PEDS Advance Access published August 5, 2005

Protein Engineering, Design & Selection
doi:10.1093/protein/gzi045

Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds

U.Bastolla[1,2,3] and Lloyd Demetrius[4]

[1]Centro de Astrobiología (INTA-CSIC), 28850 Torrejón de Ardoz and
[2]Centro de Biología Molecular 'Severo Ochoa', Cantoblanco, 28049 Madrid,
Spain and [4]Department of Organismic and Evolutionary Biology,
Harvard University, Cambridge, MA 02138, USA

[3]To whom correspondence should be addressed at the Madrid address.
E-mail: ubastolla@cbm.uam.es

Stability of the native state is an essential requirement in protein evolution and design. Here we investigated the interplay between chain length and stability constraints using a simple model of protein folding and a statistical study of the Protein Data Bank. We distinguish two types of stability of the native state: with respect to the unfolded state (unfolding stability) and with respect to misfolded configurations (misfolding stability). Several contributions to stability are evaluated and their correlations are disentangled through principal components analysis, with the following main results. (1) We show that longer proteins can fulfil more easily the requirements of unfolding and misfolding stability, because they have a higher number of native interactions per residue. Consistently, in longer proteins native interactions are weaker and they are less optimized with respect to non-native interactions. (2) Stability against misfolding is negatively correlated with the strength of native interactions, which is related to hydrophobicity. Hence there is a trade-off between unfolding and misfolding stability. This trade-off is influenced by protein length: less hydrophobic sequences are observed in very long proteins. (3) The number of disulfide bonds is positively correlated with the deficit of free energy stabilizing the native state. Chain length and the number of disulfide bonds per residue are negatively correlated in proteins with short chains and uncorrelated in proteins with long chains. (4) The number of salt bridges per residue and per native contact increases with chain length. We interpret these observations as an indication that the constraints imposed by unfolding stability are less demanding in long proteins and they are further reduced by the competing requirement for stability against misfolding. In particular, disulfide bonds appear to be positively selected in short proteins, whereas they evolve in an effectively neutral way in long proteins.

Keywords: disulfide bonds/folding thermodynamics/
protein evolution/protein folding

Introduction

The need to maintain the thermodynamic stability of the native state is an essential requirement in protein evolution. This constraint, however, is compatible with high tolerance to mutations: Only marginal stability with respect to the unfolded state, of the order of a few kilocalories per mole, is required for proper biological function and a large fraction of engineered mutations do not lead to a substantial decrease in this unfolding free energy (see, for instance, the ProTherm database; Bava *et al*., 2004).

Natural selection eliminates mutants that do not fulfil thermodynamic or functional requirements and favors the fixation of variants conferring a large enough selective advantage. However, according to the neutral theory of molecular evolution, most amino acid substitutions have been fixed in biological populations not through positive natural selection, but through random genetic drift of selectively neutral alleles (Kimura, 1983). The hypothesis that protein stability is selectively neutral above some threshold provides a natural explanation for the marginal stability of proteins (Taverna and Goldstein, 2002) and it is consistent with the strong correlation between the observed amino acid frequencies and those predictable from nucleotide frequencies (Sueoka, 1961; Lobry, 1997).

Evolution at the molecular level depends on the evolution at the population level. As predicted by the directionality principle for demographic stability, described by Demetrius and co-workers (Kowald and Demetrius, 2005; Ziehe and Demetrius, 2005), evolution at the population level may result in an increase, a decrease or a random non-directional change in demographic stability, depending on the ecological constraints the population experiences. Since changes in demographic stability are predicted to be positively correlated with changes in protein stability (Demetrius, 2002), evolutionary changes in protein stability can also result in an increase, a decrease or a random non-directional change. The latter is consistent with the notion of selectively neutral alleles in the neutral theory of molecular evolution (Kimura, 1983).

It is important to consider that the native state of a protein must be stable not only with respect to the unfolded state (unfolding stability), but also with respect to compact, incorrectly folded conformations (misfolding stability). We distinguish here between these two types of stability. For globular proteins, the two requirements may act in opposite directions. Lattice simulations suggest that stability against misfolded states sets an upper limit on protein hydrophobicity (Sandelin, 2004). Another recent study found a negative correlation between stability against unfolding and stability against misfolding in orthologous proteins, suggesting that there is a trade-off in the evolution of these two quantities (Bastolla *et al*., 2004). According to this view, protein evolution is frustrated: it cannot optimize both stabilities at the same time.

We aimed here to investigate the relationship between the thermodynamic requirements on protein evolution and chain length and composition. We adopted a computational study of protein structures in the Protein Data Bank (PDB) (Berman *et al*., 2000), using a knowledge-based effective energy function to predict thermodynamic properties. Our work is focused on the interplay between stability parameters, chain length and composition. The results are consistent with the hypothesis that thermodynamic requirements impose weaker constraints on the evolution of longer proteins. This interpretation agrees with a recent analysis of site-specific amino acid distributions as a function of chain length (Bastolla *et al*., 2005). We distinguish three types of intra-protein interactions: hydrophobic interactions, salt bridges and disulfide bonds.

In particular, we studied in detail the distribution and correlations of disulfide bonds, finding that they illustrate nicely the general pattern. Disulfide bonds are the strongest interactions in protein structures (Anfinsen and Scheraga, 1975; Betz, 1993). It has been shown that their number is negatively correlated with the content of aliphatic hydrophobic residues in the protein chain (Abkevich and Shakhnovich, 2000). This finding suggests that disulfide bonds are more frequently observed where the native state, in absence of them, would have limited stability. In a recent computational study of intracellular proteins in the genomes of hyperthermophilic Archaea, Mallick *et al*. (2002) found a correlation between the incidence of disulfide bonds and the optimal growth temperature of the cell—the higher the optimal growth temperature, the larger is the fraction of the total number of intracellular cysteines predicted to form disulfide bonds. This suggests that disulfide bonding may be under strong selection for increased stability in hyperthermophilic Archeae. The results presented here are consistent with the hypothesis that disulfide bonds tend to be positively selected in short proteins, whereas they evolve neutrally in long proteins.

Materials and methods

Protein set

We consider protein domains as parsed in the SCOP database (Murzin *et al*., 1995). We use a subset of SCOP domains having <40% pairwise sequence similarity. The structure files can be downloaded at the URL http://www.astral.berkeley.edu (Chandonia *et al*., 2004). We exclude structures determined by NMR, since they are on the average less compact and they are not suitable to the application of Equation 1 (Bastolla *et al*., 2001). We also excluded the following proteins: chains in multimeric complexes for which the interactions with other chains, evaluated through our effective energy function, contribute more than 15% to the stability; chains much less compact than expected for their length according to Equation 8; chains for which we find alternative conformations lower in energy than the native (apart some of the cases listed above, these are typically short fragments or proteins with large cofactors); and chains longer than 700 residues. In this way, we obtained a data set of 4528 protein domains. These were further divided in two subsets: 2835 domains with <200 residues and 1166 domains with $\geqslant$250 residues.

Effective energy function

We adopt a simplified description of protein structures using the contact map C_{ij}, a symmetric matrix with elements one if residues i and j are in contact and zero otherwise. The residues are considered in contact if any pair of their heavy atoms are closer than 4.5 Å and $|i - j| > 2$. The number of contacts is indicated as $C = \sum_{ij} C_{ij}$.

We approximate the effective free energy of a contact map $\boldsymbol{C}$ through a simple pairwise contact approximation:

$$E(\mathbf{C},\mathbf{A})/k_{\mathrm{B}}T = \sum_{i<j} C_{ij}U(A_i,A_j) \qquad (1)$$

where A_j is the amino acid at position i along the sequence and $U(a,b)$ is the contact interaction matrix determined by Bastolla *et al*. (2000). This free energy includes thermally averaged solvent contributions, therefore it depends in principle on the temperature and pH of the solvent. It does not include explicitly the conformational entropy of the chain.

This energy function assigns lowest energy to the experimentally known native state of almost all monomeric proteins studied by X-ray crystallography, among a large set of alternative structures generated by gapless threading and the resulting energy landscapes are typically well correlated (see below).

