doi:10.1006/jmbi.2001.4949 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **312**, 289–307

JMB

Understanding Hierarchical Protein Evolution from First Principles

Nikolay V. Dokholyan* and Eugene I. Shakhnovich

Department of Chemistry
Harvard University
12 Oxford Street, Cambridge
MA 02138, USA

**Corresponding author*

We propose a model that explains the hierarchical organization of proteins in fold families. The model, which is based on the evolutionary selection of proteins by their native state stability, reproduces patterns of amino acids conserved across protein families. Due to its dynamic nature, the model sheds light on the evolutionary time-scales. By studying the relaxation of the correlation function between consecutive mutations at a given position in proteins, we observe separation of the evolutionary time-scales: at short time intervals families of proteins with similar sequences and structures are formed, while at long time intervals the families of structurally similar proteins that have low sequence similarity are formed. We discuss the evolutionary implications of our model. We provide a ''profile'' solution to our model and find agreement between predicted patterns of conserved amino acids and those actually observed in nature.

Keywords: protein evolution; energy gap model; profile solution

Introduction

Understanding protein evolution still remains a major challenge in molecular biology.[1–16] While the mechanisms of mutations in DNA sequences that code for proteins are known,[17] the selective fixation of these mutations in proteins is far from clear. Mutations occurring in DNA are directly governed by physical-chemical processes and their fixation is subject to cellular repair mechanisms being able to protect nucleotide(s) from modifications. Mutations occurring in protein sequences may drastically alter their physical, chemical, and biological properties. Thus, in the course of evolution nature exerts pressure to preserve those amino acids that play an important role in the folding kinetics, functionality and stability of proteins. Our goal is to understand evolution from the statistical mechanics perspective.

There are several principal facts observed in proteins: (i) a protein sequence folds into a unique three-dimensional structure (there might be exceptions, e.g. prions); (ii) protein sequences are selected, i.e. a randomly chosen polypeptide most likely aggregates in solution without forming any definite three-dimensional structure; (iii) proteins taken from various species and having sequence identity (*ID*) at least $25-30\,\%$ have similar three-dimensional structures (native state)[15,17–24] and are said to belong to the same fold family; (iv) some pairs of proteins sharing the same fold have sequence similarity as low as expected for random sequences $ID \sim 8-9\,\%$;[11,25,26] (v) within the same fold family, protein sequences have only 3-4 % ''anchored'' amino acids.[9]

Here, we call homologs a set of proteins that have at least 25 % sequence similarity and are structurally similar. A set of structurally similar proteins that may have less than 25 % sequence similarity we call a group of analogous proteins or analogs. Analogs include several families of homologs and generally constitute a larger set of proteins than homologs. Known homologs and analogs are collected in the HSSP[22] and FSSP[25] databases, respectively, and are the subject of our study. Because our conclusions suggest that the concept of homology and analogy may be ill-defined in some cases, we do not follow conventional definition of (analogs) homologs as proteins (not) sharing the same ancestor.[27] Rather we define homologs as proteins that have substantial sequence similarity ($ID \geqslant 25\,\%$), while analogs are proteins that share the same fold but have low sequence similarity. This definition is natural for our considerations that focuses on physical aspects of protein evolution, such as stability and, to a lesser extent, folding kinetics.

E-mail address of the corresponding author:
dokh@wild.harvard.edu

0022-2836/01/010289–19 $35.00/0

Here, we propose a model of evolution (energy gap model) that, based on facts (i) and (ii), attempts to reproduce the rest of the remaining principal observations (iii)-(v) described above. The energy gap model is based on the design of a set of structurally identical sequences by the *Z*-score minimization.[28–31] The idea is to find the similarities in the sequences of such a set and to recover those residues that are conserved across this set. The protein folding theory[32,33] suggests that *Z*-score minimization is equivalent to maximizing the energy gap between misfolded or unfolded conformations and the native state of a protein. It has been pointed out that such maximization results in stable and fast-folding proteins.[29,34] Thus, by designing sequences that have the same fold, we attempt to mimic evolution in diversifying protein sequences for the same fold family. In addition, the energy gap model is a dynamical model, i.e. there is an implicit time-scale that allows one to follow the evolution of sequences during the design procedure. The model is discussed in detail and a profile approximation to this model is outlined below. We show that our view of evolution proposed below is consistent with the implications of the proposed model. Next, we discuss our scenario of protein evolution.

