Estimating the contribution of folding stability to non-specific epistasis in protein evolution

Pouria Dasmeh[1,2], Adrian W.R. Serohijos[1,2,*]

[1]Department of Biochemistry, [2]Cedergren Center for Bioinformatics and Genomics, University of Montreal, 2900 Edouard-Montpetit, Montreal, Quebec H3T 1J4, Canada.

*Correspondence: adrian.serohijos@umontreal.ca

Abstract

The extent of non-additive interaction among mutations or epistasis reflects the ruggedness of the fitness landscape, the mapping of genotype to reproductive fitness. In protein evolution, there is strong support for the importance and prevalence of epistasis, but whether there is a universal mechanism behind epistasis remains unknown. It is also unclear which of the biophysical properties of proteins—folding stability, activity, binding affinity, and dynamics—have the strongest contribution to epistasis. Here, we determine the contribution of selection for folding stability to epistasis in protein evolution. By combining theoretical estimates of the rates of molecular evolution and protein folding thermodynamics, we show that simple selection for folding stability implies that at least ~30% to ~60% of among amino acid substitutions would have experienced epistasis. Additionally, our model predicts substantial epistasis at marginal stabilities therefore linking epistasis to the strength of selection. Estimating the contribution of governing factors in molecular evolution such as protein folding stability to epistasis will provide a better understanding of epistasis that could improve methods in molecular evolution.

INTRODUCTION

Epistasis refers to the non-linear and non-additive interactions among mutations. In the presence of epistasis, the genetic background affects the selective advantage of a mutation and the order at which amino acids are substituted matters. The extent of epistasis reflects the ruggedness and topology of the fitness landscape, which is the multi-dimensional mapping of genomic sequence to molecular properties and to organismal fitness. Since epistasis reflects the ruggedness of the fitness landscape, a mechanistic understanding of its origin is important for reconstructing the evolutionary history of proteins (HARMS AND THORNTON 2013). Epistasis is also crucial to the predictability of microbial evolution, especially pathogenic bacteria and viruses (O'DEA *et al.* 2010; DRAGHI AND PLOTKIN 2013; SEROHIJOS AND SHAKHNOVICH 2014b; ECHAVE *et al.* 2016; BERSHTEIN *et al.* 2017; LÄSSIG *et al.* 2017). Additionally, understanding the extent and mechanism for epistasis is crucial for inference of evolutionary past, such as phylogenetic methods and various statistical tests for adaptive evolution (WEINREICH *et al.* 2013). A major shortcoming of most of the methods in molecular evolution is the assumption of additivity of mutational effects and independent evolution among sites within a gene or protein. Accounting for epistasis could enhance the accuracy and predictability of these tools (CORDELL 2002).

There are numerous examples of epistasis in proteins (STARR AND THORNTON 2016) (MITON AND TOKURIKI 2016). In the evolution of cefotaxime resistance in *E. coli* ß-lactamase, mutations that enhanced cefotaxime hydrolysis are also destabilizing and therefore only beneficial in high stable backgrounds (BERSHTEIN *et al.* 2006; WEINREICH

et al. 2006). Similar findings were observed in the evolution of the vertebrate glucocorticoid receptor (BRIDGHAM *et al.* 2009) and the nucleoprotein in human influenza viruses (GONG *et al.* 2013). In the study of site-directed mutagenesis in hepatitis C virus NS3 protease variants, the same mutations introduced to different backgrounds resulted in a broad range of fitness effects, from nearly-neutral to almost lethal (PARERA AND MARTINEZ 2014).

In general, non-commutativity of mutations can be of two types—*magnitude epistasis* and *sign epistasis*. In magnitude epistasis, the beneficial or deleterious nature of mutations remains unchanged but their magnitude is amplified or suppressed depending on genetic background. However, in sign epistasis, the beneficial/deleterious nature of mutations are interchanged, a beneficial mutations in one genetic background can become deleterious in another (WEINREICH *et al.* 2005). By comparing stability effect of all single and double mutations of IgG-binding domain of protein G, Olson et al. reported pervasive sign epistasis among different combinations of mutations (OLSON *et al.* 2014). Epistasis can also be classified as either positive or negative. *Positive epistasis* occurs when the combined effect of two mutations is higher the arithmetic sum of their individual effects, *negative epistasis* occurs when this sum is lower.

In proteins, it is helpful to distinguish between *specific epistasis* and *non-specific epistasis* (STARR AND THORNTON 2016) (**Figure 1**). Specific epistasis results from direct physical interaction of spatially close residues in 3D structure or indirect influence of residues on each other through long-range allosteric effects. Mutations occurring at spatially proximate sites will have non-additive contribution to the biophysical properties of proteins such as stability, activity, dynamics, or binding with partner proteins (STUDER

77 *et al.* 2013). If the biophysical properties determine organismal fitness, as recently
78 shown in examples of viral and microbial evolution (BERSHTEIN *et al.* 2017), the non-
79 additivity at the level of proteins translates to non-additivity at the level of fitness. As a
80 consequence, the rate and patterns of substitution in one site may be correlated with
81 that of another spatially close site (SÜEL *et al.* 2003; MORCOS *et al.* 2011; MARKS *et al.*
82 2012; POLLOCK *et al.* 2012; DICKINSON *et al.* 2013).

83 Non-specific epistasis arises from the non-linear dependence of
84 cellular/organismal fitness to biophysical properties such as folding stability (**Figure 1**).
85 Even if biophysical traits are additive and non-epistatic, the non-linear mapping of the
86 biophysical property to fitness introduces non-linear interactions among mutations at the
87 fitness level. Since Darwinian selection acts at the level of organismal fitness, non-
88 specific epistasis could affect the rate of evolution. The simplest mapping between
89 fitness and protein properties exhibits a single peak, such as shown in **Figure 1** for
90 folding stability. This single-peak and plateau-like fitness landscape has also been
91 shown for the relationship between fitness and intracellular abundance of a gene and
92 between fitness and enzyme activity (FLINT *et al.* 1981; DYKHUIZEN *et al.* 1987;
93 BERSHTEIN *et al.* 2013b; BERSHTEIN *et al.* 2013a; BERSHTEIN *et al.* 2015b; RODRIGUES *et*
94 *al.* 2016). In these simple landscapes, non-specific epistasis gives rise to the "law of
95 diminishing returns", i.e., the selective advantage of mutations decreases as the fitness
96 of the organism increases, a well-known feature of many optimization processes and in
97 adaptive trajectories in protein evolution (HARTL *et al.* 1985; CHOU *et al.* 2011; MITON
98 AND TOKURIKI 2016).

Several works have investigated the role of stability to protein epistasis (Bershtein *et al.* 2006; Wylie and Shakhnovich 2011; Ashenberg *et al.* 2013; Gong *et al.* 2013; Serohijos and Shakhnovich 2014a; Shah *et al.* 2015; Bershtein *et al.* 2017). Folding stability is a universal property of proteins that determine the fraction of proteins in the native state. Based on the simple assumption that proteins need to be folded for organisms to be viable, a quantitative fitness landscape (**Figure 1**) may be constructed from protein folding thermodynamics (protein folding fitness landscape or PFL). Folding stability directly affects the evolutionary rate of proteins (DePristo *et al.* 2005; Bloom *et al.* 2006; Pál *et al.* 2006; Serohijos *et al.* 2012; Serohijos and Shakhnovich 2014b; Echave *et al.* 2016; Bershtein *et al.* 2017). Gain-of-function mutations are on average destabilizing (Tokuriki *et al.* 2008), thus stabilizing substitutions can act as permissive or compensatory mutations that increase the likelihood of fixation of functional mutants (Weinreich *et al.* 2005; Soskine and Tawfik 2010; Gong *et al.* 2013). Additionally, several observations in molecular evolution and genomics have been explained based on PFL (Serohijos and Shakhnovich 2014b). For example, folding stability has a direct role in the genomic observation that highly abundant proteins evolve slowly (Serohijos *et al.* 2012), a consistent finding across organisms from different kingdoms of life (Drummond and Wilke 2008).

Despite these works highlighting the role of stability in protein epistasis, several questions are unanswered. First, to *what extent does selection for folding stability contribute to protein epistasis?* Kondrashov and collaborators argues that epistasis is pervasive in protein evolution and estimated that up to ~90% of amino acid substitutions

121 experienced epistasis (BREEN *et al.* 2012). However, it is not established what factors
122 could give rise to this prevalence of epistasis.

123 Here we hypothesize that this estimated fraction of epistasis is systematically
124 influenced by selection for folding stability. To elucidate the contribution of PFS to
125 epistasis, we build on the previous work of Breen *et al.* that estimated epistasis by
126 comparing two relative evolutionary rates—the average dN/dS among protein
127 orthologues and the average mutational usage (**Figure 2A**). Using a combination of
128 forward evolutionary simulations and theoretical analysis, we show that the fraction of
129 amino acid substitutions that experience epistasis due to folding stability is at least
130 ~30%, and could reach up to ~60% for proteins evolving at low stability. We also find
131 significant negative epistasis under selection for folding stability, in agreement with
132 experimental observations. The magnitude of this epistatic interactions are increased at
133 marginal folding stabilities. Since marginal stability is also the regime where purifying
134 selection is strong, our results highlight the strong coupling between epistasis and the
135 strength of selection. Altogether, our quantitative estimate of the contribution of
136 selection for folding stability to epistasis, could lead to a more mechanistic
137 understanding of this important evolutionary force.

138

139 **RESULTS**

140

141 ***Non-specific epistasis due to the protein folding fitness landscape***

142 In the absence of epistasis, the substitution rates are independent of genetic
143 background. Thus, to estimate epistasis, one approach is to compare the rates of

144 substitution of mutations with and without background specificity. Kondrashov and
145 coworkers (BREEN et al. 2012) applied this method to estimate the extent of epistasis in
146 protein evolution by comparing two rates calculated from a multiple sequence alignment
147 (MSA) of orthologs—the average pairwise substitution rate $R_{dN/dS}$ and the rate of
148 mutational usage R_u:

$$\varepsilon = 1 - \frac{R_{dN/dS}}{R_u}$$

149 (Equation 1)

150

151 $R_{dN/dS}$ is the average d*N*/d*S* (the ratio of non-synonymous substitution rate *dN* and
152 synonymous substitution rate *dS*) for all pairs of orthologues in an MSA. $R_{dN/dS}$ reflects
153 background- and lineage-specificity of amino acid substitutions (**Figure 2A**).
154 $R_u = \left(\frac{1}{L} \sum_i u_i - 1 \right) \Big/ 19$ where *u* is the number of unique amino acids in each site in an MSA
155 and is referred to as the mutational usage. *L* is the length of the protein. R_u represents
156 the ratio between observed accessible amino acid substitutions in a site, (*u*-1), and all
157 possible amino acid substitution assuming no selection, that is, (20-1)=19. Because R_u
158 is calculated per site, it is independent of background and lineage.

