Functionality and the evolution of marginal stability in proteins: Inferences from lattice simulations

Paul D. Williams[1], David D. Pollock[2] and Richard A. Goldstein[3]

[1]Department of Chemistry, University of Michigan, Ann Arbor, MI, 48109, USA; [2]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, 70803, USA; [3]Mathematical Biology, National Institute for Medical Sciences, The Ridgeway, Mill Hill, London MW7 1AA, UK

Abstract: It has been known for some time that many proteins are marginally stable. This has inspired several explanations. Having noted that the functionality of many enzymes is correlated with subunit motion, flexibility, or general disorder, some have suggested that marginally stable proteins should have an evolutionary advantage over proteins of differing stability. Others have suggested that stability and functionality are contradictory qualities, and that selection for both criteria results in marginally stable proteins, optimised to satisfy the competing design pressures. While these explanations are plausible, recent research simulating the evolution of model proteins has shown that selection for stability, ignoring any aspects of functionality, can result in marginally stable proteins because of the underlying makeup of protein sequence-space. We extend this research by simulating the evolution of proteins, using a computational protein model that equates functionality with binding and catalysis. In the model, marginal stability is not required for ligand-binding functionality and we observe no competing design pressures. The resulting proteins are marginally stable, again demonstrating that neutral evolution is sufficient for explaining marginal stability in observed proteins.

Keywords: lattice models, protein thermodynamics, molecular evolution

Introduction

It has been repeatedly observed that a high proportion of globular proteins are marginally stable under physiological conditions, with a $\Delta G_{\text{folding}}$ of about -5 to -10 kcal/mol. (Brandts 1967; Privalov and Khechinashvili 1974; Savage et al. 1993; Ruvinov et al. 1997; Vogl et al. 1997; Giver et al. 1998). This is in spite of several factors that suggest that stable proteins might have advantages over marginally stable proteins. For instance, flexible proteins may be less resistant to proteolysis (Fontana, Polverino de Laureto, and De Filippis 1993; Hubbard, Eisenmenger, and Thornton 1994; Fontana et al. 1997; Hubbard, Beynon, and Thornton 1998), denaturation (Wagner and Wuthrich 1979), detrimental conformational change (Carrell and Lomas 1997; Dobson 2001; Bucciantini et al. 2002), aggregation (Lomas and Carrell 2002), and loss of active-site integrity. In addition, binding between less stable and more flexible proteins and their corresponding ligands requires strong binding interactions. More stable proteins do not need such strong binding energies because they lose less entropy upon binding (Schulz 1979). These consequences of higher stability might be expected to increase the evolutionary success of organisms containing stabilised forms of these proteins, suggesting that highly stable proteins should be more common.

The fact that most proteins are not highly stable suggests that other factors are involved, and several hypotheses have been developed to explain this discrepancy. Most of these hypotheses centre on various reasons why marginally stable proteins would have a selective advantage over more stable proteins. For instance, it has been suggested that proteins have evolved towards marginal stability in order to function, suggesting that there is a narrow range of stability consistent with functionality (Rasmussen et al. 1992; Tsou 1998; Zavodszky et al. 1998). There are several reasons why functionality might be limited to proteins of marginal stability. Marginally stable proteins might be more flexible (Wagner and Wuthrich 1979; Tang and Dill 1998), increasing functionality by enabling the formation of strong binding interactions with specific ligands or by providing flexibility needed for conformational change (Lipscomb 1970; Artymiuk et al. 1979; Frauenfelder, Petsko, and Tsernoglou 1979; Wrba et al. 1990; Varley and Pain 1991; Daniel, Dines, and Petach 1996; Zavodszky et al. 1998; Daniel and Cowan 2000). It has also been suggested that marginal stability may be

Correspondence: Richard A Goldstein: richard.goldstein@nimr.mrc.ac.uk

advantageous because ligand binding with marginally stable proteins is comparatively difficult. Binding affinities and specificities of less stable proteins might be more readily adjusted by mutation, phosphorylation or other processes, or selectivity might be enhanced by allowing binding only in the presence of specific interactions (Dunker et al. 1998; Wright and Dyson 1999; Dunker and Obradovic 2001; Dunker and others 2001; Namba 2001). In addition, the physiological importance of marginal stability might involve considerations other than maximizing functionality or selectivity. For instance, unstable proteins might provide more rapid turnover than stable proteins.

A second class of explanations involves the hypothesis that marginal stability is the result of a trade-off between functionality and stability, that the constraints on the amino acids imposed by functionality reduce the opportunities to produce a highly-stable protein resulting in a negative correlation between functionality and stability. This could result if functionality required specific amino acids at functionally-important locations that were incompatible with high stability, so that large numbers of possible sequences, including those with high stability, would be excluded from the evolutionary dynamics. It has been observed, for instance, that mutations increasing protein stability and activity are much more rare than mutations increasing either separately (Alber and Wozniak 1985; Bryan et al. 1986; Liao, McKenzie, and Hageman 1986; Shoichet et al. 1995), although the presence of mutations that increase both (Giver et al. 1998) indicates that functionality and stability are not mutually exclusive. If this trade-off holds for all protein sequences, marginal stability is prevalent because it provides the best balance between the sequences that result in stability and functionality.

These explanations generally arise from an 'adaptationist' paradigm, which is to say that the observation of marginal stability in proteins requires an explanation of how this contributes to the reproductive fitness of the organism, either as a direct adaptation or as optimization given constraints. Random events and processes, however, are important factors in the dynamics of evolution and can influence the characters that result (Sueoka 1962; Kimura 1968; King and Jukes 1969). Gould and Lewontin have emphasised the importance of examining possible alternatives to adaptationist selection (Gould and Lewontin 1979). Specifically, they stress that the present usefulness of a character may belie its origins, so that one should avoid ascribing characters to adaptation simply due to their present use. Random events may have led to the existence of the character, which was used to advantage only later. Before adaptation can be judged the cause of the emergence of a character, other explanations must be ruled out.