Unfolding free energy

To estimate the difference in free energy ΔG between the denatured and the native state of a protein, we used the effective free energy of the native state, $E_{\mathrm{nat}} = E(\mathbf{C}_{\mathrm{nat}}, \mathbf{A})$, calculated as described above. We neglected the chain entropy of the native state with respect to that of the denatured state and we neglected the effective energy of the denatured state with respect to the native. With these approximations, the unfolding free energy per residue was estimated as

$$\Delta G/N \approx -aE(\mathbf{C}_{\mathrm{nat}}, \mathbf{A})/N - S(\mathbf{A})T \qquad (2)$$

where a is a conversion factor, T is the absolute temperature, k_{B} is the Boltzmann constant and $S(\mathbf{A})$ is the chain entropy per residue of the denatured state.

First, we neglected the sequence dependence of the chain entropy, setting $S(\mathbf{A}) = s$. The above equation, with $S(\mathbf{A})T = s$, was fitted against experimental measures of unfolding free energies for proteins that fold with two-state kinetics, i.e. intermediate states do not contribute appreciably to folding thermodynamics. The correlation coefficient was 0.89 for a set of 44 proteins (U.Bastolla, unpublished work), with the two free parameters a and s, whose fitted values were $a = 1.62 \pm 0.14$ and $s = 0.123 \pm 0.03$ cal/mol. The latter is much smaller than the typical value of side-chain entropies, which is 3.33 cal/mole per side chain (Doig and Sternberg, 1995). This discrepancy suggests that the term s also includes in an effective way stabilizing contributions not included in the effective energy function and proportional to chain length, which partially compensate the chain entropy. These may be, for instance, the main-chain hydrogen bonds that constitute secondary structure.

Then we calculated the conformational entropy per residue of the unfolded state, $S(\mathbf{A})$, using the mean side-chain entropy scale reported by Doig and Sternberg (1995) and tried to fit Equation 2 in the form $\Delta G + NTS(\mathbf{A}) = -aE(\mathbf{C}_{\mathrm{nat}}, \mathbf{A})$, with only one free parameter a. Since the unfolding free energy per residue is much smaller than the typical side-chain entropy, the quantity on the left-hand side is dominated by the chain entropy. The fit became worse than before, possibly because of compensations due to terms not included in the effective free energy, so we decided to treat the effective free energy $E(\mathbf{C}_{\mathrm{nat}}, \mathbf{A})$ and minus the chain entropy $-S(\mathbf{A})$ as two separate variables.

Stability with respect to misfolded states

An estimate of the stability with respect to misfolded states is given by the energy gap (Goldstein *et al.*, 1992, Gutin *et al.*, 1995). We use a dimensionless version of this quantity, normalizing it with the native energy (Bastolla *et al.*, 1999). The normalized energy gap of sequence **A**, $\alpha(\mathbf{A})$, is defined as

$$\alpha(\mathbf{A}) = \min\left(\frac{E(\mathbf{C}, \mathbf{A}) - E(\mathbf{C}_{\text{nat}} - \mathbf{A})}{|E(\mathbf{C}_{\text{nat}}, \mathbf{A})|[1 - q(\mathbf{C}, \mathbf{C}_{\text{nat}})]}\right) \qquad (3)$$

where the overlap $q(\boldsymbol{C}, \boldsymbol{C}_{\text{nat}}) \in \{0, 1\}$ is a measure of structural similarity with the native state and **C** is any compact conformation. A large value of the normalized energy gap guarantees that misfolded states very dissimilar from the native (low q) have much higher effective energy, so that their contribution to the Boltzmann ensemble is negligible. This implies that the energy landscape is well correlated, i.e. all structures having low energy are similar to the native one.

Contact strength and interactivity

Here we define two quantities related to the strength of non native interactions. They can be calculated from protein sequence alone. The first one is the mean energy of native and non-native contacts, $\langle U \rangle$. It is defined as the average over all possible contacts of the contact interaction strength:

$$\langle U \rangle = \frac{\sum_{i<j+2} \mathrm{U}(\mathrm{A}_i, \mathrm{A}_j) w_{ij}}{\sum_{i<j+2} w_{ij}} \qquad (4)$$

We use weights that decrease with sequence separation, $w_{ij} = 1/|i - j|$, to take into account the decay of contact probability. Similar results are obtained with uniform weights, $w_{ij} \equiv 1$ and even omitting the condition $i < j + 2$ in the sum, which makes the mean interaction energy only dependent on chain composition.

The other indicator that we calculated is the mean hydrophobicity of the protein sequence, defined as $\langle h \rangle = \sum_i h(A_i)/N$. We used the generalized hydrophobicity scale $h(a)$ called interactivity, which was obtained as the main eigenvector of the contact interaction matrix $U(a,b)$ (Bastolla *et al.*, 2005). The quantities $-\langle U \rangle$ and $\langle h \rangle$ are strongly correlated with each other ($R = 0.73$ over more than 4000 domains).

Z-score of native interactions

To quantify the extent at which native interactions are optimized with respect to the background, we use the zeta score of native interactions, defined as

$$Z_{\text{nat}} = \frac{E_{\text{nat}}/C - \langle U \rangle}{\sqrt{\langle U^2 \rangle - \langle U \rangle^2}} \qquad (5)$$

The more negative Z_{nat} is, the stronger the native contacts are with respect to average contacts in the same sequence.

Disulfide bonds

Disulfide bonds were either read directly from the SSBOND record in the PDB file or, in case this was absent, they were defined as contacts between cysteine residues where the two sulfur atoms are closer than 2.8 Å apart, with the further condition that each cysteine residue can intervene in at most one disulfide bonds.

In our energy function, contacts between cysteine residues receive the strongest interaction value. We estimated the unfolding free energy without the contribution of disulfide bonds, $-E'_{\text{nat}}/N$, by subtracting from E_{nat} the effective energetic contribution of disulfide bonds and adding for each subtracted bond an amount of energy equivalent to the average energy per native contact:

$$E'_{\text{nat}} = \left[E_{\text{nat}} - nU(CYS, CYS)\right]\left(\frac{C}{C-n}\right) \qquad (6)$$

where $C = \sum_{i<j} C_{ij}$ is the number of native contacts and n is the number of native disulfide bonds.

Salt bridges

We measured the number of salt bridges in protein structures, N_{SALT}. Salt bridges were identified as pairs of oppositely charged residues with charged groups closer than 4.0 Å apart (Kumar and Nussinov, 1999).

Structural indicators

The main structural indicator used in this study is the number of contacts per residue:

$$C/N = \frac{\sum_{ij} C_{ij}}{N} \qquad (7)$$

It is well known (see, for instance, Bastolla *et al.*, 2000) that the number of contacts per residue is approximately constant for residues in the core of globular proteins and smaller for residues at the surface, so that the total number of contacts, $C = \sum_{i<j} C_{ij}$, divided by the number of residues N scales as the surface to volume ratio:

$$C/N \approx c\left(1 - b/N^{1/3}\right) \qquad (8)$$

where the asymptotic value, c, depends on the definition of contact (in our case, $c = 3.9 \pm 0.1$) and $b \approx 1.5$, the same value as for a compact self-avoiding-walk on the lattice. Hence shorter proteins have fewer contacts per residue and the length dependence becomes weaker for long chains.

Another useful parameter to characterize the native structure is the absolute contact order (ACO), defined as the average sequence distance between residues in contacts:

$$ACO = \frac{\sum_{ij} C_{ij}|i-j|}{\sum_{ij} C_{ij}} \qquad (9)$$

Plaxco *et al.* (1998) showed that there is a strong negative correlation between the relative contact order (RCO), which is the ACO normalized by chain length, RCO = ACO/*N*, and the logarithm of the folding rate, for proteins with two-state kinetics. However, it was later realized that the logarithm of the folding rate of two-state and non-two-state folders is better fitted as a linear functional of the ACO (Ivankov *et al.*, 2003). The ACO is a more convenient indicator than the RCO because it takes into account the negative correlation between the folding rate and chain length (Thirumalai, 1995; Gutin *et al.*, 1996; Finkelstein and Badretdinov, 1997). Chain length is positively correlated with the ACO but negatively with the RCO.