159 Selection is major determinant of substitution rates, thus any estimate of
160 epistasis must normalize for the confounding role of selection. $R_{dN/dS}$ reflects selection
161 because it is simply the pairwise dN/dS between orthologs and *dN*/*dS* itself is a
162 measure of the stringency of selection. R_u also reflects selection. Without selection and
163 if all mutations are neutral, all 20 amino acids are accessible in each site, thus R_u=1.
164 Altogether, R_u represents the optimal evolutionary rate with selection but without

165 epistasis, while $R_{dN/dS}$ is the observed rate that fully reflects both selection and epistasis.
166 Hence, Equation 1 includes the normalization for the confounding effect of selection.

167 From a set of diverse proteins, Kondrashov and co-workers (BREEN *et al.* 2012)
168 calculated $R_{dN/dS}$ ~ 0.01-0.1 and R_u ~ 0.1-0.6, resulting in an estimate of epistasis to be
169 ε~0.6-0.9. This finding implies that about ~60% to ~90% of amino acid substitutions in
170 proteins experienced epistasis. Despite this inferred prevalence of epistasis in long-term
171 protein evolution, a universal mechanism, if any, is lacking.

172 To determine how much epistasis can be explained by folding stability, we
173 simulate protein sequence evolution under selection for stability, generate multiple
174 sequence alignment, and apply Equation 1. Briefly, protein sequences are evolved
175 using a Wright-Fisher sampling approach with the fitness function:

$$Fitness \equiv F_{nat} = \frac{1}{1+e^{\beta \Delta G}} \qquad \text{(Equation 2)}$$

176

177 where ΔG is the folding free energy and β is the Boltzmann constant. Equation 2
178 represents the probability that a protein is in the native state, which is required for
179 function. When a random non-synonymous mutation occurs, it changes the folding
180 stability of the wildtype by $\Delta\Delta G = \Delta G_{mut} - \Delta G_{WT}$, where ΔG_{mut} is the new stability of the
181 mutant. In this folding stability fitness landscape, the selection coefficient is

$$s_{nat} = \frac{F_{mut} - F_{WT}}{F_{WT}} \qquad \text{(Equation 3)}$$

182

183 The functional form of s_{nat} distinguishes between different background stabilities as
184 widely discussed before (CHEN AND SHAKHNOVICH 2009; GOLDSTEIN 2011; SEROHIJOS
185 AND SHAKHNOVICH 2014b). Assuming that the population is monoclonal, at each
186 mutational attempt, the probability of fixation is defined by the Kimura formula,:

$$P_{fix} = \frac{1-\exp(-2s(\Delta G,\Delta\Delta G))}{1-\exp(-2Ns(\Delta G,\Delta\Delta G))} \quad \text{(Equation 4)}$$

In our simulations, we assume an effective monoclonal population size of N_e=10^4. Our model system is *dihydrofolate reductase* (DHFR) taken from *Candida Albicans* with PDB ID=1AI9 (WHITLOW *et al.* 1997). The PDB structure is used to estimate the effect random mutations on folding stability $\Delta\Delta G$ using a physical force field (Methods).

First, we simulate 2000 independent trajectories of protein evolution all starting from the same initial ancestral sequence. This procedure mimics the divergence of orthologs from a common ancestor. We run the simulation for 10^7 generations and extracted nucleotide sequences every 10^5 generations (that is, every $10N_e$), thus creating a set of MSAs with different divergence times. Estimating epistasis on this set of MSAs controls for the dependence of both R_u and $R_{dN/dS}$ on divergence time (**Figure 2B**). We use average pairwise synonymous substitution rate as a measure of divergence time and show the range of <*dS*> up to 0.5 which is the same range of <*dS*> for the protein set in (BREEN *et al.* 2012) (**Figure 2B** and **Table S1**). R_u increases almost three folds from 0.18 (~five unique amino acids per site) to 0.6 (~13 amino unique acid per site). $R_{dN/dS}$ is slightly higher at initial divergence time due to stochasticity in *dS* because of few fixed synonymous substitutions. Estimated epistasis based on Equation 1 indeed shows a strong dependence on divergence time (**Figure 2B**). For the set of proteins studied in (BREEN *et al.* 2012) <*dS*>=0.4, which from our simulation using folding stability, the estimated epistasis is ~30%. Thus, while Kondrashov and co-workers estimated that ~60% to ~90% epistasis in protein evolution (BREEN *et al.* 2012),

we find that up to half of this estimated can be accounted for by simple selection for folding stability.

Additionally, in our simulations based on folding stability, the mutational usage is u~13, which means that each site in the MSA has an average of 13 (out of the possible 20) unique amino acids. For the set of proteins studied in (BREEN *et al.* 2012), u~8. This difference is expected because real protein sequences are under selection for other biophysical properties (e.g., protein-protein interaction, binding, dynamics) and biological function beyond folding stability. Interestingly, the difference between u=20 (i.e., full mutational usage and zero selection), u=13 (under selection for thermodynamic stability) and u=8 (in real proteins) provides a rough estimate of the contribution of protein folding stability to selection relative to other properties. That is, the drop in amino acid usage per site in an MSA, (20-13)/(20-8) or ~60% is due to selection for folding stability.

To control for the potential bias of the number of sequences in the MSA to Equation 1, we down-sample the sequences and estimated epistasis (**Figure 2C**). We perform this down-sampling procedure on the most diverged MSA corresponding to $\langle dS \rangle$ = 0.5. R_u converges to ~0.65 (u~13) while $R_{dN/dS}$ is ~0.30 at all number of sampled sequences. Therefore, epistasis converges to ~ 30% once enough sequences (>1000) are used in the MSA and the estimation of R_u and $R_{dN/dS}$.

Prevalence of negative epistasis under selection for thermodynamic stability

In the analysis above, we examine the role of epistasis on long-term protein evolution by analyzing amino acid substitutions across orthologues. Next, we analyze

short-term epistasis by focusing on the pairwise interaction between two randomly arising mutations. Such an analysis is directly comparable to results from directed mutagenesis or comprehensive deep mutational scans. Specifically, we want to determine which type of pairwise epistasis, positive or negative, dominates under selection for stability. To do so, we pick two random mutations *A* and *B* that change the folding stability of wildtype ΔG_{WT} by $\Delta\Delta G_A$ and $\Delta\Delta G_B$, respectively. These values are drawn from the distribution of random effect of mutations on folding stability $p(\Delta\Delta G_{fold})$ inferred from the collection of experimental measurements and comprehensive computational mutagenesis of globular proteins (TOKURIKI *et al.* 2007). Using Equation 1, the corresponding fitness of these two mutants are F_A and F_B. Since we focus only on non-specific epistasis, their combined effect on stability is $\Delta\Delta G_A + \Delta\Delta G_B$ with a corresponding fitness value $F_{AB} = F(\Delta\Delta G_A + \Delta\Delta G_B)$. We calculate the pairwise epistasis using the following equation:

$$\varepsilon_{pair} = \ln(F_{AB}) - \ln(F_A) - \ln(F_B) \qquad \text{(Equation 5)}$$

where F_A and F_B are fitness of each mutant. We show in **Figure 3A-B** the pairwise epistasis among random mutations as a function of background folding stability. Indeed, the most dominant form of epistasis is negative, in agreement from comprehensive mutational scans in IgG-binding domain of *Streptococcus* protein G (OLSON *et al.* 2014), substrate binding domain of yeast Hsp90 (BANK *et al.* 2015) and green fluorescent protein from *Aequorea victoria* (SARKISYAN et al. 2016). This negative epistasis is largely due to pairs of mutations that are destabilizing (**Figure 3C**). We also plotted epistasis versus the ratio of $\Delta\Delta$G values of pair mutations in **Figure S1**. Since destabilizing mutations bring the protein to the more curved part of the PFL, negative epistasis starts

at high stability compared to positive epistasis (**Figure 3B**). Altogether, the PFL features greater curvature near low folding stability, epistasis is larger in the regime where selection is stronger.

Theoretical analysis of non-specific epistasis to examine its dependence on fraction of destabilizing random mutations in a protein fold

The effect of random mutations on folding stability $\Delta\Delta G$ affects evolutionary rates (**Equation 4**), and thus can influence the estimation of epistasis. Although the shape of the distribution of $\Delta\Delta G$ due to random mutations is consistent across several types and diverse folds of proteins (TOKURIKI *et al.* 2007), there are notable differences, such as the percentage of mutations that are stabilizing or destabilizing. To explore the robustness of estimates of epistasis on the distribution of arising random $\Delta\Delta G$ in different protein folds, we resort to theoretical analysis of **Equation 1**. Specifically, rate of substitution is the product of the rate of mutation and the probability of fixation. Thus, for the non-synonymous rate $dN=(\mu N)Pfix$, where μ is the mutation rate per generation. Since there are N birth events in a generation, $N\mu$ is the number of random mutations per generation. For the synonymous mutation rate $dS=(\mu N)(1/N)$, where the factor (1/N) is the probability of fixation for neutral mutation and in our model, synonymous mutations are neutral. Thus,

$$Rate = \frac{dN}{dS} = \omega(\Delta G, \Delta\Delta G) = NP_{fix}(\Delta G, \Delta\Delta G) = N\frac{1-\exp(-2s(\Delta G, \Delta\Delta G))}{1-\exp(-2Ns(\Delta G, \Delta\Delta G))} \quad \text{(Equation 6)}$$

where we use **Equations 3** and **4** for the specific case of selection for folding stability. Equation 6 is the *dN*/*dS* for a random mutation that changes the wildtype folding stability, ΔG, by an amount $\Delta\Delta G$. The rate $R_{dN/dS}$ is calculated multiple pairs of sequences in an MSA (**Figure 1**), thus it reflects the average *dN/dS* over many backgrounds and multiple arising mutations. Thus, we can arrive at an theoretical estimate of $R_{dN/dS}$ by integrating over the distribution of wildtype folding stability ΔG and the distribution effects on folding stability due to random mutations $\Delta\Delta G$. The distribution $P(\Delta\Delta G)$ is the probability distribution of mutational effects on folding stability known from large-scale mutational studies (ALBER 1989; GUEROIS *et al.* 2002; TOKURIKI *et al.* 2007; SOSKINE AND TAWFIK 2010), and has been parameterized for proteins of different folds. The distribution of background folding stability $P(\Delta G)$ is a consequence of mutation-selection balance on the protein folding fitness landscape (TAVERNA AND GOLDSTEIN 2002b; TAVERNA AND GOLDSTEIN 2002a; BLOOM *et al.* 2007; ZELDOVICH *et al.* 2007; SEROHIJOS AND SHAKHNOVICH 2014b). This distribution has also been documented experimentally from ~4000 proteins (BAVA *et al.* 2004). For self-consistency, we numerically derive the distribution of folding stability stability $p(\Delta G)$ under mutation-selection balance using the parameters used in our sequence simulation (see Methods). Altogether, because the distributions $P(\Delta\Delta G)$ and $P(\Delta G)$ are well-determined, we can arrive at an estimate of $R_{dN/dS}$.