Consistent with these ideas, a third explanation of the observed marginal stability in proteins involves the concept of regions of sequence space and what has been termed 'designability' (Govindarajan and Goldstein 1995; Govindarajan and Goldstein 1996; Li et al. 1996; Buchler and Goldstein 1998; Shakhnovich 1998; England, Shakhnovich, and Shakhnovich 2003). The idea is that the 'sequence entropy' or volume of sequence space corresponding to any property influences whether this property is likely to result from evolutionary dynamics. This idea has been applied to the distribution of different structures (Finkelstein and Ptitsyn 1987; Govindarajan and Goldstein 1995; Govindarajan and Goldstein 1996; Li et al. 1996), the question of whether proteins fulfill the thermodynamic hypothesis (Govindarajan and Goldstein 1998), the stability of proteins, (Taverna and Goldstein 2002), ligand binding properties (Blackburne and Hirst 2001; Williams, Pollock, and Goldstein 2001; Blackburne and Hirst 2003; Bloom et al. 2004), and the ability of proteins to explore a range of different structures (Govindarajan and Goldstein 1997a; Govindarajan and Goldstein 1997b; Bornberg-Bauer and Chan 1999; Taverna and Goldstein 2000; Deeds, Dokholyan, and Shakhnovich 2003; Tiana et al. 2004; Shakhnovich et al. 2005). In previous work exploring the evolution of lattice proteins with a fitness function dependent only on stability, it was found that sequence entropy was a sufficient explanation for the observation of marginal stability in proteins, and that this effect would favor mechanisms of function consistent with marginal stability (Taverna and Goldstein 2002). In further work, we developed a simple model with fitness represented by the ability of the protein to bind a ligand, observing similar results (Williams, Pollock, and Goldstein 2001).

To examine these hypotheses in the context of protein functionality and to extend the previous

work, we simulate the evolution of proteins using a lattice-protein model of protein-folding and ligand-binding that allows the study of protein stability and function. In these simulations, the fitness function is based on diffusion-limited reaction kinetics. The model is designed so that fitness increases with binding strength, which tends to increase with stability, meaning that marginally stable proteins have no evolutionary advantage over proteins of greater stability. We also observe no design constraints or trade-offs between stability and functionality among these evolved proteins. Our evolutionary simulations, however, lead to marginally stable proteins. This indicates again, with a more realistic simulation, that random evolution is sufficient to generate marginal stability. This does not prove that marginal stability is not an adaptation, but rather demonstrates that marginal stability could result in the absence of any adaptive role, that its presence does not indicate that it plays such an adaptive role, and therefore that marginal stability is not a phenomenon that needs an explanation based on evolutionary advantage.

Methods

Protein Model

Protein models should accurately represent important and relevant aspects of real proteins yet be simple enough for rapid computational evaluation. Our model must be relatively simple indeed, as evolution simulations involve the examination of many protein sequences over a large number of generations. To examine the previously described hypotheses, the model must map protein sequence to stability in a compact state as well as the propensity for binding and acting upon a specified ligand.

The details of this model have been more thoroughly described elsewhere (Williams, Pollock, and Goldstein 2001). Each model protein consists of a chain of 16 amino acids on a 2-dimensional square lattice. While the 2-dimensional model is problematic for dynamics simulations (Shakhnovich 1997), for thermodynamic analyses involving sums over states it is more accurate at representing the appropriate number of buried vs. exposed residues. Intra-protein contacts are defined as non-sequentially-adjacent residues one lattice-unit apart in distance. Compact structures have nine contacts (the maximum number possible for a 16-residue protein) and fit in a square with four residues per side. All 802,075 possible structures are considered, of which only 69 are compact.

The free energy $G(k)$ of a sequence $\{A_1, A_2...A_{16}\}$ in conformation k is given by

$$G(k) = \sum_{r<s} \gamma(A_r, A_s) Q^k_{r,s} \tag{1}$$

where $\gamma(A_r, A_s)$ is the contact potential between amino acids A_r and A_s, and where $Q^k_{r,s}$, is 1 if residues r and s are in contact in structure k, and is otherwise 0. The contact potentials are obtained from the statistical analysis of Miyazawa and Jernigan, who developed a contact-potential matrix that describes the interactions between amino acids (Miyazawa and Jernigan 1985). Due to the nature of this statistical analysis, these potentials represent 'potentials of mean force', implicitly including hydrophobic interactions and other effects of the solvent. They therefore represent contributions to the free energy rather than enthalpy. In this matrix, the influence of covalent cysteine crosslinks is shown by the high magnitude of the Cys-Cys potential. As such binary interactions are incompatible with the contact potential as encoded in our model, and would significantly change the number and character of allowed conformations, we do not consider them in our model. To account for this, we use a modified potential matrix where the Cys-Cys potential has been replaced by the Ser-Ser potential. In addition, the values in the matrix have all been multiplied by two to counteract the effect of the limited number of two dimensions.

We use Boltzmann statistics to determine $P(k)$, the thermodynamic probability of folding into conformation k, assuming all conformations are in equilibrium:

$$P(k) = \frac{\exp\left(\frac{-G(k)}{k_B T}\right)}{\sum_{|\text{All onformation } k'} \exp\left(\frac{-G(k')}{k_B T}\right)}. \tag{2}$$

where T is the temperature and k_B is Boltzmann's constant, and again the sum in the denominator is

over all structures, both compact and extended. P(Compact) is defined as the sum of probabilities of all compact structures; the change in free energy upon folding into a compact state is then

$$\Delta G(\text{Compact}) = -k_B T \ln\left(\frac{P(\text{Compact})}{1-P(\text{Compact})}\right) \quad (3)$$

We assume that the conformation with the lowest free energy should be the native state of the protein; since we are mainly interested in compact structures, the compact structure with the lowest free energy shall be referred to as the native state (Govindarajan and Goldstein 1998).