We also measured the RCO of disulfide bridges (i.e. only corresponding to disulfide bridges are considered), which we called DRCO.

Correlations

We measured correlations between pairs of variables using the Pearson correlation coefficient, defined as

$$R(X,Y)=\frac{\langle XY\rangle-\langle X\rangle\langle Y\rangle}{\sqrt{\left(\langle X^2\rangle-\langle X\rangle^2\right)\left(\langle Y^2\rangle-\langle Y\rangle^2\right)}} \tag{10}$$

where $\langle X\rangle=\sum_{i=1}^{n}X_i/n$. For assessing whether the correlation between two variables X and Y is essentially due to a third variable Z, we measured sometimes the partial correlation coefficient, which evaluates the correlation between X and Y at Z held constant and is defined as

$$R(X,Y,Z)=-\frac{R(X,Y)-R(X,Z)R(Y,Z)}{\sqrt{\left[1-R(X,Z)^2\right]\left[1-R(Y,Z)^2\right]}} \tag{11}$$

The dependences among multiple variables were disentangled using principal component analysis (PCA). This determines the orthogonal directions in a multivariate space that account for the largest part of the variance of a set of elements embedded in this space. PCA consists in diagonalizing the correlation matrix $R(X^a,X^b)$ for the set of variables whose relationships are under study, in order to determine the coefficients through which each variable enters each principal component.

Results

We measured correlations between properties of a large set of proteins, covering a broad range of functions, oligomerization states and organisms in which they reside. Because of this complexity, the correlations that we present are generally weak, but they are strongly significant and can be used to contrast different evolutionary hypothesis. Using a stringent significance level of 0.001, correlation coefficients >0.04 are significant in the complete data set. This increases to 0.06 for the subset of short proteins (<200 residues) and 0.09 for the subset of long proteins (>250 residues).

Chain length and unfolding stability

We found that the effective free energy per residue, $-E_{\mathrm{nat}}/N$, is uncorrelated with chain length ($R = 0.01$), whereas the average strength of native contacts, defined as $-E_{\mathrm{nat}}/C$, has a negative association both with the number of contacts per residue, C/N (correlation coefficient $R = -0.33$) and with chain length ($R = -0.28$). The high significance was confirmed by a chi-squared test where the proteins were divided into nine subsets of approximately equal size according to their length ($P < 10^{-6}$), see Figure 1. The average interaction strength decreases by 15% from the shortest to the longest proteins.

These correlations can be explained by noting that the requirements for stability against unfolding are more easily met for longer proteins, owing to the increase in the number of contacts per residue with chain length. We rewrite Equation 2 as

$$-E_{\mathrm{nat}}/C=\left(\frac{N}{C}\right)\frac{1}{a}\left(\frac{\Delta G}{N}+ST\right) \tag{12}$$

Proteins are marginally stable and it is not expected that their unfolding free energy per residue $\Delta G/N$ increases with chain length: the stability requirements are not expected to depend on chain length. If, in addition, the chain entropy per residue $S(\mathbf{A})$ does not increase with chain length, the above equation implies that $-E_{\mathrm{nat}}/C$ decreases with chain length.

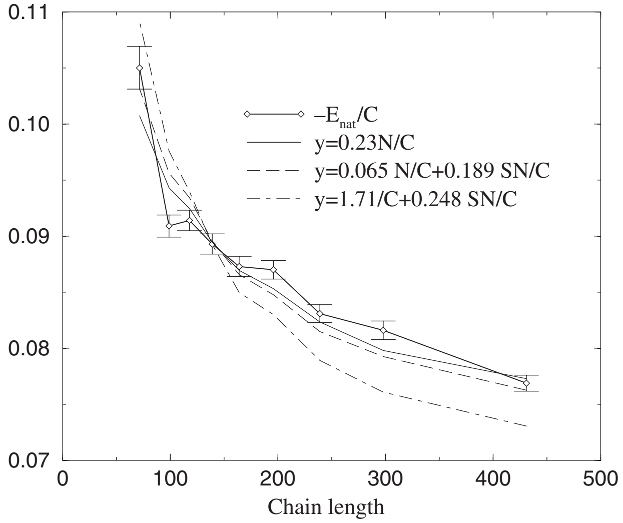

Fig. 1. Minus the effective native free energy per native contact, estimating the strength of native interactions, is found to decrease with chain length, in good agreement with Equation 12.

In agreement with Equation 12, $-E_{\mathrm{nat}}/C$ is correlated with N/C with $R = 0.38$. This correlation is stronger than for any other size-related indicator that we tested. $-E_{\mathrm{nat}}/C$ is also positively correlated with the chain entropy S, with correlation $R = 0.32$.

We attempted several mean-square fits of the strength of native contacts based on Equation 12. The average length behavior of these fitted curves is represented in Figure 1, but the goodness of the fit was evaluated using all 4528 protein domains studied. We first neglected the variations of $\Delta G/N$ and S (one free parameter), obtaining $-E_{\mathrm{nat}}/C\approx 0.23\ N/C$ with $R = 0.37$. Taking into account also the calculated value of the chain entropy per residue S, we then obtained $-E_{\mathrm{nat}}/C\approx 0.065\ N/C + 0.189\ SN/C$ with $R = 0.45$ for two free parameters. Note, however, that the coefficients of N/C and SN/C are expected from Equation 2 to be roughly equal and nevertheless the coefficient representing $\Delta G/N$ is much smaller. Assuming instead that the unfolding free energy ΔG is not correlated with chain length, we obtain a fit of Equation 12 in the form $-E_{\mathrm{nat}}/C\approx 1.71/\mathrm{C} + 0.25\ SN/C$, with $R = 0.48$ for two free parameters. The fitted value of ΔG corresponds in this case to $\Delta G\approx 7$ kcal/mol, which is in line with experimental data.

Note that this fit, which is the best and the most plausible one, is such that long proteins lay significantly above the predicted line: the minimal stability requirements in terms of $-E_{\mathrm{nat}}/C$ are predicted to decrease with chain length, but longer proteins tend to stay higher than the minimal requirements. This observation is further supported by PCA and we will come back to it later.

Chain entropy

As reported below, the amino acid composition changes with chain length so that the small amino acids Ala and Gly become more abundant in longer chains. This compositional change induces a decrease of the chain entropy of the unfolded state, S, for longer chains. This decrease is modest, 5 % in the length range examined. Since the entropy of the unfolded state contributes negatively to folding stability, we consider the variable –S, defined as minus the conformational entropy per residue. The correlation between this variable and chain length is $R = 0.15$.

This decrease in the chain entropy per residue of the unfolded state for longer chains compensates in part for the decrease in the strength of native interactions, $-E_{\text{nat}}/C$. The length behavior of the entropy can be explained considering $-E_{\text{nat}}/C$ as intermediate variable. In fact, this is negatively correlated both with chain length ($R = -0.28$) and with –S ($R = -0.32$) and the partial correlation between length and entropy at fixed $-E_{\text{nat}}/C$ is only $R = 0.065$. –S is also strongly and negatively correlated with the mean interaction energy $-\langle U\rangle$ ($R = -0.45$). This is to be expected, since more interactive residues tend to have a large conformational entropy. However, at odds with this argument, S is uncorrelated with the mean interactivity, $\langle h\rangle$.

Chain length, misfolding stability and design of native interactions

The decrease in the strength of native interactions with chain length can be attributed in part to the overall decrease in the protein interactivity, involving both native and non-native interactions, as discussed in another section, and in part to the fact that, in longer proteins, native interactions differ less in strength from non-native interactions.

To test this hypothesis, we measured the Z-score of native interactions, Equation 5. The larger $-Z_{\text{nat}}$ is, the stronger are the native interactions with respect to non-native interactions. We observed that $-Z_{\text{nat}}$ is negatively correlated with chain length ($R = -0.41$), thus confirming our hypothesis that native interactions are less optimized in longer proteins (see Figure 2).

The normalized energy gap

This behavior can be explained by considering the condition of stability against misfolding, that we evaluate through the normalized energy gap α. We first calculated this quantity by using Equation 3 on all alternative structures generated by threading the protein sequence A along all structures in the PDB with a larger or equal number of residues. The normalized energy gap evaluated in this way shows a very strong length dependence, increasing roughly as the logarithm of chain length.