Next, we seek a theoretical estimate of the mutational usage, R_u. To do so, we note that each protein sequence in an MSA corresponds to a ΔG value in the distribution $P(\Delta G)$. That is, each sequence in an MSA is a random sampling of the $P(\Delta G)$ distribution. Since R_u assumes site-independence (BREEN *et al.* 2012), we can

consider each site to follow the distribution P(ΔG). Similarly, for a given site, each amino acid is a random sampling of the P(ΔG) distribution. In the language of molecular evolution, the P(ΔG) distribution may be considered as the site-equilibrium frequency in analogy to site-independence models of amino acids. Additionally, we note that in our simple model of selection for folding stability, in the regime of very high stability (ΔG < -20 kcal/mol), more mutations are allowed and amino acid usage (**Figure S2, black line**). Thus, curating an MSA of *k* orthologous sequences is sampling the P(ΔG) distribution *k* times (**Figure S2**). From the MSA, the value *u* is an upperbound estimate of mutational usage because it counts the number of *unique* amino acid in a site irrespective of frequency. The equivalent in our sampling method is the highest *u* from the *k* sampling of the P(ΔG) distribution (**Figure S2**). We used a cut-off of P(ΔG)=0.001 and estimated R_u as the evolutionary rate at highest stability at this probability:

$$R_u = \int \omega(\Delta G = \min(\Delta G|_{P(\Delta G)=0.001}), \Delta\Delta G) P(\Delta\Delta G) \qquad \text{(Equation 7)}$$

To check the consistency between simulations and our theoretical approach, we calculated R_u for different number of sequences and also from Equation 6. **Figure S3** shows the excellent agreement (r^2=0.97, p-value < 10^{-16}) between R_u calculated from theory and simulations. The higher values of R_u is calculated form theory is because of conceivable higher d*N* within the theoretical approach. Since all the calculations in the theoretical model are done with a continuous $\Delta\Delta$G distribution, this condition is only achieved when all residues or at least a viable fraction have been mutated once in the sequence-explicit approach. Using both $R_{dN/dS}$ and R_u calculated from theory, we

arrived at $\varepsilon \sim 0.35–0.45$, which is consistent with our results from explicit sequence simulation (**Figure 2**).

The theoretical approach enables us to investigate the sensitivity of estimated epistasis to the distribution of mutational effects (**Table 1** and **Figure 5**), in particular, the fraction of mutations that are destabilizing. The values for averages and standard deviations of $P(\Delta\Delta G)$ distributions are reported values for real proteins from large-scale mutagenesis studies (TOKURIKI *et al.* 2007). We plotted the expected percentage of epistasis imposed by selection for folding stability of different globular proteins in **Figure 5** (see **Table S3** for details). For example, average and standard deviation of $P(\Delta\Delta G>0)$ (bi-gaussian distribution (see Methods)) can vary from (1.33, 0.42) ± (1.64, 0.83) in Ubiquitin to (3.02,2.29) ± (0.76, 1.12) in Human lysozyme giving rise to $\varepsilon=0.35$ and $\varepsilon=0.46$ for the two proteins within our approach, respectively. As shown in **Table 1**, minor changes in $P(\Delta\Delta G>0)$ from 0.62 to 0.72 changes percentage of epistasis from 12% to 32%. In a hypothetical protein when the fraction of stabilizing and destabilizing mutations are almost equal ε is negligible.

DISCUSSION

To what extent does the estimated ~30% epistasis agree with estimates from proteome-wide observations? Although several multiple factors that potentially contributing to epistasis, some of them beyond the property of one single gene (protein-protein interaction, centrality in a metabolic pathway, or genetic interactions), selection for folding stability have a major role, as quantified in this work. Since, folding stability is a primary selective force in protein evolution, and our model is based on a simple two-

346 state folding thermodynamics, our estimate of ~30% epistasis sets a lower limit for
347 epistasis experienced by real proteins. The higher limit for epistasis in molecular
348 evolution might be ~90% epistasis as reported by Breen et al. (BREEN *et al.* 2012).

349 Additionally, estimating epistasis in long-term protein evolution by comparing
350 $R_{dN/dS}$ and R_u from a multiple sequence alignment is simple, but not without potential
351 complications. Plotkin and co-workers (MCCANDLISH *et al.* 2013) argued that correcting
352 the approach of Breen et al. (BREEN *et al.* 2012) by using a distribution of selection
353 coefficients for nonsynonymous substitutions could yield to a lower estimate of
354 epistasis. In our approach, each nonsynonymous substitution would indeed have a
355 different selection coefficient which not only depends on the effect size of mutation but
356 also on the background stability in which it occurs. Therefore, the impact of
357 nonsynonymous mutations is modeled more realistically by our approach, and the 30%
358 estimate from folding stability is robust.

359 Furthermore, in line with two recent experimental systematic study of epistasis in
360 protein G domain 1 (GB1) (OLSON *et al.* 2014) and Hsp90 in yeast (BANK *et al.* 2015),
361 we show that negative epistasis is the major type of epistasis under selection for PFS.
362 Negative epistasis is mainly caused by sampling curved parts of fitness landscape, i.e.,
363 lower stabilities in the folding stability landscape. As illustrated in **Figure S4**, any factor
364 that increases further sampling of the curvature of fitness landscape would increase
365 epistasis. Indeed, adding other biophysical properties that could be relevant to fitness
366 such as activity, dynamics, and binding to other proteins will only increase the
367 ruggedness of the landscape, and hence the estimated epistasis.
368

METHODS

Protein evolution model and simulated phylogenetic tree

Protein sequences were evolved using a Wright-Fisher sampling approach with two fitness functions described by equations (4) and (5). In brief, codons were randomly mutated in one of the sites and once a nonsynonymous substitution arose, the relevant probability of fixation was calculated by the following equation:

$$P_{fix} = \frac{1-\exp(-2s(\Delta G,\Delta\Delta G))}{1-\exp(-2Ns(\Delta G,\Delta\Delta G))} \qquad \text{(Equation 8)}$$

while $s(\Delta G, \Delta\Delta G)$ depends on the choice of fitness function. We assumed an effective monoclonal population size of $N=10^4$. Initial sequence was taken to be that of *dihydrofolate reductase*, DHFR, taken from *Candida Albicans* with PDB ID=1AI9_A_1 (WHITLOW *et al.* 1997). As described previously, the effect of mutations on protein stability was calculated using ERIS force field. To simulate sequence divergence, we simulated 2000 trajectories all started from an ancestral sequence and sampled 100 time points every 10^5 mutational attempts. The 2000 sequences at each sampling point were regarded as orthologues used for estimation of epistasis.

To do so, we performed sequence-explicit simulations on the protein folding landscape (**Figure 1**) (Methods). To do so, we simulated dihydrofolate reductase (DHFR) sequences evolved under selection for thermodynamic stability (see Methods). DHFR is an integral protein in DNA nucleotide synthesis, as it converts dihydrofolic acid to

394 tetrahydrofolic acid (FIERKE *et al.* 1987) and therefore has a highly conserved function
395 across organisms (HECHT *et al.* 2011). Several studies have shown that selection for
396 thermodynamic stability might be the primary driving force in the evolution of DHFR
397 (BERSHTEIN *et al.* 2012; BERSHTEIN *et al.* 2015a).

398

399

400 **Evolutionary rate estimation and pair epistasis**

401

402 The pairwise rate of evolution, $R_{dN/dS}$ =d*N*/d*S* and d*S* of DHFR sequences were
403 estimated by Maximum-likelihood (ML) codon based model in codeml within PAML suite
404 (YANG 2007) . We estimated codon frequencies from the products of the average
405 observed nucleotide frequencies in three codon positions (F3X4).

406

407 **Estimating limiting distribution of protein stabilities**

408 To estimate epistasis using our numerical approach, we used distribution of mutational
409 effects on protein stability, P(ΔΔG), from the extensive study of thousands of mutants in
410 different proteins (TOKURIKI *et al.* 2007):

411 $$P(\Delta\Delta G) = p_1 \mathbb{N}(\mu_1, \sigma_1^2) + (1 - p_1)\mathbb{N}(\mu_2, \sigma_2^2) \qquad \text{(Equation 9)}$$

412 where p_1 is the weights of first distribution, μ_1, μ_2,σ_1 and σ_2 are the average values and
413 standard deviations of each Gaussian distribution found to be 0.53 ± 0.12, 0.56 ± 0.12,
414 1.96 ± 0.53, 0.90 ± 0.16 and, 1.93 ± 0.29 respectively. The distribution of background
415 stabilities, P(ΔG), however, is a limiting distribution resulting from mutational supply and
416 selection. We have previously used a numerical algorithm to obtain limiting distribution
417 of protein stabilities under mutation-selection balance (KEPP AND DASMEH 2014) . In

brief, we start with an initial distribution, $P(\Delta G)_{t=0}$, which is simply the distribution of mutational effects centered on an initial stability, ΔG_{int}. This distribution is then updated iteratively as:

$$P(\Delta G)_{t_1} = \int P(\Delta G)_{t_0} P(\Delta G_i \rightarrow \Delta G_j) d(\Delta G_i) \quad \text{(Equation 10)}$$

Here $P(\Delta G_i \rightarrow \Delta G_j)$ is the transition probability of a protein from ΔG_i to ΔG_j which is the product of arising mutations and their fixation probability:

$$P(\Delta G_i \rightarrow \Delta G_j) = P(\Delta\Delta G = \Delta G_j - \Delta G_i) P_{fix}(\Delta G_j, \Delta\Delta G) \quad \text{(Equation 11)}$$

This approach is equivalent to locally weighted random walk sampling where each stability is sampled per the known distribution at time t_n giving rise to the distribution at time t_{n+1}. Each time step in this numerical scheme is one mutational step. This procedure is continued iteratively while the convergence to a limiting distribution is reached judged by Kolmogorov-Smirinov two-sample test (**Figure S5**) .

