doi:10.1016/j.jmb.2006.08.020 J. Mol. Biol. (2006) 363, 422–432

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Point Mutations in Protein Globular Domains: Contributions from Function, Stability and Misfolding

I. E. Sánchez[1]*, J. Tejero[2,3], C. Gómez-Moreno[2,3], M. Medina[2,3] and L. Serrano[1]

[1]*European Molecular Biology Laboratory, Meyerhofstrasse 1 69117 Heidelberg, Germany*

[2]*Institute of Biocomputation and Physics of Complex Systems (BIFI), Universidad de Zaragoza, E-50009 Zaragoza Spain*

[3]*Departamento de Bioquímica y Biología Molecular y Celular Universidad de Zaragoza E-50009 Zaragoza, Spain*

**Corresponding author*

Several contrasting hypotheses have been formulated about the influence of functional and conformational properties, like stability and avoidance of misfolding, on the evolution of protein globular domains. Selection at functional sites has been suggested to be detrimental to stability or coupled to it. Avoidance of misfolding may be achieved by discarding misfolding-prone sequences or by maintaining a stable native state and thus destabilizing partially or fully unfolded states from which misfolding can take place. We have performed a hierarchical analysis of a large database of point mutations to dissect the relative contributions of function, stability and misfolding in the evolution of natural sequences. We show that at catalytic sites, selection for function overrules selection for stability but find no evidence for an anticorrelation between function and stability. Selection for stability plays a secondary role at binding sites, but is not fully coupled to selection for function. Remarkably, we did not find a selective pressure against misfolding-prone sequences in globular proteins at the level of individual positions. We suggest that such a selection would compromise native-state stability due to a correlation between the stabilities of native and misfolded states. Stabilization of the native state is the most frequent way in which natural proteins avoid misfolding.

Keywords: protein stability; function; mutation; amyloid; aggregation

Introduction

Point mutations are a crucial mechanism in the evolution of protein globular domains. Several factors have been proposed to influence which mutations actually take place in nature, such as the conservation of functional residues,[1,2] and biophysical properties, like native state stability,[3–8] and negative design against misfolded states.[9–15] The relative importance of function, stability and misfolding in protein globular domains has not been established unambiguously. Some functional sites have been reported to be suboptimal in terms of stability,[16–24] leading to the principle of function–stability trade-off.[16–24] Some sites contribute to both stability and function,[6,16,25–33] which led to the opposing hypothesis of a coupling between stability and function.[6,26,29,30] Similarly, the evolutionary relationship between stability and misfolding is not clear. Misfolding of a protein requires a total or partial unfolding event that exposes a misfolding-prone stretch to the solvent.[34] It has been proposed that natural globular proteins avoid misfolding by keeping a stable native state (and thus destabilizing partially or fully unfolded states from which misfolding can take place)[35] or by destabilizing putative misfolded states.[9–15] To our knowledge, the relative occurrence of these strategies in nature has not been addressed.

The lack of consensus on the roles of function, stability and misfolding likely indicates that proteins can choose from a variety of strategies to fulfil the different functional and conformational constraints. A full understanding of protein evolution should identify all possible strategies but also determine which are used more frequently in nature. We have performed a systematic analysis of the importance of function, stability and the prevention of amyloid fibril formation and amorphous aggregation in determining the sequence of protein globular

Present address: J. Tejero, Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA.

E-mail address of the corresponding author: ignacio.sanchez@embl.de

domains. By means of a large database of point mutations, a combination of experimental data and empirical tools and a stepwise analysis, we identified and ranked the roles of several functional and conformational properties in the evolution of natural sequences. We identify two main kinds of sequence positions. Selection for function is the main pressure at a minority of sites, sometimes overruling selection for stability. Selection for native-state stability dominates at a majority of sites and, indirectly, protects proteins against misfolding. Interestingly, the stabilities of native states, amyloid fibrils and amorphous aggregates seem to be correlated.

Results

General approach

We analyzed a data set of 2351 point mutations in 44 globular domains (see Supplementary Data). Mutations were first classified according to which of the proposed selection pressures were present at each site. We assigned a role in catalysis,[36] or binding of other molecules,[37–39] using information from structural and protein engineering studies (see Methods). Sites involved in the formation of amyloid fibrils,[10] or amorphous aggregates,[40] were identified using computer empirical tools (see Methods). We used *in vitro* measurements of native state stability[41] to model *in vivo* stability.[42,43]

As a second step, each mutation in the database was classified as "allowed" or "forbidden", depending on whether or not the mutant residue appears in homologous sequences.[44–46] In the case of allowed mutations, we used the relative occurrences of wild-type and mutant residues in the family of homologous sequences, f_{wt} and f_{mutant}, to calculate a change in evolutionary pseudo free energy upon mutation (see Methods):[3,7,8]

$$\Delta\Delta G_{evolution} = -RT \cdot \ln(f_{mutant}/f_{wt}) \qquad (1)$$

The validity of this evolutionary pseudo free energy is supported by the statistically significant correlation between $\Delta\Delta G_{evolution}$ and the experimental $\Delta\Delta G_{stability}$ for mutations not involved in either function or misfolding (see below).

The third step in the analysis was to test the correlation of the allowed *versus* forbidden status and the $\Delta\Delta G_{evolution}$ of a point mutation and the data available for each of the proposed evolutionary pressures (stability, function and misfolding). Fourth and last, we investigated which evolutionary pressure was stronger by analyzing groups of mutations where two pressures (stability and misfolding, stability and function, misfolding and function) were present.

Stability

We first considered the group of mutations not involved in either function or misfolding (see Methods). The main evolutionary pressure at these positions is expected to be the stability of the native state against unfolding.[4,7] The strong correlation between folding rate and stability makes the results presented here for stability also applicable to folding.[47] A total of 30% of the mutations in this group are forbidden (Table 1). They are, on average, more destabilizing than allowed mutations (Table 1). We observe a statistically significant correlation between $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{evolution}$ for the group of allowed mutations (Figure 1(a)). The standard deviation between predicted and observed $\Delta\Delta G$-values is 1.07 kcal/mol, significantly better than that for available methods based on sequence,[48] and close to those based on structure.[49] The *R*-value and statistical significance of the correlation depend little on secondary structure or solvent exposure at the site of mutation (data not shown), supporting the generality of the result. The *R*-value of the correlation increases with the number of homologous sequences used in the calculation of $\Delta\Delta G_{evolution}$ (data not shown), indicating that the main limitation of this approach is our sampling of the repertoire of natural sequences for each domain. Altogether, our results generalize previous studies that focused on single domains,[3,6,8,50] and demonstrate that native-state stability is the main evolutionary pressure acting on this group of mutation sites.

