that similar selective forces have shaped evolutionary rates in this prokaryotic organism.

Discussion

We have reported the most comprehensive comparative analysis to date of potential determinants of nonsynonymous (dN) and synonymous (dS) yeast gene evolutionary rates[93]. We used a previously published data set of evolutionary rates, previously used to establish an independent role for dispensability[89] to highlight the methodological improvements introduced here. We find that a single underlying component explains roughly half the variation in both dN and dS, and that this dominant component is almost entirely determined by gene expression level, protein abundance and codon bias as measured by the codon adaptation index (CAI). Our results generalize to *E. coli* despite use of a reduced set of predictors.

The predictors we included in our analysis appear to explain roughly half the variation in dN and dS. Some other predictor(s) could explain the remaining half, but this seems quite unlikely, for a variety of reasons. First, a significant portion of evolutionary-rate variations are probably random, because the evolutionary process is inherently stochastic. Second, our R^2 estimates constitute a lower bound, because the R^2's we find are attenuated by measurement noise, for example on microarray readings of gene expression[108], by systematic error, *e.g.*, in some protein-protein interactions data[104], and by time variation, for example in expression over the cell cycle[114]. Finally, the true relationship between any of the predictors we examine and dN or dS is unlikely to be perfectly linear, and deviations from linearity reduce parametric R^2. We return to the question of how much evolutionary-rate variation one can ever expect to explain in Chapter 6.

Our results point to a single dominant cause for most of the 1,000-fold variation in evolutionary rates among yeast genes, and the dominant component's three biological contributors suggest that cause is translational selection. We hypothesize that the number of translation events a gene experiences determines its evolutionary rate, and that expression, abundance and CAI are all roughly equally good predictors of the number of translation events. A causal hypothesis to explain the translation's dominant role is introduced and defended in Chapter 5.

We used principal component regression for our analysis because, as we demonstrate, the more commonly employed techniques of partial correlation analysis and multivariate regression are inapplicable by assumption (in the latter case) and prone to produce spurious effects in the presence of noisy correlated data (in both cases). By contrast, under principal component regression, the transformed predictors are orthogonal and uncorrelated, so that their relative contributions to the overall regression model can be evaluated independently and reliably.

Wall *et al.*[89] use a structural equation model to examine the influence of measurement inaccuracy on their partial correlation analysis of the effects of expression level and dispensability on dN. Given their analysis, they admit an inability to determine the relative importance of these two predictors, but conclude that dispensability has an independent effect on dN. We claim to be able to determine relative importance, and come to an opposite conclusion, for two reasons. First, a general advantage of principal component regression over partial correlation is the ability to find predictors not originally included in the analysis. We were fortunate in this case that the dominant predictor is not expression level, CAI or abundance, but rather a variable (likely the frequency of translation) that these three predictors measure with roughly equal accuracy. Partial correlation can never find such underlying variables. Second, Wall *et al.*'s structural equation model attempts to quantify how much the predictors could explain given hypothetical levels of measurement inaccuracy, but with principal component regression, we are asking how much the given predictors can explain, whatever their accuracy. Here, we were doubly fortunate. Three of our predictors (CAI, abundance and expression) triangulate on the same underlying variable, increasing accuracy essentially by measuring it in triplicate; this variable happens to explain a large portion, perhaps most, of dN's explainable variance.

How much dispensability and degree influence evolutionary rate has been a contentious issue. Regarding the former, the literature reflects disagreement over whether dispensability has any effect whatsoever on the rate of evolution, with partial correlation analyses playing a prominent evidentiary role[89,105,107]. Our analysis, which avoids problematic partial correlations, but uses the same data as in previous analyses that appeared to confirm a significant role for dispensability[89], is quite clear: dispensability neither constitutes an independent source of variation in dN nor contributes meaningfully to the dominant component that does influence dN. In the case of degree, the disagreement has pivoted on whether experimental surveys are biased toward detecting interactions more often in highly expressed proteins[104,115,116], leading to a true, but biologically irrelevant

degree–dN relationship. Our analysis shows that degree does not contribute independently, but makes a small, significant contribution to the variable dominated by expression, abundance and CAI, as expected under the expression-bias hypothesis and inconsistent with a true constraint from the number of interactions. In short, our results suggest neither degree nor dispensability make much difference in dN, and point out precisely why previous authors have been led to the opposite conclusion.

The rates dN and dS are routinely used to carry out analyses on selection, often under the assumption that $dN/dS > 1$ indicates adaptive protein evolution and $dN/dS < 1$ indicates purifying selection, and generally with the intent of quantifying functional pressures. Our results suggest that both evolutionary rates are determined by translational selection and are therefore likely poor predictors of functional selection, because translational selection by definition operates before a protein becomes functional. In yeast, dS does not measure neutral divergence, and thus, in the absence of a quantitative description of the relative strengths of selection on nonsynonymous and synonymous sites, the measure dN/dS is meaningless. We provide just such a quantitative description in Chapter 6 to explain how dN, dS and dN/dS can simultaneously decrease with expression level, a non-trivial finding which suggests that precisely the same selective force must not govern the first two measures (because then their ratio would not be expected to co-vary with both the numerator and denominator).

We have found that yeast coding sequences accumulate substitutions according to a surprisingly simple formula: more predicted translation events means slower evolution. In recent years, evidence has accumulated that translation-linked variables—in particular, expression levels—govern the evolutionary rate of proteins across all life, from bacteria[83] to fungi[82], plants[97] and animals[117] including humans[99], but translational selection has only recently been proposed as an explanation for this puzzling trend[84,94]. Our results suggest that translational selection dominates the rate of protein evolution, and by extension,

suggest that translational selection operates across the tree of life, from prokaryotes to humans. Questions remain concerning the biophysical basis of evolutionary rate variation, but we have shown that, at least in yeast, the answers may be found in translation.

Box 4.1: Comparing partial correlation, multivariate regression and principal component regression

How do the three analytical techniques considered here fare given a case where only one variable determines evolutionary rate? For each technique, what would we conclude about the number and strength of the rate determinants? Consider a simple model in which a variable X (*e.g.*, expression level) determines two other variables, a putative determinant D (*e.g.*, dispensability) and a response K (evolutionary rate), so that $D = X + \varepsilon_D$ and $K = X + \varepsilon_K$, where ε_D and ε_K are noise terms with mean 0 and variances σ_D^2 and σ_K^2. Further assume that we cannot measure X, but only a noisy correlate, $X' = X + \varepsilon_{X'}$. In this model, X is responsible for all the correlation between D and K. We let X be normally distributed with mean 0.5 and standard deviation 0.25 (so that X values span the unit interval) with the observable predictors X', D, and K equal to X plus zero-mean normally distributed noise with standard deviations of 0.3. We ran each analysis 100 times with 3,000 measurements each.

Partial correlation analysis suggests that both D and X' contribute to the rate K independently and with equal strength:

Partial correlation with K	P-value
$r_{DK\|X'} = 0.296 \pm 0.03$	$<< 10^{-9}$
$r_{X'K\|D} = 0.291 \pm 0.02$	$<< 10^{-9}$

Multivariate regression similarly suggests that both D and X' independently influence the rate K:

Predictor	% variance in K explained (R^2)	P-value
X'	16.9 ± 2	$<< 10^{-9}$
D	17.3 ± 2	$<< 10^{-9}$

Principal component regression, however, properly identifies only one component which contributes significantly to the rate K. The two components identified are $X' + D$, which measures mostly X, and $X' - D$, which measures mostly noise. Component 1 alone carries predictive value for K.

Component	% variance in K explained (R^2)	P-value
1 ($X' + D$)	21.3	$<< 10^{-9}$
2 ($X' - D$)	0	0.7

We may proceed with the confidence that we have properly identified the number and strength of the underlying determinants of K.

In general, the underlying variable represented by the dominant component is not known *a priori* and its identification requires additional insight. In this case, we know it is X, which is accurately captured by the principal component regression method, but not by the other methods. Other methods are therefore likely to lead to erroneous results when faced with the problem of trying to find true predictors within noisy data. Principal component regression, as shown here, is unlikely to do so.

Our toy model underscores a key observation: in the presence of noisy and correlated data, nonzero partial correlations and R^2 values from multivariate regression—even those with very high statistical significance—must not be taken as evidence for independent effects, as in previous studies[89,106].

Methods

Genomic data

We obtained codon adaptation indices and high-quality evolutionary rates (nonsynonymous substitutions per site dN, synonymous substitutions per site dS, and ratios dN/dS) from four-species alignments in the *Saccharomyces* genus for 3,036 *S. cerevisiae* genes [89] (their supporting information, Table 4).

Deletion-strain growth-rate data were downloaded from http://chemogenomics.stanford.edu/supplements/01yfh/files/orfgenedata.txt; the average growth rates of the homozygous deletion strains were used as dispensability measurements in our analysis. The FYI yeast protein interaction data set[112] provided interaction network hub types for 199 genes and the number of interactions for 1,379 yeast genes. The latter data set was used to compute betweenness-centrality values, which quantify the frequency with which a network node lies on the shortest path between other nodes, as described by Hahn and Kern[87]. Genomic data for *S. paradoxus* and *K. waltii* were obtained exactly as described by Drummond *et al*.[94] Genome sequences for *Escherichia coli* K12 and *Salmonella typhimurium* LT2 were obtained from TIGR[118], with orthologs identified and evolutionary rates computed exactly as described[94]. Gene expression levels for *E. coli* measured in mRNAs per cell in Luria-Bertani and M9 media were obtained from Bernstein *et al*.[119]

Statistical analysis

We used R[31] for statistical analyses and plotting. The package 'pls' was used to perform principal component regression. We log-transformed all variables except dispensability. We decided whether or not to log-transform a variable based on whether log-transformation led to a higher R^2. For those variables that contained zeros, we added a small constant before the log-transformation, as previously suggested [89]. This constant was 0.001 for dN,

dS and dN/dS, and 10^{-7} for betweenness centrality. We scaled the predictor variables to zero mean and unit variance before carrying out the principal component analysis. In all regression analyses (both against the original predictors and against the principal components), we determined statistical significance levels by starting with the full model and successively dropping the least-significant predictor until only significant predictors ($P < 0.01$) remained.

Table 4.1: Partial correlation analysis of seven putative determinants of evolutionary rate.

Variable *X*	**Correlation** $r_{X,\mathrm{dN}}$	**Partial Correlation** $r_{X,\mathrm{dN}\mid\mathrm{expr.}}$	**Variance Inflation Factor**
Gene expression level	−0.50**	0	2.8
Codon adaptation index (CAI)	−0.52**	−0.34**	2.0
Protein abundance	−0.46**	−0.26**	1.9
Gene length	0.08**	0.05*	1.3
Gene dispensability	0.23**	0.14**	1.1
Degree (# of protein-protein interactions)	−0.25**	−0.15**	2.0
Protein centrality (frequency on node-node shortest paths)	−0.10*	$-0.08^{\#}$	1.9

Significance codes: $^{\#}$, $P < 0.05$, *, $P < 0.01$; **, $P < 10^{-3}$.

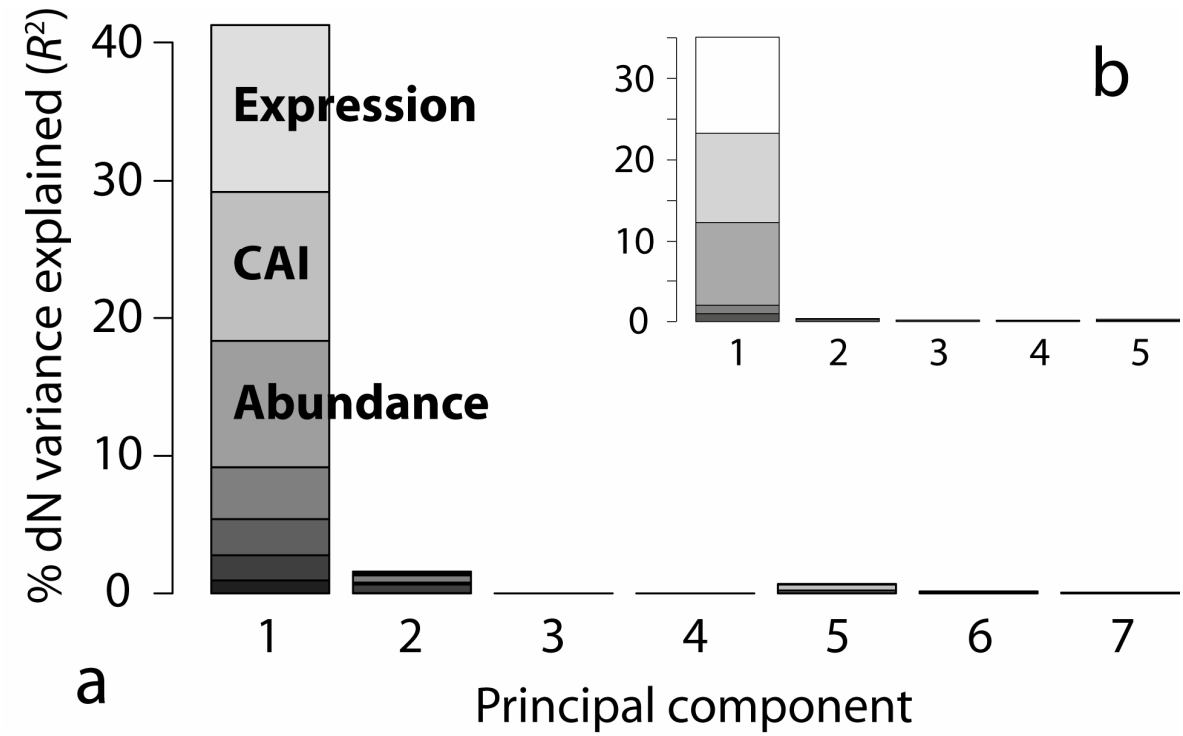


Figure 4.1: Principal component regression on the rate of protein evolution (dN) in 568 yeast genes reveals a single dominant underlying component. **a**, Of the seven principal components, only three explained a statistically significant proportion of the variation in dN. The dominant component explained 41% of the variance, while no other component explained more than 2%. Expression level, codon adaptation index, and protein abundance determined most of this dominant component (labeled), while the remaining predictors (in order from top to bottom: length, dispensability, degree, centrality) determined < 25% of the component's R^2. **b**, A larger data set (1,939 genes) excluding protein-protein interaction predictors showed the same patterns as in **a**.

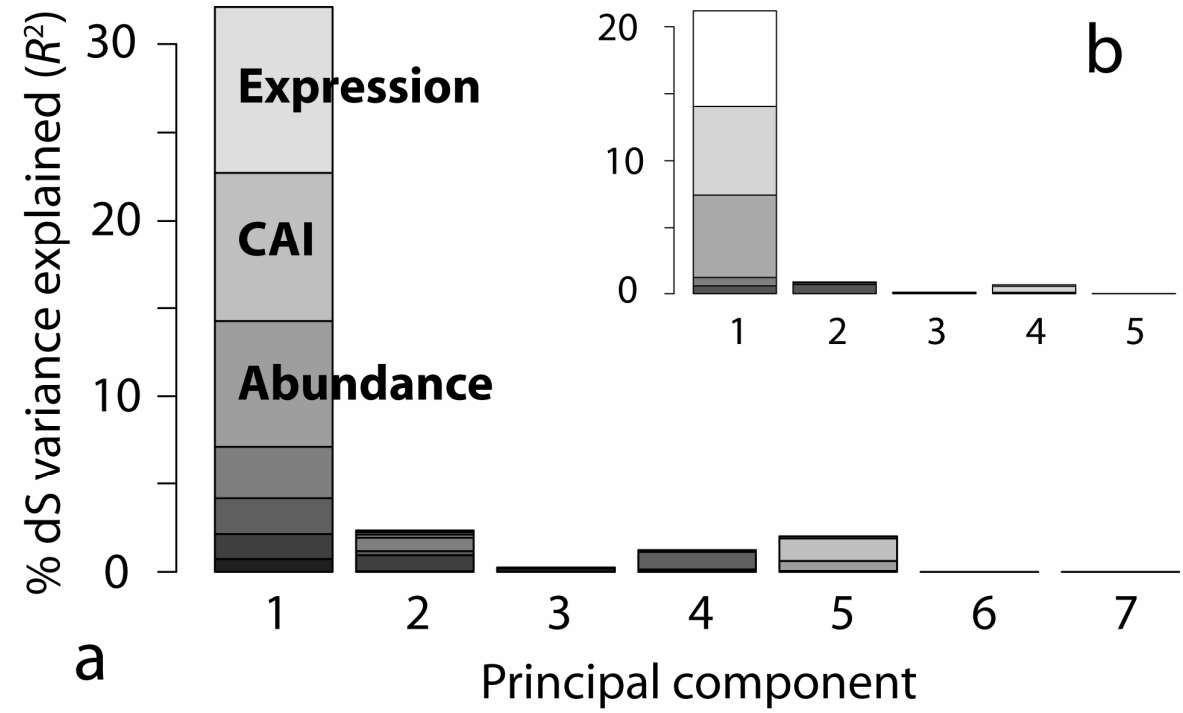


Figure 4.2: Principal component regression on the rate of synonymous-site evolution (dS) in 568 yeast genes reveals a single dominant underlying component. **a**, Seven predictor variables (see text) yielded seven principal components, of which four explained a statistically significant proportion of the variation in dS. The dominant component explained 32% of the variance, while no other component explained more than 3%. See Figure 4.1 caption for the breakdown of predictor contributions. **b**, A larger data set (1,939 genes) excluding protein-protein interaction predictors showed the same patterns as in **a**.

C h a p t e r 5

THE TRANSLATIONAL ROBUSTNESS HYPOTHESIS[4]

We must not say every mistake is a foolish one.