We conjecture that hierarchical organization of structurally similar proteins may be the result of the separation of the evolutionary time-scales, shown schematically in Figure 1. On a time-scale τ_o, a set of mutations occur that do not affect those amino acids that play crucial thermodynamical, kinetical and/or functional roles. As a result, there is little variation in sequences at the important sites of proteins. If a mutation occurs at the thermodynamically, kinetically and/or functionally important sites, it usually substitutes amino acids with close physical properties so that core, nucleus and/or functional site are not disrupted and the protein folds into its family fold, is stable in this fold, and its function is preserved. At this time, a family of homologs is born.

Rarely, at time-scale τ, correlated mutations occur[35–37] that modify several amino acids at the core, nucleus and/or functional site, so that the stability and kinetics of proteins are not altered. Such a set of mutations can drastically modify the sequence of the protein. However, within the time-scale τ_o, a family of homologs is born within which there is conservation of (already new) amino acids in the specific (important) sites of homologous proteins. Although there are alternations in the specific sites of the proteins at the time-scale τ, these sites are more preserved than the rest of the sequence. The proposed view of protein evolution is consistent with the observations of hierarchical organization of structurally similar proteins in families of homologs. Sets of families of homologs are organized, in turn, in super-families of analogs.

The time in our discussion is associated with the number of mutations that accumulate in the course of evolution. Because the rates may vary between families and even proteins, there is only a hypothetical relation of evolutionary time to physical time. Evolutionary time can be rigorously defined statistically as the number of mutations that occur in a fold family, averaged over all family members. The real time for one family may be different from that of another.

Energy Gap Model

We start with a random protein amino acid sequence and perform a Monte Carlo search for the mutation that energetically favors interactions in such a sequence. The Monte Carlo design algorithm is based on the minimization of the so-called *Z*-score, defined as:

$$Z = \frac{E_{\mathrm{NS}} - \langle E \rangle}{\sigma(E)} \tag{1}$$

which corresponds to the minimization of the energy gap between the native state, E_{NS}, of the selected sequence and the average energy, $\langle E \rangle$, of structurally unrelated conformations (decoys).[38–40] $\sigma(E)$ is the standard deviation of energies of all decoys.[41]

Since *Z*-score minimization is equivalent to maximizing the energy gap between misfolded or unfolded conformations and the native state of the protein,[32,33,38] such maximization results in stable and fast-folding proteins. The energy gap must be "significant", meaning that E_{NS} must deviate from $\langle E \rangle$ by many standard deviations σ: $E_{\mathrm{NS}} \ll \langle E \rangle - \sigma$. Many researchers have pointed out (e.g. see Shakhnovich[34]) that minimization of the *Z*-score corresponds to the stabilization of the protein in its native state.

The design proceeds as follows: (i) we select an amino acid σ_i at a random position $1 \leqslant i \leqslant N$; (ii) we substitute this amino acid by σ_i' with probability p:

$$p = \begin{cases} 1, & \text{if } \delta Z < 0 \\ \exp(-\delta Z/T_{\mathrm{des}}), & \text{if } \delta Z > 0 \end{cases} \tag{2}$$

where $\delta Z = Z(\sigma_i') - Z(\sigma)$ is the difference between the *Z*-scores of the mutated and the original proteins. We design each of $N_s = 100$ sequences by running the simulations for N_m Monte Carlo steps at some design temperature, T_{des}.