Table 1. The effect of stability, function and misfolding on the evolution of protein globular domains

Class	No. mutations	% Forbidden	$\Delta\Delta G_{evolution}$ (kcal/mol)		$\Delta\Delta G_{stability}$ (kcal/mol)		
			Allowed	P[a]	Allowed	Forbidden	P[b]
Stability	966	30	0.85±1.02	–	0.90±1.24	1.97±1.86	$<10^{-4}$
Stability + catalysis	72	74	2.17±1.29	$<10^{-4}$	0.55±2.11	0.33±3.25	0.79
Stability + binding	379	31	0.94±1.02	0.27	0.38±1.49	1.16±1.70	$<10^{-4}$
Stability + aggregation	361	20	0.91±1.08	0.48	1.17±1.45	2.50±1.72	$<10^{-4}$
Stability + amyloid	141	28	0.87±1.06	0.88	1.00±1.45	2.26±2.10	$2\cdot10^{-3}$
Stability + binding + aggregation	108	30	1.07±1.12	0.33	0.65±1.77	1.77±1.71	$3\cdot10^{-3}$
Stability + binding + amyloid	56	30	1.04±0.99	0.55	0.03±1.35	1.17±1.60	$8\cdot10^{-3}$

The effect of stability, function and misfolding on the evolution of protein globular domains according to (1) the percentage of allowed mutations, (2) average $\Delta\Delta G_{evolution}$ for allowed mutations, and (3) average $\Delta\Delta G_{stability}$ for allowed and forbidden mutations.

[a] Student's *t*-test probability that a class has the same average $\Delta\Delta G_{evolution}$ as the relevant Stability class.

[b] Student's *t*-test probability that allowed and forbidden mutations from a class have the same average $\Delta\Delta G_{stability}$.

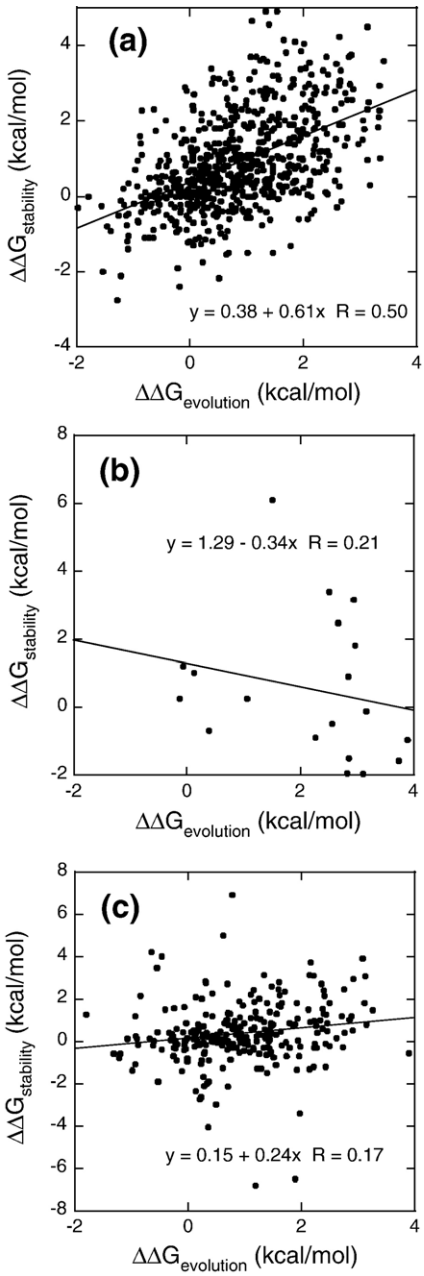

Figure 1. Correlation between $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{evolution}$ for allowed mutations involved in (a) stability, (b) stability and catalysis, and (c) stability and binding. The lines are a linear fit to the data. The slope and intercept of the fitted lines are indicated, along with the *R*-value of the correlation.

Function. Catalysis *versus* binding

Next, we considered the groups of sites involved in catalysis or binding. The results are shown in Table 1: 74% of the mutations in the group of catalytic sites are forbidden, while only 31% of the mutations in the group of binding sites are forbidden. The higher average $\Delta\Delta G_{evolution}$ for allowed mutations at catalytic sites compared to binding sites (Table 1) also indicates that mutations are much less likely to be accepted at catalytic sites. Binding sites and sites involved only in stability seem to be equally tolerant to mutation, both in terms of percentage of allowed mutations and of the average $\Delta\Delta G_{evolution}$ (Table 1). This suggests that binding and stability requirements can be met by a broader range of residues than catalysis requirements.

Misfolding

Formation of misfolded states like amyloid fibrils or amorphous aggregates could prevent the function of a globular domain.[34] Accordingly, the conservation of some residues prevents misfolding.[9,11] Several mechanisms have been suggested for the evolutionary avoidance of misfolding. For example, there is evidence for purifying selection against misfolding-prone sequences,[10,12–15] either by disrupting the misfolding core or by rendering it inactive with flanking proline residues or charged residues. Misfolding-prone sequences can be made innocuous by sequestering them in a stable native state.[34,35] It is not known to what extent natural proteins take advantage of this mechanism.

First, we have used the mutations in our dataset that change the propensity of a domain for misfolding to check for purifying selection against misfolding-prone sequences. Table 2 shows that mutations that decrease the propensity of a protein to misfold are more likely to be forbidden in nature than those that increase the propensity for misfolding. Along the same line, allowed mutations that protect against misfolding do not have a smaller $\Delta\Delta G_{evolution}$ than those that increase the propensity for misfolding (Table 2). These results do not depend on secondary structure or solvent accessibility at the site of mutation, or on the sign of $\Delta\Delta G_{stability}$ (data not shown). Our data demonstrate that globular domains do not generally avoid misfolding by discarding misfolding-prone sequences.