We model protein-ligand binding as a four-residue peptide contacting any of the four sides of a compact protein, such that maximal contact between the ligand and the face of the protein is made, as illustrated in Figure 1. The ligand may face either of two directions on any of the four sides of a conformation, so there are $69 \times 4 \times 2 = 552$ possible binding sites on a protein sequence. The free energy of a complex where the protein is in compact conformation k and the ligand is bound to site l is

$$G(k,l) = G(k) + \sum_{r}^{16}\sum_{q}^{4} \gamma(A_r, A_q)\, Q_{r,q}^{k,l} \quad (4)$$

where q is over the four locations in the peptide ligand, and $Q^{k,l}_{r,q}$ is equal to 1 if residue r in the protein is in contact with residue q in the ligand in this particular bound conformation. We use Boltzmann statistics to determine the probability that the protein binds the ligand

P^0(Bound) =

$$\frac{\sum_{k,l}\exp\left(\frac{-G(k,l)}{k_B T} + \frac{\Delta S_{\text{lig}}}{k_B}\right)}{\sum_{k^4}\exp\left(\frac{-G(k')}{k_B T}\right) + \sum_{k,l}\exp\left(\frac{-G(k,l)}{k_B T} + \frac{\Delta S_{\text{lig}}}{k_B}\right)} \quad (5)$$

where ΔS_{lig} is the concentration-dependent change in the entropy of the ligand upon binding. We represent the probability of ligand bound by P^0(Bound) to indicate that this is calculated without considering any forward reactions, that this assumes an equilibrium between the bound and unbound forms.

Under conditions when very little of the protein is bound to ligand, we can ignore the second term in the denominator and calculate the relative probability of the protein binding the ligand for a fixed concentration by multiplying P^0(Bound) by $\exp(-\Delta S_{\text{lig}}/k_B)$, yielding

$$P^0_{\text{rel}}(\text{Bound}) \equiv \exp\left(-\frac{\Delta S_{\text{lig}}}{k_B}\right) P^0(\text{Bound}) \approx \frac{\sum_{k,l}\exp\left(\frac{-G(k,l)}{k_B T}\right)}{\sum_{k'}\exp\left(\frac{-G(k')}{k_B T}\right)}. \quad (6)$$

Evolution Model

We model a population of random proteins evolving through mutation and replication. Starting with an initial population of 1000 protein sequences, we allow a fixed rate of mutations, modeled as a Poisson distribution with a mean of 20 mutations per generation. Sequences are then replicated according to their fitness, as described below. The population size is maintained at a constant level of 1000 proteins throughout the experiment.

Measure of Fitness

Many factors affect the evolutionary success of a protein, ranging from the intrinsic properties of the protein to external and indirectly related circumstances. For this paper, we are concerned only with the effects of selection for protein functionality. To study these effects, we construct a fitness function based on the rate of catalysis of bound ligands.

Assuming that product-formation is beneficial, we consider fitness directly proportional to the rate of catalysis. We model this rate with Michaelis--Menten kinetics, corresponding to reactions of the following type,

$$P + L \underset{k_{\text{uni}}}{\overset{k_D}{\rightleftarrows}} PL \xrightarrow{k_2} P + \text{Product} \quad (7)$$

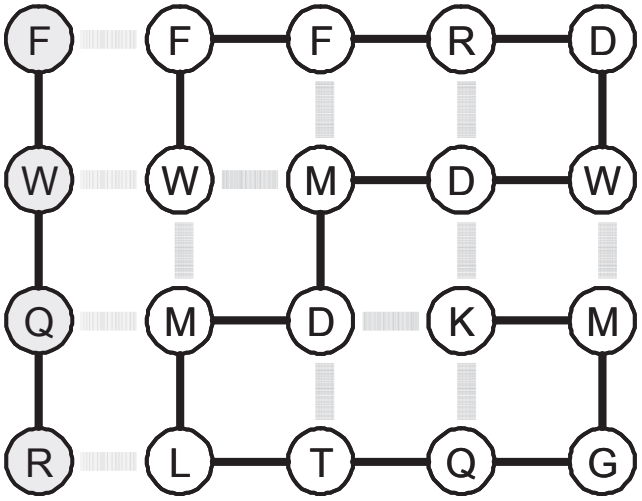

Figure 1. An example of a model protein in a compact conformation bound to a ligand, shown in grey. Covalent bonds are shown as solid lines, contact interactions as thick (intramolecular) and thin (intermolecular) stripes.

where P, L, and PL are the protein, ligand, and protein-ligand encounter complex, respectively, and k_D, k_{uni}, and k_2 are the rates of diffusional encounter, unimolecular dissociation, and catalysis, respectively. While our protein-ligand binding model is designed for the calculation of thermodynamic probabilities and cannot be used to explicitly calculate the kinetics of folding or binding processes, we assume that k_D does not depend upon the strength of binding, so that as P^0(Bound) increases, k_{uni} should decrease to satisfy the conditions for equilibrium.

We are interested in investigating the ease with which marginal stability could result in evolving proteins in the absence of any specific advantage to marginal stability. For this reason, we assume that k_2 is independent of binding strength. In this case, the rate of catalysis, and thus the fitness used in the evolution simulations, is given by

$$\text{Fitness} = \frac{1}{1+\frac{P^0(\text{Bound})_{1/2}}{P^0(\text{Bound})}} = \frac{1}{1+\frac{P^0_{\text{rel}}(\text{Bound})_{1/2}}{P^0_{\text{rel}}(\text{Bound})}} \quad (8)$$

where $P^0(\text{Bound})_{1/2}$ is equal to

$$P^0(\text{Bound})_{1/2} = \frac{k_D[L]}{k_2} \quad (9)$$

and $P^0_{\text{rel}}(\text{Bound})_{1/2} \equiv \exp(-\Delta S_{\text{lig}})P^0(\text{Bound})_{1/2}$. (See Appendix for a detailed derivation.)