Cicero

Summary

Gene expression levels are the single best known predictor of evolutionary rates. Chapter 4 reveals a single dominant constraint on yeast protein evolution that clearly aligns with gene expression levels and is consistent with the number of translation events. In this chapter, we extend our analysis to examine potential differences in functional pressures between proteins expressed at different levels. Using several sequenced yeast genomes, global expression and protein abundance data, and sets of paralogs traceable to an ancient whole-genome duplication in yeast, we rule out several confounding effects to show that expression level alone explains roughly half the variation in *Saccharomyces cerevisiae* protein evolutionary rates. To explain why highly expressed proteins evolve slowly, we hypothesize that selection to reduce the burden of protein misfolding will favor protein sequences with increased robustness to translational missense errors. Pressure for translational robustness increases with expression level and constrains sequence evolution. Genome-wide tests favor the translational robustness explanation over existing hypotheses that invoke constraints on function or translational efficiency. Our results suggest that

---

[4] Adapted from *Proceedings of the National Academy of Sciences, USA,* **102**(40), D. Allan Drummond, Jesse D. Bloom, Christoph Adami, Claus O. Wilke, and Frances H. Arnold, "Why highly expressed proteins evolve slowly," p. 14338–14343, copyright (2005).

proteins evolve at rates largely unrelated to their functions, and can explain why highly expressed proteins evolve slowly across the tree of life.

Introduction

Thirty years ago, Zuckerkandl proposed that a protein's sequence will evolve at a rate primarily determined by the proportion of its sites involved in specific functions, or its "functional density"[85]. While this proposal has gained wide acceptance[120], measurement of functional density remains problematic because residues may contribute to protein function in unpredictable ways and arduous sequence-wide saturation mutagenesis and mutant characterization studies are required to ascertain these effects.

Instead, many recent studies have focused on other, more readily obtained measures which may approximate functional density. For example, protein-protein interactions presumably constrain interfacial residues, and some reports indicate that highly interactive proteins evolve slowly[86]. The intuition that a protein's overall functional importance should amplify the fitness costs of mutations at sites which make subtle functional contributions has been captured in analyses of how a gene's functional category[82,83], its essentiality for organism survival[83,121,122], or its dispensability[89,90] correlate with evolutionary rate. In all cases, the effects under consideration explain only a small fraction (~5% or less) of the observed variation in evolutionary rate as quantified by their squared correlation coefficients r^2, and as Chapter 4 shows, many if not most of these are spurious correlations arising from pervasive influence of gene expression level[93].

Expression level's disproportionate influence remains unexplained[82-84,101,105,123]. Indeed, significant questions have persisted about whether expression level truly determines evolutionary rate, because highly expressed proteins may possess unique structural or functional features which constrain their sequences. Paralogous gene pairs resulting from a whole-genome duplication (WGD) event, such as in the lineage of *Saccharomyces cerevisiae*[124], minimize such differences: homology ensures a similar structure, and the majority of yeast paralogs shows little, if any, difference in function[125]. Analyses of evolutionary rates among paralogs have to date confirmed only a small independent role for expression level. Among a set of 185 yeast paralog pairs, evolutionary rate and expression level in mRNA molecules per cell correlated ($r^2 = 0.341$), but the correlation of rate and

expression differences between members of a paralogous pair was much smaller (r^2 = 0.046), and no significant tendency for the higher-expressed paralog to evolve slower was found[82]. A recent study which proved the whole-genome duplication in yeast[124] analyzed patterns of paralog evolutionary rates and concluded that they supported a widely cited model of evolution by gene duplication[126] in which one duplicate gene retains the ancestral function and evolves slowly, while the other evolves rapidly and acquires a new function. Such behavior would obscure the influence of other variables such as expression level on paralog evolutionary rates.

Recently, several resources have become available that allow a more thorough analysis of these issues: a set of 900 *S. cerevisiae* paralogs derived from gene synteny and traceable to the whole-genome duplication event[124], a global measurement of yeast protein abundances[103], and several additional yeast genome sequences[124,127]. Here, using this new information, we examine the strength, independence and physical basis of expression-based constraints on protein sequence evolution. We carry out a systematic analysis designed to answer several questions. How strongly does expression constrain yeast protein evolution after controlling for structure and function? What role does functional differentiation play compared to gene expression in predicting the relative evolutionary rates of duplicate genes? And, what do these correlations reveal about underlying causes of evolutionary rate differences? We introduce a novel hypothesis to explain why highly expressed proteins evolve slowly, and test this explanation against other causal hypotheses using genome-wide data. Finally, we explore whether the selective pressure we propose increases functional density, and examine the biological costs underlying it.

Results

Expression level and evolutionary rate

Using genome-wide measurements of expression level (mRNA molecules per cell) and evolutionary rate (the number of nonsynonymous substitutions per site, dN) in *S. cerevisiae*, we confirm that expression level strongly predicts protein evolutionary rate. Figure 5.1a shows that expression level alone explains between a quarter and a third of the uncorrected variance in dN for 4,255 *S. cerevisiae* proteins with *S. bayanus* orthologs and measured expression levels (Pearson $r^2_{\text{dN-expr}} = 0.28$, $P << 10^{-9}$) and for the 580 paralogs (290 pairs) ($r^2_{\text{dN-expr}} = 0.31$, $P << 10^{-9}$). We find that the strongest simple relationship linking dN and expression is a power law (linear on a log-log scale) and that evolutionary rates span three orders of magnitude. Expression level affects evolutionary rates of duplicated and non-duplicated genes similarly.

Structural or functional differences between proteins with differing expression levels may systematically bias the dN–expression relationship. If the power-law relationship observed across paralogs holds between paralogs in a pair, the ratio of paralog expression levels should correlate linearly with the ratio of evolutionary rates on a log-log scale. Figure 5.1b confirms this prediction ($r^2_{\text{dN-expr}} = 0.29$, $P << 10^{-9}$), and demonstrates that a more limited previous analysis[82] underestimated this relationship's strength by more than six-fold.

Measurement noise attenuates correlations, possibly obscuring the strength of the relationships we have examined. For example, yeast gene expression levels measured by different groups correlate with coefficients of only 0.39 to 0.68[108]. We therefore first examined the dependence of relative inter-paralog evolutionary rate on the degree of expression level disparity, and found a dramatic association (Figure 5.1c). For all 290 pairs, in 192 cases the higher-expressed protein evolved slower ($P < 10^{-7}$, binomial test). Among the 19 pairs for which expression differs by at least 18-fold, all of the higher-expressed paralogs have evolved slower and $r^2_{\text{dN-expr}} = 0.67$. The dN–expression correlation

can also be corrected for attenuation, allowing us to determine how much of the explainable variation in dN—variation not due to measurement noise—can be attributed to expression level. Spearman's correction for attenuation in a squared correlation coefficient is $r^2_{\text{corr}} = r^2_{\text{dN-expr}} / (r_{\text{dN}} r_{\text{expr}})$. We found that the correlation between two independent measurements of yeast gene expression using the same commercial oligonucleotide array was $r_{\text{expr}} = 0.72$ (Pearson's r, 5,555 genes), and the correlation between dNs we measured using orthologs in *S. bayanus* to those measured using *S. paradoxus* orthologs was $r_{\text{dN}} = 0.92$ (4,208 genes), yielding an overall $r^2_{\text{corr}} = 0.47$ for the 580 paralogs and $r^2_{\text{corr}} = 0.42$ for all 4,255 genes.

Repeating these analyses using CAI as an expression-level proxy (see *Methods*) led to similar conclusions (Supplemental Material and Fig. 5.S1).

These analyses lead us to conclude that expression level accounts for up to half of the explainable variation in yeast protein evolutionary rates, even when considering only proteins with similar structures and functions.

Functional divergence of gene duplicates and evolutionary rate

Are the disparate evolutionary rates in paralogous proteins a result of acquisition of new function ("neofunctionalization") in one paralog[124,126], or do they simply reflect expression differences? Both explanations predict asymmetric paralog evolutionary rates measured against a pre-duplication relative. However, only the expression level explanation predicts that asymmetric rates will continue indefinitely, which can be measured using a post-duplication relative in which the genomic upheavals following whole-genome duplication (massive gene loss, genome rearrangements, neofunctionalization) have long since quieted.

For *S. cerevisiae*, the pre-duplication relative *K. waltii*, which diverged >100 million years ago, allows evaluation of evolutionary rates relative to a single gene descended directly from the ancestral duplicated gene[124] (Fig. 5.2). *S. paradoxus*, at present the closest relative

of *S. cerevisiae* with a sequenced genome, with a divergence time of ~5 million years ago[127], provides a suitable post-duplication relative (Fig. 5.2).

We found unique *S. paradoxus* orthologs and measured expression levels for both paralogs in 73 of the 115 paralog pairs claimed to strongly support Ohno's functional divergence model[124] (as above, we excluded ribosomal proteins). In 64 of 73 cases (88%), the faster-evolving paralog relative to *K. waltii* has also evolved faster relative to *S. paradoxus*, even though roughly 100 million years have elapsed since the duplication event. (Using codon adaptation index [CAI] as a proxy for expression level, 74 of 84 pairs [88%] showed the same pattern.) In 48 of 52 pairs (92%) in which expression differs at least twofold, the higher-expressed paralog evolves slower. Finally, as Figure 5.1 shows, duplicated genes obey the same evolutionary rate–expression relationship as the rest of the genome, and relative expression between paralogs predicts their relative evolutionary rates.

In sum, we find little evidence that functional differentiation causes disparate evolutionary rates among duplicate genes, and plentiful evidence for the influence of expression level. A categorical consideration of neofunctionalization models is beyond our scope; we simply note that relative expression level cannot be ignored in evolutionary analyses of gene duplicates.

Causal hypotheses

Having established the strong and apparently independent correlation of expression level with evolutionary rate, we now turn to our central question: Why do highly expressed proteins evolve slowly? We will first attend to hypotheses offering a unified mechanistic explanation for most or all of expression level's effect, and only then address the possibility that expression level merely aggregates many independent effects to create the illusion of a single cause. In considering unified explanations, we begin by eliminating all the effects considered in the *Introduction*: previous analyses have already established that essentiality, dispensability, recombination rate, functional category, amino acid biosynthetic cost, and number or type of protein-protein interactions explain roughly 0–5% of evolutionary rate variation, while expression level accounts for more than 30%.

As Table 5.1 shows, the nonparametric correlation between expression and dN is twice as strong as that between expression and the rate of synonymous-site evolution (dS). Nucleotide-level pressures such as transcription-associated mutation or DNA repair, or selection on mRNA structure or stability, cannot be the primary explanation for why highly expressed proteins evolve slowly, because they predict an equal expression-linked constraint on dS and dN.

We now consider three hypotheses for why highly expressed proteins evolve slowly. The first, most concisely phrased by Rocha and Danchin[83], posits that each protein molecule contributes a small amount to organism fitness by performing its function, so mutations which reduce two proteins' functional output (*e.g.*, catalytic rate) equally will have fitness effects weighted by the number of molecules of each protein in the cell, or their abundances, causing the more abundant protein to evolve slower. We call this the "functional loss" hypothesis. Note that a highly expressed protein (whose encoding gene is transcribed at high levels) can have a low abundance (if the mRNA is translated infrequently or the protein is rapidly turned over), and vice versa. The second hypothesis, due to Akashi[84,101], holds that because increased expression level leads to selection for synonymous codons that are translated faster or more accurately, nonsynonymous mutations to translationally less efficient codons may be evolutionarily disfavored, slowing the rate of amino acid sequence change. We call this the "translational efficiency" hypothesis.

We advance a third hypothesis based on a simple observation: to reduce the number of proteins which misfold due to translation errors, selection can act both on the nucleotide sequence, to increase translational accuracy by optimizing codon usage[110], and on the amino acid sequence, to increase the number of proteins which fold properly despite mistranslation (Fig. 5.3). We call this increased tolerance for translational missense errors "translational robustness." At the canonical ribosomal error rate of five errors per 10,000 codons translated[128], approximately 19% of average-length yeast proteins (415 amino acids) contain a missense error, and these errors may cause misfolding[129]. Proteins vary in their tolerance for amino acid substitutions[11], providing the necessary raw material for

evolution, while misfolded-protein aggregation and toxicity[129,130] and production of non-functional protein[131] impose burdens on most cellular metabolisms, providing selective pressure. So long as translationally robust sequences are comparatively rare, an assumption we examine in detail in Chapter 6, intensified selection pressure resulting from increased expression level will slow the rate of amino acid substitution in higher-expressed proteins.

These three hypotheses differ in important ways. The functional loss hypothesis points to loss of protein function as the key cost constraining evolution. The translational efficiency hypothesis states that the protein sequence is constrained as a side effect of selection on the mRNA sequence. And the translational robustness hypothesis instead implicates the direct costs of misfolded proteins, independent of function. These hypotheses make testable and opposing predictions, which we now consider.

Functional loss versus translational robustness

Given two proteins with differing abundances $A > a$, measured in protein molecules per cell, but oppositely differing expression levels $x < X$, measured in mRNA molecules per cell, the functional loss hypothesis predicts $\mathrm{dN}_{Ax} < \mathrm{dN}_{aX}$: the more abundant protein will evolve slower. By contrast, the translational robustness hypothesis states that fitness costs are dominated by translation-error-induced misfolding, leading to the opposite prediction ($\mathrm{dN}_{Ax} > \mathrm{dN}_{aX}$), because despite *Ax*'s higher abundance, *aX*'s higher expression level suggests more frequent translation and turnover[132].

We tested these competing predictions using a recent global analysis of protein abundance in yeast[103]. Ten thousand unique pairs of yeast proteins for which one member had a higher expression level and a lower abundance than the other were assembled at random. In 5,579 of 10,000 pairs, the more abundant but lower-expressed protein evolved faster ($\mathrm{dN}_{Ax} > \mathrm{dN}_{aX}$, $P << 10^{-9}$, binomial test) consistent with translational robustness but contradicting the functional loss hypothesis. When we sampled pairs with at least a twofold difference in each measure, limiting the influence of measurement noise, 5,430 of 10,000 pairs showed the same pattern ($P << 10^{-9}$). Among synteny-derived paralog pairs,

25 of 48 showed the same pattern (not significant), as did 7 of 8 pairs with twofold differences ($P < 0.05$). Using CAI as an expression proxy (see *Methods*), 6,262 of 10,000 pairs ($P << 10^{-9}$) and 17 of 20 paralog pairs ($P < 0.002$) also showed the same pattern. These results suggest that the number of translation events, a correlate of expression level and CAI, is a better predictor of relative protein evolutionary rates than the number of functional protein molecules, a suggestion in accordance with the results obtained in Chapter 4 by principal component regression.

The functional loss hypothesis rests on the supposition that protein molecules contribute roughly the same amount to organism fitness through their biological function, so that less-abundant proteins are less important to organism fitness. We find this assumption difficult to accept on biochemical grounds. Protein abundance seems to depend mainly on substrate or target availability, which has no obvious relationship to fitness contribution. For example, most gene regulatory proteins and DNA polymerases have only a few hundred targets and correspondingly low cellular abundances, yet play crucial cellular roles. While cells seem unlikely to invest in synthesis of high-abundance proteins without a comparably high return, the inference that low-expression proteins generate low fitness returns does not follow. Accordingly, under the functional loss hypothesis, we should expect low-expression proteins to span the range of evolutionary rates, while high-expression proteins evolve under a more uniformly tight constraint. Instead, in yeast, the slowest-evolving low-expression proteins evolve an order of magnitude more rapidly than their highly expressed counterparts (Fig. 5.1a). This pattern again supports translational robustness, which supposes that, while folded proteins may confer widely varying fitness benefits, misfolded polypeptides impose similar costs.

Translational efficiency versus translational robustness

Pressure to retain translationally efficient preferred codons will constrain synonymous evolution (dS) and, as a consequence, protein evolution (dN). Pressure for translationally efficient amino acids[84] would bias amino acid preferences at aligned positions in high- and low-expression paralogs. By contrast, translational robustness predicts that the dS and dN

constraints reflect two independent points of selection (Figure 5.3) and that no consistent translational preference for either codons or amino acids is required to explain the dN trend.

To assess the protein-level constraint attributable to selection for preferred codons, which is strongest at functionally important and conserved sites[110], we computed evolutionary rates using the portions of genes consisting only of unpreferred codons. Because those sites most constrained by codon preference are removed in these reduced genes, the codon preference hypothesis predicts that the correlation of expression level with dS and dN should vanish. Translational robustness hypothesizes a direct constraint on the amino acid sequence, so the dN–expression correlation should remain strong while the dS–expression correlation vanishes, essentially an impossibility if synonymous-site selection for translational efficiency governs protein evolution. Using sets of aligned *S. cerevisiae*–ortholog genes (see *Methods*), we discarded all aligned codons except those where the "relative adaptiveness"[133] of the *S. cerevisiae* codon was less than 0.5. We then recomputed dN, dS and their expression correlations using these reduced genes, discarding genes with fewer than 30 codons or dS values of 3.0 or larger.

Table 5.1 shows that after removal of preferred codons, the reduced genes showed only slightly reduced dN–expression correlations, while the dS–expression correlations all became insignificant or, in the case of *S. paradoxus*, reversed direction. We found similar results using CAI as an expression proxy (Table 5.S2). These results demonstrate that expression-linked synonymous selection is concentrated at sites bearing preferred codons and that sites showing no such selection still show strong protein-level constraint, consistent with selection for translational robustness.

Translational efficiency selection on amino acids predicts asymmetric substitution of one amino acid for another in highly expressed proteins. If two amino acids x and y have efficiencies $x < y$, then at aligned positions in paralogs where both x and y occur, y should disproportionately appear in the higher-expressed paralog. We tabulated these pairwise frequencies in the 580 paralogs analyzed in Figure 5.1 and assessed statistical significance using a binomial test with the false-discovery-rate correction for multiple tests[134]. All

residue pairs appeared in our data set, but no pairs showed asymmetries at the 1% or 5% levels.