Is misfolding then prevented by stabilizing the native state? The data in Table 1 show that forbidden mutations at sites involved in misfolding are, on average, more destabilizing than allowed mutations (Table 1). Figure 2(a) and (b) show the correlation between $\Delta\Delta G_{evolution}$ and $\Delta\Delta G_{stability}$ for allowed mutations involved in misfolding. The *R*-value and slope of the correlation are very similar to those for mutations involved only in stability. These results indicate that stability is the main selection pressure at misfolding-related sites, and the most frequent way in which natural proteins avoid misfolding.

Misfolding *versus* stability

Why do globular domains not avoid sequences prone to misfolding? Aggregation-prone regions are more common in globular domains than in intrinsically unstructured proteins, which suggests that the propensities of a sequence to fold and aggregate are correlated.[35] In this case, avoiding misfolding-prone

Table 2. Comparison of mutations that increase or decrease the propensity of a domain to misfold

Class	No. mutations	% Forbidden	$\Delta\Delta G_{evolution}$ (kcal/mol) Allowed	P^a	$\Delta\Delta G_{stability}$ (kcal/mol)	P^b
Stability + aggregation plus	115	7	0.84±1.03	–	0.93±1.59	–
Stability + aggregation minus	175	30	1.18±1.07	0.02	1.96±1.56	$<10^{-4}$
Stability + amyloid plus	28	4	0.89±0.98	–	0.40±1.18	–
Stability + amyloid minus	49	47	1.26±1.23	0.22	2.33±2.22	$<10^{-4}$
Stability + binding + aggregation plus	34	22	1.08±0.83	–	0.14±1.47	–
Stability + binding + aggregation minus	60	35	1.24±1.25	0.52	1.51±1.85	$3\cdot10^{-4}$
Stability + binding + amyloid plus	10	0	1.26±1.01	–	−0.51±2.05	–
Stability + binding + amyloid minus	17	61	1.25±1.11	0.97	1.34±1.56	0.01

Comparison of mutations that increase or decrease the propensity of a domain to misfold according to (1) the percentage of allowed mutations, (2) average $\Delta\Delta G_{evolution}$ for allowed mutations, (3) average $\Delta\Delta G_{stability}$.

[a] Student's *t*-test probability that the average $\Delta\Delta G_{evolution}$ for mutations that increase and decrease the propensity of a domain to misfold are the same.

[b] Student's *t*-test probability that the average $\Delta\Delta G_{stability}$ for mutations that increase and decrease the propensity of a domain to misfold are the same.

sequences would be detrimental to stability. We have analyzed the correlation between protein stability and formation of misfolded states by comparing $\Delta\Delta G_{stability}$ for mutations that increase and decrease the propensity of a domain to form amyloid fibrils and amorphous aggregates. The results indicate that mutations that decrease the tendency to misfold are more destabilizing than those that increase it (Table 2), regardless of involvement in function (Table 2), secondary

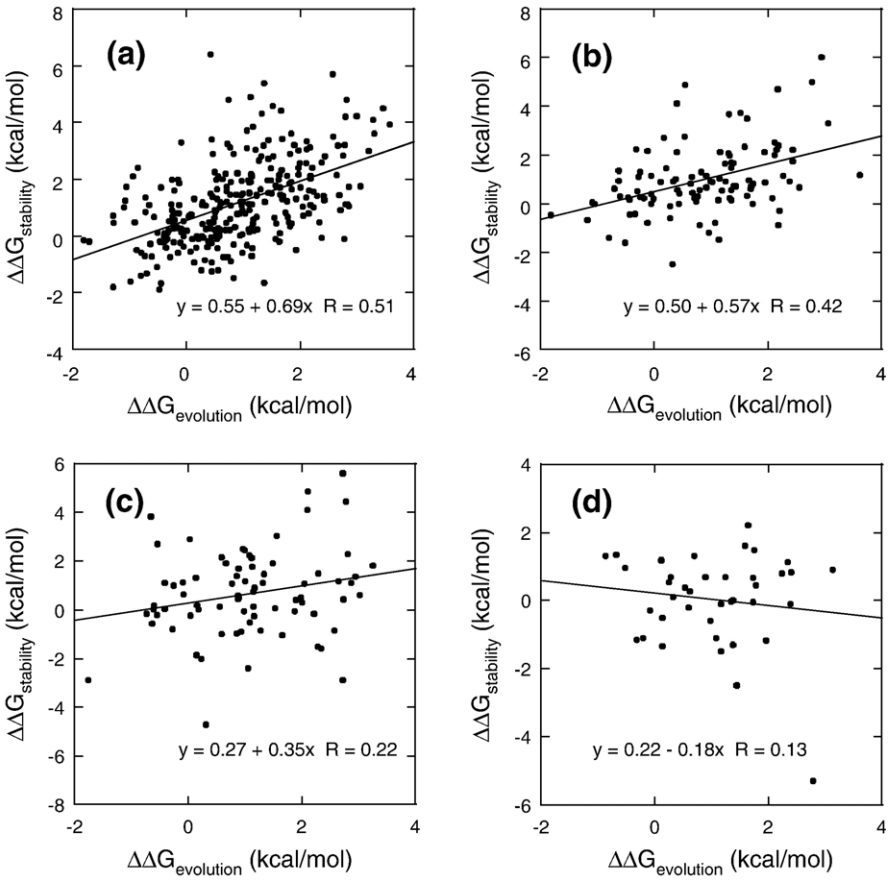

Figure 2. The correlation between $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{evolution}$ for allowed mutations involved in (a) stability and amorphous aggregation, (b) stability and amyloid formation, (c) stability, binding and amorphous aggregation, and (d) stability, binding and amyloid formation. The lines are a linear fit to the data. The slope and intercept of the fitted lines are indicated, along with the *R*-value of the correlation.

structure or solvent accessibility at the site of mutation (data not shown). Our data point at a link between native-state stability and the stability of both amyloid fibrils and amorphous aggregates. We suggest that avoiding misfolding-prone sequences could compromise native-state stability.