Computation of $\langle E \rangle$ and $\sigma(E)$ is straightforward:

$$\langle E \rangle = \frac{1}{2} \sum_{i \neq j} U(\sigma_i, \sigma_j) f_{ij} \tag{3}$$

and:

$$\begin{aligned} \sigma^2(E) &= \langle (E - \langle E \rangle)^2 \rangle \\ &= \frac{1}{2} \sum_{i \neq j} f_{ij}(1 - f_{ij}) U^2(\sigma_i, \sigma_j) + \mathcal{O}(f_{ij}^2) \end{aligned} \tag{4}$$

where f_{ij} is the frequency of a contact between

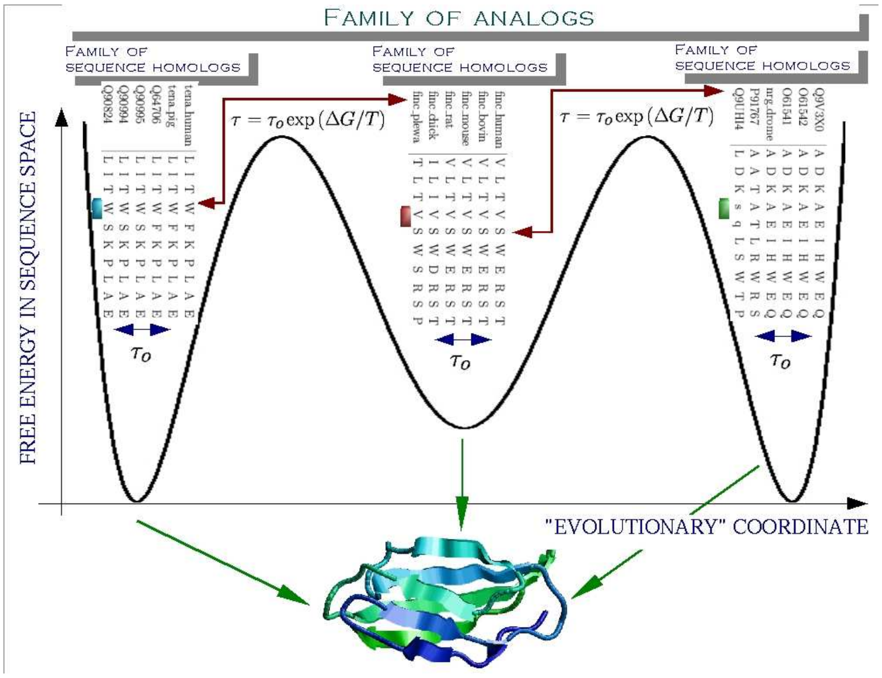

Figure 1. A schematic representation of the evolutionary processes that result in conservation patterns of amino acids. For a given family of folds, e.g. Ig in this diagram, there are several alternative minima (three) in the hypothetical free energy landscape in the sequence space as a function of the "evolutionary" reaction coordinate (e.g. time). Each of these minima is formed by mutations in protein sequences at time-scales, τ_o, that do not alter the protein's thermodynamically and/or kinetically important sites, forming families of homologous proteins. Transitions from one minimum to another occur at time-scales $\tau = \tau_0 \exp(\Delta G/T)$. At time-scale τ, mutations occur that would alter several amino acids at the important sites of the proteins in such a way that the protein properties are not compromised. At time-scale τ, the family of analogs is formed. In three minima we present three families of homologs (1TEN, 1FNF, and 1CFB) each comprised of six homologous proteins. We show ten positions in the aligned proteins: from 18 to 28. It can be observed that at position 4 (marked by blocks) in each of the families presented in the diagram amino acids are conserved within each family of homologs, but vary between these families. This position corresponds to position 21 in Ig fold alignment (to 1TEN) and is conserved (see Figure 12(a)).