As a control, we performed the same test using synonymous codons, and found that 21 codon pairs showed significant asymmetries at the 1% level, invariably favoring the codon with higher relative adaptiveness in the higher-expressed paralog (Table 5.S1). Of the 21 favored codons, 17 were unique and encoded 13 of the 18 amino acids with synonymous codons.

Our results offer no support for translational efficiency selection on amino acids, but confirm such selection on synonymous codons, though with little consequence for dN. Although translational efficiency selection may constrain amino acid sequences to some degree, it cannot explain why highly expressed yeast proteins evolve slowly.

Expression level is a master causal variable

We now consider the possibility that many variables (e.g., dispensability, number of protein-protein interactions, amino acid biosynthetic cost, codon preference, recombination rate) independently exert small but cumulatively severe constraining effects on protein sequence evolution, and expression level's influence derives from its relationships to each of these variables. While such a possibility cannot be ruled out, several observations make it unlikely. First, expression level is a major determinant of most of the candidate variables: high expression causes decreased dispensability[102], causes more experimentally detected interactions[104], increases pressure for cheaper proteins and higher translational efficiency[101], and, through increased transcription, causes exposed chromatin structures that are hotspots for recombination. No reverse mechanisms have been proposed by which these variables cause genes to become highly expressed. Second, as noted in Chapter 4, the degree to which variables not linked to translation appear to influence evolutionary rate vanishes after controlling for expression level.

Discussion

We have provided evidence that expression level is the dominant determinant of evolutionary rate in *S. cerevisiae* genes. Our results show that *i*) expression level explains roughly half the variation in gene evolutionary rates; *ii*) expression level affects evolutionary rates of duplicated and singleton genes similarly; *iii*) once variability in expression level is accounted for, the higher-expressed member of a paralog pair is disproportionately likely to evolve slower; *iv*) asymmetric evolutionary rates in duplicated genes persist over tens of millions of years, consistent with expression-level differences but not neofunctionalization; and *v*) expression level appears to influence evolutionary rate through the number of translation events rather than cellular protein abundance, constraining the protein sequence directly rather than through translational efficiency selection.

We have introduced a general hypothesis to explain why highly expressed proteins evolve slowly: selection against the expression-level-dependent cost of misfolded proteins favors rare protein sequences which fold properly despite translation errors (Fig. 5.3). Tests comparing the opposing predictions of this translational robustness hypothesis to two previously advanced alternative hypotheses show that genome-wide yeast data support the predictions of translational robustness and contradict the alternatives. Our hypothesis contradicts the intuitive notion that highly expressed proteins evolve slowly because they are more functionally important, perhaps explaining why more direct measures of functional importance, such as essentiality and dispensability, explain far less variation in evolutionary rates. The hypothesis also provides an explanation for the widely observed correlation between dN and dS[123]: Figure 5.3 indicates how one cost (misfolding) can be counteracted in two ways (translational accuracy, slowing dS, and translational robustness, slowing dN).

Would more translationally robust proteins have a higher functional density[85]? Consider URA5 and URA10 (orotate phosphoribosyltransferases 1 and 2), paralogs with similar functions which differ 60+-fold in expression and 6-fold in evolutionary rate. Do we

expect URA5 to have a larger proportion of its residues involved in specific functions? The translational robustness hypothesis suggests not. Instead, functionally unconstrained residues may be more carefully selected to preserve the protein's native structure after missense substitutions in URA5 than in URA10. These residues would contribute to fitness not by aiding in URA5's function, but by preventing the burdensome misfolding of mistranslated polypeptides. Thus, the fitness density of a protein, the proportion of residues under meaningful natural selection, can be larger than the functional density, and directly determines the rate of sequence evolution.

Functional constraints slow evolution at certain sites; our results suggest that these constraints operate on a sequence-wide background rate determined largely by expression. Expression patterns as well as levels may impose additional constraints if highly expressed proteins have unique cellular localization or cell-cycle expression profiles.

How large are the costs underlying translational robustness? We can make a crude general estimate. As mentioned above, roughly 19% of average-length yeast proteins will contain a missense error at typical ribosomal error rates. For diverse proteins, 20–65% of amino acid substitutions lead to inactivation[11,64], generally due to misfolding[11]. Consequently, 4–12% of a typical protein species would be expected to misfold due to missense errors. Because yeast protein abundances span five orders of magnitude[103], the fitness impact of error-induced misfolding could range widely. If we assume a 5% misfolding rate, the number of misfolded protein molecules ranges from negligible, as for the ~3 misfolded molecules to generate the measured cellular complement of 64 molecules of DSE4 (endo-1,3-β-glucanase), to potentially devastating, as for the ~63,000 misfolded molecules required to generate 1.26 million molecules per cell of the H^+-transporting P-type ATPase PMA1[103]. The latter misfolded species would be more abundant than 97% of yeast proteins[103]. We have neglected protein turnover, a further cost multiplier. (We have also neglected the misfolding of error-free proteins; a likely biophysical mechanism for increasing translational robustness will also mitigate stochastic misfolding [see below].) Protein misfolding generates highly toxic species capable of killing cells in a concentration-

dependent manner[135], so increased translational robustness in highly expressed proteins may reflect pressure for survival as well as efficiency.

Can selection for accuracy through codon preference eliminate (or make negligible) such error-induced misfolding costs? While codon preference cannot counter mistranslation due to misacylation of tRNAs and transcription errors, both of which occur at frequencies approaching those of missense errors[128], experimental measurements of a 4- to 9-fold reduction in missense errors from preferred codons have been reported[136]. Assuming all preferred codons are translated 10-fold more accurately than nonpreferred codons, how much accuracy improvement can we expect? Randomly selecting codons produces genes containing ~35% preferred codons, while the most highly expressed genes have >80% preferred codons (only 9 of the 4,255 yeast genes we analyzed contain >90% preferred codons). Even if translational error-rate measurements reflect the worst case of codon-randomized genes, the maximum accuracy gain in the most optimized genes is roughly five-fold. In the case of PMA1 (86% preferred codons), such a reduction would still leave thousands of misfolded proteins from this single gene to burden the cell. While that level of misfolding may represent the "cost of doing business" for the cell, such an argument assumes that mutant versions of PMA1 carried by evolutionary competitors tolerate equivalent numbers of translation errors and generate similar costs. Because a protein's tolerance to substitutions can in some cases be significantly altered with a single mutation[11], we suspect this assumption is rarely justified. Given variability in misfolding, natural selection will then favor those mutants whose costs undercut their competitors'.

A counterintuitive prediction of the translational robustness hypothesis is that selection for proteins that are more tolerant to amino acid change yields underlying genes that appear less tolerant to nucleotide change (because they evolve slowly). How is this result possible? Consider a hypothetical allele of PMA1 for which only 0.1% (~1,000 molecules) of translated proteins misfold due to errors. A nonsynonymous genetic mutation yielding a functionally equivalent mutant protein that misfolds 5% of the time, producing ~50,000 potentially toxic proteins, would be evolutionarily disfavored relative to the wildtype due to increased misfolding costs without showing any functional difference. Thus, the wildtype,

despite encoding a highly robust protein which retains function after most mutations, will appear mutationally fragile over evolutionary time. A striking example of this robust-molecule/fragile-gene behavior may be found in ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco), perhaps the most abundant protein on Earth and a rigidly conserved, generally essential enzyme for which genetic studies have nonetheless been hampered by the difficulty of finding inactivating missense mutations[137].

How might translational robustness manifest itself biophysically? We can offer a speculation which we return to in Chapter 6. Because most substitutions destabilize the native structure of a protein (cf. Chapter 1), modest increases in thermodynamic stability broaden the spectrum of substitutions a protein can tolerate before misfolding[11], increasing fitness so long as function is not compromised. Pressure for increased stability in highly expressed proteins would restrict the set of evolutionarily viable sequences[24] and slow sequence evolution as a consequence.

Materials and Methods

Gene sequences

Genome sequences for *Saccharomyces cerevisiae*, *S. kudriavzevii*, *S. paradoxus*, *S. mikatae* and *S. bayanus* were obtained from the *Saccharomyces* Genome Database (ftp://genome-ftp.stanford.edu/). The genome sequence of *Kluyveromyces waltii* was obtained from ref. [124], supplemental information.

Identification of orthologs and paralogs

900 paralogous *S. cerevisiae* genes identified by synteny[124] were downloaded. Of these pairs, 290 (580 genes) were non-ribosomal proteins with a measured expression level[138] and an ortholog in *S. bayanus*, and were used in our analysis. We excluded ribosomal proteins from all analyses because they tend to be highly expressed and slow-evolving and could skew our results.

Orthologs for *S. cerevisiae* genes in members of the *Saccharomyces* genus were found by the reciprocal shortest distance (RSD) algorithm[77] with a protein-protein BLAST[51] *E*-value cutoff of 10^{-20}, 80% minimum alignable residues, and distances computed as the number of nonsynonymous substitutions per site, dN, using PAML (see below). RSD yielded 4,255 non-ribosomal *S. cerevisiae* genes with *S. bayanus* orthologs and a measured expression level, 2,790 genes with *S. mikatae* orthologs, 4,407 with *S. paradoxus* orthologs and 2,984 with *S. kudriavzevii* orthologs. The *S. paradoxus* ortholog set was expanded to include *S. cerevisiae* matches reported by Kellis *et al*.[127]

Expression level data

We used gene expression data measured in mRNA molecules per cell by Holstege *et al.* [138]. To estimate variability in expression level data, we used normalized fluorescence data collected using the same commercial oligonucleotide array by Cho *et al.*[114] with mean expression levels computed as described[108]. Because laboratory growth media and temperatures may not reflect evolutionarily relevant environmental conditions, potentially

distorting expression profiles, we repeated all analyses using each gene's codon adaptation index (CAI)[133] as an expression-level proxy[89] (see Supplemental Material and Figure 5.S1). We assume that species closely related to *S. cerevisiae* have similar expression profiles.

Measurement of evolutionary rates

Orthologous gene alignments were constructed from protein sequences aligned using CLUSTAL W[139]. The number of nonsynonymous and synonymous substitutions per site, dN and dS, were estimated by maximum likelihood using the PAML[140] program `codeml` operating on codons with a 60-free-parameter model for codon frequencies.

Statistical analysis

We used R[31] for statistical analysis and plotting. To compute correlations on log-transformed dN data, we applied the transformation $f(k) = \log(k + 0.001)$ as in a previous study[89] to avoid excluding zeros.

Supplemental Material

We repeated each of our analyses using a gene's codon adaptation index (CAI) as a proxy for its expression level, allowing us to expand the coverage of our tests and to eliminate the dependence of expression measurements on particular growth conditions. Figure 5.S1 shows that all the trends we identified in Figure 5.1 remain highly significant using the CAI proxy. For the 325 pairs with *S. bayanus* orthologs, 224 of the higher-CAI paralogs evolved slower than their lower-CAI counterpart ($P << 10^{-9}$, binomial test). Table 5.S1 demonstrates that elimination of preferred codons obliterates the negative correlation between CAI and dS across multiple species.

As discussed in the main text, we examined asymmetries between the frequencies of synonymous codons (x and y) at aligned positions in two paralogs. We counted the number of times x appeared in the lower-expressed paralog while y appeared at the aligned position

in the higher-expressed paralog, #(x, y), and the number of times y appeared in the lower-expressed paralog while x appeared at the aligned position in the higher-expressed paralog, #(y, x). Deviations from chance were assessed by the binomial test with the false-discovery-rate correction for multiple tests [134]. Codons favored at the 1% level are reported in Table 5.S2. In all cases, the codon with higher relative adaptiveness [133] is favored in higher-expressed paralogs. At the 5% level, 38 significant pairs were found, corresponding to 25 unique favored codons which encode 15 of 18 amino acids with synonymous codons. At either significance level, no amino acid pairs showed asymmetries when subjected to the same test.

Table 5.1: Evolutionary rate vs. expression correlations (Kendall's τ) relative to four yeast species for *S. cerevisiae* genes, including and excluding preferred codons.

	All codons		Codons with relative adaptedness < 0.5	
Ortholog (# of genes)	dN–expr.[†]	dS–expr.	dN–expr.	dS–expr.
S. bayanus (2,614)	$\tau = -0.300$***	$\tau = -0.181$***	$\tau = -0.273$***	$\tau = -0.010$
S. mikatae (2,102)	$\tau = -0.335$***	$\tau = -0.163$***	$\tau = -0.302$***	$\tau = -0.009$
S. paradoxus (4,383)	$\tau = -0.340$***	$\tau = -0.153$***	$\tau = -0.303$***	$\tau = +0.046$**
S. kudriavzevii (2,193)	$\tau = -0.340$***	$\tau = -0.162$***	$\tau = -0.314$***	$\tau = -0.004$

[†]Significance codes: *, $P < 10^{-2}$; **, $P < 10^{-4}$; ***, $P < 10^{-6}$.

Table 5.S1: Evolutionary rate vs. CAI correlations (Kendall's τ) relative to four yeast species for *S. cerevisiae* genes, including and excluding preferred codons.

	All codons		Codons with relative adaptedness < 0.5	
Ortholog (# of genes)	dN–CAI[†]	dS–CAI	dN–CAI	dS–CAI
S. bayanus (2,613)	$\tau = -0.268$***	$\tau = -0.096$***	$\tau = -0.233$***	$\tau = +0.099$***
S. mikatae (2,108)	$\tau = -0.321$***	$\tau = -0.050$*	$\tau = -0.281$***	$\tau = +0.107$***
S. paradoxus (4,656)	$\tau = -0.326$***	$\tau = -0.068$***	$\tau = -0.277$***	$\tau = +0.146$***
S. kudriavzevii (2,340)	$\tau = -0.281$***	$\tau = -0.050$*	$\tau = -0.245$***	$\tau = +0.112$***

[†]Significance codes: *, $P < 10^{-2}$; **, $P < 10^{-4}$; ***, $P < 10^{-6}$.

Table 5.S2: Significant asymmetries in synonymous codon usage between high- and low-expressed paralogs at aligned positions reflects relative adaptedness.

Codon		Amino	#(low,high)		$P^{\dagger}$	Rel. adaptedness	
x	y	acid	#(x,y)	#(y,x)		x	y
GCA	GCC	A	281	215	*	0.015	0.316
GCA	GCT	A	479	312	***	0.015	1.000
GAG	GAA	E	1081	869	**	0.016	1.000
GGC	GGT	G	463	350	*	0.020	1.000
GGG	GGT	G	306	187	**	0.004	1.000
GGA	GGT	G	552	346	***	0.002	1.000
CAT	CAC	H	391	294	*	0.245	1.000
ATA	ATC	I	364	266	*	0.003	1.000
AAA	AAG	K	1315	1130	*	0.135	1.000
CTT	TTA	L	314	241	*	0.006	0.117
TTA	TTG	L	855	730	*	0.117	1.000
AAT	AAC	N	972	830	*	0.053	1.000
CCT	CCA	P	554	433	*	0.047	1.000
CCG	CCA	P	254	172	*	0.002	1.000
AGG	CGT	R	74	42	*	0.003	0.137
AGG	AGA	R	511	377	**	0.003	1.000
ACA	ACT	T	413	315	*	0.012	0.921
GTA	GTT	V	309	236	*	0.002	1.000
GTA	GTC	V	171	117	*	0.002	0.831
GTG	GTT	V	322	212	**	0.018	1.000
TAT	TAC	Y	886	692	**	0.071	1.000

†Binomial probability with false-discovery-rate correction for multiple tests. Significance codes: *, $P < 10^{-2}$; **, $P < 10^{-4}$; ***, $P < 10^{-6}$.

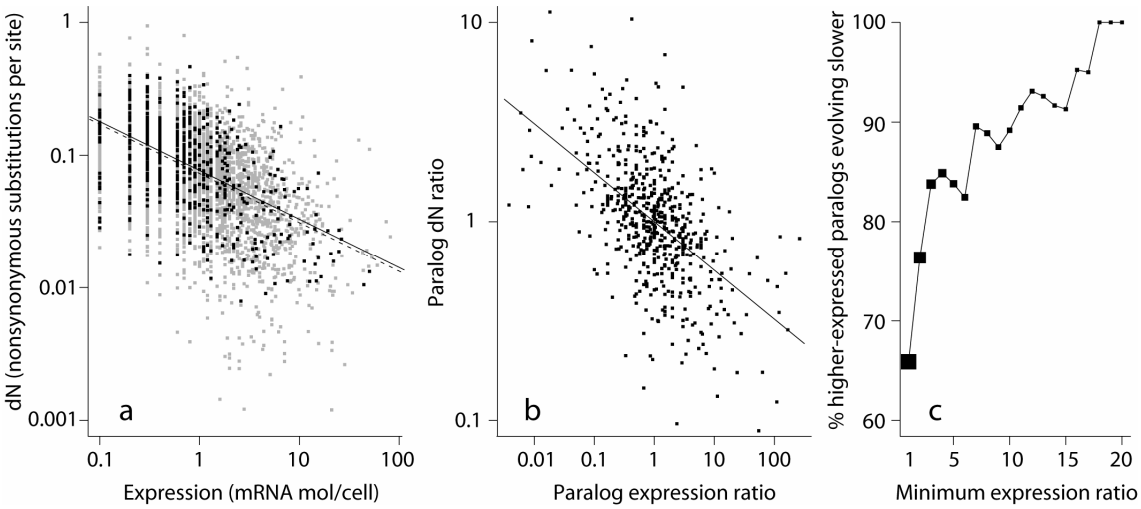


Figure 5.1. Expression level governs gene and paralog evolutionary rates in *S. cerevisiae*. **a**, Highly expressed proteins evolve more slowly, and paralogs mirror the genome-wide pattern. Evolutionary rates measured relative to *S. bayanus* for 4,255 *S. cerevisiae* genes (▒) and 580 paralogous genes (■) correlate with expression levels. Lines show best log-log linear fit. For all genes (dotted line), $r^2 = 0.28$, $P << 10^{-9}$; for paralogs (solid line), $r^2 = 0.31$, $P << 10^{-9}$. **b**, Within a paralog pair, the ratio of expression levels correlates with the ratio of evolutionary rates ($r^2 = 0.29$, $P << 10^{-9}$), as predicted from the log-log linear relationship in **a**. Each pair generates two ratio points, making the plot symmetrical. **c**, Relative expression level determines relative evolutionary rate. The percentage of pairs in which the higher-expressed paralog evolves slower are shown as a function of minimum paralog pair expression ratio (■). Point areas are proportional to the number of included pairs.