Amyloid fibrils *versus* amorphous aggregates

We have investigated the relationship between the abilities of a sequence to form amyloid fibrils and amorphous aggregates. To check for potential correlations, we calculated the frequency with which a site in our database is involved in formation of amyloid fibrils, amorphous aggregates or both (Table 3). If the two propensities were not correlated, the frequency of co-occurrence would be the product of the individual frequencies. If the two propensities occurred in a correlated manner, the observed frequency would be higher than the product of the individual frequencies, and *vice versa*. Interestingly, a site in our database is involved in formation of both amyloid fibrils and amorphous aggregates 1.7±0.03 times more often than expected if these two events were independent (Table 3), indicating that these two propensities are correlated. The sites involved in both amyloid formation and amorphous aggregation give us a chance to examine this correlation from the mutational viewpoint. There are 115 sites in which the mutation changes the propensity of the domain for both amyloid fibril formation and amorphous aggregation. In 80% of the cases, the two propensities change in the same direction, showing that effects of a mutation on amyloid fibril formation and amorphous aggregation are usually coupled. These data suggest an association between the formation of amyloid fibrils and amorphous aggregates.

Table 3. Analysis of the co-occurrence of different conformational and functional features at the same site

Class	Observed frequency	Expected frequency	Observed/expected
Stability + aggregation + amyloid	0.113	0.066	1.72±0.04
Binding + aggregation	0.075	0.081	0.93±0.04
Binding + amyloid	0.053	0.051	1.04±0.02
Catalysis + aggregation	0.0034	0.0111	0.31±0.07
Catalysis + amyloid	0.0026	0.0070	0.37±0.07

The observed frequency corresponds to the actual co-occurrence of a set of features at the same mutation site in our database. The expected frequency is the product of the frequencies of occurrence of each of the features considered. If the features co-occur in an uncorrelated manner, the quotient between observed and expected frequencies will be 1. If the features co-occur in a correlated manner, the quotient between observed and expected frequencies will be >1. If the features co-occur in an anticorrelated manner, the quotient between observed and expected frequencies will be <1. The standard deviation of the observed/expected ratios were calculated using five random subsets of size 75% of the full database (see Methods).

Function *versus* stability

There are two conflicting hypotheses on the relationship between protein stability and function. The first one states that there is a trade-off between function and stability.[16–24] It is not known whether this function–stability relationship takes the form of an anticorrelation between activity and stability, or whether functional sites are simply suboptimal in terms of stability. On the other hand, it has been pointed out that some functional residues can contribute to both function and stability, giving a positive function–stability correlation.[6,26,29,30] Altogether, there is no consensus on the evolutionary relationship between function and stability.

Our large database allows us to test which is the most common relationship between function and stability in naturally occurring globular domains, and to differentiate between positions involved in catalysis and in binding. If there was an anticorrelation between activity and stability at functional sites, allowed mutations would be more destabilizing than forbidden mutations, and $\Delta\Delta G_{evolution}$ and $\Delta\Delta G_{stability}$ for allowed mutations would be anticorrelated. If there was a general correlation between function and stability, allowed mutations would be less destabilizing than forbidden mutations, and $\Delta\Delta G_{evolution}$ and $\Delta\Delta G_{stability}$ for allowed mutations would be positively correlated, as for sites involved only in stability. Neither set of predictions is met by the proteins in our database. In catalytic positions, allowed and forbidden mutations are equally destabilizing (Table 1), and $\Delta\Delta G_{evolution}$ and $\Delta\Delta G_{stability}$ for allowed mutations are not correlated (Figure 1(b)). Thus, while selection for catalysis generally overrules selection for stability, the function–stability trade-off does not imply a negative correlation between these two properties. In the case of binding positions, forbidden mutations are more destabilizing than allowed mutations (Table 1). Thus, a residue can be incorporated into a binding site only if it does not drastically reduce the stability of the domain. On the other hand, the correlation between $\Delta\Delta G_{evolution}$ and $\Delta\Delta G_{stability}$ for allowed mutations is much weaker than for residues involved only in stability (Figure 1(c)). We conclude that, although stability constraints are present at these sites, they have a secondary role in determining the frequencies of allowed residues. Thus, there is neither an anticorrelation between activity and stability at binding positions nor a full coupling between function and stability, but rather a weak selection for stability at binding sites.

Function *versus* misfolding

The relationship between protein function and misfolding remains largely unexplored.

As a first step, we calculated the frequency with which a site in our database is involved in misfolding, function or both (Table 3). If the two propensities were not correlated, the observed frequency of co-occurrence would be the product

of the individual frequencies. If the two propensities occurred in a correlated manner, the observed frequency would be higher than the product of the individual frequencies, and *vice versa*. The results shown in Table 3 indicate that the observed frequency at which a site in our database is involved in both binding and misfolding is very close to the product of the individual frequencies, indicating that these two events are not correlated. On the other hand, the observed frequency at which a site in our database is involved in both catalysis and misfolding is only about one-third of the product of the individual frequencies, indicating that these two events are anticorrelated. This last result may be due to the abundance of charged residues at catalytic sites.[2]

Second, we have examined how misfolding due to sites involved in binding is prevented. We first looked for negative selection against misfolding-prone sequences. For this group of sites, mutations that decrease the propensity of a protein to misfold are more likely to be forbidden than those that increase the propensity for misfolding (Table 2) and allowed mutations that protect against misfolding do not have a smaller $\Delta\Delta G_{evolution}$ than those that increase the propensity for misfolding (Table 2). Thus, there is no selection against misfolding at binding sites. The next question was whether misfolding due to binding sites is prevented by selecting for stability at these sites. As for sites involved only in binding, forbidden mutations are more destabilizing than allowed mutations (Table 1), but the correlation between $\Delta\Delta G_{evolution}$ and $\Delta\Delta G_{stability}$ for allowed mutations is weak (Figure 2(c) and (d)). Thus, selection for stability at sites involved in binding and misfolding is not stronger than at sites involved only in binding. We propose that misfolding due to sites involved in function is prevented through the contribution of non-functional sites to stability.