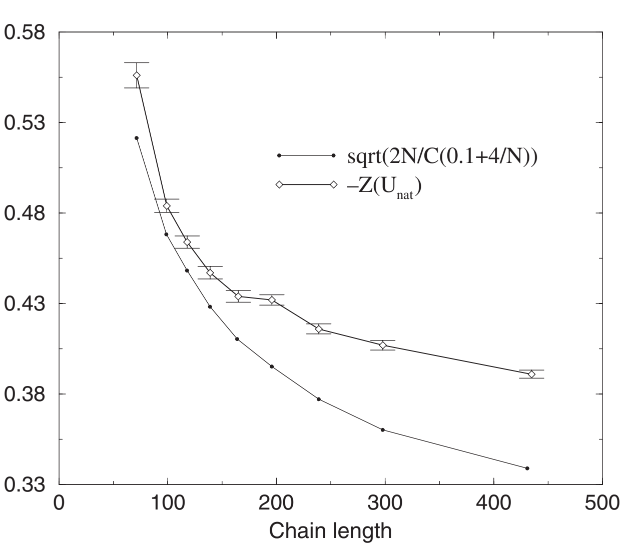

Fig. 2. Absolute Z-score of native interactions and minimum value of the absolute Z-score obtained through a computation based on the REM (see text), as a function of chain length.

This length dependence is essentially due to the fact that the number of alternative structures m generated by threading decreases for increasing chain length N, since the protein sequence can be threaded only on structures that are longer than N. This leads to underestimation of the minimal energy of alternative structures and consequently overestimation of α with respect to its true value when all possible decoys are used. This artifact is not important when comparing the folding properties of proteins of similar length, but it is essential in the present study.

Therefore, in this study we did not estimate α through threading, but resorted to a calculation based on the random energy model (REM) (Derrida, 1981; Shakhnovich and Gutin, 1989). In the framework of the REM, it is possible to estimate the minimal energy of alternative structures as

$$E_{\text{REM}} \approx \langle U\rangle C - \sigma_U\sqrt{2C\,\log(m_N)} \qquad (13)$$

where m_N is the number of independent contact matrices for a protein of length N. The minimal energy estimated in this way, E_{REM}, is in very good agreement with the value found by threading, E_{min}, when m_N is set equal to the number of structures generated through threading: E_{min} (1.003 ± 0.009)E_{REM} − (0.0016 ± 0.0012), with $R = 0.96$.

The number of all possible independent decoys is equal to the number of contact matrices with N residues compatible with steric constraints and hydrogen bonding of core residues. This number is expected to increase exponentially with N. We therefore make the ansatz

$$\log(m_N) \approx AN + B \qquad (14)$$

The value of the exponent A is not known, but we expect that the condition that allowed structure must have regular secondary structure reduces the exponent A to a very small value. It is easy to obtain an order-of-magnitude estimate, since the exponent for the number of maximally compact self-avoiding walks on a regular lattice has been calculated in a mean field approximation as $A' = \log(z) - 1$, where z is the coordination number) of the lattice (Orland *et al.*, 1985). On the cubic lattice, $z = 6$ and $a' = 0.792$. The number of steps equals in this case the number of secondary structure elements, which is linearly correlated with chain length as $N_{\text{elements}} \approx 0.145N$, corresponding to seven residues per secondary structure element on average. This leads to the estimate $A = 0.145 \times 0.79 = 0.115$.

We obtained a rough evaluation of the two parameters A and B by imposing two conditions on our data: (1) the estimated minimal effective energy should be similar to the observed value for short proteins, for which the number of decoys generated is large; and (2) the estimated minimal effective energy should be higher than the native energy for most proteins. We found that these conditions are satisfied by parameters in the range $A \approx 0.1$, $B \approx 4$. The evaluated value of the exponent A agrees very well with the theoretical estimate $A \approx 0.115$. With these parameters, only 105 domains out of 4528 have negative α'. The average normalized energy gap is 0.3, which is significantly larger than for random sequences, α' is not correlated with chain length and it is correlated with the estimate obtained through threading, α, with $R = 0.49$. These results support the plausibility of our new computation of the normalized energy gap.

The above equations allow the calculation of the normalized energy gap as a function of the zeta score $Z_{\rm nat}$, Equation 5, and the standard deviation of the contact interactions, $\sigma_U = \sqrt{\langle U^2\rangle - \langle U\rangle^2}$:

$$\alpha' = \frac{\sigma U}{(1-q)|E_{\rm nat}/C|}\left(-Z_{\rm nat} + \sqrt{\frac{2}{C}\log(m_N)}\right) \qquad (15)$$

The condition that α' should be positive establishes a minimal absolute value of the zeta score as

$$-Z_{\rm nat} > \sqrt{\frac{2(AN+B)}{(C/N)N}} \qquad (16)$$

Therefore, since C/N increases with N, the minimal allowed $-Z_{\rm nat}$ is predicted to be negatively correlated with chain length. This correlation was indeed observed for the actual value of $-Z_{\rm nat}$ in our data set. Figure 2 shows $|Z_{\rm nat}|$ together with its theoretical lower bound, Equation 16, using the parameters A and B determined above. The two quantities behave very similarly, supporting our argument that the relaxation of the lower bound is the reason why $|Z_{\rm nat}|$ is found to decrease with chain length.

Also in this case, as in Figure 1, we can see that long proteins lay significantly above the predicted line: the minimal stability requirements in terms of $-Z_{\rm nat}$ are predicted to decrease with chain length, but longer proteins tend to stay higher than the minimal requirements. We will come back on this observation when discussing the PCA.

Stability with respect to misfolding measured by α' and stability with respect to the unfolding measured by $-E_{\rm nat}/N$ are only weakly correlated (R = 0.13). The correlation is slightly stronger between α' and $-E_{\rm nat}/C$ (R = 0.17). However, their correlation is essentially due to the variable $-Z_{\rm nat}$, which is positively correlated with both α' (R = 0.74) and with $-E_{\rm nat}/C$ (R = 0.63). If we calculate the partial correlation coefficient of α' and $-E_{\rm nat}/C$ at fixed $-Z_{\rm nat}$, we find a rather strong negative correlation, $R = -0.51$. This negative correlation is easily explained considering the variable $-\langle U\rangle$, Equation 4, the average strength of non-native and native contacts. This quantity is related to the average hydrophobicity of the protein sequence. As expected, $-E_{\rm nat}/C$ and $-\langle U\rangle$ are strongly correlated, with R = 0.65. Moreover, interactive sequences tend also to assign low energy to misfolded configurations, so that the normalized energy gap α' tends to be reduced ($R = -0.46$). PCA confirms and further clarifies this trade-off between unfolding and misfolding stability mediated by hydrophobicity.

Chain length and protein composition

Apart from the reduced optimization of native interactions with respect to the sequence background, the decrease in the strength of native interactions with chain length can also be due to the change in amino acid composition, affecting both native and non-native interactions. To address this question, we examined the association between amino acid frequencies and chain length. This was already studied by White (1992) several years ago and recently by Sandelin (2004), in this case limited to hydrophobic amino acids. Our present results are in qualitative agreement with previous ones and are included for completeness.

Confirming the previous results of White (1992), we observed that the cumulative frequency of positively charged amino acids, Lys and Arg, decreases with chain length, whereas the cumulative frequency of negatively charged amino acids does not vary significantly, because the significant decrease in Glu is compensated by the increase in Asp. Despite the decrease in positively charged residues, the number of salt bridges per residue increases with chain length (see below).

Both the small amino acids Ala and Gly increase significantly in content with chain length. Since these amino acids are weakly interacting, this result is in line with the observed decrease in the average interaction strength. Since their side chain has no conformational entropy, the increase in Ala and Gly content partly explains the decrease in the conformational entropy with chain length.

The frequency of Cys decreases strongly with chain length, as we discuss below in our analysis of disulfide bonds.

The most strongly interacting amino acids with aliphatic (Val, Ile, Leu, Met) and aromatic (Phe, Tyr, Trp) side chains do not show a significant association with chain length, either individually or in combination, in agreement with results of White (1992) and Sandelin (2004). Nevertheless, the frequency of hydrophobic residues versus length has a maximum, which is significant ($P < 10^{-4}$). The decrease at short chains is probably due to their smaller hydrophobic core.