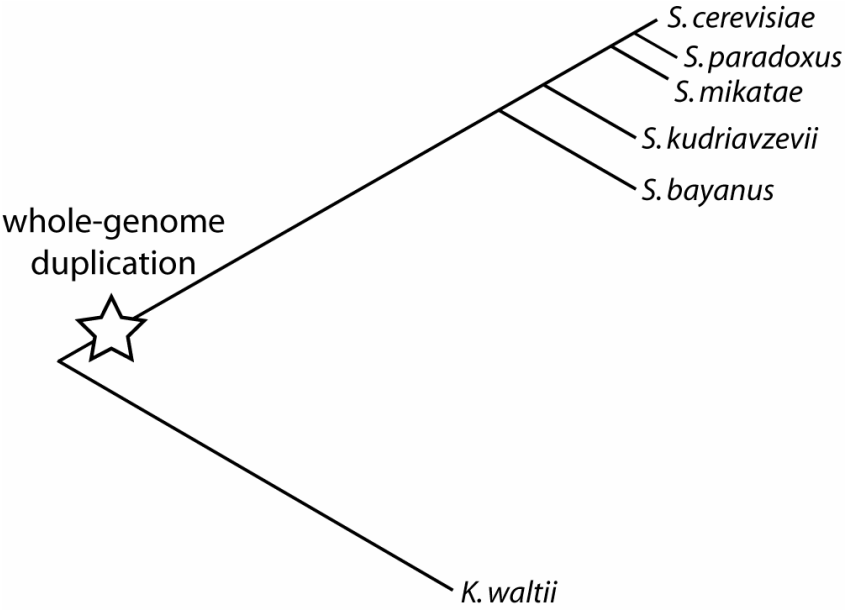


Figure 5.2. Phylogenetic relationships between analyzed yeast species. Relationships follow ref. [141], branch lengths indicate nucleotide substitution distances from ref. [79], and the indicated time of the whole-genome duplication follows ref. [124].

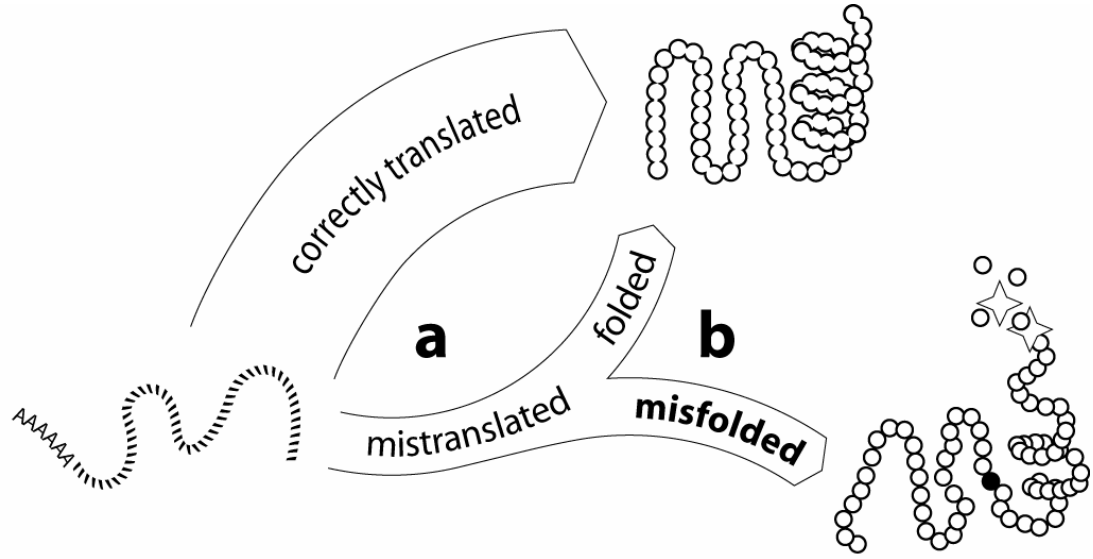


Figure 5.3. Translational selection against the cost of misfolded proteins can act at two distinct points. Messenger RNA (left) may be translated without errors to produce a folded protein (top); if an error is made, the resulting protein may still fold properly, or may misfold and undergo degradation (right). Selection can act at **a** to increase the proportion of error-free proteins through codon preference (translational accuracy), and also at **b** to increase the proportion of proteins that fold despite errors (translational robustness). We neglect misfolding of error-free proteins (see text).

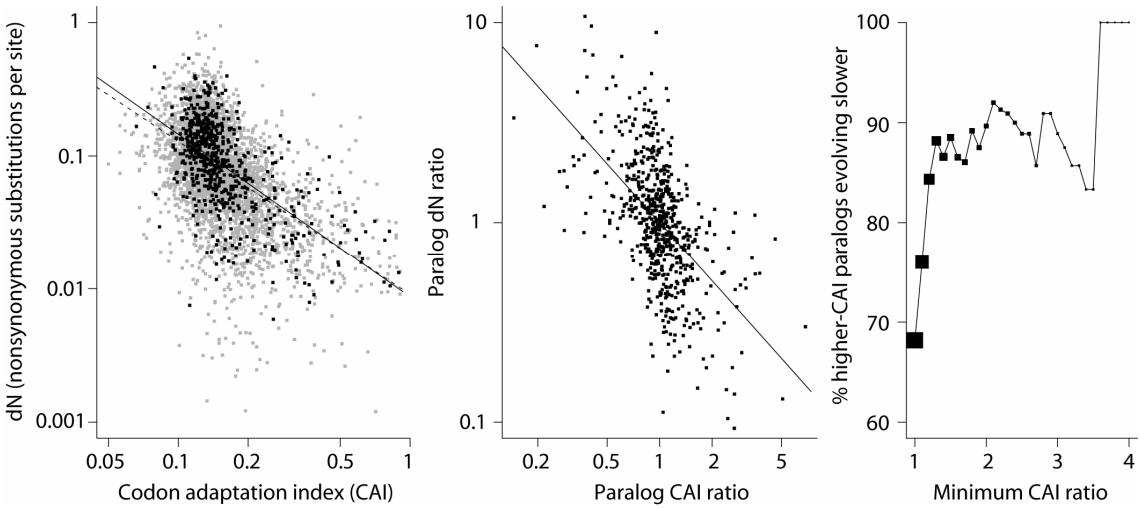


Figure 5.S1. Estimating expression level with the codon adaptation index (CAI) reveals evolutionary rate relationships similar to those found using more direct microarray measurements. **a**, Highly expressed proteins evolve slowly, and paralogs mirror the genome-wide pattern. Evolutionary rates measured relative to *S. bayanus* for 4,534 *S. cerevisiae* genes (■) and 650 paralogous genes (■) correlate with CAI. Lines show best log-log linear fit. For all genes (dotted line), $r^2 = 0.27$, $P << 10^{-9}$; for paralogs (solid line), $r^2 = 0.38$, $P << 10^{-9}$. **b**, Within a paralog pair, the ratio of expression levels correlates with the ratio of evolutionary rates ($r^2 = 0.31$, $P << 10^{-9}$), as predicted from the log-log linear relationship in **a**. Each pair generates two ratio points, making the plot symmetrical. **c**, Relative CAI governs relative evolutionary rate. The percentage of pairs in which the higher-expressed paralog evolves slower are shown as a function of minimum paralog pair CAI ratio (■). Point areas are proportional to the number of included pairs.

C h a p t e r 6

MISFOLDING DOMINATES GENOME EVOLUTION

Know when to fold 'em.

Kenny Rogers

Summary

Mistranslation generates misfolded proteins[129] which form inherently toxic aggregates,[129,135] leading to natural selection against misfolding. Extensive retrospective evidence suggests that expression level determines a large proportion of evolutionary rate variation between proteins[82,83,93,94] implicating selection on translation acting before a protein becomes functional[93] (cf. Chapters 4 & 5). To study prospectively the evolutionary effects of mistranslation-induced protein misfolding, we evolved a population of simulated organisms whose genomes consisted of hundreds of genes encoding model proteins expressed at levels spanning four orders of magnitude. Protein misfolding imposed the only fitness cost. Strikingly, a large number of previously studied intergenic patterns arose, from major trends (highly expressed proteins evolve slowly) to more subtle relationships (synonymous and nonsynonymous evolution evolutionary rates are correlated), which matched all those observed in the yeast *Saccharomyces cerevisiae* with high accuracy. Our model allowed us to trace the cause of slowed protein evolution to selection for translational robustness. Contrary to basic intuitions from Chapters 1–3, but consistent with predictions derived from genomic data in Chapter 5, we find the slowest-evolving genes have the highest tolerance for mutations. On the basis of simulation results, we predict a novel trend linking intragenic evolutionary rate correlations to expression level which we subsequently find in the *S. cerevisiae* genome. Our results unify multiple

disparate evolutionary trends, provide explanations for several puzzling relationships, and confirm widely credited but largely untested theories in molecular evolution.

Introduction

Why do highly expressed proteins evolve slowly? As suggested in Chapters 4 and 5, extraordinary consistency in genomic trends argues against multiple independent pressures (such as breadth of tissue expression,[97,117] which does not apply to microbes[82]), and favors a general force that can operate across taxa. Protein misfolding burdens all life, as evidenced by the universally conserved heat-shock response,[142] and misfolded proteins of all stripes form inherently cytotoxic aggregates.[135] All organisms produce some misfolded protein during translation because the fidelity of the ribosome has limits[143] and errors in proteins often cause misfolding.[129] Selection against mistranslation-induced protein misfolding is therefore a force with sufficient generality to explain the dependence of evolutionary rate on expression level across taxa.

The connection between expression-linked translational selection and evolutionary rate, and the plausibility of its competing explanations, rests upon correlational evidence. Experimental studies have not been forthcoming, for reasons that are not difficult to understand. More than fifty years after the discovery of the ribosome, despite the efforts of multiple groups[128,136,144-151], we possess estimates of translational accuracy at only a handful of codons, rarely by a consistent protocol, frequently in starving cells[128], and only for errors that have unusual properties (*e.g.*, arginine-to-cysteine in cysteine-free bacterial flagellin[150]). Because misfolded proteins have short half-lives, protein misfolding upon mistranslation is exceedingly difficult to measure[129] and considerable disagreement remains regarding even the bulk amount of defective protein that undergoes rapid degradation[152-154]. Turning to the fitness effects of mistranslation, experimental studies remain anecdotal[145,155], with the exception of the general observation that hyper-accurate mutants grow slowly and are rarely found in nature[145]. Measuring growth-rate fitness is of limited utility in examining questions of translational selection: evolution can easily resolve fitness differences invisible to our assays. Compound all the above difficulties with the slow and stochastic nature of the evolutionary trends in question, which necessitates observation of large ensembles of genes over long intervals, and it is tempting to ask whether a different approach can create insight into the source of these trends.

Synthetic biology[156] provides an attractive alternative for two main reasons. First, synthetic evolution experiments have none of the limitations listed above[157], in principle allowing access to every mutation at every site along the line of descent, accurate fitness measurements, true replicates, and, crucially, the power to observe evolution over millions of generations. Second, even for the most widely accepted explanations in translational selection (codon usage biased for translational accuracy[110], expression-linked reduction in synonymous substitutions due to codon bias[158]), it is not known whether the observed trends actually follow from their supposed causes, a deficit easily addressed by a synthetic approach.

We therefore endeavored to construct a model system in which mistranslation of genes expressed at different levels generated costly misfolded proteins, to see if trends inferred from genomic data would arise and, if so, to examine causality in a way presently impossible with biological systems.

We constructed a large-scale simulation in which a population of 1,000 organisms evolved for many generations. These organisms possessed genomes consisting of 650 coding nucleotide sequences (genes) expressed at 13 different levels spanning four orders of magnitude. Each gene was essential and encoded a model polypeptide capable of thermodynamically driven folding. Protein misfolding imposed the only fitness cost, either through the lethal loss of wildtype folding or the growth-rate burden of mistranslation-induced misfolding (Fig. 6.1). The simulation, parallelized according to a simple scheme, proceeded for a total of 97.5 million generations, until each gene had experienced 150,000 generations of evolution.

To simulate regulated expression, polypeptides translated from a gene were folded until a target number of folded proteins, the gene's expression level, was obtained. Error-free mRNAs were translated at an error rate producing missense errors in 15% of low-expression proteins, approximating the per-protein error rate inferred for average-length yeast proteins.[94] The translation error spectrum was implemented as described[159] such that only single-nucleotide misreading errors occurred, from most- to least-frequent, at the third,

first, and second codon positions; we neglected frameshifts. To model the 4- to 9-fold difference in translational accuracy observed between codon synonyms,[136] codons designated optimal for yeast[160] were translated 6-fold more accurately.

To model protein folding and misfolding, simulated genes encoded short (25-residue) polypeptides which fold to a lowest-free-energy maximally compact structure on a square lattice. Side-by-side experiments and simulations have established that these lattice proteins are an accurate and tractable model of relevant trends in protein thermodynamics and mutational tolerance[11,14,19] and they allow the rapid, exact folding necessary to make long-run evolution possible. If a polypeptide adopted the natively encoded wildtype structure as its lowest free-energy conformation with a free energy of unfolding ΔG of at least 5 kcal/mol, it was designated properly folded, and misfolded otherwise. The entire simulation required folding approximately 10^{10} proteins, and was repeated five times under various conditions.

The likelihood of reproduction was proportional to organism fitness (Wright-Fisher sampling). To assess fitness, we imposed the following constraints: 1) equal changes in the amount of misfolded protein must produce equal fitness disadvantages s (simulating nonspecific toxicity); 2) fitness is a monotonically decreasing function of the amount of misfolded protein; 3) the fitness associated with no misfolding is 1 (arbitrary scaling). Only one fitness function satisfies these constraints (Box 6.1):

$$\text{fitness}(m) = e^{-cm}, \tag{6.1}$$

where c is a positive constant and m is the amount of misfolded protein. We chose $c = 0.001$ as a convenient reference so that the population-size-scaled fitness disadvantage $N \times s = -1$ when one additional protein misfolds. The number of misfolded proteins m generated while expressing x folded proteins was estimated by folding all possible polypeptides generated by translation and weighting each outcome by its probability to obtain the mean fraction folded p_{folded}; then $m = x\,(1-p_{\text{folded}})/p_{\text{folded}}$. This estimate of the amount of mistranslation-induced misfolding was highly accurate (see *Methods* and Fig.

6.S1) and was crucial in making a biologically relevant range of expression levels computationally feasible.

After evolution, we tabulated various commonly used quantities, including the number of nonsynonymous and synonymous substitutions per site along the line of descent to the most-recent common ancestor (evolutionary rates dN and dS, often denoted K_a and K_s), their ratio (dN/dS, often denoted ω) and the fraction of optimal codons per gene (F_{op}). Figure 6.2 compares genome-wide trends from the yeast *Saccharomyces cerevisiae* with our simulation and demonstrates a striking correspondence. In both natural and simulated genome evolution, higher expression was accompanied by slower protein evolution (fewer substitutions per nonsynonymous site, dN), slower synonymous evolution (dS), a decreased dN/dS ratio, and strongly biased codon usage.

Other relationships between these five variables, some not involving expression level, have been noted, and some remain unexplained. For example, in *Drosophila*, dN and dS correlate,[161] as do F_{op} and dN[123]; in humans, dN/dS correlates with dS.[162] We computed the correlation matrix for dN, dS, dN/dS, F_{op} and expression level, obtaining all 10 pairwise correlations for yeast and 10 for our simulation, all highly significant. In every case, our simulation produced correlations of the proper sign and magnitude which linearly correlated with those of yeast, $r = 0.98$, $P < 10^{-6}$ (Fig. 6.3). (The three unexplained correlations also appeared in both yeast and our simulation.) We then attempted to understand how and why these biological trends arose in the simulation, taking advantage of our access to the entire lineage.

By tracking the fates of 10,000 individual polypeptides translated from each of the evolved genes, we could dissect of adaptations to increased expression level. The rate of misfolding was modest in all cases, with averages ranging from 8.4% to 0.56% from lowest to highest expression level. We found that highly expressed genes counteracted protein misfolding costs in three main ways (Table 6.1): by increasing translational accuracy through biased synonymous codon usage, by reducing truncation errors, and by increasing translational robustness, the propensity of the encoded protein to fold properly despite mistranslation.[94]

Highly expressed proteins evolved high mutational tolerance, gaining the ability to withstand almost all mistranslation-induced substitutions without misfolding—93% of highest-expression proteins versus 43% of lowest-expression proteins folded properly despite mistranslation (Table 6.1). Yet they simultaneously appeared intolerant to mutations, accumulating nonsynonymous changes nearly an order of magnitude more slowly than their low-expression counterparts over evolutionary time. These paradoxical observations, recently predicted[94] and theoretically explored[163], demonstrate that increased tolerance to mutations, long thought to predict faster evolution, can do just the opposite after selection has acted.