Discussion

Our results, summarized in Table 4, provide information about the relative importance of function, stability and misfolding in the evolution of natural globular domains, and about the relationship

Table 4. Summary of the main evidence and conclusions of this work

Evolutionary pressure		Evidence	Conclusion
Stability		$\Delta\Delta G_{stability}$ (forbidden) > $\Delta\Delta G_{stability}$ (allowed) $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{evolution}$ correlate	Stability is the main evolutionary pressure in the absence of other factors
Function (catalysis *versus* binding)		% Forbidden (catalysis) > % forbidden (binding) $\Delta\Delta G_{evolution}$ (catalysis) > $\Delta\Delta G_{evolution}$ (binding)	Catalysis may be less tolerant to mutation than binding
Misfolding		% Forbidden (misfolding minus) > % forbidden (misfolding plus) $\Delta\Delta G_{evolution}$ (misfolding minus) ≈ $\Delta\Delta G_{evolution}$ (misfolding plus) $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{evolution}$ correlate	No selection against misfolding-prone sequences. Misfolding avoided by selection for stability
Misfolding *versus* stability		$\Delta\Delta G_{stability}$ (misfolding minus) > $\Delta\Delta G_{stability}$ (misfolding plus)	Stabilities of native and misfolded states are correlated. Avoiding misfolding-prone sequences compromises stability
Amyloid fibrils *versus* amorphous aggregates		Amyloid fibrils and amorphous aggregates co-occur in a correlated manner. Mutational effects on amyloid fibrils and amorphous aggregates correlate	Stabilities of amyloid fibrils and amorphous aggregates seem to be correlated
Function *versus* stability	Catalysis	$\Delta\Delta G_{stability}$ (forbidden) ≈ $\Delta\Delta G_{stability}$ (allowed) $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{evolution}$ do not correlate	Strongly destabilizing residues allowed at catalytic sites. Allowed residues not selected for stability
	Binding	$\Delta\Delta G_{stability}$ (forbidden) > $\Delta\Delta G_{stability}$ (allowed) $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{evolution}$ correlate weakly	Strongly destabilizing residues forbidden at binding sites. Allowed residues weakly selected for stability
Function *versus* misfolding	Catalysis	Catalysis and misfolding co-occur in an anticorrelated manner	Catalytic site composition destabilizes misfolded states
	Binding	Binding and misfolding co-occur in an uncorrelated manner. % Forbidden (misfolding minus) > % forbidden (misfolding plus) $\Delta\Delta G_{evolution}$ (misfolding minus) ≈ $\Delta\Delta G_{evolution}$ (misfolding plus) $\Delta\Delta G_{stability}$ and $\Delta\Delta G_{evolution}$ do not correlate	Binding site composition does not destabilize misfolded states. No selection against misfolding-prone sequences. Misfolding due to binding sites is avoided by selection for stability at other sites

between globular structures, amyloid fibrils and amorphous aggregates. We will now discuss some of the most informative pieces of evidence.

First of all, it is apparent that a majority of positions in a protein globular domain are selected for stability. The probability that a protein with a random sequence is folded is very low,[51] of the order of 10^{-11}. Thus, any polypeptide having a well-defined globular structure must be the subject of a strong selection and its sequence is, from this point of view, nearly optimal in terms of stability. This considered, it is obvious that some proteins are more stable than others. Full optimization of stability may sometimes be unnecessary, such as for proteins from mesophilic organisms compared to their thermophilic counterparts.[52] In other cases it may even be detrimental to function by impairing flexibility or the desired levels of degradation.[52] We hypothesize that this variability in the selection for stability should be reflected at the sequence level.

Functional positions involved in catalysis and binding show very different patterns of point mutations. Both the percentage of forbidden residues and the distribution of $\Delta\Delta G_{\text{evolution}}$-values indicate that amino acid changes at catalytic sites are allowed much less often than at non-functional sites (Table 4). Selection for catalysis also completely overrules selection for stability (Table 4). On the other hand, binding positions seem to be much more tolerant to mutation (Table 4), with a percentage of forbidden residues and a distribution of $\Delta\Delta G_{\text{evolution}}$ values very similar to those of non-functional sites. This suggests that not all amino acids involved in binding are essential for function, which would allow strongly destabilizing mutations to be discarded at binding sites (Table 4). Alternatively, this apparent tolerance to mutation may come from different sequences in the alignment being selected for binding to different target molecules. If the target diversity in the alignment is high, binding positions will appear to be tolerant to mutation even if selection for each target is as strong as selection for catalysis. Detailed binding data for a substantial fraction of sequences in an alignment would be necessary to clarify this point.

Another interesting result is that there is neither an anticorrelation between activity and stability at functional sites nor a strong coupling between selection for function and for stability (Table 4). Selection for stability is absent at catalytic sites, but there is no general selection against it either. Selection for stability is present at binding sites, although it is weak and not fully coupled to selection for function. Such incomplete optimization for stability at functional sites may be necessary for proper function of many proteins,[52] and explains the success of computational methods for the prediction of functional residues.[53–56] Finally, in the absence of detailed data on the functional consequences of most mutations in our study, we speculate that the dominant selection pressure at functional positions is binding to biological ligands.[6,57]

Our data suggest that naturally occurring protein globular domains avoid misfolding primarily by sequestering misfolding-prone stretches into a stable native-state structure (Table 4). As a consequence, misfolding-prone sequences are often found in globular domains.[15,35] Misfolding of newly synthesized and transiently unfolded proteins due to such sequences is probably prevented by chaperones.[15] Our analysis, which considers sequence positions in isolation, does not provide evidence for an evolutionary pressure against misfolding-prone sequences. Interestingly, previous work considering multiple positions at a time did find misfolding-prone sequences to be underrepresented in sequence databases,[10,12,13] and/or surrounded by gatekeeping residues.[10,15] We suggest that selection against misfolding in natural globular domains may act only at groups of residues and not at the single-residue level in order to minimize its impact on native-state stability (Table 4).

Native states, amyloid fibrils and amorphous aggregates are distinct states.[58] However, some sequences have been observed to fold or assemble into more than one of these states,[10,34,58] depending on protein concentration and medium conditions. According to our data, catalytic residues are detrimental to the stability of both globular structures and misfolded states (Table 4) and the effects of point mutations on the formation of globular domains, amyloid fibrils and amorphous aggregates are correlated (Table 4). These results imply that the stabilities of these disparate structures are linked. We can explain this link by the common set of physicochemical principles that formation of any ordered structure must fulfil,[35,40,59] such as burial of hydrophobic residues or hydrogen bonding.

Global analysis of the data shown in Table 4 reveals two different regimes for the evolution of globular domains. In positions involved in function, selection for binding and catalysis is the dominating force. The remaining positions in a domain ensure native-state stability and, with it, protection against misfolding. Protein globular domains have a high degree of structural cooperativity, which makes many amino acids in a domain not essential for folding.[60,61] Our results show that naturally occurring proteins take advantage of this redundancy by allowing some positions to evolve according mainly to functional criteria and using the rest to keep a stable native structure.