monomers i and j in a set of decoys, i.e.:

$$f_{ij} = \langle \Delta_{ij} \rangle \tag{5}$$

We estimate frequencies of contacts by making two assumptions about the set of decoys: (1) the distribution, $P(\ell = |i-j|; i, j)$ of the contact distances, $\ell = |i-j|$, between various amino acids at the positions i and j is universal among globular proteins; and (2) the actual frequency of contacts between various amino acids, i and j, is only a function of the absolute value of the length of contacts, $|i-j|$, and is equal to the distribution of the contact lengths, i.e.:

$$f_{ij} = f_{|i-j|} = f(\ell) \tag{6}$$

The distribution $P(\ell)$ is then:

$$P(\ell) = \frac{f(\ell)}{\sum_{\ell=1}^{N} f(\ell)} \tag{7}$$

Both assumptions, (1) and (2), are motivated by the fact that the variety of protein structures known to date samples adequately the conformational space of proteins under study.

In order to estimate frequencies, f_{ij}, according to equation (7) we compute the distribution of contacts of length $\ell = |i-j|$ in the ensemble of approximately 10^3 representative globular proteins in the Protein Data Bank (PDB).[42,43] The distribution shown in Figure 2 is obtained using C^{β}-representation of proteins. The contacts are defined by equation (20).

The estimation of contact frequencies f_{ij} is one of the key ingredients to protein design. An alternative approach based on sampling of homopolymer conformations appears to be less efficient, so we omit it in the present study. Nevertheless, due to its importance and possible potential for other studies, we discuss this approach in the Appendix.

After we obtain N_s number of designed sequences, we compute the probability of an amino acid σ_k to be in the kth position, $P_Z(\sigma_k)$, as the frequency of occurrence of this amino acid:

$$P_Z(\sigma_k) = N(\sigma_k)/N_s \tag{8}$$

where $N(\sigma_k)$ is the total number of occurrences of an amino acid σ_k at the position k. Next, using equation (22) we compute the sequence entropy, $S_Z(k)$.

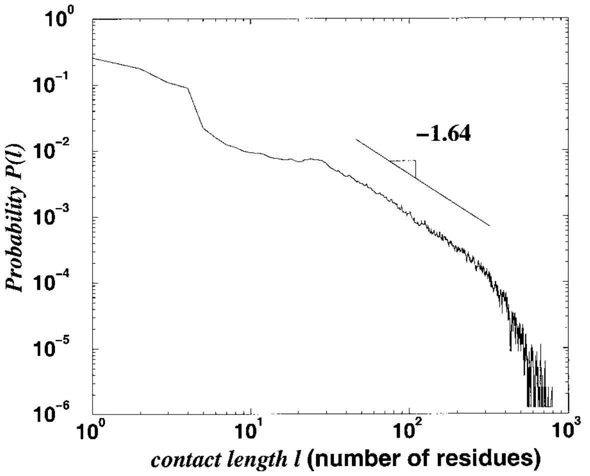

Figure 2. The double-logarithmic plot of the probability distribution $P(\ell)$ that any two amino acids, positioned ith and jth along the protein chain so that their contact length $|i-j|$ is equal to ℓ, are in physical contact, as a function of the contact length ℓ. Residues positioned ith and jth along the protein chain are defined to be in physical contact if their corresponding C^{β} (C^{α} in the case of Gly) are separated by less than 7.5 Å (equation (20)). $P(\ell)$ is computed for the ensemble of approximately 10^3 representative globular proteins in PDB.[42,43] The parallel line in the range of length $20 \leqslant \ell \leqslant 200$ indicates the power-law behavior of $P(\ell)$ in this region, $P(\ell) \sim \ell^{-1.64}$. The region $5 \leqslant \ell \leqslant 20$ is specific to proteins and has been discussed in detail by Berezovsky *et al.*[100]