We tested multiple hypotheses for why highly expressed proteins evolve slowly using data from our simulation. A major biological trend, reproduced by our simulations, properly focuses attention on proteins and not just genes: the ratio dN/dS is virtually always less than 1 (all but one simulated gene, that with the lowest dS) and declines with increasing expression level (Fig. 6.2c), revealing stronger protein- than nucleotide-level constraints that grow even more lopsided for highly expressed genes. This trend is particularly remarkable because dS also declines with expression level (Fig. 6.2b), a well-understood consequence of codon adaptation[111] (Fig. 6.2d) which, all else equal, would cause dN/dS to increase.

In our simulation, mutation rates and protein structure were held constant, and proteins experienced no functional pressures, excluding such pressures as causes for slowed protein evolution. We proposed that selection for rare translationally robust sequences which fold properly despite mistranslation constrains evolution in highly expressed proteins;[94] we observed such adaptation in this simulation (Table 6.1). We then repeated the simulation, this time preventing evolution of translational robustness by forcing all mistranslated proteins to misfold, resulting in selection for translational accuracy alone. Highly expressed proteins evolved slowly (dN declined with increasing expression level), but this time in a way inconsistent with biological data: dN/dS frequently exceeded 1 (39 genes) and did not change with expression level ($r = 0.02$, $P = 0.68$). We also repeated the entire simulation under conditions designed to produce selection for translational robustness

alone: a fixed proportion (85%) of polypeptides was forced to translate properly, and the encoded protein sequence was subjected to all single amino-acid substitutions to assess mistranslation-induced misfolding. Again, highly expressed proteins evolved slowly, but did not match the biology: dN/dS declined with expression level ($r = -0.76$, $P << 10^{-9}$) but the biological relationships between F_{op}, dS and expression (Fig. 6.2b,d) vanished.

In short, under the conditions we studied, any pressure against misfolding slowed dN at high expression levels, translational accuracy selection was responsible for recreating biologically observed trends in dS and F_{op}, and translational robustness selection was necessary and sufficient to recreate protein-level constraints reflected in dN/dS trends. These findings accord with a previous analysis in yeast which established that the dS–expression relationship (Fig. 6.2b) was limited to optimal codons but the dN–expression relationship (Fig. 6.2a) was not.[94] Both analyses support the view that synonymous evolution is slowed by codon bias and that nonsynonymous evolution reflects primarily a protein-level constraint, conclusions also reached independently for *Drosophila*.[98]

If translational accuracy selection on the nucleotide sequence suffices to slow protein evolution, but the biological data suggest a dominant protein-level constraint consistent with translational robustness, we may ask whether robustness selection dominated when both accuracy and robustness adaptations were possible. Since fitness in our simulation depends only on the fraction of misfolded (or, equivalently, of folded) proteins, and folded proteins are either accurately translated or fold despite mistranslation, translational accuracy and robustness are the only possible responses to selection. By comparing the accuracy and robustness of evolved genes to a large random sampling of genes encoding folded proteins, we can gauge the strength of selection on each adaptation when they act together. We generated at random 150,000 genes encoding folded proteins and recorded the distributions of translation outcomes in terms of fractions folded, accurate, and robust (folded despite mistranslation) (Fig. 6.4a). Random genes yielding a high fraction folded tended to display elevated accuracy and very high robustness (Fig. 6.4b). Using the random distributions to quantitatively estimate the strength of selection (see *Methods*) on traits of the evolved genes, we found that model genes were virtually always selected more

strongly for robustness than accuracy (Fig. 6.4c), with pressure 36-fold stronger in highly expressed model genes (1,000 molecules or more) (Fig. 6.4c).

We thus arrive at our central causal question: why is translational robustness, a protein-level constraint directly associated with slowed protein evolution, favored so heavily over translational accuracy to reduce mistranslation-induced misfolding? As neither adaptation is under direct selection in our model, any adaptive bias must reflect a bias in the underlying composition of sequence space: the selection strength on random genes, conditional on a low misfolding rate, skews toward robustness. Genes evolving neutrally, under pressure only to maintain a threshold fraction folded, should gravitate toward rare high-robustness sequences at the expense of high accuracy simply by seeking the means of their conditional distributions.

To test this hypothesis, we evolved the gene sequences obtained after selecting for translational accuracy alone (above) under the neutral constraint that they maintain a fraction folded of at least 0.975 (Fig. 6.4b), a modest level attained by virtually all model genes with expression levels of 1,000 proteins or more (Table 6.1). Figure 6.4d shows that the mean robustness and accuracy values indeed rapidly stabilize at the conditional-mean values derived from random sequences, confirming our prediction. The answer to our causal question, then, is that fundamental properties of the space of all genes encoding properly folded proteins determines the ultimate balance of robustness and accuracy, of protein-level and nucleotide-level constraints. Mutation-selection balance determines the acceptable rate of misfolding at a given expression level, and then neutral evolution determines the balance of accuracy and robustness, leading to a robustness-dominated constraint on the protein sequence.

We then turned our attention to two related open problems in evolutionary genomics: why are dN and dS correlated both between genes[123,161,164-166] and within them[167-170]? Both relationships arose in our simulations: a strong intergenic dN–dS $r = 0.62$, $P << 10^{-9}$, and a weak intragenic dN–dS $r = 0.11$, $P << 10^{-9}$. As the only non-random independent variable distinguishing simulated genes, expression level must mediate the intergenic relationship in

our model, and our results (cf. Figure 6.2) provide strong support for an intergenic dN–dS relationship arising from expression-dependent nucleotide-level pressure for translational accuracy and protein-level pressure for translational robustness.[94] (The partial correlation of dN with dS controlling for expression level remains highly significant, partial $r = 0.46$, $P << 10^{-9}$, a spurious result typical of this analytical method[93] [cf. Chapter 4].) Yet it is not obvious how expression can create an intragenic correlation, because all sites within a gene are expressed at the same level. While some workers have implicated effects linked to mutational biases[171] or positive selection[172], Bernardi and colleagues[170,173] found that elements of protein structure covaried with both nonsynonymous and synonymous substitutions: in the GP63 gene of *Leishmania*, residues in the metalloprotease core underwent fewer amino acid replacements, experienced fewer synonymous-site changes, and maintained higher usage of optimal codons than surface residues[173]. These observations prompted the hypothesis that dN and dS are linked through structurally constrained amino acids and synonymous sites selected for translational accuracy[169,173]. The relationship between substitutions and lattice protein structure in our simulations agreed with this pattern. Codons encoding core residues accumulated fewer than half the substitutions of surface residues, consistent with real proteins[174] (4,996 vs. 10,590 overall; 2,270 vs. 5,369 nonsynonymous; 2,726 vs. 5,221 synonymous). Substitutions were distributed nonrandomly over the genes in accordance with structural constraints (Figure 6.5a). Aggregating substitutions into surface or core categories produced a striking increase in the intragenic dN–dS relationship, surface $r = 0.53$, core $r = 0.49$, both $P << 10^{-9}$, indicating that stochastic variation was responsible for the seemingly weak relationship observed initially. If translational accuracy selection linked intragenic dS to dN, then their relationship should strengthen with increasing expression level, and it did: high-expression proteins showed clearer covariation along the sequence than their low-expression counterparts (Fig. 6.5b,c) and r_{dN-dS} correlated with expression level, $r = 0.22$, $P << 10^{-9}$, a relationship easily seen on average (Fig. 6.6a).

We reasoned that, while genes of mammals and other metazoa were used to establish most dN–dS relationships above, the hypothesized forces should apply equally well to yeast.

Moreover, our simulation allows us to predict that dN and dS will correlate within sequences and that this relationship will strengthen with expression level. To maximize the number of substitutions accumulated in the analysis, we collected 1,374 *S. cerevisiae* genes for which orthologs have been identified in six additional yeast species whose evolutionary tree is known[141] and estimated the number of synonymous and nonsynonymous substitutions for each codon over the entire tree[175]. These substitutions were then correlated to find r_{dN-dS} within each sequence. As predicted, r_{dN-dS} correlated with expression level, $r = 0.26$, $P << 10^{-9}$. When we aggregated r_{dN-dS} values by expression as before, the trend is not only striking (Fig. 6.6b), but quantitatively predicted by our simulation results. Both sets of data show a puzzling negative r_{dN-dS} on average at very low expression levels.

These results constitute strong evidence that selective pressure to preserve protein folding at the translational level creates the correlation between dN and dS within sequences. We predict that a relationship between dN and dS will reliably appear only in genes under significant selection for translational accuracy, and thus will be linked strongly to gene expression. Indeed, GP63 is a highly abundant cell-surface protein[176].

Our model also sheds light on an interesting issue in our own species. Recently, a "highly unexpected" correlation ($r^2 = 0.1$, $r \sim 0.32$) between the selective strength (K_a/K_s, equivalent to dN/dS) and K_s (dS) was reported among human genes with mouse orthologs,[162] inconsistent with current paradigms in coding sequence evolution.[162,177] This correlation arises in our simulations (dN/dS–dS $r = 0.13$, $P < 0.001$), and we know precisely why: both dN/dS and dS correlate negatively with expression level (Fig. 6.2b,c), because of translational selection for robustness and accuracy, respectively, and thus they correlate positively with each other. We suggest that this correlation should be expected, modulo the fragility inherent in its definition, in most organisms subject to translational selection. Indeed, we find it in yeast (Fig. 6.3). In humans, all the requisite forces are already known: translational selection also acts on synonymous sites[178] and highly expressed genes evolve slowly.[95,99] We hypothesize that a single force, translational

selection against the expression-level-dependent costs of protein misfolding, is sufficient to create all these relationships.

Our results support predictions that selective pressure favoring translationally robust proteins can result in indirect pressure for increased thermodynamic stability.[94,163] Previous studies have shown that in both lattice proteins and real proteins, increased stability confers increased tolerance of mutations.[11,12,19] The highest-expressed model proteins indeed evolved increased stability relative to the lowest-expressed (mean $\Delta G = 6.29$ versus 5.17 kcal/mol). Considering only evolved proteins possessing stabilities within a standard deviation of the low-expression mean ($\Delta G \leq 5.35$ kcal/mol), the proportion of proteins folding despite mistranslation still rose with expression level from 38% to 63% (compare to Table 6.1), indicating that high stability was beneficial but not required for increased translational robustness. Although in our model increased stability is virtually always beneficial, high stability only evolved under strong selective pressure at high expression levels, because highly stable proteins were otherwise too rare to persist even under low mutational pressure.[24] Whether stability competes intrinsically or merely statistically with biological activity remains a point of active research,[12,179] and our results suggest that differentially expressed paralogous proteins with similar biological activities may provide a vast set of test cases.

Our approach of evolving an entire simulated genome mutation by mutation, folding hundreds of millions of proteins over tens of millions of generations, transforms our evolutionary inquiry from a retrospective, comparative study into a prospective, exact one. For example, most genome evolution studies, including the present work on yeast, require sequence-conservation-dependent ortholog identification in another species and subsequent inference of evolutionary rates, leading to the possibility of multiple compounded methodological biases in reported evolutionary rate trends. By contrast, simulation permits inference-free recording of evolutionary rates; precise proteome-wide measurement of expression, misfolding and thermodynamic stability; identification of optimal codons by their translational accuracy rather than frequency biases; and perfect knowledge of replication and translational error rates. Under these conditions, the simulation

recapitulates and thus confirms much of the evolutionary biology derived from retrospective studies. As a simple example, our study is the first to demonstrate that selection favoring translational accuracy can in fact produce the qualitative expression-linked codon bias pattern (Fig. 6.2d), and the concomitant constraint on synonymous evolution (Fig. 6.2b), observed in a real genome.

It is reasonable to ask how details of our model might influence our findings. We employ a crude model of protein folding, a choice necessitated by the enormous number of folding events required for observing long-run evolution at a nontrivial population size. We believe the critical feature of this model is its accordance with biophysical data on mutational tolerance and thermodynamic stability,[11,12,14,19] and predict that other models with similar properties will produce similar results. Choice of fitness function might influence outcomes, but the biological assumptions whose consequences we study left no choice: they dictate the fitness function (Box 6.1). We show that the underlying distributions of accuracies and robustnesses play a pivotal role in shaping relative selective pressures against misfolding at the nucleotide and protein levels. These distributions are unknown for real genes but in our simulation are completely determined by the protein folding model and parameters drawn from biological measurements.

Our central result is that an extremely simple force (costly misfolded proteins) can produce, and thus explain, a large number of previously studied biological patterns imprinted on genomes across the tree of life, some linked to gene expression level (correlations of dN, dS, dN/dS and codon bias with expression) and some seemingly independent (correlations of dN, dS, dN/dS and codon bias with each other), both between and within gene sequences, all at the same time. Additionally, our simulation suggests a number of testable biological predictions for future genome- and proteome-wide studies: highly expressed proteins will succumb to mistranslation-induced misfolding less often than low-expression proteins; among related proteins, evolutionary rate will predict relative misfolding rates and thermodynamic stabilities, and slower-evolving proteins will tolerate a wider spectrum of mutations before misfolding.

Protein misfolding is certainly not the only cause of the relationships considered here. But multiple lines of evidence suggest that it may well be the most important determinant of the strength of those relationships—with the exception of noise. Our controlled simulation sets expectations on how much variation in key variables is likely to have any explanation at all in real genomes. Because signs of translational selection that are glaring in microbes[133] weaken in metazoa[180] and can be extremely subtle in mammals,[178] other factors are sometimes assumed to play a larger role in evolutionary rate variation in higher organisms. However, we find that even when expression level is the only independent variable, all measurements are exact, and no bias exists in the number of genes toward low expression levels where trends are weaker, we cannot explain much more variance in our simulation than in yeast (Fig. 6.3), underscoring the role of truly random variation, *e.g.* due to reduced effective population sizes. (These simulated organisms do have very short genes, of potential importance because variability of gene-wide average properties depends on length. Many questions about the effect of gene length on evolutionary rate and codon bias remain unresolved.[88,180])

We speculate that protein misfolding, a general fitness cost operating from bacteria to humans, may unify the study of other broad patterns in molecular evolution. For example, breadth of gene expression across tissues predicts evolutionary rate better than expression level in plants[97] and mammals[78], and we speculate that expression breadth simply better predicts the burden of misfolding in multicellular organisms. The expanding use of synthetic evolutionary biology[156,157,181] as a complement to traditional approaches may yield causal insights not readily accessible to retrospective analyses and laboratory evolution.

Box 6.1: A unique fitness function describes protein misfolding costs.

Let the fitness of an organism $f(m) > 0$ be a monotonically decreasing function of the amount of protein misfolding m with a continuous first derivative f' and $f(0) = 1$. Assume that misfolded protein is nonspecifically toxic, such that any change Δm in the amount of misfolded protein produces the same fitness disadvantage $s = f(m + \Delta m) / f(m) - 1$. We claim these assumptions determine $f(m)$ up to a constant.

Proof: We consider $\Delta m > 0$ without loss of generality. Consider two genes expressing amounts m_1 and m_2 of misfolded protein. Then:

$$s_1 = s_2$$

$$\frac{f(m_1 + \Delta m)}{f(m_1)} - 1 = \frac{f(m_2 + \Delta m)}{f(m_2)} - 1$$

$$\ln \frac{f(m_1 + \Delta m)}{f(m_1)} = \ln \frac{f(m_2 + \Delta m)}{f(m_2)}$$

$$g(m_1 + \Delta m) - g(m_1) = g(m_2 + \Delta m) - g(m_2) \qquad \text{(substitute } g(x) = \ln f(x)\text{)}$$

$$g'(x_1) = g'(x_2) \qquad \text{(mean value thm., } m_i \le x_i \le m_i + \Delta m\text{)}$$

$$g'(m) = c \qquad \text{(}c\text{ constant)}$$

$$g(m) = cm + d \qquad \text{(}d\text{ constant)}$$

$$f(m) = e^{cm+d} \qquad \text{(back out } g(x) = \ln f(x)\text{)}$$

$$f(m) = e^{-cm} \qquad \text{(}f(0) = 1\text{, monotonic decrease.)}$$

Note that in this model, polypeptides have no production cost, and misfolding does not impede the synthesis of a full complement of properly folded proteins. The only cost is the toxicity of misfolded proteins produced during synthesis.

Methods

Simulation

The fitness function dictated by our biological assumptions (Eq. 6.1) allowed us to efficiently parallelize genome evolution. Assuming no genetic linkage and a low mutation rate, an overall evolutionary competition between N organisms having n genes expressed at different (but fixed) levels is equivalent to parallel competitions within n populations of N one-gene, one-expression-level individuals. We carried out n=650 such sub-simulations of N=1,000 genes, with 50 distinct genes at each of 13 expression levels evenly spaced on a log scale from 10 to 100,000. Initial genes were chosen at random by choosing a random sequence encoding a lattice protein (see below) that adopted a target structure with free energy of unfolding (stability) of at least 0 kcal/mol, and hill-climbing until the stability exceeded 5 kcal/mol. During evolution, the initial sequences equilibrated for 50,000 generations, and recording of evolutionary data (see below) then proceeded until the most-recent common ancestor of the final population had a birth time at least 100,000 generations after the end of equilibration. Fitness was converted into reproductive success by Wright-Fisher sampling using non-overlapping generations. Fitness costs (see below) were derived from absolute numbers of misfolded proteins (Eq. 6.1) and so had the same meaning within and between sub-simulations, making evolutionary rates between sub-simulations directly comparable. A mutation rate of 0.00001 changes per nucleotide (μ=0.00075 per gene) per generation ($N\mu$=0.75) was held constant across all sub-simulations. A full simulation run required approximately one month of computing time on a 2.0GHz Pentium 4 PC with 0.5 GB of RAM. Simulation code was written in C++.