Conclusions and Outlook

The main results of our hierarchical analysis on protein globular domains are the following: (1) Selection for function at catalytic sites overrules selection for stability, but does not select against it. Selection for function at binding sites has to fulfil weak stability requirements. Non-functional sites in a protein are selected for stability. (2) Selection pressure against misfolding is not targeted against

misfolding-prone sequences but towards native-state stability. (3) The stabilities of native states, amyloid fibrils and amorphous aggregates seem to be correlated. (4) There is considerable redundancy of the roles of individual amino acids in stability and perhaps in binding, but not in catalysis. Taken together, these data may help us understand the effects of disease-related mutations in protein globular domains.[62]

The main limitations of our approach are that it considers sequence positions in isolation and that it does not differentiate between domains or between homologous sequences in an alignment. We therefore expect that future studies will benefit from second-order sequence analysis,[10,15,63,64] and the consideration of domain-specific[65,66] and organism-specific features.[67,68] We suggest that extension of this study to intrinsically disordered proteins[35] may help us understand the evolution of their sequences, structures and functions.

Methods

We included in our database protein globular domains with at least 25 point mutations of known $\Delta\Delta G_{\text{stability}}$,[41] and 40 unique homologous sequences in the corresponding Pfam alignment.[46] The high degree of divergence in Pfam alignments ensures that the observed patterns of mutation are determined only by domain fitness and not by the genetic code.[69] We performed all calculations excluding sequences with less than 20%, 30%, 40%, 50% and 60% identity with the one for which stability data were available. Although the actual figures varied, qualitatively the results did not change (data not shown), indicating that our approach is valid for both closely related and highly divergent sequences. Sequences in a Pfam alignment are structurally similar in terms of solvent accessibility, secondary structure and side-chain dihedral angles.[70] The location of the functional residues is also fairly well conserved.[1,2,71] The solvent accessibility at the site of mutation was calculated using DSSP,[72] and normalized using empirical maximum values.[73] The secondary structure at the site of mutation was determined using DSSP.[72] The seven-state assignment done by the software was collapsed into a three-state assignment as follows: α, π and 3_{10} helices are "helix", strands and beta bridges are "beta", and coils, bends and hydrogen bonded turns are "coil".

We defined two types of functional residue. First, we identified subsets of five or less residues essential to the function of a protein ("catalytic residues") using the Catalytic Site Atlas[36] and mutagenesis studies. Second, we identified larger groups of residues with a collective functional role ("binding residues") using mutagenesis studies and structural information.[74] The programs LIGPLOT,[37] NUCPLOT,[38] and iMolTalk (4 Å cutoff)[39] were used for protein–ligand complexes, protein–nucleic acid complexes and protein–protein complexes, respectively.

Involvement of a mutation site in amyloid fibril formation was tested by matching sequences in the alignment with a sequence pattern derived from saturation mutagenesis experiments with a designed hexapeptide.[10] For a mutation at a given position, we first identified sequences having the wild-type or mutant residues and then tested whether they matched the pattern at the site of mutation. If more than 10% of sequences having either the wild-type or mutant residues at the site of mutation matched the pattern, the site was considered to be involved in the formation of amyloid fibrils. The changes induced by a mutation on the propensity to form amyloid fibrils were estimated by taking sequences with the wild-type residue at the site of mutation and substituting the wild-type residue by the mutant residue. The percentage of sequences matching the pattern was calculated before and after substitution, and the sign of the difference was taken as the change in the propensity to form amyloid fibrils. Involvement of a mutation site in amorphous aggregation was determined in an analogous way, except that the TANGO software was used to test for the existence of aggregation sites.[40]

The *in vivo* change in stability upon mutation $\Delta\Delta G_{\text{stability}}$ was approximated to that observed *in vitro*.[42,43] The stability data used here have been measured by many different groups and at different conditions, which raises the question of their reliability. Recently, three groups measured independently $\Delta\Delta G_{\text{stability}}$ for 28 mutants of an SH3 domain.[75] The average standard deviation of $\Delta\Delta G_{\text{stability}}$ was approximately 0.4 kcal/mol, which we can take as a realistic measure of experimental uncertainty. The range of $\Delta\Delta G_{\text{stability}}$ for the mutants considered in this work is close to 8 kcal/mol (Figure 1(a)), significantly larger than the uncertainty in the measurements. Thus, our stability data are adequate for the correlation analysis we performed.

Mutations in the database were classified as allowed or forbidden, depending on whether the mutant residue appears homologous sequences.[46] By doing this, we assume that the sequence space available for a family of proteins is similar to the sequence space available for point mutants of a given member of the family. Such a correspondence has been observed in directed evolution experiments with several proteins.[44,45] Many factors, such as the size of the sequence family being studied, may influence the percentage of mutations that are found to be allowed or forbidden for a particular domain. However, we are not considering the percentage of forbidden residues of a certain class of sites in absolute terms but in comparison with other classes. Our implicit assumption is that, to a first approximation, the effects of the evolutionary pressures considered here (stability, function and misfolding) are independent of the evolutionary pressures that were excluded.

The change in evolutionary pseudo free energy upon mutation (equation (1)) was calculated using empirical weights for the sequences in the alignment,[76] and a pseudotemperature T of 298 K. We used the slope of the $\Delta\Delta G_{\text{stability}}$ *versus* $\Delta\Delta G_{\text{evolution}}$ plot[8] to calculate a pseudotemperature. We obtain a value of 182 K for mutations involved only in stability, in good agreement with previous estimates.[8]

We have checked for potential pairwise correlations between function, aggregation and amyloid fibril formation. For each pair of features, we calculated the frequency with which a site in our database is involved in feature A, feature B or both. If the two propensities were not correlated, the frequency of co-occurrence would be the product of the individual frequencies. If the two propensities occurred in a correlated manner, the observed frequency would be higher than the product of the individual frequencies, and *vice versa*. The effect of database size in the propensity ratios was estimated in the following way. First, we picked at random five subsets of sites representing 25%, 50% and 75 of the full database. The propensity ratios for each subset were used to calculate the average ratio and its standard deviation. In all cases, the average ratio was within error of the ratio

calculated for the whole dataset (data not shown). The standard deviation decreases when the size of the database is increased (data not shown). The standard deviation calculated with the 75 % dataset was taken as an upper limit for the standard deviation of the full dataset.

Data analysis was performed using Kaleidagraph (Synergy software), Excel (Microsoft) and in-house Perl scripts.