To test whether the changes in protein composition produce a decrease of inter-residue interaction strength with chain length, we used the interactivity scale defined by Bastolla *et al.* (2005), which accounts for the main component of our effective interaction matrix. The mean interactivity $\langle h\rangle$ can be calculated from the protein composition alone (see Materials and methods). Like the hydrophobic content, the mean interactivity also has a maximum as a function of chain length, which is once again attained for chains between 200 and 300 residues long (see Figure 3, bottom left). A chi-squared test with nine length bins yields $\chi^2 = 113$, confirming the high significance. Both shorter and longer proteins have significantly lower

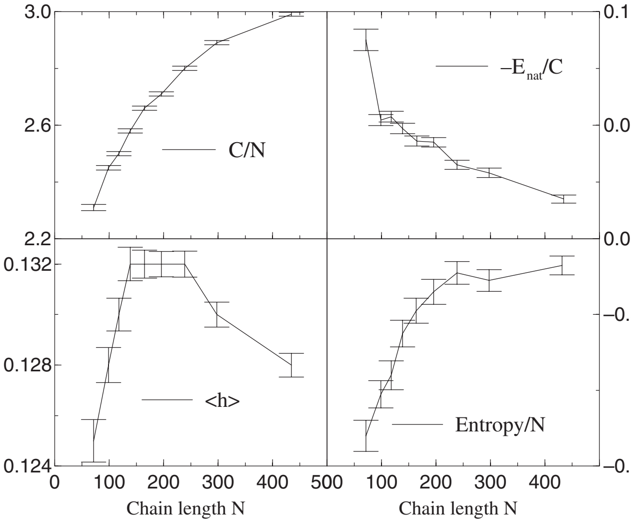

Fig. 3. Relationship between protein energetics and chain length. Mean and standard error of the mean are shown for each length bin. Top left: increase in the number of contacts per residue C/N with domain length. Top right: minus effective native free energy per native contact, estimating the strength of native interactions. Bottom left: mean interactivity of the protein sequence, taking into account both native and non-native interactions. Bottom right: conformational entropy per residue of the unfolded state.

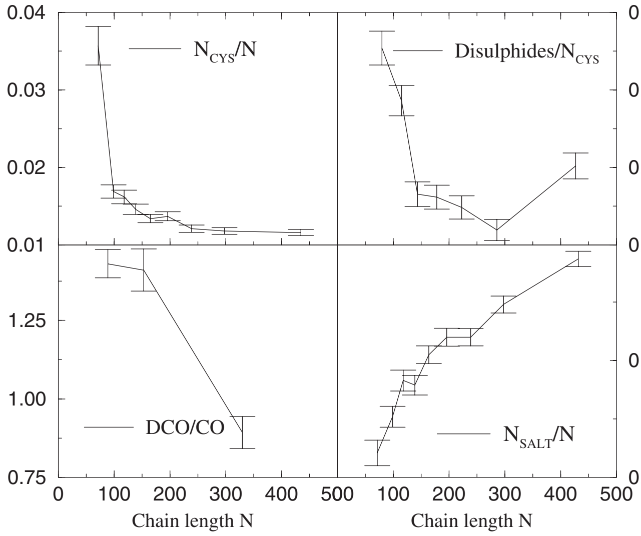

Fig. 4. Relationship between chain length and disulfide bonds. Mean and standard error of the mean are shown for each bin. Top left: the frequency of cysteine residues decreases with chain length. Top right: the frequency of disulfide bonds per cysteine residue also decreases with chain length. As a result, there is a strong decrease in disulfide bridges with chain length. Bottom left: DRCO divided by RCO. The loops between disulfide-bonded residues are significantly longer than for other contacts in short but not in long proteins. Bottom left: in contrast with disulfide bonds, the frequency of salt bridges per residue increases with chain length.

interactivity and for proteins with >250 residues the mean interactivity is negatively correlated with chain length (R = −0.21). The same qualitative conclusions were obtained considering the average interaction strength $-\langle U\rangle$ instead of interactivity.

Disulfide bonds and folding stability

Disulfide bonds are present in about 24% of the proteins in our data set. They are not randomly distributed among protein structures. The number of intra-chain disulfide bonds is negatively correlated with the unfolding free energy per residue that the protein would have in the absence of disulfide bonds, $-E'_{\mathrm{nat}}/N$, Equation 6. The correlation coefficient is $R = -0.31$ ($P < 10^{-4}$). A chi-squared analysis confirms that the difference in unfolding free energy among classes with different numbers of disulfide bonds is highly significant ($\chi^2 = 92$, six degrees of freedom).

We also measured the average length of loops between disulfide-bonded residues, normalized by chain length (DRCO). This variable is related to the reduction in the entropy of the denatured state, which is thought to be the main mechanism by which disulfide bonds stabilize the native state (Anfinsen and Scheraga, 1975; Betz, 1993). Also this variable is negatively correlated with $-E'_{\mathrm{nat}}/N$ ($R = -0.26$, $P < 10^{-4}$).

The number of disulfide bonds per residue is also positively correlated with the normalized energy gap α' ($R = 0.20$). This may be in part explained because disulfide bonds per residues are associated with weaker interactivity $\langle h\rangle$ ($R = -0.21$), which implies larger normalized energy gaps. Therefore, these results suggest that the increase in the number of disulfide bonds in long proteins allows the reduction of hydrophobic interactions, decreasing non-native interactions and enhancing stability against misfolding. This hypothetical mechanism is supported by the PCA.

Disulfide bonds and chain length

Investigating the relationship between disulfide bonds and chain length allows us to distinguish between the relative contribution of genetic drift and natural selection in the evolution of disulfide bonds. In case disulfide bonds are mainly fixed through random drift, we expect that their number is correlated with the number of cysteines in the protein chain, which in turn is expected to be correlated with chain length, since by hypothesis it results from a random mutation process. This leads to the expectation that the number of disulfides divided by chain length is unrelated to chain length if their evolution is driven by random drift.

Instead, the number of disulfide bonds per residue decreases with chain length for short proteins (see Figure 4). For long proteins, this number does not appear to vary systematically with chain length. The hypothesis that the average number of disulfide bonds per residue is the same for all lengths yields $\chi^2 = 190$ with six degrees of freedom, which is associated with a vanishingly small probability. Dividing the data set into short (<200 residues) and long chains, we find that the number of disulfides per residue is not correlated with chain length for long chains (the correlation coefficient is $R = 0.028$), whereas for short chains the correlation is negative and highly significant ($R = -0.24$, $P < 10^{-4}$).

The increase in the frequency of disulfide bonds for short proteins is in part a consequence of the increase of cysteine frequency for short proteins (White, 1992); see Figure 4, top left. Nevertheless, this is not sufficient to account for the whole increase in disulfide bond frequency. For short proteins, most cysteine residues form intra-chain disulfide bonds, the exceptions being mostly metalloproteins, where the metal–cysteine clusters play an important stabilizing role. In contrast, for long proteins the frequency of cysteine residues is almost independent of chain length and the fraction of cysteine residues that form disulfide bonds is much smaller, as shown in Figure 4.

Chains shorter than 200 residues also favor relatively longer disulfide loops than long chains (the DRCO is 0.35 and 0.16, respectively). Since longer loops imply a stronger entropy reduction in the unfolded state, this is another indication that disulfide bonds lead to a relatively greater enhancement of stability in shorter proteins. This results is in part due to the general tendency of the RCO to decrease with chain length, but also the normalized variable DRCO/RCO is significantly larger in short chains (mean 1.41, standard error of the mean 0.04) than in long chains (mean 0.91, standard error of the mean 0.06), where it is compatible with equality between DRCO and RCO.

These results are summarized in Figure 4. Taken together, they strongly reject the hypothesis that the fixation of disulfide bonds is driven by random drift for short chains, whereas they are consistent with this hypothesis for long chains.