Fitness measurement

The magnitudes of expression levels examined rendered folding each expressed protein computationally intractable. Because the translation error spectrum[159] and codon composition were known precisely at the time of translation, the expected number of

misfolded proteins could be estimated by attempting to fold all possible translation outcomes and weighting each by its probability. The only approximation concerned folding after multiple translation errors (very rare events), such that if two single errors preserved folding, the double error was assumed to also, and otherwise was assumed to induce misfolding. This implementation allowed translational outcomes and fitness to be accurately estimated after folding ~150 proteins on average, independent of expression level. Stochastic fluctuations in misfolding were thereby excluded. Estimated fitnesses over all expression levels reproduced those obtained by individually translating and folding 10,000 polypeptides with correlations of >0.99 (Fig. 6.S1).

Protein folding model

Folding of lattice proteins was implemented as described.[11,30] Briefly, we used an alphabet of 20 amino acids forming 25-residue chains whose nearest-neighbor non-bonded interactions contributed additive energies as tabulated in Table 3 of ref. [25], allowing energies, and thus the thermodynamic partition function and free energy of unfolding, to be calculated for all 1,081 maximally compact conformations not related by symmetry.

Genomic data and evolutionary calculations

Genomic data for *Saccharomyces cerevisiae* and *Saccharomyces bayanus* were obtained from the *Saccharomyces Genome Database*.[182] Orthologs were identified by the reciprocal-shortest-distance method;[77] protein sequences aligned using MUSCLE[183] were used to align nucleotide sequences, and evolutionary quantities dN, dS, and the number of synonymous and nonsynonymous sites were calculated by maximum likelihood using the PAML[140] program `codeml` operating on codons under the F3×4 model for codon frequencies. (For the simulation, synonymous and nonsynonymous substitutions were counted as they occurred along the line of descent. The number of synonymous sites for a gene was determined by adding up the fraction of possible synonymous mutations at each site. The number of nonsynonymous sites was then the total number of sites minus the

number of synonymous sites.) To prevent omission of zeros in log-log plots, substitution counts (*e.g.*, dN times the number of nonsynonymous sites) were incremented by 1 (Laplace estimation). The fraction of optimal codons F_{op} was computed exactly as described[180] using optimal codons as defined for yeast.[160] Intragenic rates of dN and dS in the simulation were computed by summing nonsynonymous substitutions per codon for all genes expressed at the same level to obtain two lists of 325 numbers (25 codons × 13 expression levels) and computing correlations between these lists. For yeast data, intragenic dN–dS correlations were computed on 1,374 *S. cerevisiae* genes with orthologs in six *Saccharomyces*-genus species (*S. paradoxus*, *bayanus*, *mikatae*, *castelli*, *kudriavzevii*, and *kluyveri*) identified by other groups, downloaded from the Saccharomyces Genome Database[184], as follows. Seven-way alignments were constructed using MUSCLE (see below), ancestral sequences in the tree for each 7-member group were reconstructed using PAML (same settings as above and `RateAncestor=1`), and synonymous (*s*) and nonsynonymous (*n*) substitutions per codon for the whole reconstructed tree were estimated by the method of Suzuki and Gojobori[175] using PAML. Pearson correlations between *n* and *s* were computed for each gene.

Gene expression data

We used *S. cerevisiae* gene expression levels measured in mRNA molecules per cell at log phase by Holstege et al.[138]

Selection strength measurements

Selection strength on a quantitative trait of a protein-coding gene was defined as the negative $\log_{10}$-likelihood of finding genes with at least the observed trait level among a large random sample of genes encoding folded proteins. The sample was generated by a blind-ant random walk[18] in which each codon step sampled all neighboring folded proteins with equal probability.

Statistical analysis

R[31] was used for statistical analysis and plotting. All correlations are Spearman rank correlations unless otherwise noted.

Table 6.1: Translation outcomes reflect adaptation to misfolding costs.

Expression level	% accurate	% robust (folded despite mistranslation)	% truncated	% folded
10	85.1 (2.2)	43.1 (19.0)	0.33 (0.2)	91.6 (3.0)
100	85.7 (2.5)	69.2 (14.0)	0.32 (0.2)	95.7 (1.8)
1,000	89.2 (2.2)	87.2 (5.6)	0.19 (0.1)	98.7 (0.49)
10,000	91.5 (1.7)	90.2 (12.0)	0.09 (0.06)	99.3 (0.71)
100,000	91.1 (2.1)	92.1 (9.0)	0.08 (0.06)	99.4 (0.50)

Results are measurements [mean (s.d.)] of translation outcomes from 10,000 simulated polypeptides translated from each of 50 genes at each expression level.

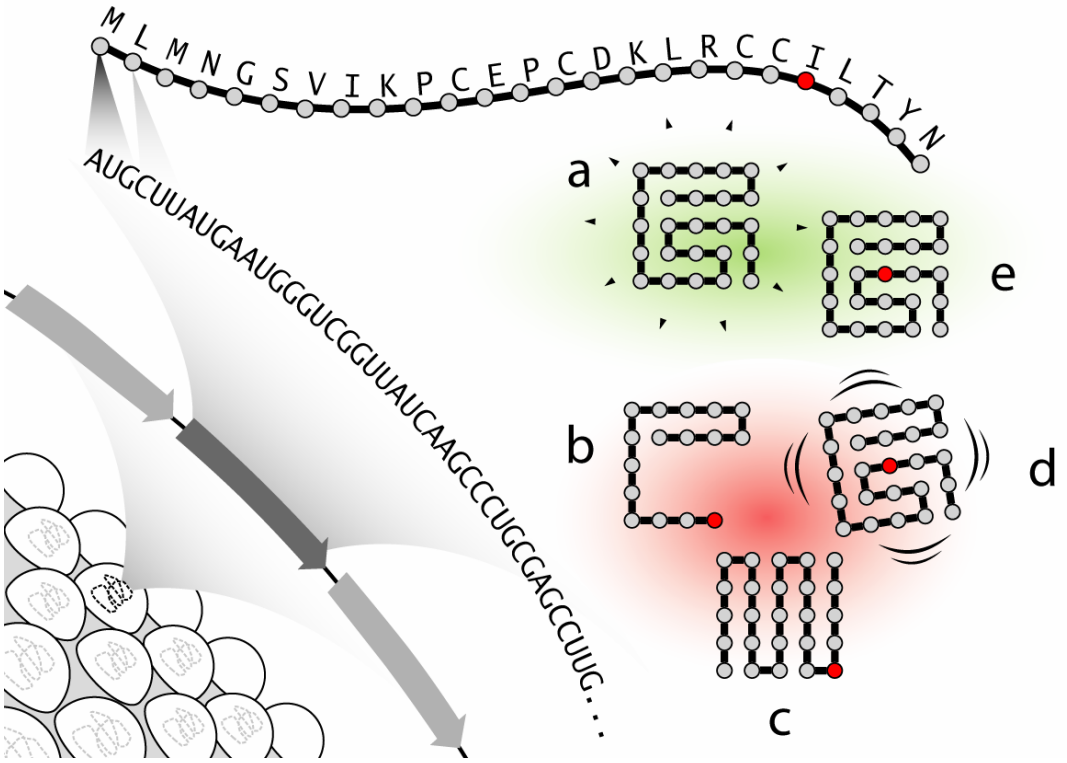


Figure 6.1. Overview of whole-genome evolutionary simulation protocol. Simulated organisms (cartooned in lower left) evolved over millions of generations; each had a genome of 650 essential genes which expressed computationally foldable lattice proteins at widely varying levels. In each generation, genes were transcribed without errors into mRNA which was then translated with occasional errors (red residues). Error-free proteins (**a**) folded properly (when translated from non-lethal alleles), while mistranslated proteins (**b**–**e**) met one of two fates. Most proteins retained wildtype folding (**a**,**e**), many despite missense errors (**e**). Misfolding resulted from nonsense errors (**b**) and missense errors that caused the sequence to adopt a non-native conformation (**c**) or destabilized the protein's native structure beyond a threshold stability (**d**). Misfolded proteins (**b**–**d**) imposed a fitness burden.

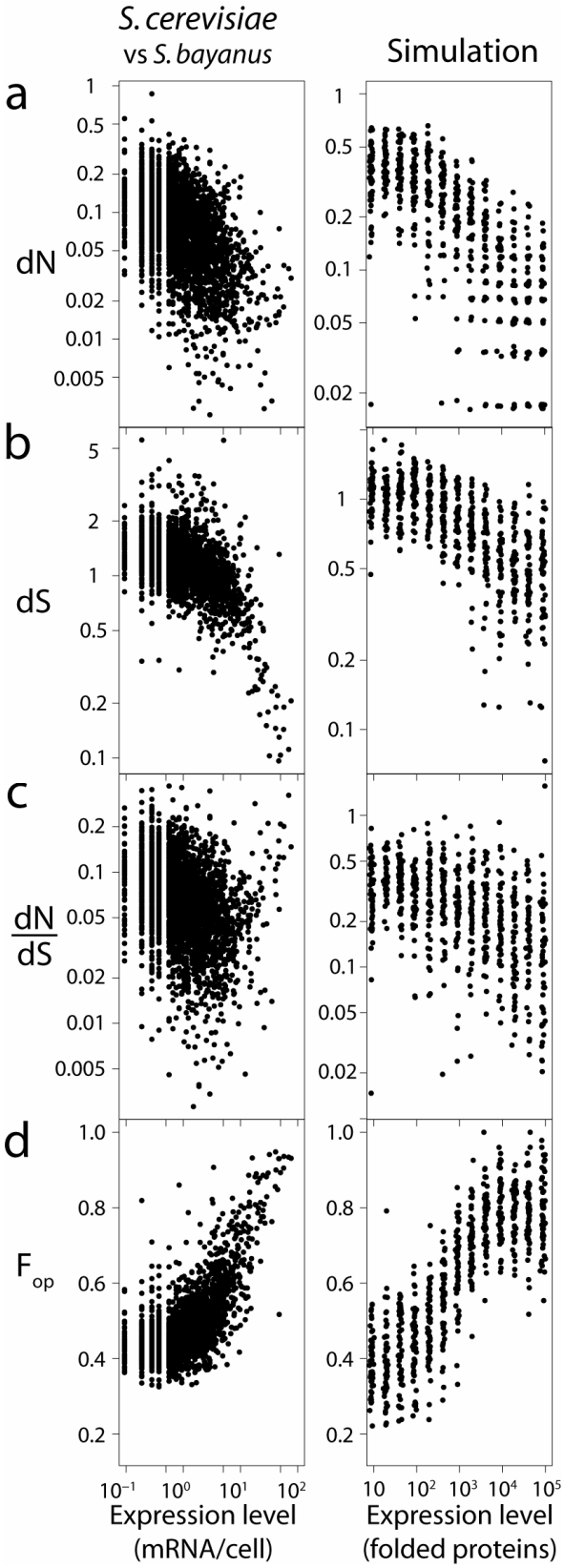


Figure 6.2. A model genome evolved under selection against protein misfolding reproduces multiple sequence evolution trends from yeast. Left, genome-wide expression-linked trends in *S. cerevisiae*, with *S. bayanus* orthologs used for evolutionary rate estimates and mRNA molecules per cell[138] as a measure of expression level. Right, simulation, with data taken from the line of descent (no estimation) and target number of folded proteins as the measure of expression. **a**, Substitutions per nonsynonymous site (dN) decreases with expression level. **b**, Substitutions per synonymous site (dS) decreases with expression. **c**, dN/dS decreases with expression. **d**, Fraction of optimal codons (F_{op}) increases with expression. Average F_{op} over the line of descent are shown for the simulation, and final F_{op} values follow a similar pattern. Small offsets were added to the plotted expression levels to allow overlapping points to be visually distinguished. Correlation coefficients for all relationships are displayed in Fig. 6.3.

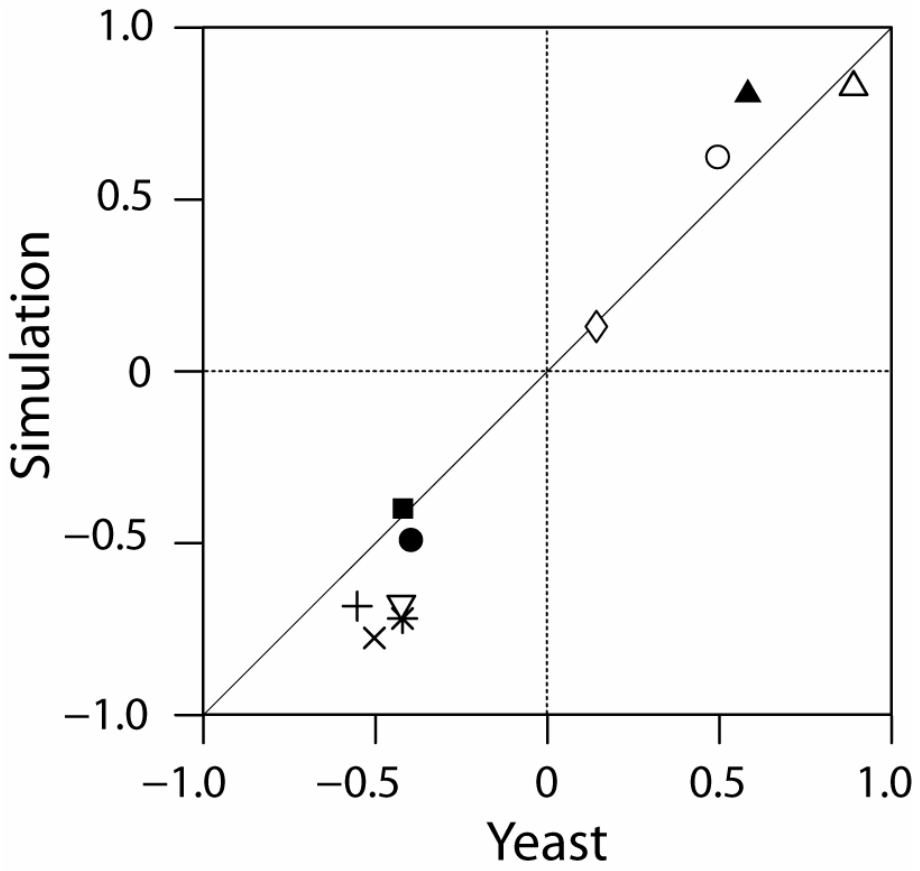


Figure 6.3. All ten pairwise correlations between dN, dS, dN/dS, F_{op} and expression level in *S. cerevisiae* and a simulated genome are similar (linear $r = 0.98$, $P < 10^{-6}$) and highly significant ($P < 10^{-8}$ unless otherwise noted): +, F_{op}–dN; ×, expression–dN; *, expression–dS; ∇, F_{op}–dS; •, expression–dN/dS; ■, F_{op}–dN/dS; ◊, dS–dN/dS ($P < 0.001$); ○, dN–dS; ▲, expression–F_{op}; Δ, dN–dN/dS. Solid line indicates perfect correlation.

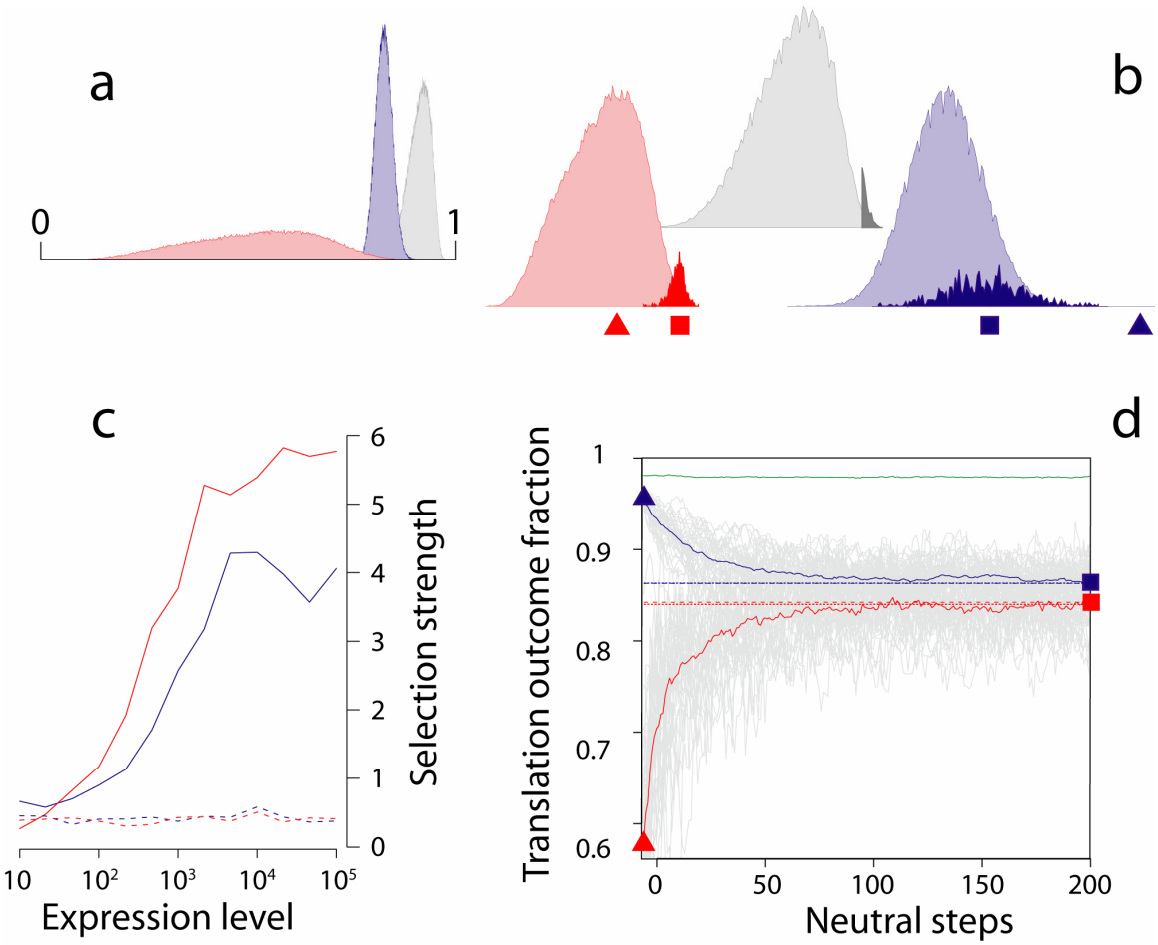


Figure 6.4. Why highly expressed model proteins evolved slowly. **a**, Overall densities of translation outcomes for 150,000 random sequences encoding folded proteins: fraction folded (gray), fraction accurately translated (blue) and fraction folded despite mistranslation [fraction robust] (red). **b**, Distributions scaled to equal variance. To obtain a fraction folded p_{folded} of at least 0.975 (magnified sub-distributions) required moderately increased accuracy and dramatically increased robustness. **c**, Selection for robustness greatly exceeds selection for accuracy in model genes. The strength of selection on

robustness (red) and accuracy (blue) for evolved sequences is shown as a function of expression level, with random sequences encoding folded proteins (dashed lines) for comparison. A difference of 1 in selection strength corresponds to a 10-fold difference in the probability of observing each trait level by chance, so the average difference of 1.56 in genes with expression levels greater than 1,000 molecules reflects $10^{1.56}$=36-fold stronger selective pressure. **d**, Neutral evolution of accuracy-optimized genes, under constraint to maintain a post-translation p_{folded} of least 0.975 (green line), results in rapid convergence of mean robustness and accuracy (red and blue lines; dashed lines show long-run averages over steps 500 to 1000; gray lines show unaveraged traces) to the mean values for random sequences having p_{folded} = 0.975 (dotted lines, cf. **a**), corresponding to sequences with elevated accuracy and very high robustness (**b**,**c**).