Estimating the effect of point mutations on protein folding stability

To estimate the effect of site variations in mutational effects, i.e., $\Delta\Delta G = \Delta G_{mutated} - \Delta G_{pre\text{-}mutated}$, we took $\Delta\Delta G$ values calculated for Dihydrofolate reductase (SEROHIJOS AND SHAKHNOVICH 2014a). In brief, $\Delta\Delta G$s are calculated using the flexible-back bone method of the ERIS algorithm (YIN *et al.* 2007). All side chains within 10Å of the mutated site were optimized and all dihedrals were relaxed to minimize backbone strain. This approach will give us a (Sequence length $\times$ 20 aa) matrix of $\Delta\Delta G$ values which is used for estimating the contribution of single sites to non-specific epistasis.

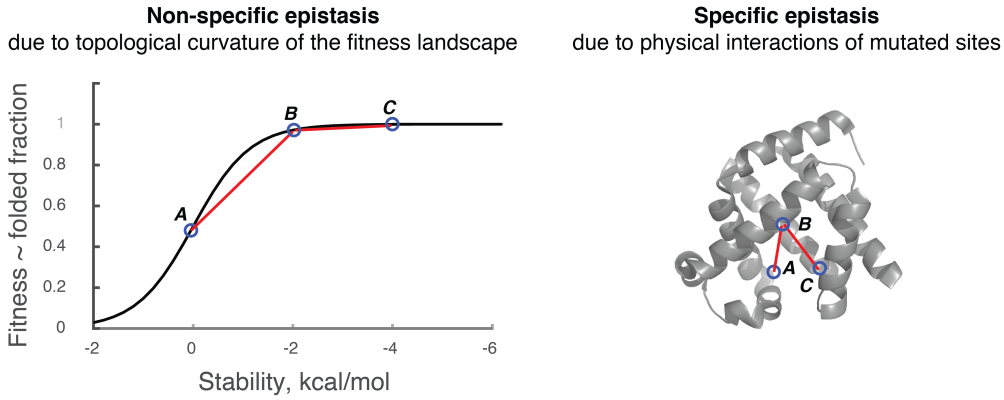

Figure 1. Non-specific epistasis. Non-specific epistasis is defined as non-linear relationship between mutational effects on the structure and protein property, e.g., protein stability and between protein properties and cellular fitness. In the figure, evolution is epistatic with respect to the order of A, B and C mutations even though they might be located far from each other on 3D structure.

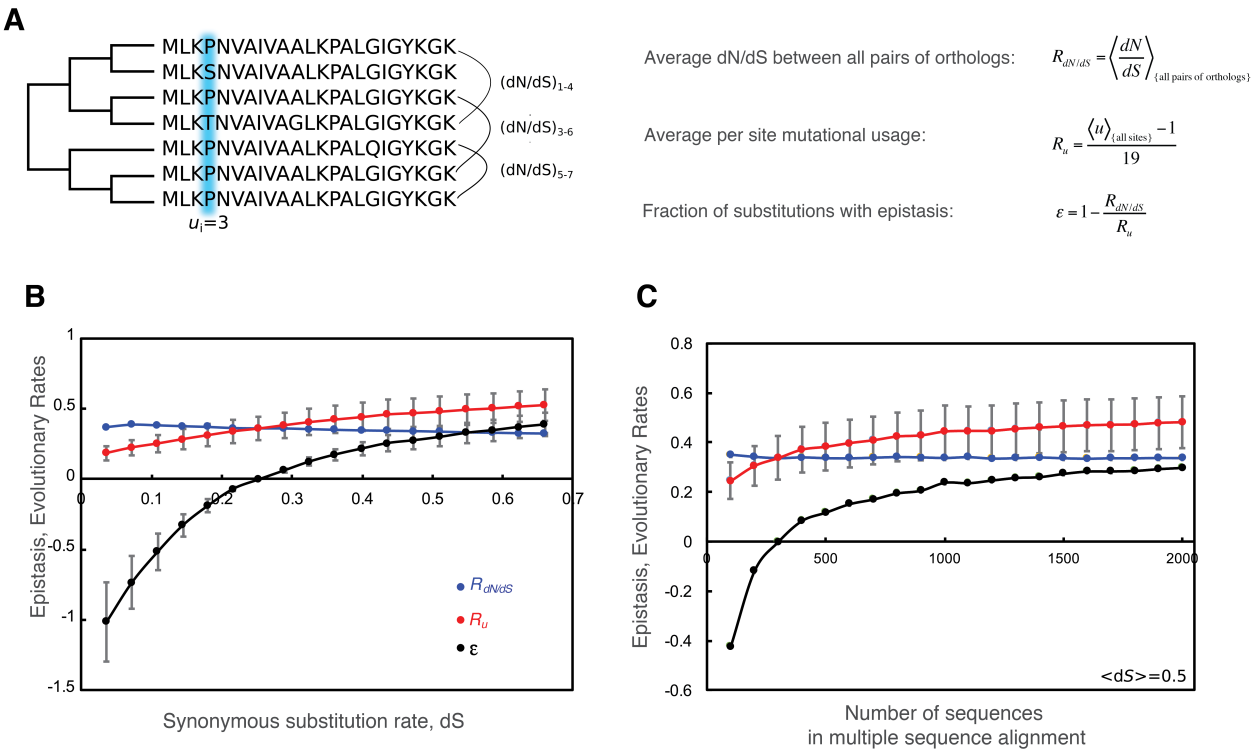

Figure 2. Protein sequences evolved under selection for thermodynamic stability show on average 30% epistasis. (A) Orthologous protein sequences are used to estimate epistatic evolution as the between rate of evolution over all possible backgrounds estimated from mutational usage, R_u, and pairwise rate of evolution, $R_{dN/dS}$. **(B)** $R_{dN/dS}$ and Ru are plotted as a function of sequence divergenc. **(C)** Pairwise rate of evolution, $R_{dN/dS}$ and the rate calculated from mutational usage, R_u, are plotted for different number of sampled sequences corresponding to the MSA at <d*S*>=0.5 in panel B. In both B and C, epistasis is calculated as the percentage difference between the two rates, $R_{dN/dS}$ and Ru.

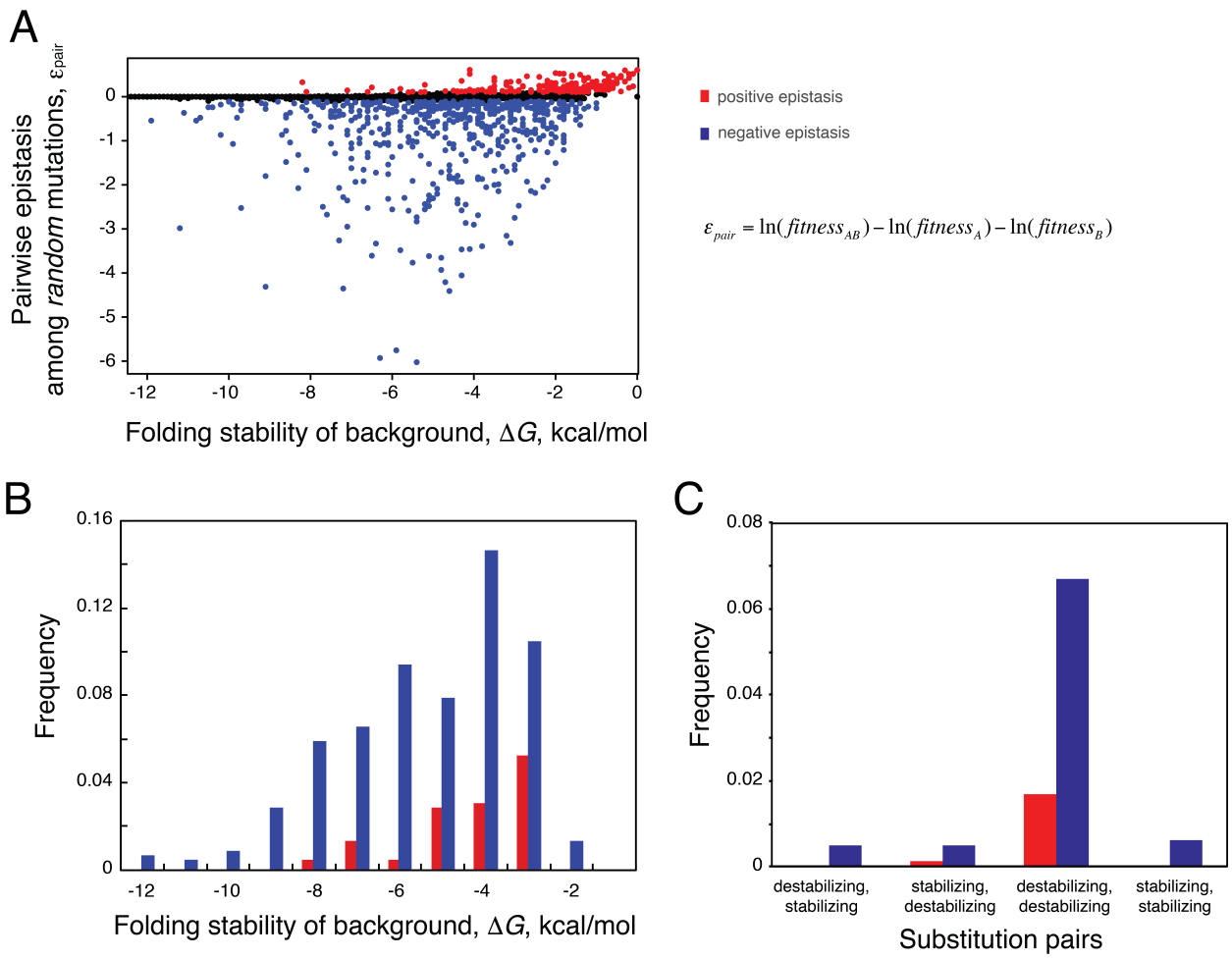

Figure 3. Pervasive negative epistasis under selection for thermodynamic stability.
(A) Pair epistasis between two randomly chosen mutations *A* and *B* as a function of the folding stability of the background. Positive epistasis ε_{pair}>0.1 are shown in red; Negative epistasis ε_{pair}<-0.1 are shown in blue. **(B)** Frequency of positive and negative epistasis with respect to wildtype stability. **(C)** Frequency of positive and negative epistasis with respect to effect of mutations on stability.