Acknowledgements

I.E.S. is the recipient of an EMBO Long Term Fellowship. We thank P. Beltrao, M. Bueno, A. Esteras-Chopo, J. Fernández-Recio, M. López de la Paz and J. Mendes for scientific discussions, A. Möglich for setting up PearPC, and R. Godoy-Ruiz and J. M. Sánchez-Ruiz for communicating unpublished results.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.08.020

References

1. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
2. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121.
3. Steipe, B., Schiller, B., Pluckthun, A. & Steinbacher, S. (1994). Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* **240**, 188–192.
4. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
5. Lehmann, M., Kostrewa, D., Wyss, M., Brugger, R., D'Arcy, A., Pasamontes, L. & van Loon, A. P. (2000). From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng.* **13**, 49–57.
6. Di Nardo, A. A., Larson, S. M. & Davidson, A. R. (2003). The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J. Mol. Biol.* **333**, 641–655.
7. Steipe, B. (2004). Consensus-based engineering of protein stability: from intrabodies to thermostable enzymes. *Methods Enzymol.* **388**, 176–186.
8. Godoy-Ruiz, R., Perez-Jimenez, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. (2005). A stability pattern of protein hydrophobic mutations that reflects evolutionary structural optimization. *Biophys. J.* **89**, 3320–3331.
9. Steward, A., Adhya, S. & Clarke, J. (2002). Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. *J. Mol. Biol.* **318**, 935–940.
10. Lopez de la Paz, M. & Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA*, **101**, 87–92.
11. Parrini, C., Taddei, N., Ramazzotti, M., Degl'Innocenti, D., Ramponi, G., Dobson, C. M. & Chiti, F. (2005). Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure (Camb)*, **13**, 1143–1151.
12. Broome, B. M. & Hecht, M. H. (2000). Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J. Mol. Biol.* **296**, 961–968.
13. Schwartz, R., Istrail, S. & King, J. (2001). Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein Sci.* **10**, 1023–1031.
14. Richardson, J. S. & Richardson, D. C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA*, **99**, 2754–2759.
15. Rousseau, F., Serrano, L. & Schymkowitz, J. W. (2006). How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol.* **355**, 1037–1047.
16. Meiering, E. M., Serrano, L. & Fersht, A. R. (1992). Effect of active site residues in barnase on activity and stability. *J. Mol. Biol.* **225**, 585–589.
17. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl Acad. Sci. USA*, **92**, 452–456.
18. Schreiber, G., Buckle, A. M. & Fersht, A. R. (1994). Stability and function: two constraints in the evolution of barstar and other proteins. *Structure*, **2**, 945–951.
19. Garcia, C., Nishimura, C., Cavagnero, S., Dyson, H. J. & Wright, P. E. (2000). Changes in the apomyoglobin folding pathway caused by mutation of the distal histidine residue. *Biochemistry*, **39**, 11227–11237.
20. Greene, L. H., Grobler, J. A., Malinovskii, V. A., Tian, J., Acharya, K. R. & Brew, K. (1999). Stability, activity and flexibility in alpha-lactalbumin. *Protein Eng.* **12**, 581–587.
21. Lee, K. N., Park, S. D. & Yu, M. H. (1996). Probing the native strain iin alpha1-antitrypsin. *Nature Struct. Biol.* **3**, 497–500.
22. Haruki, M., Noguchi, E., Nakai, C., Liu, Y. Y., Oobatake, M., Itaya, M. & Kanaya, S. (1994). Investigating the role of conserved residue Asp134 in *Escherichia coli* ribonuclease HI by site-directed random mutagenesis. *Eur. J. Biochem.* **220**, 623–631.
23. Zhi, W., Srere, P. A. & Evans, C. T. (1991). Conformational stability of pig citrate synthase and some active-site mutants. *Biochemistry*, **30**, 9281–9286.
24. Zhang, J., Liu, Z. P., Jones, T. A., Gierasch, L. M. & Sambrook, J. F. (1992). Mutating the charged residues in the binding pocket of cellular retinoic acid-binding protein simultaneously reduces its binding affinity to retinoic acid and increases its thermostability. *Proteins: Struct. Funct. Genet.* **13**, 87–99.
25. Di Nardo, A. A., Korzhnev, D. M., Stogios, P. J., Zarrine-Afsar, A., Kay, L. E. & Davidson, A. R. (2004). Dramatic acceleration of protein folding by stabilization of a nonnative backbone conformation. *Proc. Natl Acad. Sci. USA*, **101**, 7954–7959.
26. Quirk, D. J., Park, C., Thompson, J. E. & Raines, R. T. (1998). His…Asp catalytic dyad of ribonuclease A: conformational stability of the wild-type, D121N, D121A, and H119A enzymes. *Biochemistry*, **37**, 17958–17964.