Fitness increases monotonically with increasing P^0_{rel}(Bound), and approaches the maximum value asymptotically. This asymptotic domain represents the diffusion-limited nature of Michaelis-Menten kinetics - at a certain point, better ligand-binding will not result in faster catalysis.

In our evolution runs, proteins are selected for their ability to bind a specific ligand. The sixteen four-residue permutations of glutamate and lysine can form the strongest binding interactions of any ligand, while the optimal binding interaction of a polyalanine ligand is the weakest possible. We performed evolution runs using the ligands AAAA, EEEE, and EKEK, to examine the influence ligand-choice has on evolution. Real proteins have a wide range of k_2 values, and act on ligands of varying concentration, diffusion rates, and ΔS_{lig}. To account for this variety, we performed simulations with a range of values for $P^0_{\text{rel}}(\text{Bound})_{1/2}$. For AAAA, $P^0_{\text{rel}}(\text{Bound})_{1/2}$ was 1.25, 12.5, 50, and 100, for EEEE it was 1.25, 100, 400, 3750, 7500 and 15,000, and for EKEK it was 1.25, 100, 400, and 3750.

In addition to performing evolution experiments, we also optimise proteins for maximum fitness. Beginning with a random sequence and a specified ligand, we perform hill-climbing walks on the fitness landscape. The steps made on this landscape are random point mutations of the protein sequence; a mutation is accepted if it results in an increase in the fitness, that is to say, an increase in P^0_{rel}(Bound). This walk is continued until the protein sequence resulting in a local fitness maximum is reached. By calculating the fitness for all single-point-mutants of the sequence, we ensure that the sequence is indeed at a local-maximum. We performed 1000 optimization runs for each ligand. We also perform similar hill-climbing walks designed to maximise P(Compact), independent of any fitness based on ligand-binding or catalysis.

Results

The results of a typical evolution run, showing population-weighted averages of P(Compact) and P^0_{rel}(Bound) are illustrated in Figure 2. The data for the first 10,000 generations are the results of one of the fifty experimental runs

with ligand EEEE and $P^0_{rel}(Bound)_{½} = 15{,}000$. This run has been extended for an additional 5000 generations with a lower $P^0_{rel}(Bound)_{½}$ of 1.25. $\langle P^0_{rel}(Bound)\rangle$ increases steadily and rapidly for several hundred generations, then fluctuates until generation 10,000. This behaviour is due to the semineutral relationship between fitness and $P^0_{rel}(Bound)$ described in equation 9. Initially, proteins with higher values of $P^0_{rel}(Bound)$ have a selective advantage and are more successful at reproducing, resulting in the increase of $\langle P^0_{rel}(Bound)\rangle$. As proteins with very high values of $P^0_{rel}(Bound)$ emerge and become established in the population, the selective advantage of higher binding probability diminishes, proteins become more equally fit, and the makeup of population becomes more subject to random factors than to fitness effects. After $P^0_{rel}(Bound)_{½}$ decreases at generation 10,000, $\langle P^0_{rel}(Bound)\rangle$ decreases rapidly, not due to selection for weaker binding affinity, but due to the larger number of mutations that decrease rather than increase binding affinity. For both values of $P^0_{rel}(Bound)_{½}$, but especially for $P^0_{rel}(Bound)$=15,000 P(Compact) is approximately equal to P(Native State), indicating that one compact conformation is dominant at equilibrium. It is also generally true that the simulations produce a single dominant binding site, which dominates $P^0_{rel}(Bound)$.

Figure 3 illustrates the physical properties of the final generations of the evolution experiments for two different ligands, AAAA (3a-3b) and EEEE (3c-3d), compared with the proteins that result from optimization through hill-climbing. Results for different values of $P^0_{rel}(Bound)_{½}$ are represented as different colours on the graphs. To examine the stability distribution of the evolved and optimised proteins, we plot cumulative distributions of their levels (calculated with Equation 4) in Figures 3a and 3c. These results are population-weighted, so that common, well-represented proteins influence the distribution more than poorly-represented transients. Most proteins have ΔG(Compact) > -1, indicating that proteins are at most marginally stable, consistent with observations of real proteins. This is more clearly evident compared with the distribution of ΔG(Compact) values for proteins that have been optimised for stability with a hill-climbing algorithm.

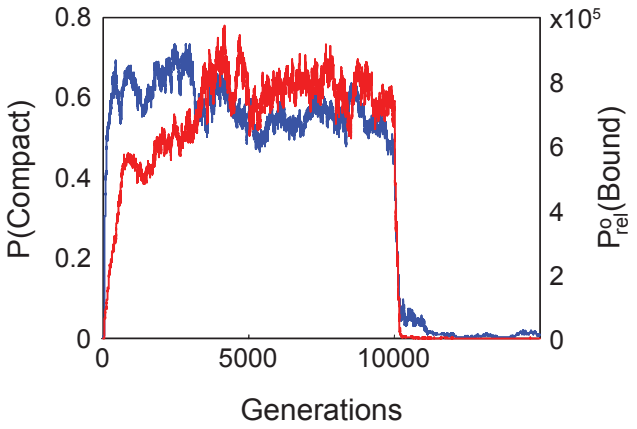

Figure 2. Extended typical evolution run with fitness based on ability to bind and catalyze EEEE, showing the effects of changing $P^0_{rel}(Bound)_{½}$. The value of $P^0_{rel}(Bound)_{½}$ in the first 10,000 generations equals 15,000, and for the last 5000 generations $P^0_{rel}(Bound)_{½}$ equals 1.25. The values plotted are population-weighted averages. Blue line: $\langle P(\text{Compact})\rangle$, the probability that a protein is in a compact state. Red line: $\langle P^0_{rel}(Bound)\rangle$, the relative probability that a protein binds a ligand at any site in any conformation. $\langle P(\text{Native State})\rangle$, the probability of the compact structure with minimum free energy, is indistinguishable from $\langle P(\text{Compact})\rangle$ throughout the simulation.