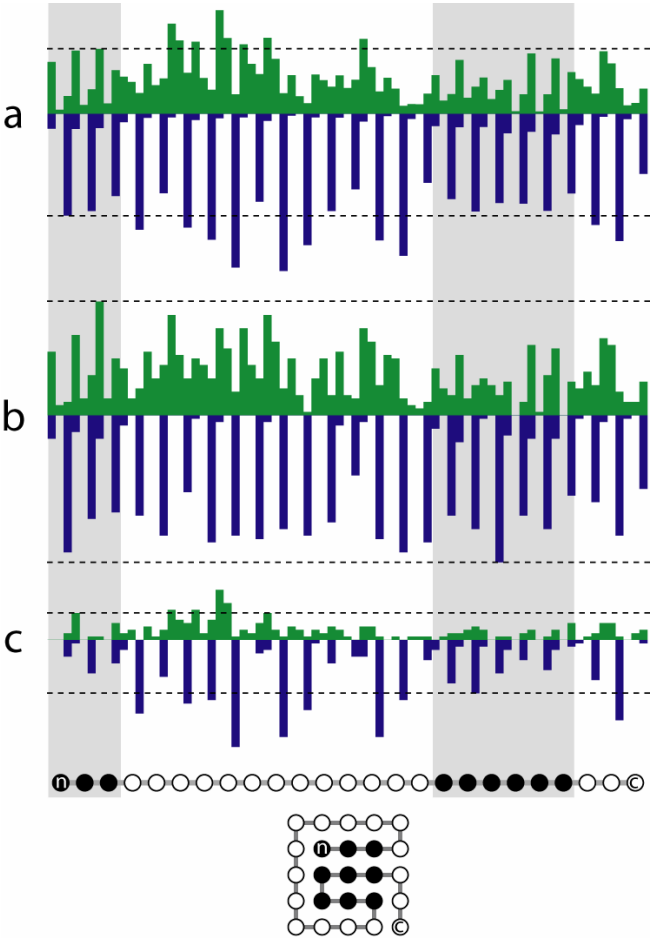


Figure 6.5: Sequence conservation patterns in simulated genes reflect structural constraints and differ with expression level. **a**, Aggregate substitution counts over all 650 evolved genes show nonrandom patterns across the gene. Nonsynonymous substitutions (top, green bars) and synonymous substitutions (blue bars) cluster in regions encoding surface residues (light background; cf. structure, bottom) such that many surface sites evolved faster than the fastest-evolving core site (dashed lines). Long synonymous bars correspond to third-position mutations. **b**, Patterns established for all genes were weak or nonexistent for low-expression genes (50 genes expressing 10 folded proteins). Core and surface sites could not be distinguished by relative rates as in the aggregate case (dotted lines). Histogram bars lengths are adjusted so that overall graph height matches that in **a**. **c**, High-expression genes (50 genes expression 10^5 folded proteins) evolved slower, but showed all the patterns observed in **a**. Histogram bars are in proper proportion to those in **b**. Synonymous substitutions are markedly higher in surface regions.

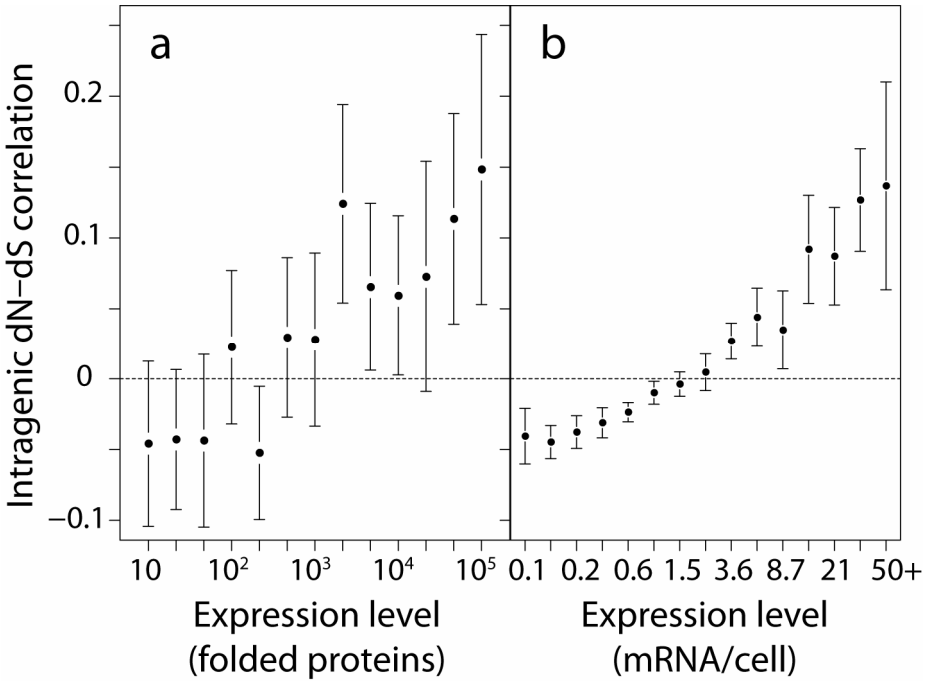


Figure 6.6: Intragenic nonsynonymous-synonymous correlations predicted from simulation results are present and numerically similar in yeast. Correlations between dN and dS along the codon sequence are aggregated by expression level for 650 simulated genes (**a**) and 1,374 yeast genes with measured expression levels (**b**). Bins are evenly spaced on a log scale.

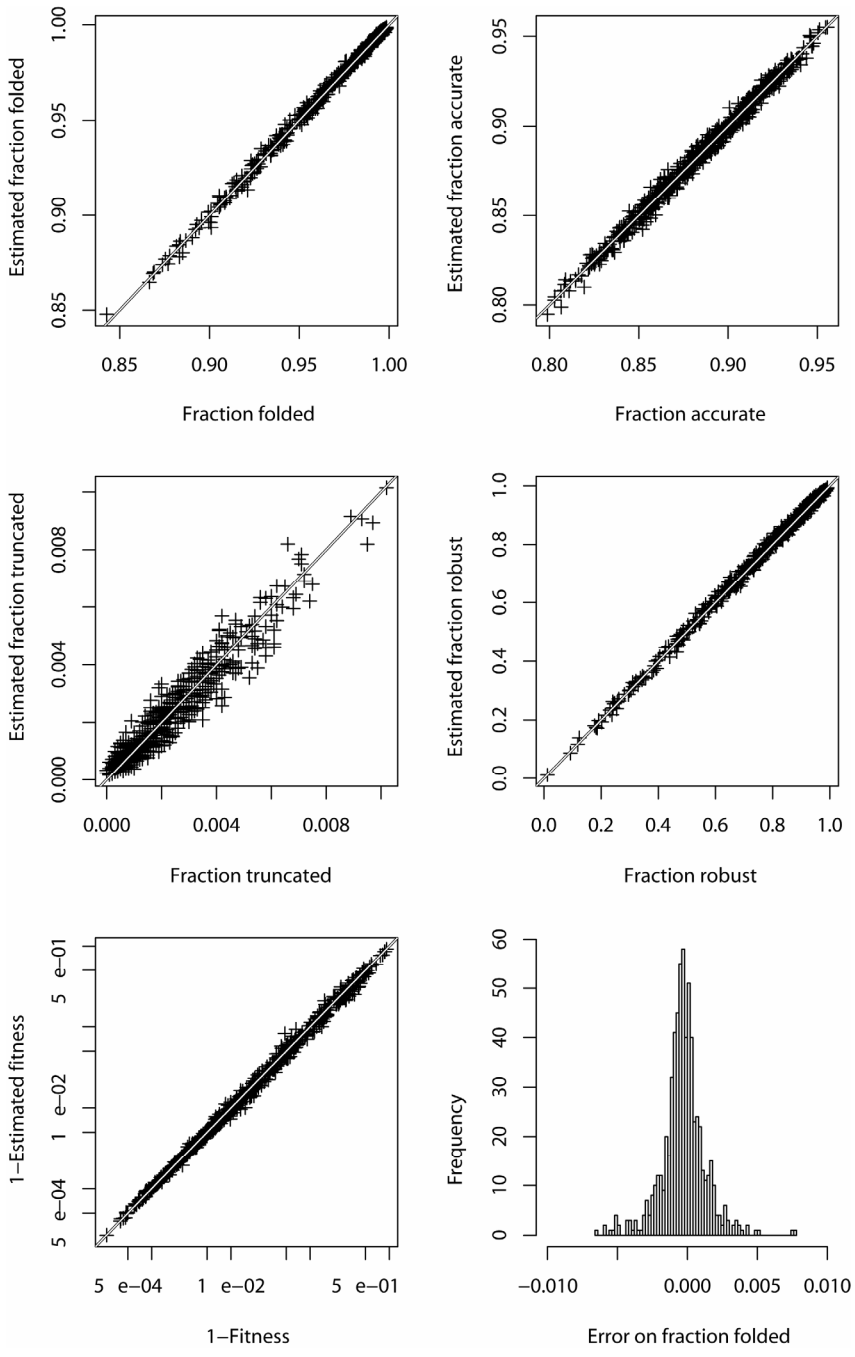


Figure 6.S1: Estimates of translation outcomes based on the translational error spectrum closely match actual results of individual translations. Each of the 650 genes evolved by the end of the simulation were translated 10,000 times with stochastic outcomes governed by the translational error spectrum described in the main text, and the fractional outcomes were compared to those predicted by the approximation scheme used during genome evolution (see *Methods*). Spearman correlations exceeded 0.99 in all cases.

REFERENCES

1. Arnold, F.H. Design by directed evolution. *Accounts of Chemical Research* **31**, 125-131 (1998).
2. Collins, C.H., Yokobayashi, Y., Umeno, D. & Arnold, F.H. Engineering proteins that bind, move, make and break DNA. *Curr Opin Biotechnol* **14**, 371-8 (2003).
3. Maynard, J. & Georgiou, G. Antibody engineering. *Annu Rev Biomed Eng* **2**, 339-76 (2000).
4. Endelman, J.B., Silberg, J.J., Wang, Z.-G. & Arnold, F.H. Site-directed protein recombination as a shortest-path problem. *Protein Engineering Design and Selection* **17**, 589-594 (2004).
5. Meyer, M.M. et al. Library analysis of SCHEMA-guided protein recombination. *Protein Sci* **12**, 1686-93 (2003).
6. Otey, C.R. et al. Structure-Guided Recombination Creates an Artificial Family of Cytochromes P450. *PLoS Biol* **4**, e112 (2006).
7. Maynard Smith, J. Natural selection and the concept of a protein space. *Nature* **225**, 563-564 (1970).
8. Shortle, D. & Lin, B. Genetic analysis of staphylococcal nuclease: identification of three intragenic "global" suppressors of nuclease-minus mutations. *Genetics* **110**, 539-55 (1985).
9. Pakula, A.A., Young, V.B. & Sauer, R.T. Bacteriophage lambda cro mutations: effects on activity and intracellular degradation. *Proc Natl Acad Sci U S A* **83**, 8829-33 (1986).
10. Loeb, D.D. et al. Complete mutagenesis of the HIV-1 protease. *Nature* **340**, 397-400 (1989).
11. Bloom, J.D. et al. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci U S A* **102**, 606-611 (2005).
12. Bloom, J.D., Labthavikul, S., Otey, C.R. & Arnold, F.H. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **103**, 5869-5874 (2006).
13. Wells, J.A. Additivity of mutational effects in proteins. *Biochemistry* **29**, 8509-17 (1990).
14. Wilke, C.O., Bloom, J.D., Drummond, D.A. & Raval, A. Predicting the tolerance of proteins to random amino acid substitution. *Biophysical Journal* **89**, 3714-3720 (2005).
15. Suzuki, M. et al. Tolerance of different proteins for amino acid diversity. *Mol Divers* **2**, 111-8 (1996).
16. Daugherty, P.S., Chen, G., Iverson, B.L. & Georgiou, G. Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc Natl Acad Sci U S A* **97**, 2029-34 (2000).
17. Shafikhani, S., Siegel, R.A., Ferrari, E. & Schellenberger, V. Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization. *Biotechniques* **23**, 304-10 (1997).
18. van Nimwegen, E., Crutchfield, J.P. & Huynen, M. Neutral evolution of mutational robustness. *Proc Natl Acad Sci U S A* **96**, 9716-20 (1999).
19. Drummond, D.A., Silberg, J.J., Wilke, C.O. & Arnold, F.H. On the conservative nature of intragenic recombination. *Proc Natl Acad Sci U S A* **102**, 5380-5385 (2005).
20. Kauffman, S. *Origins of Order: Self-Organization and Selection in Evolution*, 709 (Oxford University Press, New York, NY, 1993).

21. Zaccolo, M. & Gherardi, E. The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1 beta-lactamase. *J Mol Biol* **285**, 775-83 (1999).
22. Otey, C.R. et al. Structure-Guided Recombination Creates an Artificial Family of Cytochromes P450. *PLoS Biology* **in press**(2006).
23. Giver, L., Gershenson, A., Freskgard, P.O. & Arnold, F.H. Directed evolution of a thermostable esterase. *Proc Natl Acad Sci U S A* **95**, 12809-13 (1998).
24. Taverna, D.M. & Goldstein, R.A. Why are proteins marginally stable? *Proteins* **46**, 105-9 (2002).
25. Miyazawa, S. & Jernigan, R.L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256**, 623-44 (1996).
26. Arnold, F.H. Enzyme engineering reaches the boiling point. *Proc Natl Acad Sci U S A* **95**, 2035-6 (1998).
27. Serrano, L., Day, A.G. & Fersht, A.R. 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* **233**, 305-12 (1993).
28. Ostermeier, M. Synthetic gene libraries: in search of the optimal diversity. *Trends Biotechnol* **21**, 244-7 (2003).
29. Voigt, C.A., Kauffman, S. & Wang, Z.G. Rational evolutionary design: the theory of in vitro protein evolution. *Adv Protein Chem* **55**, 79-160 (2000).
30. Wilke, C.O. Molecular clock in neutral protein evolution. *BMC Genet* **5**, 25 (2004).
31. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299-314 (1996).
32. Georgiou, G. Analysis of large libraries of protein mutants using flow cytometry. *Adv Protein Chem* **55**, 293-315 (2001).
33. Kunichika, K., Hashimoto, Y. & Imoto, T. Robustness of hen lysozyme monitored by random mutations. *Protein Eng* **15**, 805-9 (2002).
34. Wilke, C.O. & Adami, C. Interaction between directional epistasis and average mutational effects. *Proc R Soc Lond B Biol Sci* **268**, 1469-74 (2001).
35. Elena, S.F. & Lenski, R.E. Epistasis between new mutations and genetic background and a test of genetic canalization. *Evolution Int J Org Evolution* **55**, 1746-52 (2001).
36. Kondrashov, A.S. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435-40 (1988).
37. Francisco, J.A., Campbell, R., Iverson, B.L. & Georgiou, G. Production and fluorescence-activated cell sorting of Escherichia coli expressing a functional antibody fragment on the external surface. *Proc Natl Acad Sci U S A* **90**, 10444-8 (1993).
38. Sun, F. The polymerase chain reaction and branching processes. *J Comput Biol* **2**, 63-86 (1995).
39. Weiss, G. & von Haeseler, A. Modeling the polymerase chain reaction. *J Comput Biol* **2**, 49-61 (1995).
40. Weiss, G. & von Haeseler, A. A coalescent approach to the polymerase chain reaction. *Nucleic Acids Res* **25**, 3082-7 (1997).
41. Bornberg-Bauer, E. & Chan, H.S. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci U S A* **96**, 10689-94 (1999).