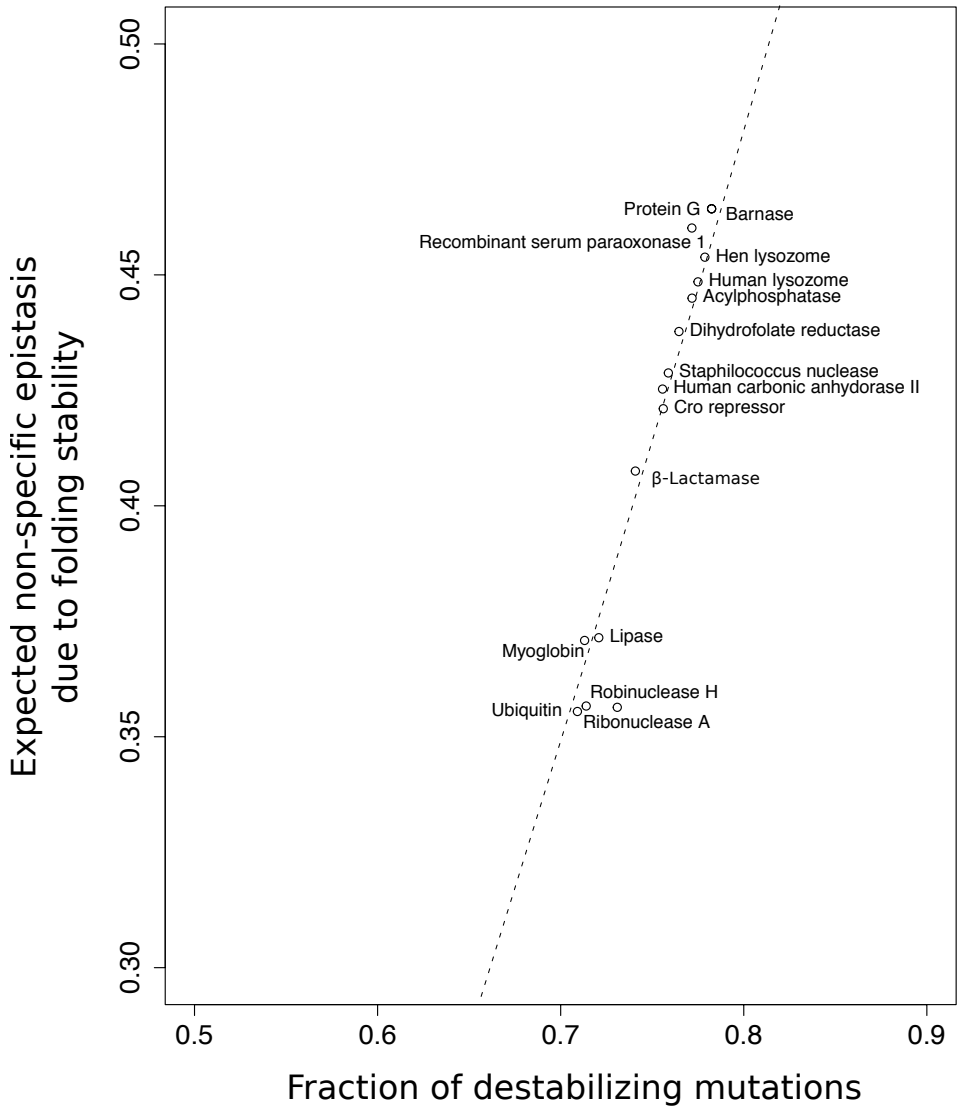

Figure 4. Expected epistasis imposed by folding stability increases by the fraction of destabilizing mutations. Dashed curve shows expected epistasis by folding stability when distribution of mutational effects on protein folding stability shifts towards destabilizing mutations. Expected epistasis for different globular proteins (filled circles) are shown. The line fitted to data has the equation: *epistasis*=1.52×(*fraction of destabilizing mutations*) – 0.72. Fraction of destabilizing mutations is based on reference (TOKURIKI *et al.* 2007).

Table 1. Sensitivity of estimated epistasis to parameters of distribution of mutational effects on protein folding stability.

Remark	μ_1 [a]	μ_2 [b]	σ_1 [c]	σ_2 [d]	$\mathbf{P}(\Delta\Delta G > 0)$ [e]	$\mathbf{P}(\Delta\Delta G < 0)$ [f]	ε
Average globular proteins (default)	0.54	2.05	0.98	1.91	0.71	0.29	0.35
Half-variance	0.54	2.05	0.49	0.95	0.85	0.15	0.50
Half-mean	0.27	1.03	0.98	1.91	0.60	0.40	0.17
Zero mean	0	0	0.98	1.91	0.46	0.54	0.00

a: average value of the first Gaussian distribution. b: average value of the second distribution. c: standard deviation of the first Gaussian distribution. d: standard deviation of the second Gaussian distribution. e: fraction of destabilizing mutations. f: fraction of stabilizing mutations. (See Equation 8 for details). Population size is 10^4. Gray cells show are the ones different from default parameters shown in the first row.

References

Alber, T., 1989 Mutational effects on protein stability. Annual review of biochemistry 58: 765-792.

Ashenberg, O., L. I. Gong and J. D. Bloom, 2013 Mutational effects on stability are largely conserved during protein evolution. Proc Natl Acad Sci U S A 110: 21071-21076.

Bank, C., R. T. Hietpas, J. D. Jensen and D. N. Bolon, 2015 A systematic survey of an intragenic epistatic landscape. Molecular biology and evolution 32: 229-238.

Bava, K. A., M. M. Gromiha, H. Uedaira, K. Kitajima and A. Sarai, 2004 ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic acids research 32: D120-D121.

Bershtein, S., J.-M. Choi, S. Bhattacharyya, B. Budnik and E. Shakhnovich, 2015a Systems-level response to point mutations in a core metabolic enzyme modulates genotype-phenotype relationship. Cell reports 11: 645-656.

Bershtein, S., W. Mu, A. W. Serohijos, J. Zhou and E. I. Shakhnovich, 2013a Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. Molecular cell 49: 133-144.

Bershtein, S., W. Mu, A. W. Serohijos, J. Zhou and E. I. Shakhnovich, 2013b Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. Mol Cell 49: 133-144.

Bershtein, S., W. Mu and E. I. Shakhnovich, 2012 Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. Proceedings of the National Academy of Sciences 109: 4857-4862.

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

Bershtein, S., A. W. Serohijos, S. Bhattacharyya, M. Manhart, J. M. Choi *et al.*, 2015b Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. PLoS Genet 11: e1005612.

Bershtein, S., A. W. Serohijos and E. I. Shakhnovich, 2017 Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. Current Opinion in Structural Biology 42: 31-40.

Bloom, J. D., S. T. Labthavikul, C. R. Otey and F. H. Arnold, 2006 Protein stability promotes evolvability. Proceedings of the National Academy of Sciences 103: 5869-5874.

Bloom, J. D., A. Raval and C. O. Wilke, 2007 Thermodynamics of neutral protein evolution. Genetics 175: 255-266.

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

Bridgham, J. T., E. A. Ortlund and J. W. Thornton, 2009 An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461: 515-519.

Chen, P., and E. I. Shakhnovich, 2009 Lethal mutagenesis in viruses and bacteria. Genetics 183: 639-650.

536 Chou, H.-H., H.-C. Chiu, N. F. Delaney, D. Segrè and C. J. Marx, 2011 Diminishing
537 returns epistasis among beneficial mutations decelerates adaptation. Science
538 332**:** 1190-1192.
539 Cordell, H. J., 2002 Epistasis: what it means, what it doesn't mean, and statistical
540 methods to detect it in humans. Human molecular genetics 11**:** 2463-2468.
541 DePristo, M. A., D. M. Weinreich and D. L. Hartl, 2005 Missense meanderings in
542 sequence space: a biophysical view of protein evolution. Nature Reviews
543 Genetics 6**:** 678-687.
544 Dickinson, B. C., A. M. Leconte, B. Allen, K. M. Esvelt and D. R. Liu, 2013 Experimental
545 interrogation of the path dependence and stochasticity of protein evolution using
546 phage-assisted continuous evolution. Proceedings of the National Academy of
547 Sciences 110**:** 9007-9012.
548 Draghi, J. A., and J. B. Plotkin, 2013 Selection biases the prevalence and type of
549 epistasis along adaptive trajectories. Evolution 67**:** 3120-3131.
550 Drummond, D. A., and C. O. Wilke, 2008 Mistranslation-induced protein misfolding as a
551 dominant constraint on coding-sequence evolution. Cell 134**:** 341-352.
552 Dykhuizen, D. E., A. M. Dean and D. L. Hartl, 1987 Metabolic flux and fitness. Genetics
553 115**:** 25-31.
554 Echave, J., S. J. Spielman and C. O. Wilke, 2016 Causes of evolutionary rate variation
555 among protein sites. Nature Reviews Genetics.
556 Fierke, C. A., K. A. Johnson and S. J. Benkovic, 1987 Construction and evaluation of
557 the kinetic scheme associated with dihydrofolate reductase from Escherichia coli.
558 Biochemistry 26**:** 4085-4092.
559 Flint, H. J., R. W. Tateson, I. B. Barthelmess, D. J. Porteous, W. D. Donachie *et al.*,
560 1981 Control of the flux in the arginine pathway of Neurospora crassa.
561 Modulations of enzyme activity and concentration. Biochem J 200**:** 231-246.
562 Goldstein, R. A., 2011 The evolution and evolutionary consequences of marginal
563 thermostability in proteins. Proteins: Structure, Function, and Bioinformatics 79**:**
564 1396-1407.
565 Gong, L. I., M. A. Suchard and J. D. Bloom, 2013 Stability-mediated epistasis constrains
566 the evolution of an influenza protein. Elife 2**:** e00631.
567 Guerois, R., J. E. Nielsen and L. Serrano, 2002 Predicting changes in the stability of
568 proteins and protein complexes: a study of more than 1000 mutations. Journal of
569 molecular biology 320**:** 369-387.
570 Harms, M. J., and J. W. Thornton, 2013 Evolutionary biochemistry: revealing the
571 historical and physical causes of protein properties. Nature Reviews Genetics 14**:**
572 559-571.
573 Hartl, D. L., D. E. Dykhuizen and A. M. Dean, 1985 Limits of adaptation: the evolution of
574 selective neutrality. Genetics 111**:** 655-674.
575 Hecht, D., J. Tran and G. B. Fogel, 2011 Structural-based analysis of dihydrofolate
576 reductase evolution. Molecular phylogenetics and evolution 61**:** 212-230.
577 Kepp, K. P., and P. Dasmeh, 2014 A model of proteostatic energy cost and its use in
578 analysis of proteome trends and sequence evolution. PLoS One 9**:** e90504.
579 Lässig, M., V. Mustonen and A. M. Walczak, 2017 Predicting evolution. Nature Ecology
580 & Evolution 1**:** 0077.