There are several possible explanations for the relationship between overall protein stability and probability of binding a ligand. Thermodynamically, we would expect that a protein with higher stability would, on average, bind more strongly than unstable proteins, due to the entropy penalty when a less-stable protein binds a ligand. Alternatively, the 'optimization given constraints' model suggests that there might be a negative correlation between protein stability and binding, as the residues that optimized stability might not be the same as the residues that optimised binding interactions. The positive correlation between binding probability and protein stability shown in Figures 3b and 3d, demonstrates that the thermodynamic effect dominates any possible trade-offs between stability and binding.

Further evidence of the lack of trade-off between ligand binding and protein stability is provided by considering the proteins that have been optimised for binding by the hillclimbing algorithm. We observe no correlation between the stability of the proteins that result and the strength of their binding interactions with the peptide ligand, as would be expected if strong binding interactions were incompatible with high protein stability.

In general, the properties of evolved proteins fall within a range of values, but the ligand and

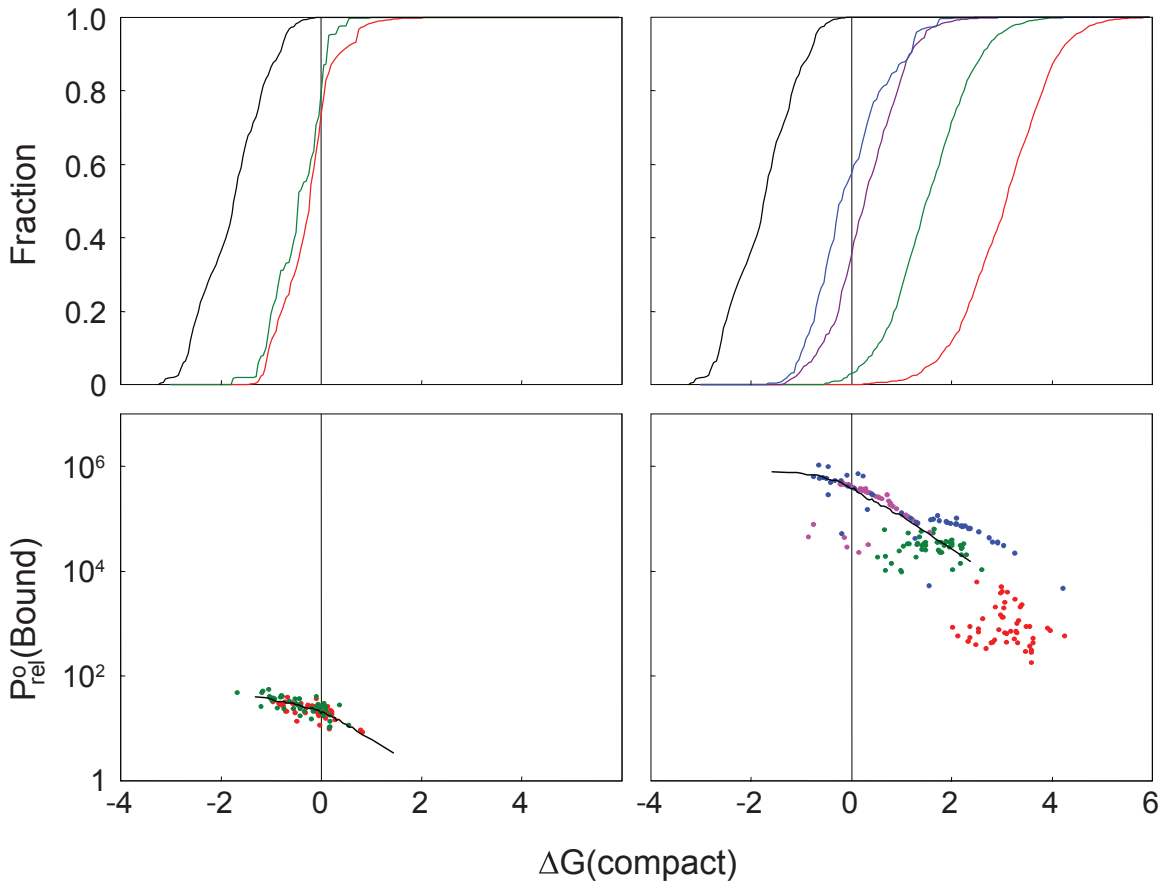

Figure 3. The properties of evolved and optimised proteins, colour-coded by the value of $P^0_{rel}(\text{Bound})_{½}$:$P^0_{rel}(\text{Bound})_{½}$ =1.25 (red), 100 (green), 3750 (magenta), and 15,000 (blue). Not all values of $P^0_{rel}(\text{Bound})_{½}$ are included in each plot. In the left column (panels **a, b**) the ligand is AAAA, in the right (panels **c, d**), EEEE. **a** and **c**. Cumulative distribution of protein conformational stability. The fraction of final-generation (or optimized) proteins with a AG(Compact) less than the value on the x-axis is plotted on the y-axis. The properties of proteins optimised for compaction are represented in black. **b** and **d**. Scatter 22 plot of relative binding probability vs. protein conformational stability. Each point represents the population average of one simulation at the final generation. Black line represents the averages of approximately 8000 proteins optimised to bind the ligand. Simulations with similar values of ΔG(Compact) were binned together into 50 groups, and average values of ΔG(Compact) and P^0_{rel}(Bound) computed for proteins in each group.

the $P^0_{rel}(\text{Bound})_{½}$ affect the properties of resulting proteins within these bounds. As can be seen in Figure 3, among proteins evolved to bind AAAA there are no discernible changes in the distributions of protein stability, binding probability, and binding interaction strength with different values of $P^0_{rel}(\text{Bound})_{½}$, but among proteins evolved to bind EEEE these quantities tend to increase with $P^0_{rel}(\text{Bound})_{½}$. In addition, the variation in protein properties is higher for proteins evolved with EEEE, especially at lower values of $P^0_{rel}(\text{Bound})_{½}$. Proteins evolved to bind EKEK have similar properties to proteins evolved to bind EEEE.