Chain length and salt bridges

In addition to disulfide bonds, we also distinguished salt bridges from other kinds of interactions for two reasons: (1) contact energy functions such as ours cannot estimate electrostatic interactions accurately, since their contribution to the free energy of unfolding depends on a delicate balance between electrostatic interactions and desolvation penalty, whose net effect is not always stabilizing (Bosshard *et al*., 2004); and (2) salt bridges differ from hydrophobic interactions in that they are highly specific and they differ from disulfide bonds in that they are typically short range (Kumar and Nussinov, 1995), therefore their role in stabilizing protein folding is expected to be distinct.

We observed that the occurrence of salt bridges depends strongly on chain length. As the length increases, salt bridges represent an increasing fraction of the inter-residue contacts. This fraction, N_{SALT}/C, increases by 27% in the length range examined. The correlation between N_{SALT}/C and chain length is $R = 0.10$, which is small but significant. The trend of the number of salt bridges per residue, N_{SALT}/N, is much stronger, since there is an additional factor C/N. It increases by 61% of its value in the whole range and it has correlation coefficient $R = 0.23$ with chain length. The number of salt bridges, normalized either by N or by C, is also strongly associated with the RCO, with a negative correlation coefficient $R = -0.18$ and -0.25, respectively. This is in part a consequence of the negative correlation between RCO and chain length and in part (partial correlation coefficient at fixed length, $R = -0.15$) it is a consequence of the fact that salt bridges are frequent in contacts at short sequence distance (Kumar and Nussinov, 1995), which are more common in structures with low RCO.

The number of salt bridges per residue is also negatively correlated with the number of disulfide bonds per residue ($R = -0.19$), which is only in part explained by their opposite length dependence (partial correlation coefficient $R = -0.16$). It is positively correlated with the mean interaction energy, $-\langle U \rangle$, with $R = 0.16$, and uncorrelated with the mean native energy per contact, therefore it is negatively correlated with $-Z_{\mathrm{nat}}$, with $R = -0.23$. In other words, proteins with better designed hydrophobic interactions tend to have fewer salt bridges.

Summarizing (see Table I), we see that salt bridges tend to substitute disulfide bonds in longer proteins, in particular in proteins with short-range contacts (smaller RCO), and they occur more frequently in sequences with large interactivity and poor design of hydrophobic interactions.

Principal component analysis

The measures of protein stability and the interactions that contribute to it are strongly correlated with each other. We used PCA in order to disentangle these contributions. The dominant eigenvectors that emerge from this analysis can be interpreted as the most common directions in stability and interaction space along which protein evolution can move. We interpret the first three directions as follows: (1) the interactivity direction, along which there is a trade-off between stability against misfolding and stability against unfolding; (2) the size direction, along which minimal stability requirements change as a consequence of changes in protein size; and (3) the design direction, along which both types of stability improve.

In this analysis, we considered eight variables that represent the most important determinants of protein stability with little redundancy. Two variables represent protein size and topology, that determine the structural constraints acting in protein evolution: the number of contacts, per residue C/N, strongly correlated with chain length; and the absolute contact order (ACO), an indicator of native topology, which is also strongly positively correlated with chain length. Two variables estimate two kinds of protein stability: stability against unfolding is estimated through the native energy per residue, $-E_{\mathrm{nat}}/N$ (however, this variable does not take into account the conformational entropy), and stability with respect to misfolding is estimated through the normalized energy gap α', evaluated through Equations 15 and 14. (The value of α' depends on the estimated parameters A and B, but we found that qualitatively results emerging from PCA are robust with respect to the choice of these parameters. For a control, we also examined the normalized energy gap obtained through threading, which is overestimated more and more for longer proteins, for which less decoys are generated. This artifact produces a spurious correlation between α and length that can be effectively eliminated through PCA, since this correlation basically only contributes in the size direction. The qualitative description of the other directions is the same as obtained through α'.) Last, we considered four types of interactions contributing to protein stability: the mean interaction strength calculated with our energy function, $-\langle U \rangle$, which mainly takes into account hydrophobicity (similar qualitative results are obtained considering the mean interactivity $\langle h \rangle$ instead); the number of disulfide bonds per residues, N_{DIS}/N; the number of salt bridges per residue, N_{SALT}/N; and minus the conformational entropy of the unfolded state, $-S$.

We represent in Figure 5 the three components whose contribution to the total variance is larger than the average. The coefficients with which the eight variables enter the first three PCs are plotted in the figure as bars.

First component: interaction strength

The first principal component (PC) accounts for 28% of the total variance. Its main contribution comes from the interaction

Table I. Correlation coefficients between the main variables studied (see text): below the diagonal, subset of short domains ($N > 200$, 2835 domains); above the diagonal, subset of long domains ($N \leqslant 250$, 1166 domains); only highly significant values ($P > 0.001$) are shown

	N	C/N	ACO	$\langle h \rangle$	$-\langle U \rangle$	$-Z_{\mathrm{nat}}$	$-E_{\mathrm{nat}}/N$	α'	S	Dis./N	Salt/N
N	–	0.33	0.61	−0.13		−0.19				0.09	0.13
C/N	0.60	–	0.25	−0.15	−0.14	−0.14		0.17	0.19	0.09	
ACO	0.56	0.38	–	−0.21	−0.17	−0.10	−0.15			0.12	
$\langle h \rangle$	0.18	0.17		–	0.79		0.70	−0.32	−0.11	−0.18	
$-\langle U \rangle$	0.07		−0.17	0.71	–		0.80	−0.39	−0.55	−0.14	0.19
$-Z_{\mathrm{nat}}$	−0.43	−0.44	−0.17	−0.29	−0.16	–	0.52	0.84		0.20	0.13
$-E_{\mathrm{nat}}/N$	0.10	−0.12	−0.12	0.55	0.71	0.42	–	0.17	−0.42		0.13
α'	−0.10		0.12	−0.40	−0.49	0.78	0.17	–	0.16	0.24	−0.12
$-S$	0.15	0.11	0.22	−0.09	−0.43		−0.26	0.18	–		−0.30
Dis./N	−0.24	−0.20		−0.21		0.32		0.20		–	−0.14
Salt/N	0.16	0.21			0.14	−0.19	0.08	−0.15	−0.22	−0.18	–

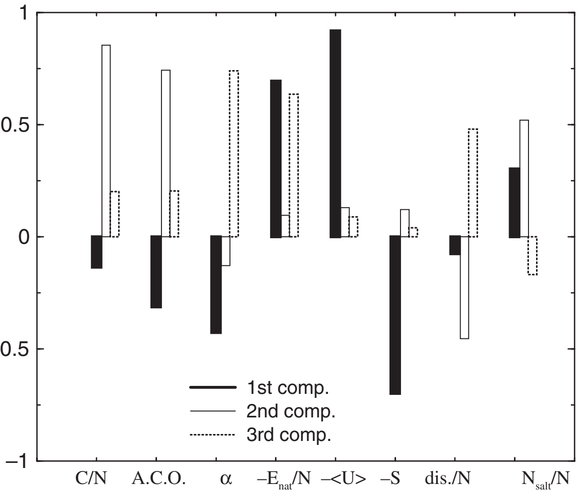

Fig. 5. Summary of PCA. Along the horizontal axis are plotted the eight variables examined, representing as bars their contributions to the three dominant PCs. All 4519 domains are used in the analysis.

strength, $-\langle U \rangle$. Along this direction, the stability against unfolding, estimated through $-E_{\text{nat}}/N$, increases, whereas the stability against misfolding, estimated through the normalized energy gap, decreases. Therefore, there is a trade-off between these two kinds of stability.

This balance is influenced by the protein structure. In fact, this component is negatively correlated with the ACO ($R = -0.19$). We believe that this correlation is due to the influence of the ACO on the folding rate (Ivankov *et al.*, 2003): proteins with larger ACO fold more slowly, therefore they are expected to undergo stronger selection for improved stability against misfolding, with the consequence of favoring this kind of stability with respect to stability against unfolding.