Profile Solution

We develop a profile solution to the energy gap model that provides a rationale for conservatism patterns caused by selection for stability. Our solution is of equilibrium evolution that maintains stability and other properties achieved at an earlier, prebiotic stage. To this end we propose that stability selection accepts only those mutations that keep the energy of the native protein, E, below a certain threshold E_c necessary to maintain an energy gap.[8,29,44,45] The requirement to maintain an energy threshold for the viable sequences makes the equilibrium ensemble of sequences analogous to a microcanonical ensemble. In analogy with statistical mechanics, a more convenient and realistic description of the sequence ensemble is a canonical ensemble, whereby strict requirements on the energy of the native state is replaced by a "soft" evolutionary pressure that allows energy fluctuations from sequence to sequence but makes sequences with high energy in the native state unlikely. In the canonical ensemble of sequences, the probability of finding a particular sequence, $\{\sigma\}$, in the ensemble follows the Boltzmann distribution:[8,29,44,46]

$$P(\{\sigma\}) = \frac{\exp(-E\{\sigma\}/T)}{Z} \tag{9}$$

where T is the effective temperature of the canonical ensemble of sequences that serves as a measure of evolutionary pressure and $Z = \Sigma_{\{\sigma\}} \exp(-E\{\sigma\}/T)$ is the partition function taken in sequence space.

Next, we apply a profile approximation that replaces all multiparticle interactions between amino acids with interaction of each amino acid with an effective field Φ acting on this amino acid from the rest of the protein, so that each amino acid experiences the exact field of its neighbors. This approximation presents $P(\{\sigma\})$ in a multiplicative form as $\Pi_{k=1}^{N} p(\sigma_k)$ of probabilities to find an amino acid σ at position k.[47] $p(\sigma_k)$ also obeys Boltzmann statistics:

$$p(\sigma_k) = \frac{\exp(-\Phi(\sigma_k)/T)}{\sum_{\sigma} \exp(-\Phi(\sigma_k)/T)} \tag{10}$$

The profile potential $\Phi(\sigma_k)$ is the effective potential energy between amino acid σ_k and all amino acids interacting with it, i.e.:

$$\Phi(\sigma_k) = \sum_{i \neq k}^{N} U(\sigma_k, \sigma_i)\Delta_{ki} \tag{11}$$

The potential Φ is similar in spirit to the protein profile introduced by Bowie *et al.*[48] to identify protein sequences that fold into a specific 3D structure.

For each member, m, of the fold family (FSSP database[25]) presented in Figure 1, we compute the profile probability, $p_m(\sigma_k)$, using equation (10). This probability, $p_m(\sigma_k)$, for each fold family member corresponds to the frequency of amino acids, σ_k, at positions, k, for a given family of homologs. Then, we compute the average profile probability over all members of the fold family:

$$p_{\mathrm{P}}(\sigma_k) = \frac{1}{N_s} \sum_{m=1}^{N_s} p_m(\sigma_k) \tag{12}$$

This quantity corresponds to the $P_{\mathrm{acr}}(\sigma_k)$.[35] Equations (10)-(12), along with the properly selected energy function, U, make it possible to predict probabilities of all amino acid types and sequence entropy $S_{\mathrm{P}}(k)$ at each position k:

$$S_{\mathrm{P}}(k) = -\sum_{\sigma} p_{\mathrm{P}}(\sigma_k) \ln p_{\mathrm{P}}(\sigma_k) \tag{13}$$

from the native structure of a protein. The summation is taken over all possible values of σ.

If stability selection is a factor in the evolution of proteins and our model captures it, then we should observe a correlation between the predicted profile based sequence entropies, $S_{\mathrm{P}}(k)$, and actual sequence entropies $S_{\mathrm{acr}}(k)$ in real proteins. Thus, the question is: can we find such T, so that the predicted conservatism profile $S_{\mathrm{P}}(k)$ matches the real one $S_{\mathrm{acr}}(k)$?

By varying the values of the temperature T in the range $0.1 \leqslant T \leqslant 4.0$, we minimize the distance,

$D^2 \equiv \Sigma_{k=1}^{N} (S_P(k) - S_{acr}(k))^2$, between the predicted and observed conservatism profiles. We exclude from this sum such positions in structurally aligned sequences that have more than 50% gaps in the structural (FSSP) alignment. We denote by T_{sel} the temperature that minimizes D.