These differences are due to the nature of the ligand. A mutation can increase $P^0_{rel}(\text{Bound})_{½}$ in two ways: by increasing the complementarity of a binding site or by increasing the probability of a compact state with a favourable binding site. The binding strength of AAAA is of lesser magnitude than that of EEEE, reflecting the difference in the strength of optimal binding interaction. This means that AAAA can only form weak binding interactions; as soon as binding faces evolve that form these interactions, the only way to increase $P^0_{rel}(\text{Bound})_{½}$ is to increase the probability of being in the conformationally correct state. Not all proteins evolved to bind AAAA have formed an optimal-interaction binding site, but binding interactions in most proteins are close to optimal. Proteins evolving to bind EEEE can form stronger binding interactions with a wider variety of binding faces, and a wider variety of successful proteins results. The results for EKEK are similar to

those for EEEE. Proteins optimised for ligand binding tend to be relatively stable and have relatively strong binding interactions compared with evolved proteins; in contrast, evolved proteins have higher variation in both ΔG(Bound) and ΔG(Compact).

Discussion

Figure 3 shows that most proteins resulting from our evolutionary experiments are not highly stable, consistent with observations about real proteins. In fact, most of the resulting proteins tend to have low or marginal stability levels, depending on ligand and value of $P^0_{\text{rel}}(\text{Bound})_{½}$.

As described in the Introduction, there are two different possible reasons generally given for the marginal stability of real proteins - either a specific selective advantage of marginal stability, or a co-optimization of the conflicting qualities of stability and other aspects of functionality. In our model, we have eliminated the first possibility by construction, in that proteins that bind more strongly have a higher fitness. We also observe a positive correlation between stability and fitness, so the theory of conflicting design pressures does not apply in our model. With both of these possible explanations inappropriate for these simulations, we still observe marginally stable proteins.

This provides evidence that sequence entropy, the third possible explanation, is sufficient to explain the observation of marginal stability in biological proteins. We are not ruling out the other proposed explanations, but we can explain the observed properties of evolved proteins without them. The existence of marginal stability in real proteins implies neither an evolutionary advantage to marginal stability nor a trade-off between stability and binding strength. The most parsimonious explanation for marginal stability does not include either of these two mechanisms.

The protein models used in this study are smaller than realistic proteins, and thus might better represent the area surrounding an active site more than an entire protein. Of the three explanations used to explain marginal stability - evolutionary advantage of marginal stability, negative correlation between stability and fitness due to design constraints, and sequence entropy - we would expect the first explanation to be independent of protein size, the second explanation to be especially appropriate around the active site, while the third explanation would involve the entire protein, as the protein generally is required to be folded in order to bind and catalyse a ligand. The fact that we do not observe evidence for the second explanation while the third explanation seems to be adequate in these smaller models suggests that it should also be adequate when a more realistically-sized protein is considered.

The effect of the underlying nature of protein sequence space may have been more important in early protein evolution. Most proteins with random amino-acid sequences are highly unstable (in our model, and likely in the real world), so the first existing proteins would likely have been unstable as well. Some degree of stability was likely beneficial, so mutations that lead to increased stability were accepted. At a certain point, the fitness gains of higher stability were counter-acted by the effect of sequence entropy, and thus protein stability did not increase further.

The resulting stability depends upon the ligand, as well as $P^0_{\text{rel}}(\text{Bound})_{½}$. Specific predictions can be made on the basis of this analysis. For instance, we would expect that observed protein stabilities should depend upon the corresponding value of $P^0_{\text{rel}}(\text{Bound})_{½} = \exp(-\Delta S_{\text{lig}}/k_B)k_D[L]/k_2$, in that ligands with higher values of $P^0_{\text{rel}}(\text{Bound})_{½}$ would likely correspond to more stable proteins. This would be the case for smaller ligands (faster k_D), and slower catalysis (slower k_2). In fact, one might expect that protein stabilities could lessen with time as the catalytic steps became more optimised, reducing the value of $P^0_{\text{rel}}(\text{Bound})_{½}$. The current analysis also suggests that highly-sticky proteins (strong binding strength) would correspond to proteins with less stability, as the selective pressure on stability would be reduced. While there are obvious examples of this, such as calmodulin, further investigation is required to see if this is a general principle.

Appendix: Derivation of the fitness function

To derive the fitness function, we start with equation 8. In the Michaelis-Menten model, the rate of change of $[PL]$ is assumed to be zero, so the steady state concentration of PL is

$$[PL] = \frac{k_D [P][L]}{k_{\text{uni}} + k_2}. \quad (10)$$

The total concentration of protein, $[P]_T$, is equal to $[P]+[PL]$. Solving for $[P]$ and $[PL]$ in terms of $[P]_T$ yields

$$[P] = \frac{k_{\text{uni}} + k_2}{k_{\text{uni}} + k_2 + k_D [L]} [P]_T \quad (11)$$

$$[PL] = \frac{k_{\text{D}} [L]}{k_{\text{uni}} + k_2 + k_{\text{D}} [L]} [P]_T. \quad (12)$$

The rate of production of the final product is

$$V = \frac{d[\text{Product}]}{dt} = k_2 [PL] = \frac{k_2 k_D [L]}{k_{\text{uni}} + k_2 + k_D [L]} [P]_T. \quad (13)$$