Marks, D. S., T. A. Hopf and C. Sander, 2012 Protein structure prediction from sequence variation. Nature biotechnology 30: 1072-1080.

McCandlish, D. M., E. Rajon, P. Shah, Y. Ding and J. B. Plotkin, 2013 The role of epistasis in protein evolution. Nature 497: E1-E2.

Miton, C. M., and N. Tokuriki, 2016 How mutational epistasis impairs predictability in protein evolution and design. Protein Science.

Morcos, F., A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks *et al.*, 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108: E1293-E1301.

O'Dea, E. B., T. E. Keller and C. O. Wilke, 2010 Does mutational robustness inhibit extinction by lethal mutagenesis in viral populations? PLoS Comput Biol 6: e1000811.

Olson, C. A., N. C. Wu and R. Sun, 2014 A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Current Biology 24: 2643-2651.

Pál, C., B. Papp and M. J. Lercher, 2006 An integrated view of protein evolution. Nature Reviews Genetics 7: 337-348.

Parera, M., and M. A. Martinez, 2014 Strong epistatic interactions within a single protein. Molecular biology and evolution: msu113.

Pollock, D. D., G. Thiltgen and R. A. Goldstein, 2012 Amino acid coevolution induces an evolutionary Stokes shift. Proceedings of the National Academy of Sciences 109: E1352-E1359.

Rodrigues, J. V., S. Bershtein, A. Li, E. R. Lozovsky, D. L. Hartl *et al.*, 2016 Biophysical principles predict fitness landscapes of drug resistance. Proc Natl Acad Sci U S A.

Sarkisyan, K. S., D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin *et al.*, 2016 Local fitness landscape of the green fluorescent protein. Nature.

Serohijos, A. W., Z. Rimas and E. I. Shakhnovich, 2012 Protein biophysics explains why highly abundant proteins evolve slowly. Cell reports 2: 249-256.

Serohijos, A. W., and E. I. Shakhnovich, 2014a Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. Molecular biology and evolution 31: 165-176.

Serohijos, A. W., and E. I. Shakhnovich, 2014b Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. Current opinion in structural biology 26: 84-91.

Shah, P., D. M. McCandlish and J. B. Plotkin, 2015 Contingency and entrenchment in protein evolution under purifying selection. Proceedings of the National Academy of Sciences 112: E3226-E3235.

Soskine, M., and D. S. Tawfik, 2010 Mutational effects and the evolution of new protein functions. Nature Reviews Genetics 11: 572-582.

Starr, T. N., and J. W. Thornton, 2016 Epistasis in protein evolution. Protein Science.

Studer, R. A., B. H. Dessailly and C. A. Orengo, 2013 Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochemical Journal 449: 581-594.

Süel, G. M., S. W. Lockless, M. A. Wall and R. Ranganathan, 2003 Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature structural & molecular biology 10: 59-69.

Taverna, D. M., and R. A. Goldstein, 2002a Why are proteins marginally stable? Proteins: Structure, Function, and Bioinformatics 46: 105-109.

Taverna, D. M., and R. A. Goldstein, 2002b Why are proteins so robust to site mutations? Journal of molecular biology 315: 479-484.

Tokuriki, N., F. Stricher, J. Schymkowitz, L. Serrano and D. S. Tawfik, 2007 The stability effects of protein mutations appear to be universally distributed. Journal of molecular biology 369: 1318-1332.

Tokuriki, N., F. Stricher, L. Serrano and D. S. Tawfik, 2008 How protein stability and new functions trade off. PLoS Comput Biol 4: e1000002.

Weinreich, D. M., N. F. Delaney, M. A. DePristo and D. L. Hartl, 2006 Darwinian evolution can follow only very few mutational paths to fitter proteins. science 312: 111-114.

Weinreich, D. M., Y. Lan, C. S. Wylie and R. B. Heckendorn, 2013 Should evolutionary geneticists worry about higher-order epistasis? Curr Opin Genet Dev 23: 700-707.

Weinreich, D. M., R. A. Watson, L. Chao and R. Harrison, 2005 Perspective: sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59: 1165-1174.

Whitlow, M., A. J. Howard, D. Stewart, K. D. Hardman, L. F. Kuyper *et al.*, 1997 X-ray Crystallographic Studies of Candida albicans Dihydrofolate Reductase HIGH RESOLUTION STRUCTURES OF THE HOLOENZYME AND AN INHIBITED TERNARY COMPLEX. Journal of Biological Chemistry 272: 30289-30298.

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

Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution 24: 1586-1591.

Yin, S., F. Ding and N. V. Dokholyan, 2007 Eris: an automated estimator of protein stability. Nature methods 4: 466-467.

Zeldovich, K. B., P. Chen and E. I. Shakhnovich, 2007 Protein stability imposes limits on organism complexity and speed of molecular evolution. Proceedings of the National Academy of Sciences 104: 16152-16157.

Supplementary information for:

Estimating the contribution of folding stability to non-specific epistasis in proteins

Pouria Dasmeh[1,2], Adrian W.R. Serohijos[1,2,*]

[1]Departement de Biochimie, [2]Centre Robert Cedergren en Bioinformatique et Génomique, Université de Montréal, 2900 Edouard-Montpetit, Montreal, Quebec H3T 1J4, Canada.

**Correspondence:* adrian.serohijos@umontreal.ca

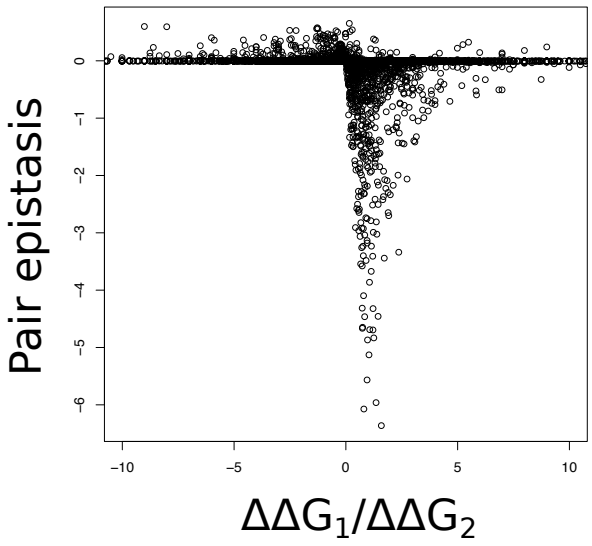

Figure S1. Pair epistasis (**Equation 8** in the main text) versus the ratio of stability effect of two pair substitutions.

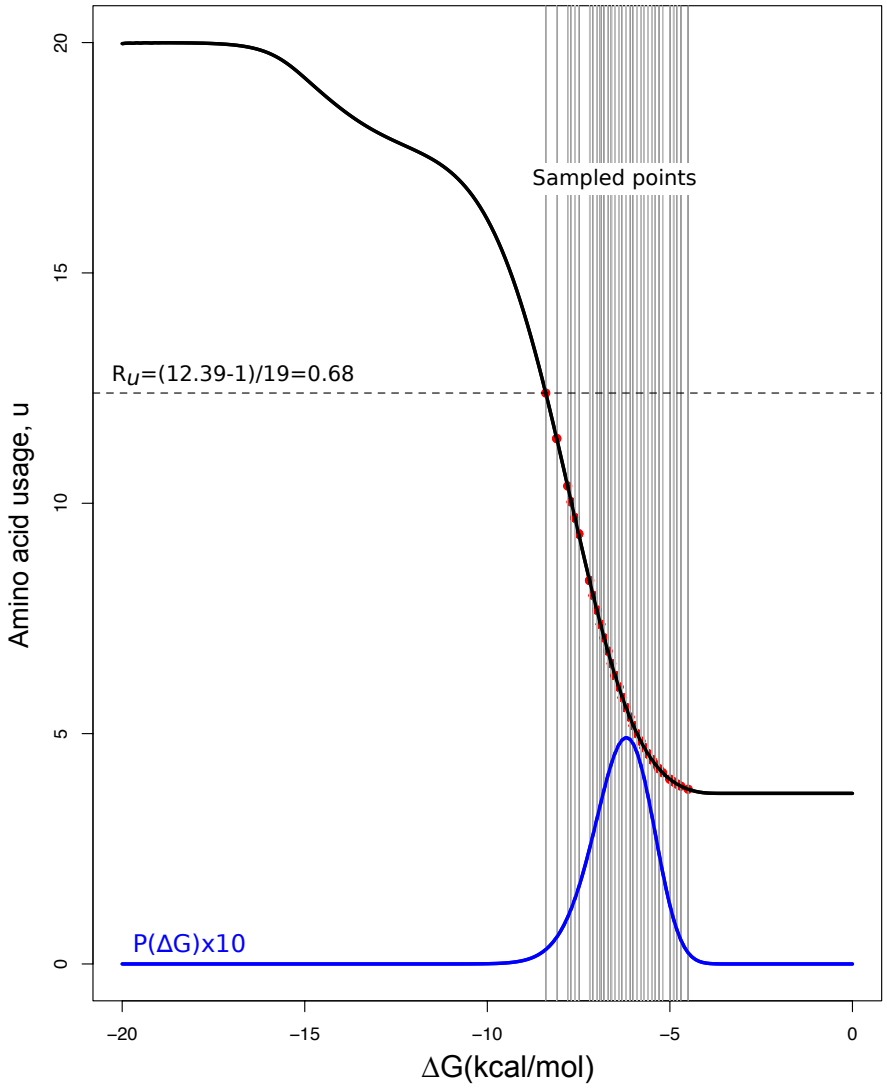

Figure S2. Fraction of fixed amino acids at different stabilities. Fixation is counted when $N \times P_{fix} \geq 1$. Red points represent 100 sampled stabilities and the vertical lines show the corresponding fraction of fixed amino acids. R_u is calculated as the fraction of fixed amino acids at the highest sampled stability. The blue plot shows $10 \times P(\Delta G)$. The factor 10 is multiplied by $P(\Delta G)$ to scale the distribution for the presentation purpose.