Looking at mechanisms of protein stability other than interactivity, we see that the conformational entropy of the unfolded state, $-S$, is negatively correlated with the mean interaction strength and decreases with this component. Therefore, the increase in unfolding stability is actually smaller than the estimate based only on $-E_{\text{nat}}/N$. The number of disulfide bonds per residue also decreases, in line with the fact that this quantity is correlated with the deficit in unfolding free energy. We interpret the increase in disulfide bridges at low interactivity as a compensation of the otherwise reduced unfolding stability. Last, the number of salt bridges per residue contributes positively to the enhanced stability. In fact, this quantity is positively correlated with both the ACO and the interaction strength, to which it contributes positively. We will discuss salt bridges further in next section.

Second component: size

The second PC accounts for 23% of the total variance and it is mainly related to protein size, since it receives its main contribution from *C/N* and the ACO. Protein stability hardly varies along this component. The normalized energy gap α' is uncorrelated with it and the effective free energy per residue $-E_{\text{nat}}/N$ is weakly negatively correlated, but its decrease is compensated for by the increase of the entropy term $-S$. Disulfide bonds and salt bridges strongly co-vary with this component, the former decreasing and the latter increasing in importance, whereas the mean interaction strength $-\langle U \rangle$ is uncorrelated with it. As we will discuss later, we interpret the decrease in disulfide bridges with chain length as an indication of the fact that they are positively selected in short proteins, whereas they evolve in an effectively neutral way in long proteins.

Third component: design

The third PC accounts for 14.5% of the variance and receives its main contribution from the normalized energy gap. Both types of stability improve along this direction. Since the average interaction strength and the chain entropy hardly vary along this direction, the improvement has to be attributed to the better design of native interactions. In line with this, if we include $-Z_{\text{nat}}$ in the analysis (not shown), this variable gives a strong positive contribution to this component. This component is weakly positively correlated with the two structural indicators *C/N* and ACO. However, the correlation becomes stronger if we distinguish between short and long domains: the correlations become $R = 0.34$ and 0.36, respectively, for short domains and $R = 0.61$ and R 0.39 for long domains.

Concerning the stabilization mechanisms, this PC is almost uncorrelated with the mean interactivity $-\langle U \rangle$ and with $-S$, positively correlated with the number of disulfide bonds per residue and negatively correlated with the number of salt bridges per residue. Therefore, it appears that most of the stabilization is achieved by optimizing the design of generic interactions.

We interpret this component as a direction along which native interactions are more efficiently designed, as a consequence of enhanced selection for protein stability. Since this component is positively correlated with both *C/N* (in particular for long domains) and with the ACO, large values of this component appear to be associated with more slowly folding proteins. We believe that these proteins undergo stronger selective pressure for enhanced stability against misfolding.

Discussion

We have found in this work that some of the most relevant factors stabilizing the native states of proteins provide significantly different contributions to protein stability for chains of different length. We examined a set of variables representing two structural indicators, the number of contacts per residue *C/N* and the ACO; two types of protein stability, against unfolding ($-E/N$) and against misfolding (α), and four types of interactions: contact interactions, mainly based on hydrophobicity, disulfide bonds, salt bridges and the conformational entropy of the unfolded state. We also considered a measure of the design of native interactions with respect to non-native ones, $-Z_{\text{nat}}$.

To rationalize and separate the roles of these contributions, we performed a PCA. This allows the identification of three main directions, along which native interactions may have evolved as a result of the stability constraints specific to their structure. Along the first direction, stability against unfolding increases at the expense of stability against misfolding. Along the second direction, the two kinds of stabilities do not change, but stability requirements vary in response to protein structure, mainly its size. Along the third direction, both kinds of stability increase, mainly through a better design of native interactions. We will discuss these three evolutionary directions separately.

The hydrophobicity dimension: trade-off between unfolding and misfolding stability

The native state of a protein must be stable with respect to the unfolded state (unfolding stability) and with respect to compact, incorrectly folded conformations (misfolding stability). These two kinds of stabilities are negatively correlated in families of orthologous proteins, sharing the same structure and function but differing in thermodynamic properties (Bastolla *et al.*, 2004). Here their association is further confirmed by the negative correlation between the strength of native interactions and the normalized energy gap for proteins with different structure, along the first principal component, expressing the variation of interactivity at constant chain length (see Figure 5).

The association between the two kinds of stability can be easily understood: A very interactive protein with many hydrophobic residues would be very stable against unfolding, but it would not fold spontaneously in a unique state, because also alternative configurations would have comparably low free energy. It would behave almost like a homopolymer or a random heteropolymer, whose normalized energy gap is expected to vanish. Consistently, Sandelin (2004) has shown through lattice simulations that the requirement of non-vanishing energy gap imposes a limit on the maximum hydrophobicity of model proteins, which is more stringent for longer proteins.

There is experimental support for this view: proteins with low hydrophobicity and high charge are unfolded in native conditions (Uversky, 2002a), implying that some minimal hydrophobicity value is required for proteins to fold. On the other hand, very hydrophobic sequences tend to adopt partially folded configurations (Uversky, 2002b), which sometimes are off-pathway and may trigger aggregation.

We believe that stability against misfolding is an important selective property, favoring correct folding and preventing aggregation. This is also a central requirement in some models of protein evolution (Bastolla *et al.*, 1999, 2003). The trade-off that we have demonstrated implies that protein evolution has to mediate between stability against unfolding and stability against misfolding. It was shown recently that a mutation bias favoring or depressing more hydrophobic mutations can shift this balance (Bastolla *et al.*, 2004). The results presented here indicate that protein length and topology can also shift this balance: longer proteins with larger ACO tend to favor misfolding stability over unfolding stability. It is possible that this is due to their slower folding rate, that exerts a stronger selective pressure against misfolding.

The chain length dimension: weakening of stability requirements

The second principal component is mainly correlated with protein size. Along this direction the two measures of protein stability hardly change, but native interactions become on average weaker with chain length (smaller $-E_{\mathrm{nat}}/C$) and less optimized with respect to non-native ones (smaller $-Z_{\mathrm{nat}}$), meaning that in longer proteins native interactions are more similar to background non-native interactions. Since our energy function is mainly based on hydrophobic interactions, this result refers mainly to this type of interactions.

The conclusion that the hydrophobic profile of the protein sequence is less optimized for its structure in longer proteins than in shorter ones was also reached in a recent study of site-specific amino acid distributions. In this study, amino acid distributions were predicted to have a Boltzmann form, whose parameters represent the intensity of site-specific selection for structural stability. The absolute values of these selection parameters were found to decrease with chain length (Porto *et al.*, 2004), implying that in longer protein it is more frequent to find residues that are less fit to their structural environment. Consistently, Sandelin noticed that the fraction of core positions increases with chain length (see Equation 8), but the frequency of hydrophobic residues is almost constant with chain length (as we have shown here, it even slightly decreases for $N > 300$). Therefore, longer proteins have on average a larger fraction of buried polar residues (Sandelin, 2004).

We interpret this reduced optimization of the hydrophobic profile in the sense that stability constraints become easier to fulfil for longer proteins, as expected from a simple model of protein folding, Equations 12 and 16. Since longer proteins have a larger number of native contacts per residue, they can counterbalance more easily the reduction in chain entropy upon folding, which is proportional to the number of residues. In this context, the tendency of the average native interactions to decrease with chain length is a manifestation of the general tendency of proteins to attain only marginal stability, as demonstrated in the neutral model of Taverna and Goldstein (2002). Moreover, the requirement of stability against misfolding imposes that the absolute value of the Z-score of native interactions must be above a minimum value. This condition, Equation 16, is easier to satisfy for longer chains. Therefore, the drift of the actual Z-scores towards their minimal allowed value is another manifestation of the tendency of proteins to reach marginal stability.

An exception to this picture is the increase with chain length of the number of salt bridges, both per residue and per contact. One possible explanations is that salt bridges are positively selected to compensate for the reduction in hydrophobic interactions and disulfide bonds: only these two types of interactions sense the weakening of stability requirements with chain length. However, this interpretation does not explain why salt bridges should be preferred for longer proteins. We think that this preference is due to the peculiarity of salt bridges: they are highly specific interactions and they are typically short range. As such, they are expected to accelerate the folding rate of long proteins, which would otherwise fold very slowly. Therefore, our interpretation is that salt bridges in longer proteins are preferred not in order to increase the stability of the folded state, but in order to make folding faster and more specific. As a preliminary test of this interpretation, we looked at the number of salt bridges as a function of the distance cut-off used in this definition. When this cut-off is reduced to 2.5 Å, corresponding to the most strongly interacting salt bridges, the length dependence completely disappears: long proteins are not enriched in the strongest salt bridges, but only in the weak ones.