We can relate this to terms calculated from our protein model by expressing P^0(Bound), the relative thermodynamic probability that the protein binds a ligand, in terms of protein and ligand concentrations. $P^0(\text{Bound}) = [PL]/[P]_T$, under conditions when there is no forward reaction, or k_2=0. Under these conditions, equation 13 becomes,

$$P^0(\text{Bound}) = \frac{k_D [L]}{k_{\text{uni}} + k_D [L]}. \quad (14)$$

Solving for k_{uni} yields

$$k_{\text{uni}} = \frac{k_D [L](1 - P^0(\text{Bound}))}{P^0(\text{Bound})}. \quad (15)$$

Substituting this expression into equation 14 yields

$$V = \frac{k_D [L][P]_T}{1 + \frac{k_D [L]}{k_2 P^0(\text{Bound})}}. \quad (16)$$

We assume that k_D, $[P]_T$, and $[L]$ remain relatively constant, and as fitness is relative and thus not changed by a multiplicative factor, the final fitness function is then:

$$\text{Fitness} = \frac{1}{1 + \frac{P^0(\text{Bound})_{1/2}}{P^0(\text{Bound})}} \quad (17)$$

where $P^0(\text{Bound})_{1/2} = k_D[L]/k_2$ is the value of P^0(Bound) where the fitness is half the maximum fitness.

Acknowledgements

Thanks to Darin Taverna for helpful discussions and Todd Raeker for computational assistance. Financial support was provided by the Medical Research Council and by NIH grant R01 LM005770.

References

Alber, T., and J. A. Wozniak. 1985. A genetic screen for mutations that increase the thermal-stability of phage-T4 lysozyme. *Proc. Natl. Acad. Sci. USA* 82:747--750.

Artymiuk, P. J., C. C. F. Blake, D. E. P. Grace, S. J. Oatley, D. C. Phillips, and M. J. E. Sternberg. 1979. Crystallographic studies of the dynamic properties of lysozyme. *Nature* 280:563--568.

Blackburne, B., and J. Hirst. 2001. Evolution of functional model proteins. *J ChemPhys* 115.

Blackburne, B. P., and J. D. Hirst. 2003. Three-dimensional functional model proteins: structure function and evolution. *J Chem Phys* 119: 3453--3460.

Bloom, J. D., C. O. Wilke, F. H. Arnold, and C. Adami. 2004. Stability and the evolvability of function in a model protein. *Biophysical Journal* 86:2758-2764.

Bornberg-Bauer, E., and H. S. Chan. 1999. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *PNAS* 96:10689-10694.

Brandts, J. F. 1967. Heat effects on proteins and enzymes. Pp. 25--72 in A. H. Rose, ed. Thermobiology. Academic Press Inc., New York.

Bryan, P. N., M. L. Rollence, M. W. Pantoliano, J. Wood, B. C. Finzel, G. L. Gilliland, A. J. Howard, and T. L. Poulos. 1986. Proteases of enhanced stability: characterization of a thermostable variant of subtilisin. *Proteins: Struct. Funct. Genet.* 1:326--334.

Bucciantini, M., E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. Dobson, and M. Stefani. 2002. Inherent toxicity of aggregates 24 implies a common mechanism for protein misfolding diseases. *Nature* (London) 416:507--511.

Buchler, N. E. G., and R. A. Goldstein. 1998. Effect of Alphabet Size and Foldability Requirements on Protein Structure Designability. *Proteins: Struct.*, *Funct.*, *Genet.* 34:113-124.

Carrell, R. W., and D. A. Lomas. 1997. Conformational Disease. *Lancet* 350:134--138.

Daniel, R. M., and D. A. Cowan. 2000. Biomolecular stability and life at high temperatures. *Cell. Mol. Life Sci.* 57:250-264.

Daniel, R. M., M. Dines, and H. H. Petach. 1996. The Denatruation and Degradation of Stable Enzymes at High Temperatures. *Biochem. J.* 317:1-11.

Deeds, E. J., N. V. Dokholyan, and E. I. Shakhnovich. 2003. Protein evolution within a structural space. *Biophysical Journal* 85: 2962-2972.

Dobson, C. M. 2001. The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B* 356:133--145.

Dunker, A. K., E. Garner, S. Guilliot, P. Romero, K. Albrecht, J. Hart, Z. Obradovic, C. Kissinger, and J. E. Villafranca. 1998. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pacific Symp. Biocomputing* **3**:473--484.

Dunker, A. K., and Z. Obradovic. 2001. The protein trinity--linking function and disorder. *Nat. Biotechnol.* **19**:805--806.

Dunker, A. K., and others. 2001. Intrinsically Disordered Protein. *J. Mol. Graphics Modell.* **19**:26-59.

England, J. L., B. E. Shakhnovich, and E. I. Shakhnovich. 2003. Natural selection of more designable folds: a mechanism for thermophilic adaptation. *PNAS* **100**:8727-8731.

Finkelstein, A. V., and O. B. Ptitsyn. 1987. Why do globular proteins fit the limited set of folding patterns. *Prog. Biophys. Mol. Biol.* **50**: 171--190.

Fontana, A., P. Polverino de Laureto, and V. De Filippis. 1993. Molecular aspects of proteolysis of globular proteins *in* W. van den Tweel, A. Harder, and M. Buitelear, eds. Protein Stability and Stabilization. Elsevier Science Publ., Amsterdam.

Fontana, A., M. Zambonin, P. Polverino de Laureto, V. De Filippis, A. Clementi, and E. Scaramella. 1997. Probing the conformational state of apomyoglobin by limited proteolysis. *J. Mol. Biol.* **266**:223--230.

Frauenfelder, H., G. A. Petsko, and D. Tsernoglou. 1979. Temperature dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* **280**:558-563.

Giver, L., A. Gershenson, P. O. Freskgard, and F. H. Arnold. 1998. Directed evolution of a thermostable esterase. *Proc. Nat. Acad. Sci. USA* **95**:12809-12813.

Gould, S. J., and R. C. Lewontin. 1979. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London*, *Series B* **205**:581-598.