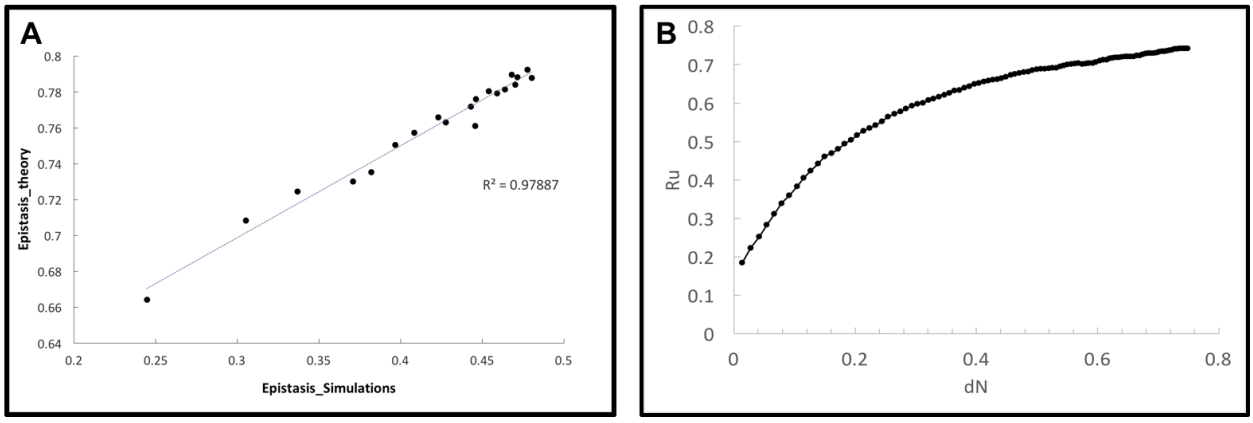

Figure S3. **(A)** Calculated epistasis by the theoretical approach versus simulation (R2=0.97 and p-value<10^{-16} using Wilcoxon signed-rank test). **(B)** R_u versus normalized nonsynonymous substitution rate by nonsynonymous sites, d*N*.

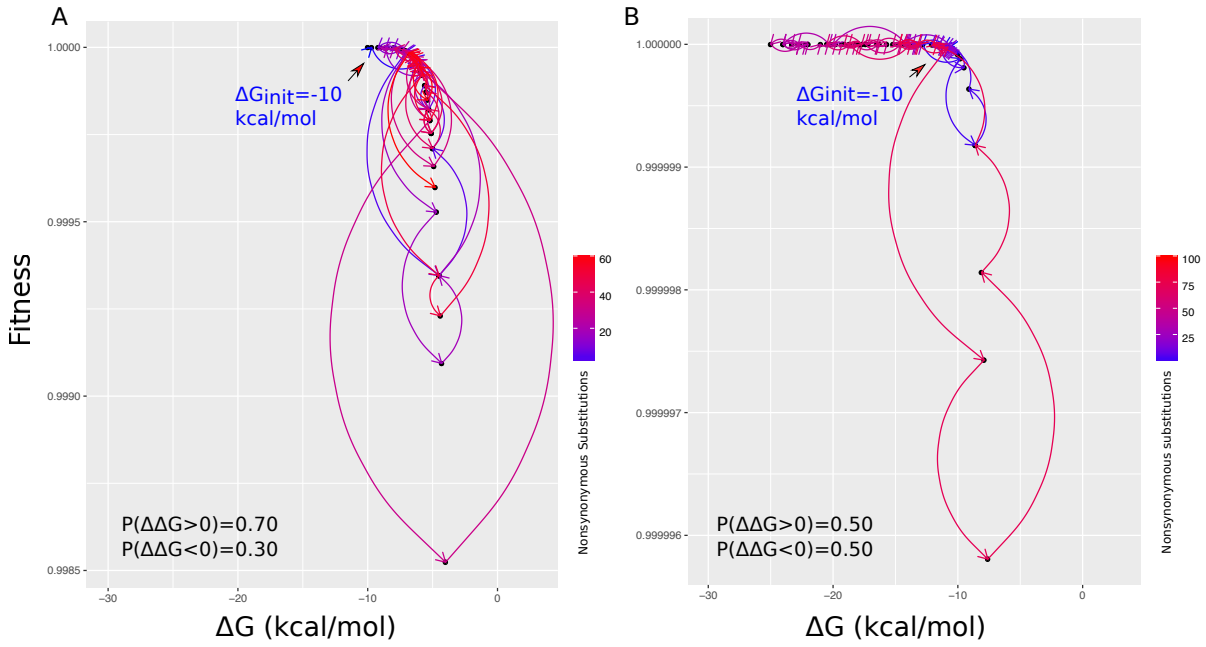

Figure S4. Destabilization shifts protein to more curved parts of fitness landscape and hence higher epistasis. **(A)** Evolutionary trajectory of an ancestral protein with stability of -10 kcal/mol when fraction of destabilizing and stabilizing mutations are 0.7 and 0.3 respectively. **(B)** the same plot in A with the difference in the fraction of destabilizing and stabilizing mutations to be equal. In both plots, orders of fixed mutations are depicted from blue to red. Fitness is proportional to F_{nat}, number of folded copies of proteins per cell.

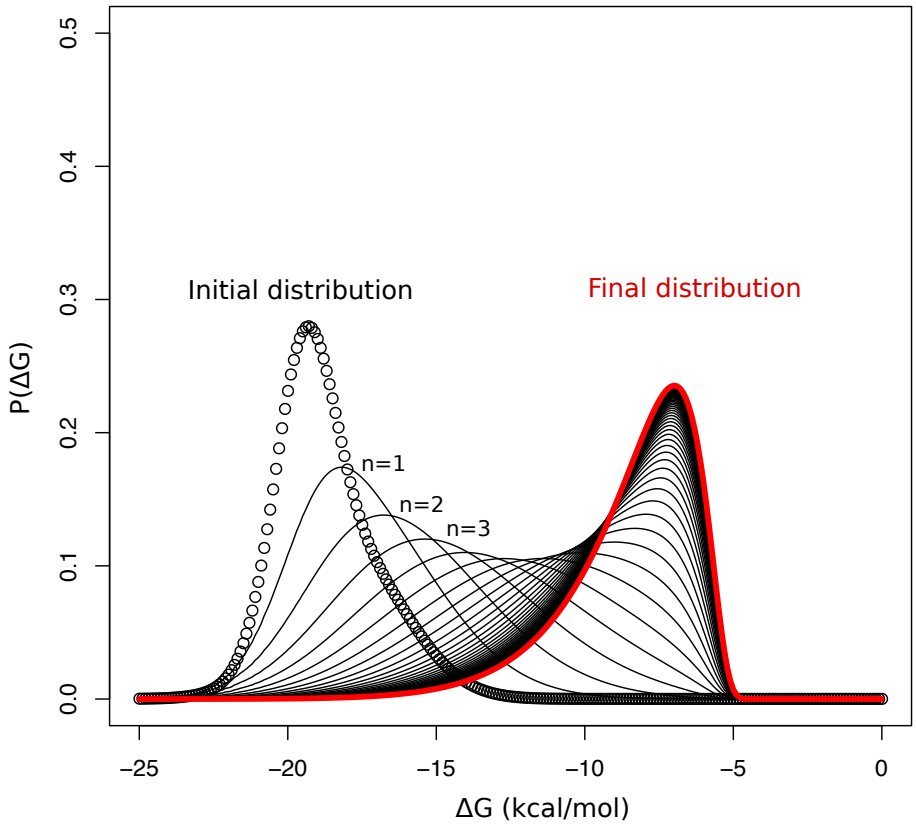

Figure S5. Evolution of distribution of protein stabilities within the numerical approach (equations 8 and 9 in the main text). The initial distribution is the distribution of mutational effects on a WT protein with ΔG = -20 kcal/mol without selection. The mutational steps are shown as n=1, n=2 and n=3 representing distribution of protein stabilities after one, two and three mutations respectively. The final distribution is shown in red which is the balance between mutation (i.e., causing destabilization) and selection for stabilizing mutations.

Table S1. Mutational usage, u, dN, dS, R_u and $R_{dN/dS}$ for 2000 simulated trajectories sampled at 100 time points (each time point= 10N where N is the population size).