The proposed profile solution has a dual role. On one hand, it allows us to understand the selective temperature scale, T_{sel}, which is the measure of evolutionary optimization. On the other hand, the correlation coefficient between $S_P(k)$ and $S_{acr}(k)$ does not vary strongly in the range of T_{sel} from 0.19 to 0.34, thus, allowing one to use the effective temperature of $T_{sel} = 0.25$ to predict the actual conservatism profiles of proteins (see Table 1).

Results and Discussion

We study five folds: immunoglobulin fold (Ig), oligonucleotide-binding fold (OB), Rossman fold (R), α/β-plait (α/β-P), and TIM-barrel fold (TIM). The three-dimensional structures of the representative proteins of these five folds are shown in Figure 3: (a) tenascin (third fibronectin type III repeat), pdb:1TEN; (b) major cold shock protein 7.4 (Cspa (Cs 7.4)) of *Escherichia coli*, pdb:1MJC; (c) chemotactic protein CheY, pdb:3CHY; (d) acylphosphatase (common type) from bovine testis, pdb:2ACY; (e) endo-β-*N*-acetylglucosaminidase F1, pdb:2EBN. We compute the correlation coefficient[47] between values of $S_P(k)$, obtained at T_{sel}, and $S_{acr}(k)$ for all five folds. The results are summarized in Table 1. The plots of $S_P(k)$ and $S_{acr}(k)$ *versus* k, as well as their scatter plots, are shown in Figures 12-16, below.

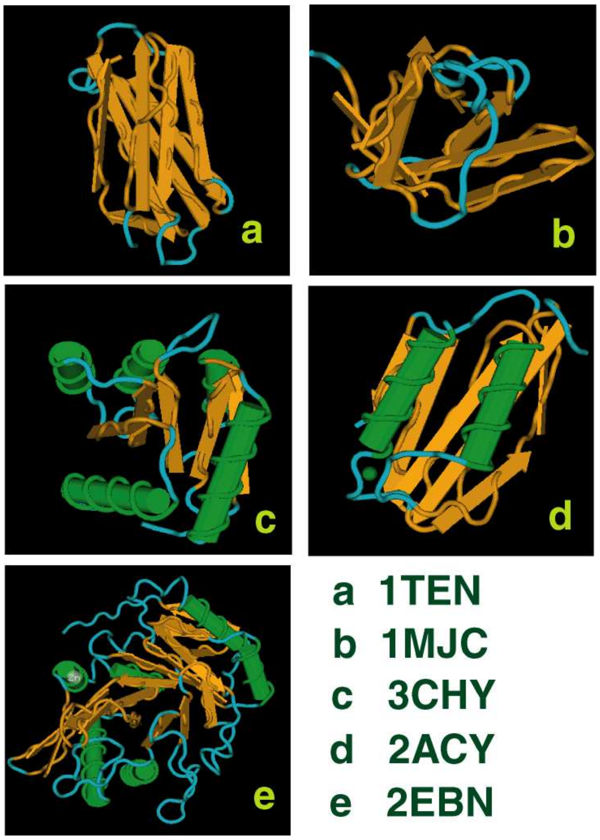

Figure 3. Three-dimensional structures of the representative proteins of the five folds under study (Ig, OB, R, α/β-P and TIM folds): (a) 1TEN protein; (b) 1MJC; (c) 3CHY; (d) 2ACY; and (e) 2EBN.

Energy gap model

We find that correlation between $S_Z(k)$ and $S_{acr}(k)$ strongly depends on the number of mutations, N_m, we introduce during design of a protein. This fact is in accord with our view (see Figure 1) of protein evolution. On a short timescale, $\tau_o \sim 10^2$ Monte Carlo steps, mutations rarely alter amino acids with specific important properties such as participation in stabilization of proteins and/or in the nucleation processes in folding kinetics of proteins. These mutations diversify the family, m, of homologs, $\mathcal{M}_h^m$. On a larger scale, $\tau \gg \tau_o$, correlated mutations[35–37] modify core and/or nucleus site(s) of the proteins without compromising their stability, folding rates or function(s). Thus, at the time-scale τ, evolution moves from one family of homologs to another, diversifying the underlying family of analogs, $\mathcal{M}_a, \cup_m \mathcal{M}_h^m \subseteq \mathcal{M}_a$. The ensemble of analogs is still much smaller