42. Fromant, M., Blanquet, S. & Plateau, P. Direct random mutagenesis of gene-sized DNA fragments using polymerase chain reaction. *Anal Biochem* **224**, 347-53 (1995).
43. Flajolet, P., Gardy, F. & Thimonier, L. Birthday paradox, coupon collectors, caching algorithms, and self-organizing search. *Discrete Applied Mathematics* **39**, 207-229 (1992).
44. Moore, G.L. & Maranas, C.D. Modeling DNA mutation and recombination for directed evolution experiments. *J Theor Biol* **205**, 483-503 (2000).
45. Zaccolo, M., Williams, D.M., Brown, D.M. & Gherardi, E. An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J Mol Biol* **255**, 589-603 (1996).
46. Goodman, M.F. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu Rev Biochem* **71**, 17-50 (2002).
47. Crameri, A., Raillard, S.A., Bermudez, E. & Stemmer, W.P. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288-91 (1998).
48. Andrews, T.D. & Gojobori, T. Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen Neisseria meningitidis. *Genetics* **166**, 25-32 (2004).
49. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368-76 (1981).
50. Kimura, M. & Ohta, T. On the stochastic model for estimation of mutational distance between homologous proteins. *Journal of Molecular Evolution* **2**, 87-90 (1972).
51. Altschul, S. et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *The FASEB journal* **12**, A1326 (1998).
52. Dooner, H.K. & Martinez-Ferez, I.M. Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. *Plant Cell* **9**, 1633-46 (1997).
53. Fu, H., Zheng, Z. & Dooner, H.K. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc Natl Acad Sci U S A* **99**, 1082-7 (2002).
54. Jakobsen, I.B. & Easteal, S. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput Appl Biosci* **12**, 291-5 (1996).
55. Sawyer, S. Statistical tests for detecting gene conversion. *Mol Biol Evol* **6**, 526-38 (1989).
56. Smith, J.M. Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126-9 (1992).
57. Feil, E.J. et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci U S A* **98**, 182-7 (2001).
58. Feil, E.J., Maiden, M.C., Achtman, M. & Spratt, B.G. The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. *Mol Biol Evol* **16**, 1496-502 (1999).
59. Millman, K.L., Tavare, S. & Dean, D. Recombination in the ompA gene but not the omcB gene of Chlamydia contributes to serovar-specific differences in tissue tropism, immune surveillance, and persistence of the organism. *J Bacteriol* **183**, 5997-6008 (2001).
60. Zhou, J., Bowler, L.D. & Spratt, B.G. Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of Neisseria meningitidis and commensal Neisseria species. *Mol Microbiol* **23**, 799-812 (1997).

61. Rajalingam, R., Parham, P. & Abi-Rached, L. Domain shuffling has been the main mechanism forming new hominoid killer cell Ig-like receptors. *J Immunol* **172**, 356-69 (2004).
62. Nossal, G.J. The double helix and immunology. *Nature* **421**, 440-4 (2003).
63. Chan, H.S. & Bornberg-Bauer, E. Perspectives on protein evolution from simple exact models. *Applied Bioinformatics* **1**, 121-144 (2002).
64. Guo, H.H., Choe, J. & Loeb, L.A. Protein tolerance to random amino acid change. *Proc Natl Acad Sci U S A* **101**, 9205-10 (2004).
65. Voigt, C.A., Martinez, C., Wang, Z.G., Mayo, S.L. & Arnold, F.H. Protein building blocks preserved by recombination. *Nat Struct Biol* **9**, 553-8 (2002).
66. Saraf, M.C., Horswill, A.R., Benkovic, S.J. & Maranas, C.D. FamClash: a method for ranking the activity of engineered enzymes. *Proc Natl Acad Sci U S A* **101**, 4142-7 (2004).
67. Otey, C.R. et al. Functional evolution and structural conservation in chimeric cytochromes p450: calibrating a structure-guided approach. *Chem Biol* **11**, 309-18 (2004).
68. Lutz, S., Ostermeier, M., Moore, G.L., Maranas, C.D. & Benkovic, S.J. Creating multiple-crossover DNA libraries independent of sequence identity. *Proc Natl Acad Sci U S A* **98**, 11248-53 (2001).
69. Ostermeier, M., Shim, J.H. & Benkovic, S.J. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat Biotechnol* **17**, 1205-9 (1999).
70. Taverna, D.M. & Goldstein, R.A. Why are proteins so robust to site mutations? *J Mol Biol* **315**, 479-84 (2002).
71. Li, H., Tang, C. & Wingreen, N.S. Nature of driving force for protein folding: A result from analyzing the statistical potential. *Physical Review Letters* **79**, 765-768 (1997).
72. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94 (1999).
73. Cui, Y., Wong, W.H., Bornberg-Bauer, E. & Chan, H.S. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci U S A* **99**, 809-14 (2002).
74. Drummond, D.A., Iverson, B.L., Georgiou, G. & Arnold, F.H. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *J Mol Biol* **350**, 806-816 (2005).
75. Guex, N. & Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714-23 (1997).
76. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics*, (Oxford University Press, New York, 2000).
77. Wall, D.P., Fraser, H.B. & Hirsh, A.E. Detecting putative orthologs. *Bioinformatics* **19**, 1710-1711 (2003).
78. Pal, C., Papp, B. & Lercher, M.J. An integrated view of protein evolution. *Nat Rev Genet* **7**, 337-48 (2006).
79. Kurtzman, C.P. & Robnett, C.J. Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Research* **3**, 417-432 (2003).
80. Sarich, V.M. & Wilson, A.C. Immunological time scale for hominid evolution. *Science* **158**, 1200-3 (1967).
81. Liberles, D.A. & Wayne, M.L. Tracking adaptive evolutionary events in genomic sequences. *Genome Biol* **3**, REVIEWS1018 (2002).

82. Pál, C., Papp, B. & Hurst, L.D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927-31 (2001).
83. Rocha, E.P. & Danchin, A. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* **21**, 108-16 (2004).
84. Akashi, H. Translational selection and yeast proteome evolution. *Genetics* **164**, 1291-303 (2003).
85. Zuckerkandl, E. Evolutionary Processes and Evolutionary Noise at the Molecular Level I. Functional Density in Proteins. *J Mol Evol* **7**, 167-183 (1976).
86. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. & Feldman, M.W. Evolutionary rate in the protein interaction network. *Science* **296**, 750-2 (2002).
87. Hahn, M.W. & Kern, A.D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* **22**, 803-6 (2005).
88. Marais, G. & Duret, L. Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. *J Mol Evol* **52**, 275-80 (2001).
89. Wall, D.P. et al. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A* **102**, 5483-8 (2005).
90. Hirsh, A.E. & Fraser, H.B. Protein dispensability and rate of evolution. *Nature* **411**, 1046-9 (2001).
91. Yang, J., Gu, Z. & Li, W.H. Rate of protein evolution versus fitness effect of gene deletion. *Mol Biol Evol* **20**, 772-4 (2003).
92. Zhang, J. & He, X. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* **22**, 1147-55 (2005).
93. Drummond, D.A., Raval, A. & Wilke, C.O. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* **23**, 327-337 (2006).
94. Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O. & Arnold, F.H. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* **102**, 14338-14343 (2005).
95. Krylov, D.M., Wolf, Y.I., Rogozin, I.B. & Koonin, E.V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* **13**, 2229-35 (2003).
96. Popescu, C.E., Borza, T., Bielawski, J.P. & Lee, R.W. Evolutionary rates and expression level in Chlamydomonas. *Genetics* **172**, 1567-76 (2006).
97. Wright, S.I., Yau, C.B., Looseley, M. & Meyers, B.C. Effects of Gene Expression on Molecular Evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Mol Biol Evol* **21**, 1719-26 (2004).
98. Bierne, N. & Eyre-Walker, A. Variation in synonymous codon use and DNA polymorphism within the Drosophila genome. *J Evol Biol* **19**, 1-11 (2006).
99. Subramanian, S. & Kumar, S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168**, 373-81 (2004).
100. Fraser, H.B. Modularity and evolutionary constraint on proteins. *Nat Genet* **37**, 351-2 (2005).
101. Akashi, H. Gene expression and molecular evolution. *Curr Opin Genet Dev* **11**, 660-6 (2001).
102. Gu, Z. et al. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63-6 (2003).
103. Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature* **425**, 737-41 (2003).

104. Bloom, J.D. & Adami, C. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evol Biol* **3**(2003).
105. Pál, C., Papp, B. & Hurst, L.D. Genomic function: Rate of evolution and gene dispensability. *Nature* **421**, 496-7; discussion 497-8 (2003).
106. Lemos, B., Bettencourt, B.R., Meiklejohn, C.D. & Hartl, D.L. Evolution of Proteins and Gene Expression Levels are Coupled in Drosophila and are Independently Associated with mRNA Abundance, Protein Length, and Number of Protein-Protein Interactions. *Mol Biol Evol* (2005).
107. Hirsh, A.E. & Fraser, H.B. Rate of evolution and gene dispensability: reply. *Nature* **421**, 497-8 (2003).
108. Coghlan, A. & Wolfe, K.H. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast* **16**, 1131-45 (2000).
109. Mandel, J. Use of the singular value decomposition in regression analysis. *The American Statistician* **36**, 15-24 (1982).
110. Akashi, H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**, 927-35 (1994).
111. Hirsh, A.E., Fraser, H.B. & Wall, D.P. Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol Biol Evol* **22**, 174-7 (2005).
112. Han, J.D. et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88-93 (2004).
113. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**, 13-34 (1985).
114. Cho, R.J. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65-73 (1998).
115. Fraser, H.B. & Hirsh, A.E. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evol Biol* **4**, 13 (2004).
116. Bloom, J.D. & Adami, C. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. *BMC Evol Biol* **4**, 14 (2004).
117. Duret, L. & Mouchiroud, D. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* **17**, 68-74 (2000).
118. Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. & White, O. The Comprehensive Microbial Resource. *Nucleic Acids Res* **29**, 123-5 (2001).
119. Bernstein, J.A., Khodursky, A.B., Lin, P.H., Lin-Chao, S. & Cohen, S.N. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A* **99**, 9697-702 (2002).
120. Graur, D. & Li, W.-H. *Fundamentals of Molecular Evolution*, (Sinauer Associates, Inc., Sunderland, MA, 2000).
121. Hurst, L.D. & Smith, N.G. Do essential genes evolve slowly? *Curr Biol* **9**, 747-50 (1999).
122. Jordan, I.K., Rogozin, I.B., Wolf, Y.I. & Koonin, E.V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* **12**, 962-8 (2002).
123. Marais, G., Domazet-Loso, T., Tautz, D. & Charlesworth, B. Correlated evolution of synonymous and nonsynonymous sites in Drosophila. *J Mol Evol* **59**, 771-9 (2004).

124. Kellis, M., Birren, B.W. & Lander, E.S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617-24 (2004).
125. Seoighe, C. & Wolfe, K.H. Yeast genome evolution in the post-genome era. *Curr Opin Microbiol* **2**, 548-54 (1999).
126. Ohno, S. *Evolution by Gene Duplication*, (Allen and Unwin, London, 1970).
127. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-54 (2003).
128. Parker, J. Errors and alternatives in reading the universal genetic code. *Microbiol Rev* **53**, 273-98 (1989).
129. Goldberg, A.L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**, 895-9 (2003).
130. Ellis, R.J. & Pinheiro, T.J. Medicine: danger--misfolding proteins. *Nature* **416**, 483-4 (2002).
131. Dong, H., Nilsson, L. & Kurland, C.G. Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. *J Bacteriol* **177**, 1497-504 (1995).
132. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4**(2003).
133. Sharp, P.M. & Li, W.H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-95 (1987).
134. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289-300 (1995).
135. Bucciantini, M. et al. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**, 507-11 (2002).
136. Precup, J. & Parker, J. Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem* **262**, 11351-5 (1987).
137. Spreitzer, R.J. Genetic dissection of Rubisco structure and function. *Annual Review of Plant Physiology and Plant Molecular Biology* **44**, 411-434 (1993).
138. Holstege, F.C. et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717-28 (1998).
139. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
140. Yang, Z.H. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555-556 (1997).
141. Rokas, A., Williams, B.L., King, N. & Carroll, S.B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798-804 (2003).
142. Lindquist, S. & Craig, E.A. The heat-shock proteins. *Annu Rev Genet* **22**, 631-77 (1988).
143. Hopfield, J.J. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc Natl Acad Sci U S A* **71**, 4135-9 (1974).
144. Dong, H. & Kurland, C.G. Ribosome mutants with altered accuracy translate with reduced processivity. *J Mol Biol* **248**, 551-61 (1995).
145. Kurland, C.G. Translational accuracy and the fitness of bacteria. *Annu Rev Genet* **26**, 29-50 (1992).

146. Kurland, C.G., Hughes, D. & Ehrenberg, M. Limitations of translational accuracy. in *Escherichia Coli and Salmonella Typhimurium: Cellular and Molecular Biology*, Vol. 1 (ed. Neidhardt, F.C., et al.) 979-1004 (ASM Press, Washington D.C., 1996).
147. Parker, J. Mistranslated protein in Escherichia coli. *J Biol Chem* **256**, 9770-3 (1981).
148. Parker, J. et al. Codon usage and mistranslation. In vivo basal level misreading of the MS2 coat protein message. *J Biol Chem* **258**, 10007-12 (1983).
149. Stansfield, I. et al. Missense translation errors in Saccharomyces cerevisiae. *J Mol Biol* **282**, 13-24 (1998).
150. Edelmann, P. & Gallant, J. Mistranslation in E. coli. *Cell* **10**, 131-7 (1977).
151. Hall, B. & Gallant, J. Defective translation in RC - cells. *Nat New Biol* **237**, 131-5 (1972).
152. Princiotta, M.F. et al. Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity* **18**, 343-54 (2003).
153. Schubert, U. et al. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* **404**, 770-4 (2000).
154. Vabulas, R.M. & Hartl, F.U. Protein synthesis upon acute nutrient restriction relies on proteasome function. *Science* **310**, 1960-3 (2005).
155. Carlini, D.B. Experimental reduction of codon bias in the Drosophila alcohol dehydrogenase gene results in decreased ethanol tolerance of adult flies. *J Evol Biol* **17**, 779-85 (2004).
156. Adami, C. Digital genetics: unravelling the genetic basis of evolution. *Nat Rev Genet* **7**, 109-18 (2006).
157. Lenski, R.E., Ofria, C., Pennock, R.T. & Adami, C. The evolutionary origin of complex features. *Nature* **423**, 139-44 (2003).
158. Bierne, N. & Eyre-Walker, A. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* **165**, 1587-97 (2003).
159. Freeland, S.J. & Hurst, L.D. The genetic code is one in a million. *J Mol Evol* **47**, 238-48 (1998).
160. Sharp, P.M. & Cowe, E. Synonymous codon usage in Saccharomyces cerevisiae. *Yeast* **7**, 657-78 (1991).
161. Comeron, J.M. & Kreitman, M. The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection or relaxed constraints? *Genetics* **150**, 767-75 (1998).
162. Wyckoff, G.J., Malcom, C.M., Vallender, E.J. & Lahn, B.T. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet* **21**, 381-5 (2005).
163. Wilke, C.O. & Drummond, D.A. Population genetics of translational robustness. *Genetics* **in press**(2006).
164. Smith, N.G. & Hurst, L.D. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**, 1395-402 (1999).
165. Makalowski, W. & Boguski, M.S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci U S A* **95**, 9407-12 (1998).
166. Mouchiroud, D., Gautier, C. & Bernardi, G. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. *J Mol Evol* **40**, 107-13 (1995).

167. Alvarez-Valin, F., Jabbari, K. & Bernardi, G. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. *J Mol Evol* **46**, 37-44 (1998).
168. Alvarez-Valin, F., Jabbari, K., Carels, N. & Bernardi, G. Synonymous and nonsynonymous substitutions in genes from Gramineae: intragenic correlations. *J Mol Evol* **49**, 330-42 (1999).
169. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3-17 (2000).
170. Chiusano, M.L. et al. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. *Gene* **238**, 23-31 (1999).
171. Wolfe, K.H. & Sharp, P.M. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J Mol Evol* **37**, 441-56 (1993).
172. Lipman, D.J. & Wilbur, W.J. Interaction of silent and replacement changes in eukaryotic coding sequences. *J Mol Evol* **21**, 161-7 (1984).
173. Alvarez-Valin, F., Tort, J.F. & Bernardi, G. Nonrandom spatial distribution of synonymous substitutions in the GP63 gene from Leishmania. *Genetics* **155**, 1683-92 (2000).
174. Goldman, N., Thorne, J.L. & Jones, D.T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445-58 (1998).
175. Suzuki, Y. & Gojobori, T. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* **16**, 1315-28 (1999).
176. El-Sayed, N.M. & Donelson, J.E. African trypanosomes have differentially expressed genes encoding homologues of the Leishmania GP63 surface protease. *J Biol Chem* **272**, 26742-8 (1997).
177. Akashi, H. Faculty of 1000 Biology, 19 August 2005, http://www.f1000biology.com/article/15946765/evaluation. (2005).
178. Comeron, J.M. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167**, 1293-304 (2004).
179. DePristo, M.A., Weinreich, D.M. & Hartl, D.L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* **6**, 678-687 (2005).
180. Duret, L. & Mouchiroud, D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A* **96**, 4482-7 (1999).
181. Tiana, G., Dokholyan, N.V., Broglia, R.A. & Shakhnovich, E.I. The evolution dynamics of model proteins. *J Chem Phys* **121**, 2381-9 (2004).
182. Dolinski, K., *et al. Saccharomyces* Genome Database, ftp://genome-ftp.stanford.edu/ (Accessed 8/3/2004).
183. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).
184. Dolinski, K., *et al. Saccharomyces* Genome Database, ftp://genome-ftp.stanford.edu/. (2006).