Time point	u	dN	dS	R_u	$R_{dN/dS}$
1	4.4948	0.0134	0.0362	0.1839	0.3701
2	5.2396	0.0279	0.0720	0.2231	0.3871
3	5.7865	0.0415	0.1087	0.2519	0.3821
4	6.3802	0.0544	0.1447	0.2832	0.3761
5	6.9167	0.0667	0.1801	0.3114	0.3706
6	7.4427	0.0785	0.2163	0.3391	0.3629
7	7.8229	0.0910	0.2526	0.3591	0.3604
8	8.2760	0.1040	0.2890	0.3829	0.3598
9	8.6927	0.1152	0.3244	0.4049	0.3552
10	9.0469	0.1266	0.3614	0.4235	0.3502
11	9.3906	0.1385	0.4002	0.4416	0.3460
12	9.7500	0.1498	0.4362	0.4605	0.3434
13	9.9271	0.1613	0.4733	0.4698	0.3409
14	10.1250	0.1721	0.5108	0.4803	0.3369
15	10.3854	0.1824	0.5487	0.4940	0.3324
16	10.5729	0.1932	0.5866	0.5038	0.3293
17	10.8177	0.2034	0.6240	0.5167	0.3260
18	11.0260	0.2136	0.6599	0.5277	0.3237
19	11.1510	0.2235	0.7013	0.5343	0.3186
20	11.3073	0.2338	0.7405	0.5425	0.3158
21	11.4948	0.2440	0.7802	0.5524	0.3128
22	11.7135	0.2542	0.8185	0.5639	0.3106
23	11.8646	0.2647	0.8563	0.5718	0.3091
24	11.9740	0.2749	0.8960	0.5776	0.3068

25	12.1250	0.2843	0.9401	0.5855	0.3024
26	12.2656	0.2936	0.9784	0.5929	0.3000
27	12.3646	0.3027	1.0224	0.5981	0.2960
28	12.4010	0.3118	1.0686	0.6001	0.2917
29	12.5365	0.3207	1.1159	0.6072	0.2874
30	12.6146	0.3295	1.1715	0.6113	0.2812
31	12.7083	0.3381	1.2283	0.6162	0.2753
32	12.7969	0.3473	1.2925	0.6209	0.2687
33	12.8906	0.3552	1.3587	0.6258	0.2614
34	13.0052	0.3637	1.4270	0.6319	0.2549
35	13.0365	0.3720	1.5076	0.6335	0.2467
36	13.1510	0.3802	1.5994	0.6395	0.2377
37	13.2292	0.3885	1.6938	0.6436	0.2293
38	13.3385	0.3966	1.8081	0.6494	0.2193
39	13.3958	0.4045	1.9316	0.6524	0.2094
40	13.4531	0.4124	2.0540	0.6554	0.2008
41	13.4948	0.4200	2.2002	0.6576	0.1909
42	13.5417	0.4276	2.3312	0.6601	0.1834
43	13.5833	0.4348	2.4917	0.6623	0.1745
44	13.6250	0.4416	2.6533	0.6645	0.1664
45	13.6979	0.4490	2.8194	0.6683	0.1593
46	13.7813	0.4566	3.0083	0.6727	0.1518
47	13.8229	0.4636	3.1939	0.6749	0.1452
48	13.8802	0.4707	3.3855	0.6779	0.1390
49	13.9271	0.4779	3.5768	0.6804	0.1336
50	13.9427	0.4856	3.8040	0.6812	0.1277
51	14.0260	0.4922	3.9935	0.6856	0.1233

52	14.0573	0.4994	4.1985	0.6872	0.1190
53	14.0833	0.5063	4.3951	0.6886	0.1152
54	14.0990	0.5131	4.6142	0.6894	0.1112
55	14.1146	0.5192	4.8394	0.6902	0.1073
56	14.1406	0.5253	5.0564	0.6916	0.1039
57	14.1458	0.5319	5.2396	0.6919	0.1015
58	14.1979	0.5382	5.4282	0.6946	0.0992
59	14.2552	0.5441	5.6198	0.6976	0.0968
60	14.2917	0.5501	5.7863	0.6996	0.0951
61	14.3177	0.5565	5.9731	0.7009	0.0932
62	14.3490	0.5621	6.1320	0.7026	0.0917
63	14.3698	0.5683	6.3050	0.7037	0.0901
64	14.3281	0.5746	6.5007	0.7015	0.0884
65	14.3385	0.5801	6.6560	0.7020	0.0872
66	14.3698	0.5856	6.8187	0.7037	0.0859
67	14.3802	0.5921	6.9661	0.7042	0.0850
68	14.4271	0.5972	7.1060	0.7067	0.0840
69	14.4792	0.6034	7.2342	0.7094	0.0834
70	14.5260	0.6090	7.3925	0.7119	0.0824
71	14.5365	0.6149	7.5238	0.7124	0.0817
72	14.5990	0.6204	7.6460	0.7157	0.0811
73	14.6354	0.6249	7.7511	0.7177	0.0806
74	14.6615	0.6300	7.8720	0.7190	0.0800
75	14.6458	0.6350	7.9788	0.7182	0.0796
76	14.6667	0.6402	8.0832	0.7193	0.0792
77	14.6927	0.6456	8.1622	0.7207	0.0791
78	14.6875	0.6501	8.2432	0.7204	0.0789

79	14.6979	0.6548	8.3070	0.7209	0.0788
80	14.6979	0.6596	8.3776	0.7209	0.0787
81	14.7344	0.6649	8.4354	0.7229	0.0788
82	14.7552	0.6691	8.5072	0.7240	0.0787
83	14.7969	0.6737	8.5862	0.7262	0.0785
84	14.8385	0.6783	8.6573	0.7283	0.0784
85	14.8646	0.6832	8.7226	0.7297	0.0783
86	14.8594	0.6879	8.7751	0.7294	0.0784
87	14.8698	0.6924	8.8326	0.7300	0.0784
88	14.8958	0.6973	8.8836	0.7314	0.0785
89	14.9219	0.7016	8.9306	0.7327	0.0786
90	14.9479	0.7063	8.9804	0.7341	0.0787
91	14.9583	0.7105	9.0179	0.7346	0.0788
92	14.9740	0.7146	9.0494	0.7355	0.0790
93	14.9948	0.7191	9.0898	0.7366	0.0791
94	15.0156	0.7233	9.1252	0.7377	0.0793
95	15.0625	0.7273	9.1542	0.7401	0.0794
96	15.0833	0.7316	9.1820	0.7412	0.0797
97	15.0885	0.7358	9.2186	0.7415	0.0798
98	15.0885	0.7403	9.2497	0.7415	0.0800
99	15.0938	0.7445	9.2742	0.7418	0.0803
100	15.0885	0.7483	9.3081	0.7415	0.0804

Table S2. Mutational usage, u, dN, dS, R_u and $R_{dN/dS}$ for different number of sampled sequences in simulations and theory.

Number of sequences	u	dN	dS	dNdS_simul	dNdS_theory	Ru_simul	Ru_theory
100	5.6458	0.1738	0.4993	0.3482	0.4256	0.2445	0.6643
200	6.7969	0.1742	0.5103	0.3413	0.4252	0.3051	0.7085
300	7.3958	0.1720	0.5105	0.3369	0.4227	0.3366	0.7247
400	8.0417	0.1727	0.5090	0.3394	0.4229	0.3706	0.7303
500	8.2552	0.1720	0.5104	0.3370	0.4239	0.3819	0.7354
600	8.5365	0.1723	0.5122	0.3365	0.4238	0.3967	0.7505
700	8.7552	0.1722	0.5086	0.3385	0.4247	0.4082	0.7575
800	9.0365	0.1722	0.5058	0.3405	0.4240	0.4230	0.7660
900	9.1250	0.1727	0.5089	0.3394	0.4230	0.4276	0.7632
1000	9.4167	0.1724	0.5108	0.3375	0.4240	0.4430	0.7719
1100	9.4635	0.1729	0.5084	0.3401	0.4240	0.4454	0.7613
1200	9.4740	0.1719	0.5122	0.3356	0.4243	0.4460	0.7762
1300	9.6250	0.1726	0.5123	0.3369	0.4241	0.4539	0.7806
1400	9.7188	0.1728	0.5098	0.3390	0.4237	0.4589	0.7793
1500	9.8125	0.1716	0.5098	0.3366	0.4227	0.4638	0.7815
1600	9.8906	0.1714	0.5113	0.3352	0.4245	0.4679	0.7898
1700	9.9323	0.1719	0.5107	0.3366	0.4239	0.4701	0.7842
1800	9.9583	0.1720	0.5110	0.3366	0.4233	0.4715	0.7883
1900	10.0729	0.1723	0.5105	0.3375	0.4236	0.4775	0.7925
2000	10.1250	0.1721	0.5108	0.3369	0.4235	0.4803	0.7880

Table S3. Parameters of distribution of mutational effects for different globular proteins.

Protein	μ_1 [a]	μ_2 [b]	σ_1 [c]	σ_2 [d]	P_1 [e]	P(ΔΔG>0)	% epistasis
Recombinant serum paraoxonase 1	0.51	0.91	1.93	1.82	0.42	0.77	0.33
Lipase	0.57	1.18	2.27	2.05	0.67	0.72	0.25
β-Lactamase	0.58	1.11	2.36	1.84	0.67	0.74	0.29
Human carbonic anhydorase II	0.5	0.85	1.8	1.99	0.39	0.76	0.29
Dihydrofolate reductase	0.54	0.73	1.53	1.78	0.39	0.76	0.29
Robinuclease H	0.49	0.98	2.12	1.98	0.69	0.71	0.24
Myoglobin	0.31	0.84	1.53	1.57	0.48	0.71	0.25
Staphilococcus nuclease	0.58	0.68	1.3	1.71	0.3	0.76	0.27
Human lysosome	0.76	1.12	3.02	2.29	0.7	0.77	0.34
Hen lysosome	0.81	1.16	3	2.43	0.68	0.78	0.34
Ribonuclease A	0.59	0.83	1.76	2.56	0.55	0.73	0.21
Barnase	0.59	0.84	2.08	1.93	0.51	0.78	0.34
Acylphosphatase	0.56	0.8	1.93	1.83	0.56	0.77	0.32
Ubiquitin	0.42	0.83	1.33	1.64	0.52	0.71	0.22
Protein G	0.59	0.84	2.08	1.93	0.51	0.78	0.34
Cro repressor	0.55	0.74	1.33	1.58	0.48	0.76	0.27

a: average value of the first Gaussian distribution. b: average value of the second distribution. c: standard deviation of the first Gaussian distribution. d: standard deviation of the second Gaussian distribution. e: fraction of destabilizing mutations. f: fraction of stabilizing mutations. (See Equation 8 for details). Population size is 10^4. Gray cells show are the ones different from default parameters shown in the first row. e: weight of first Gaussian distribution (see Equation 7 in the main text). Data is taken from (TOKURIKI *et al.* 2007).

References

Tokuriki, N., F. Stricher, J. Schymkowitz, L. Serrano and D. S. Tawfik, 2007 The stability effects of protein mutations appear to be universally distributed. Journal of molecular biology 369**:** 1318-1332.