27. Jackson, S. E. & Fersht, A. R. (1994). Contribution of residues in the reactive site loop of chymotrypsin inhibitor 2 to protein stability and activity. *Biochemistry*, **33**, 13880–13887.
28. Chatani, E., Tanimizu, N., Ueno, H. & Hayashi, R. (2001). Structural and functional changes in bovine pancreatic ribonuclease a by the replacement of Phe120 with other hydrophobic residues. *J. Biochem. (Tokyo)*, **129**, 917–922.
29. Schindler, T., Perl, D., Graumann, P., Sieber, V., Marahiel, M. A. & Schmid, F. X. (1998). Surface-exposed phenylalanines in the RNP1/RNP2 motif stabilize the cold-shock protein CspB from *Bacillus subtilis*. *Proteins: Struct. Funct. Genet.* **30**, 401–406.
30. Hillier, B. J., Rodriguez, H. M. & Gregoret, L. M. (1998). Coupling protein stability and protein function in *Escherichia coli* CspA. *Fold. Des.* **3**, 87–93.
31. Kragelund, B. B., Poulsen, K., Andersen, K. V., Baldursson, T., Kroll, J. B., Neergard, T. B. *et al.* (1999). Conserved residues and their role in the structure, function, and stability of acyl-coenzyme A binding protein. *Biochemistry*, **38**, 2386–2394.
32. Eberhardt, E. S., Wittmayer, P. K., Templer, B. M. & Raines, R. T. (1996). Contribution of a tyrosine side chain to ribonuclease A catalysis and stability. *Protein Sci.* **5**, 1697–1703.
33. Fetrow, J. S., Spitzer, J. S., Gilden, B. M., Mellender, S. J., Begley, T. J., Haas, B. J. & Boose, T. L. (1998). Structure, function, and temperature sensitivity of directed, random mutants at proline 76 and glycine 77 in omega-loop D of yeast iso-1-cytochrome *c*. *Biochemistry*, **37**, 2477–2487.
34. Rochet, J. C. & Lansbury, P. T., Jr (2000). Amyloid fibrillogenesis: themes and variations. *Curr. Opin. Struct. Biol.* **10**, 60–68.
35. Linding, R., Schymkowitz, J., Rousseau, F., Diella, F. & Serrano, L. (2004). A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342**, 345–353.
36. Porter, C. T., Bartlett, G. J. & Thornton, J. M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl. Acids Res.* **32**, D129–D133.
37. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8**, 127–134.
38. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (1997). NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucl. Acids Res.* **25**, 4940–4945.
39. Diemand, A. V. & Scheib, H. (2004). iMolTalk: an interactive, internet-based protein structure analysis server. *Nucl. Acids Res.* **32**, W512–W516.
40. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnol.* **22**, 1302–1306.
41. Kumar, M. D., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H. & Sarai, A. (2006). ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucl. Acids Res.* **34**, D204–D206.
42. Parsell, D. A. & Sauer, R. T. (1989). The structural stability of a protein is an important determinant of its proteolytic susceptibility in *Escherichia coli*. *J. Biol. Chem.* **264**, 7590–7595.
43. Ghaemmaghami, S., Fitzgerald, M. C. & Oas, T. G. (2000). A quantitative, high-throughput screen for protein stability. *Proc. Natl Acad. Sci. USA*, **97**, 8296–8301.
44. Barlow, M. & Hall, B. G. (2002). Predicting evolutionary potential: *in vitro* evolution accurately reproduces natural evolution of the tem beta-lactamase. *Genetics*, **160**, 823–832.
45. Cochran, J. R., Kim, Y. S., Lippow, S. M., Rao, B. & Wittrup, K. D. (2006). Improved mutants from directed evolution are biased to orthologous substitutions. *Protein Eng.* **19**, 245–253.
46. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S. *et al.* (2004). The Pfam protein families database. *Nucl. Acids Res.* **32**, D138–D141.
47. Sanchez, I. E. & Kiefhaber, T. (2003). Origin of unusual phi-values in protein folding: evidence against specific nucleation sites. *J. Mol. Biol.* **334**, 1077–1085.
48. Capriotti, E., Fariselli, P. & Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucl. Acids Res.* **33**, W306–W310.
49. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.
50. Wang, Q., Buckle, A. M., Foster, N. W., Johnson, C. M. & Fersht, A. R. (1999). Design of highly stable functional GroEL minichaperones. *Protein Sci.* **8**, 2186–2193.
51. Keefe, A. D. & Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, **410**, 715–718.
52. Jaenicke, R. & Zavodszky, P. (1990). Proteins under extreme physical conditions. *FEBS Letters*, **268**, 344–349.
53. Ota, M., Kinoshita, K. & Nishikawa, K. (2003). Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.* **327**, 1053–1064.
54. Chelliah, V., Chen, L., Blundell, T. L. & Lovell, S. C. (2004). Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **342**, 1487–1504.
55. Cheng, G., Qian, B., Samudrala, R. & Baker, D. (2005). Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucl. Acids Res.* **33**, 5861–5867.
56. Chelliah, V., Blundell, T. L. & Fernandez-Recio, J. (2006). Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. *J. Mol. Biol.* **357**, 1669–1682.
57. Chakrabarti, R., Klibanov, A. M. & Friesner, R. A. (2005). Computational prediction of native protein ligand-binding and enzyme active site sequences. *Proc. Natl Acad. Sci. USA*, **102**, 10153–10158.
58. Rousseau, F., Schymkowitz, J. & Serrano, L. (2006). Protein aggregation and amyloidosis: confusion of the kinds? *Curr. Opin. Struct. Biol.* **16**, 118–126.
59. Lopez De La Paz, M., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C. M., Hoenger, A. & Serrano, L. (2002). De novo designed peptide-based amyloid fibrils. *Proc. Natl Acad. Sci. USA*, **99**, 16052–16057.
60. Gassner, N. C., Baase, W. A. & Matthews, B. W. (1996). A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc. Natl Acad. Sci. USA*, **93**, 12155–12158.

61. Brown, B. M. & Sauer, R. T. (1999). Tolerance of Arc repressor to multiple-alanine substitutions. *Proc. Natl Acad. Sci. USA*, **96**, 1983–1988.
62. Chasman, D. & Adams, R. M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307**, 683–706.
63. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
64. Larson, S. M., Di Nardo, A. A. & Davidson, A. R. (2000). Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J. Mol. Biol.* **303**, 433–446.
65. England, J. L., Shakhnovich, B. E. & Shakhnovich, E. I. (2003). Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc. Natl Acad. Sci. USA*, **100**, 8727–8731.
66. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA*, **102**, 14338–14343.
67. Bastolla, U., Moya, A., Viguera, E. & van Ham, R. C. (2004). Genomic determinants of protein folding thermodynamics in prokaryotic organisms. *J. Mol. Biol.* **343**, 1451–1466.
68. Tartaglia, G. G., Pellarin, R., Cavalli, A. & Caflisch, A. (2005). Organism complexity anti-correlates with proteomic beta-aggregation propensity. *Protein Sci.* **14**, 2735–2740.
69. Benner, S. A., Cohen, M. A. & Gonnet, G. H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.* **7**, 1323–1332.
70. Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**, 1811–1826.
71. Littler, S. J. & Hubbard, S. J. (2005). Conservation of orientation and sequence in protein domain–domain interactions. *J. Mol. Biol.* **345**, 1265–1279.
72. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
73. Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Genet.* **20**, 216–226.
74. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
75. de los Rios, M. A., Muralidhara, B. K., Wildes, D., Sosnick, T. R., Marqusee, S., Wittung-Stafshede, P. *et al.* (2006). On the precision of experimentally determined protein folding rates and phi-values. *Protein Sci.* **15**, 553–563.
76. Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574–578.

Edited by J. E. Ladbury

(Received 21 June 2006; received in revised form 25 July 2006; accepted 8 August 2006)
Available online 12 August 2006