Govindarajan, S., and R. A. Goldstein. 1998. On the Thermodynamic Hypothesis of Protein Folding. *Proc. Nat. Acad. Sci. USA* **95**:5545-5549.

Govindarajan, S., and R. A. Goldstein. 1997a. Evolution of Model Proteins on a Foldability Landscape. *Proteins* **29**:461-466.

Govindarajan, S., and R. A. Goldstein. 1997b. The Foldability Landscape of Model Proteins. *Biopolymers* **42**:427-438.

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

Govindarajan, S., and R. A. Goldstein. 1995. Searching for foldable protein structures using optimized energy functions. *Biopolymers* **36**:43--51.

Hubbard, S. J., R. J. Beynon, and J. M. Thornton. 1998. Assesment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng.* **11**:349--359.

Hubbard, S. J., F. Eisenmenger, and J. M. Thornton. 1994. Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Sci.* **3**:757--768.

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

King, J. L., and T. H. Jukes. 1969. Non-Darwinian Evolution. *Science* **164**:788-798.

Li, H., R. Helling, C. Tang, and N. Wingreen. 1996. Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science* **273**: 666-669.

Liao, H., T. McKenzie, and R. Hageman. 1986. Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. *Proc. Natl. Acad. Sci. USA* **83**:576--580.

Lipscomb, W. N. 1970. Structure and mechanism in the enzymatic activity of Carboxypeptidase A and relations to chemical sequence. *Acc. Chem. Res.* **3**:81-89.

Lomas, D. A., and R. W. Carrell. 2002. Serpinopathies and the conformational dementias. *Nature Reviews Genetics* **3**:759-768.

Miyazawa, S., and R. L. Jernigan. 1985. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromol.* **18**:534-552.

Namba, K. 2001. Roles of partly unfolded conformations in macromolecular selfassembly. *Genes to Cells* **6**:1-12.

Privalov, P. L., and N. N. Khechinashvili. 1974. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J. Mol. Biol.* 86:665-684.

Rasmussen, B. F., A. M. Stock, D. Ringe, and G. A. Petsko. 1992. Crystalline ribonuclease A loses function below the dynamical transition at 220K. *Nature* 357:423-424.

Ruvinov, S., L. Wang, B. Ruan, O. Almog, G. L. Gilliland, E. Eisenstein, and P. N. Bryan. 1997. Engineering the independent folding of the subtilisin BPN' prodomain: analysis of two-state folding versus protein stability. *Biochemistry* 36:10414-10421.

Savage, H. J., C. J. Elliot, C. M. Freeman, and J. L. Finney. 1993. Lost hydrogen-bonds and buried surface-area: Rationalizing stability in globular -proteins. Journal of the *Chemical Society Faraday Transactions* 89:2609-2617.

Schulz, G. E. 1979. Nucleotide Binding Proteins. Pp. 79--94 in M. Balaban, ed. Molecular Mechanism of Biological Recognition. Elsevier/ North-Holland Biomedical Press, New York.

Shakhnovich, B. E., E. Deeds, C. Delisi, and E. Shakhnovich. 2005. Protein structure and evolutionary history determine sequence space topology. *Genome Research* 15:385-392.

Shakhnovich, E. I. 1998. Protein design: a perspective from simple tractable models. *Folding and Design* 3:R45-R58.

Shakhnovich, E. I. 1997. Theoretical studies of protein folding thermodynamics and kinetics. *Curr Opin Struct Biol* 7:29-40.

Shoichet, B. K., W. A. Baase, R. Kuroki, and B. W. Matthews. 1995. A relationship between protein stability and protein function. *Proc. Nat. Acad. Sci. USA* 92:452-456.

Sueoka, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA* 48:582-592.

Tang, K. E. S., and K. A. Dill. 1998. Native protein fluctuations: the conformationalmotion temperature and the inverse correlation of protein flexibility with protein stability. *J. Biomol. Struct. Dyn.* 16:397-411.

Taverna, D., and R. A. Goldstein. 2000. The distribution of structures in evolving protein populations. *Biopolymers* 53:1-8.

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

Tiana, G., B. E. Shakhnovich, N. V. Dokholyan, and E. I. Shakhnovich. 2004. Imprint of evolution on protein structures. *Proceedings of the National Academy of Sciences of the United States of America* 101:2846-2851.

Tsou, C. L. 1998. Active site flexibility in enzyme catalysis. *Annal. N. Y. Acad. Sci.* 864:1-8.

Varley, P. G., and R. H. Pain. 1991. Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate kinases from yeast and \extit Thermus thermophilus. *J. Mol. Biol.* 220:531-538.

Vogl, T., C. Jatzke, H. J. Hinz, J. Benz, and R. Huber. 1997. Thermodynamic stability of annexin V E17G: Equilibrium parameters from an irreversible unfolding raction. *Biochemistry* 36:1657-1668.

Wagner, G., and K. Wuthrich. 1979. Correlation between the amide proton exchange rates and the denaturation temperatures in globular proteins related to the basic pancreatic trypsin inhibitor. *J. Mol. Biol.* 130:31-37.

Williams, P. D., D. D. Pollock, and R. A. Goldstein. 2001. Evolution of functionality in lattice proteins. *J. Mol. Graphics Modell.* 19:150-156.

Wrba, A., A. Schweiger, V. Schultes, R. Jaenicke, and P. Zavodszky. 1990. Extremely thermostable D-glyceraldehyde-3-phosphate dehydrogenase fromthe eubacterium \extitThermotoga maritima. *Biochemistry* 29:7584-7592.

Wright, P. E., and H. J. Dyson. 1999. Intrinsically Unstructured Proteins: Reassessing the protein structure-function paradigm. *J. Mol. Biol.* 293:321-331.

Zavodszky, P., K. Jozsef, A. Svingor, and G. A. Petsko. 1998. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Nat. Acad. Sci. USA* 98:7406-7411.