Table 1. The values of the correlation coefficient r for the linear regression of $S_P(k)$ and $R(k)$ *versus* S_{acr} for Ig, OB, R, α/β-P, and TIM folds and the corresponding optimal values of the temperature $T = T_{sel}$ for the $S_P(k)$ *versus* S_{acr} linear regression

		Representative protein		Correlation coefficient, r			
Fold	N_s	PDB code[40,41]	N	$R(k)$ *versus* $S_{acr}(k)$	$S_P(k)$ *versus* $S_{acr})k)$	T_{sel}	$S_P(k)$ *versus* $S_{acr}(k)$ ($T_{sel} = 0.25$)
Ig	51	1TEN	89	0.57	0.63	0.34	0.57
OB	18	1MJC	69	0.67	0.69	0.19	0.69
R	166	3CHY	128	0.74	0.71	0.25	0.71
α/β-P	29	2ACY	98	0.45	0.54	0.26	0.53
TIM	49	2EBN	285	0.54	0.50	0.23	0.50

The last column corresponds to the correlation coefficient for the studied folds at a fixed selective temperature $T = 0.25$. To obtain the rates of mutations, $R(k)$, we perform Z-score design of the sequences for $t_d = 10^8$ Monte Carlo steps at $T_{des} = 0.25$.

than the ensemble, $\mathcal{M}_o$, of all possible sequences ($\mathcal{M}_a \subseteq \mathcal{M}_o$), which is of the size 6^N (in a six-letter alphabet), for $N = 100$ residue protein this number is of the order of 10^{80}. These results are in agreement with the theoretical predictions[5,32,34,50–52] that there is a large number (of the order $e^{1.9N}$)[34], of fast folding sequences with a given native structure and pronounced stability gaps $\Delta = E_{\mathrm{NS}} - \langle E \rangle$.

It is important that for the small number of mutations we find correlation between entropies of the designed sequence, $S_Z(k)$, and the empirically observed one, $S_{\mathrm{acr}}(k)$. This correlation depends on the input random number, indicating that the selected sequences constitute a family of homologs, $\mathcal{M}_h^m$, that is closer or more distant to an original sequence family of homologs, $\mathcal{M}_h^0$ (both $\mathcal{M}_h^m$ and $\mathcal{M}_h^0$ belong to a given family of analogs, $\mathcal{M}_a$). Here, we present the results for the selected ensembles of the designed sequences, $\mathcal{M}_h^m$, after being optimized during N_{m} mutations. More important than the correlation between $S_Z(k)$ and $S_{\mathrm{acr}}(k)$, we find that the profiles of $S_Z(k)$ and $S_{\mathrm{acr}}(k)$ are in visible concert with each other.

The temperature-dependence of the Z-score exhibits a sharp transition at $T = T_{\mathrm{c}} \approx 0.25$ (Figure 4) for all studied proteins. Above T_{c}, protein design results in unstable sequences, while at temperatures much lower than T_{c} many of the residues "freeze" in their original states. Thus, we select T_{c} as our design temperature.