Protein evolution and disulfide bonds

The general scenario described above is well illustrated through the example of disulfide bonds, which are the strongest interactions in protein structures.

In analyzing disulfide bonds, it is important to distinguish intracellular and extracellular proteins. Disulfide bonds are expected to be found almost exclusively in the latter ones, since the chemical environment inside a typical cell is reducing (Gilbert, 1990) and cysteine residues in intracellular proteins are generally found in their reduced form with free sulfhydryl groups (Mallick *et al*., 2002). Correlations between extracellularity and protein size or other topological or thermodynamic indicators might therefore confuse the analysis of disulfide bonds. However, the PDB files do not contain automatically readable information about the cellular location of the protein and we were not able to distinguish intracellular and extracellular proteins in our large-scale analysis.

Our results are consistent with the hypothesis that disulfide bonds evolve in an effectively neutral way in long proteins, but not in short proteins, where they are much more frequent than expected under neutral evolution. In fact, shorter proteins have a larger frequency of cysteine residues, a larger frequency of those form disulfide bonds and their average relative loop length (DRCO) is larger than for other types of contacts (RCO). This is consistent with the hypothesis that short proteins experience a stronger selection for strong stabilizing interactions, owing to their reduced number of contacts per residue.

The enrichment of disulfides in short chains is complementary to the finding that proteins of thermophilic bacteria, facing more severe problems of unfolding stability, are also enriched in disulfide bonds (Mallick *et al*., 2002).

We also observed a negative correlation between the number of disulfide bonds and the predicted unfolding free energy per residue that the protein would have in the absence of them. We can decompose the latter in two factors as $\left(-E'_{\mathrm{nat}}/C\right)\cdot\left(C/N\right)$. For short proteins, this negative correlation is in large part due to the increase in disulfide bonds per residue for decreasing *C*/*N* ($R = -0.20$), which implies a deficit in unfolding free energy that may enhance the selection pressure for disulfide bonds. For long proteins, this correlation is weaker and it is entirely due to the negative correlation between $-E'_{\mathrm{nat}}/C$ and disulfide bonds per residue, since their correlation with *C*/*N* is now very small and positive in sign. This is another manifestation of the tendency of naturally evolved proteins to attain only marginal stability.

The design dimension: enhancing protein stability

Both unfolding stability and misfolding stability increase together along the third principal component. This is achieved without a large increment in the strength of hydrophobic interactions and with only a small increment in the contribution of disulfide bonds and a small decrease in the contribution of salt bridges. The key to this success is a better design of hydrophobic interactions, through an increased $-Z_{\mathrm{nat}}$. This is an example of the general fact that protein stability can be improved through evolution when it is needed.

The grounds for this enhanced stability are of course functional, but they are also in part structural. Proteins with better designed native interactions tend to be longer and to have larger ACO, as demonstrated by the third principal component. This does not contrast with the observations presented above. The scenario that these results suggest is the following: minimal stability requirements are more easily satisfied for longer proteins, but longer proteins also tend to remain above these minimal stability thresholds more than short proteins do. Both principles are exemplified in Figures 1 and 2.

Acknowledgements

UB acknowledges financial support in the last stages of this work from the I3P programme of the Spanish CSIC, co-funded by the European Social Fund, and from the project FIS2004-05073-C04-04 of the Spanish Ministry of Education and Science.

References

Abkevich,V.I. and Shakhnovich,E.I. (2000) *J. Mol. Biol.*, **300**, 975–985.

Alm,E. and Baker,D. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 11305–11310.

Anfinsen,C. and Scheraga,H. (1975) *Adv. Protein Chem.*, **29**, 205–299.

Bastolla,U., Roman,H.E. and Vendruscolo,M. (1999) *J. Theor. Biol.*, **200**, 49–64.

Bastolla,U., Farwer,J., Knapp,E.W. and Vendruscolo,M. (2001) *Proteins*, **44**, 79–96.

Bastolla,U., Porto,M., Roman,H.E. and Vendruscolo,M. (2003) *J. Mol. Evol.*, **56**, 243–254.

Bastolla,U., Moya,A., Viguera,E. and van Ham,R.C.H.J. (2004) *J. Mol. Biol.*, **343**, 1451–1466.

Bastolla,U., Porto,M., Roman,H.E. and Vendruscolo,M. (2005) *Proteins*, **58**, 22–30.

Bava,K.A., Gromiha,M.M., Uedaira,H., Kitajima,K. and Sarai,A. (2004) *Nucleic Acids Res.*, **32**, D120–D121.

Betz,S. (1993) *Protein Sci.*, **2**, 15551–15558.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G. Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Bosshard,H.R., Marti,D.N. and Jelesarov,I. (2004) *J. Mol. Recognit.*, **17**, 1–16.

Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) *Nucleic Acids Res.*, **32**, D189–D192.

Demetrius,L. (1995) *Protein Eng.*, **8**, 791–800.

Demetrius,L. (2002) *J. Theor. Biol.*, **217**, 397–411.

Derrida,B. (1981) *Phys. Rev. B*, **24**, 2613.

Finkelstein,A.V. and Badretdinov (1997) *Fold. Des.*, **2**, 115–121.

Galzitskaya,O.A. and Finkelstein,A.V. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 11299–11304.

Gilbert,H.F. (1990) *Adv. Enzymol. Relat. Areas Mol. Biol.*, **63**, 69–172.

Goldstein,R., Luthey-Schulten,Z.A. and Wolynes,P.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 4918–4922.

Gutin,A.M., Abkevich,V.I. and Shakhnovich,E.I. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 1282–1286.

Gutin,A.M., Abkevich,V.I. and Shakhnovich,E.I. (1996) *Phys. Rev. Lett.*, **77**, 5433–5436.

Hobohm,U. and Sander,C. (1994) *Protein Sci.*, **3**, 522–524.

Ivankov,D.N., Garbuzynskiy,S.O., Alm,E., Plaxco,K.W., Baker,D. and Finkelstein,A.V. (2003) *Protein Sci.*, **12**, 2057–2062.

Kimura,M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kowald,A. and Demetrius,L. (2005) *Proc. R. Soc. Lond. B*, in press.

Kumar,S. and Nussinov,R. (1999) *J. Mol. Biol.*, **293**, 1241–1255.

Lobry,J.R. (1997) *Gene*, **205**, 309–316.

Mallick,P., Boutz,D.R., Eisenberg,D. and Yeates,T.O. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 9679–9684.

Muñoz,V. and Eaton,W.A. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 11311–11316.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.

Orland,H., Itzykson,C. and de Dominicis,C. (1985) *J. Phys. Lett.*, **46**, L–353.

Plaxco,K.W., Simons,K.T. and Baker,D. (1998) *J. Mol. Biol.*, **277**, 985–994.

Porto,M., Roman,H.E., Vendruscolo,M. and Bastolla,U. (2005) *Mol. Biol. Evol.*, **22**, 630–638 [erratum: *Mol. Biol. Evol.*, **22**, 1165].

Sandelin,E. (2004) *Biophys. J.*, **86**, 23–30.

Shakhnovich,E.I. and Gutin,A.M. (1989) *Biophys. Chem.*, **f 187**.

Sueoka,N. (1961) *Proc. Natl Acad. Sci. USA*, **47**, 469–478.

Taverna,D.M. and Goldstein,R.A. (2002) *Proteins*, **46**, 105–109.

Thirumalai,D. (1995) *J. Phys.*, **5**, 1457–1469.

Uversky,V.N. (2002a) *Eur. J. Biochem.*, **269**, 2–12.

Uversky,V.N. (2002b) *FEBS Lett.*, **514**, 181–183.

White,S.H. (1992) *J. Mol. Biol.*, **227**, 991–995.

Ziehe,M. and Demetrius,L. (2005) *Proc. R. Soc. Lond. B*, in press.

Received December 9, 2004; revised and accepted June 21, 2005

Edited by David Thirumalai