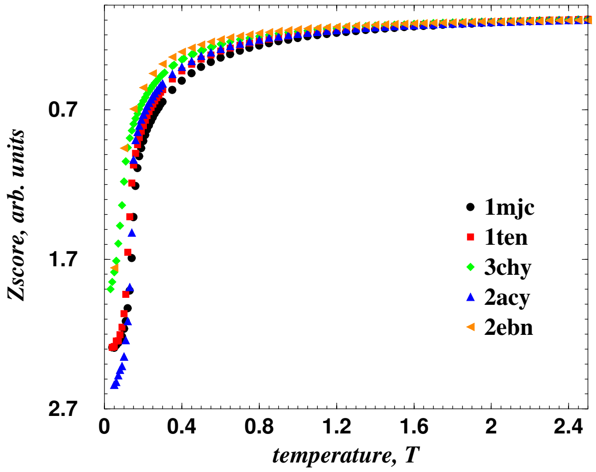

Figure 4. The temperature-dependence of the Z-score (equation (1)) for five representative proteins (1MJC, 1TEN, 3CHY, 2ACY, and 2EBN) of the five folds under study. For each fold and for a given temperature T, we compute the Z-score after performing sequence design runs during 10^5 Monte Carlo steps. We average the Z-score over 100 sequence design runs. Due to the normalization of contacts' frequencies extracted from the PDB database (equation (7)), the scales of the computed Z-scores are different from the correctly normalized ones, defined by equation (1). There is a sharp transition of $Z(T)$ at $T = T_{\mathrm{c}} \approx 0.25$ for all studied proteins. Above T_{c}, protein design results in unstable sequences, while at temperatures much lower than T_{c} many of the residues "freeze" in their original states.

Degree of divergence of sequences

To assess the degree of similarity or divergence of sequences in the course of Z-score design at various time-scales, we compute the distribution of hamming distances at these time-scales. The hamming distance,[53] also known as the p-distance,[54] $Hd(\{\sigma\}^{(1)}, \{\sigma\}^{(2)})$, between two sequences, $\{\sigma\}^{(1)}$ and $\{\sigma\}^{(2)}$, is defined as the number of distinct amino acids at equal positions in these two sequences divided by the length of the sequences, N:

$$Hd(\{\sigma\}^{(1)}, \{\sigma\}^{(2)}) = \frac{1}{N}\sum_{i=1}^{N}[1 - \delta(\sigma_k^{(1)} - \sigma_k^{(2)})] \qquad (14)$$

Hamming distance has a simple interpretation - it is the degree of divergence between two sequences: when Hd is equal to 1, the sequences have no amino acids in common, when Hd is equal to 0, the sequences are exactly the same.

We compute the distribution of hamming distances between all designed sequences for two design times (a) $t_{\mathrm{d}} = 10^3 \gg \tau_{\mathrm{o}}$ (Figure 5(a)) and (b) $t_{\mathrm{d}} = 10^2 \sim \tau_{\mathrm{o}}$ Monte Carlo steps (Figure 5(b)), where τ_{o} is a characteristic time-scale. We use 1MJC family of homologs (OB fold) as an example throughout this subsection. We also perform similar analysis with other folds and the results are qualitatively the same (not shown).

In the computation of the distribution in case (b) we omit all sequences with sequence similarity less than $ID = 55\,\%$ to mimic sequence collection in the HSSP database. This threshold sequence similarity, ID, is chosen so that the hamming distance distribution derived from the actual sequences in the HSSP database (Figure 5(d)) is similar to ours. Given that we use a six-letter alphabet, the correspondence between ID used in HSSP and our ID is not well defined. Because in (b) we select only sequences with minimal threshold similarity, ID, there are no events with $Hd > 0.55$ in Figure 5(b). In addition, the events with low $Hd \to 0$ are over-represented in our simulations, since we do not account for additional pressure that exists in real protein sequences due to function or kinetics. Therefore, the distribution in our simulations (Figure 5(b)) has a more pronounced tail $Hd \to 0$ than that in real proteins (Figure 5(d)).

At the long time-scales (Figure 5(a)) we find that most of the sequences are divergent from each other with average $\langle Hd \rangle \approx 0.7$. We observe the same result by computing the distribution of the hamming distances between all analogs belonging to the OB fold family present in the FSSP database (Figure 5(c)). The only difference between simulated (Figure 5(a)) and observed (Figure 5(c)) distributions of hamming distances is the tail present in the simulated distribution corresponding to the sequences with a significant degree of similarity. This tail is due to the fact that we compare all sequences with all sequences, thus, effectively

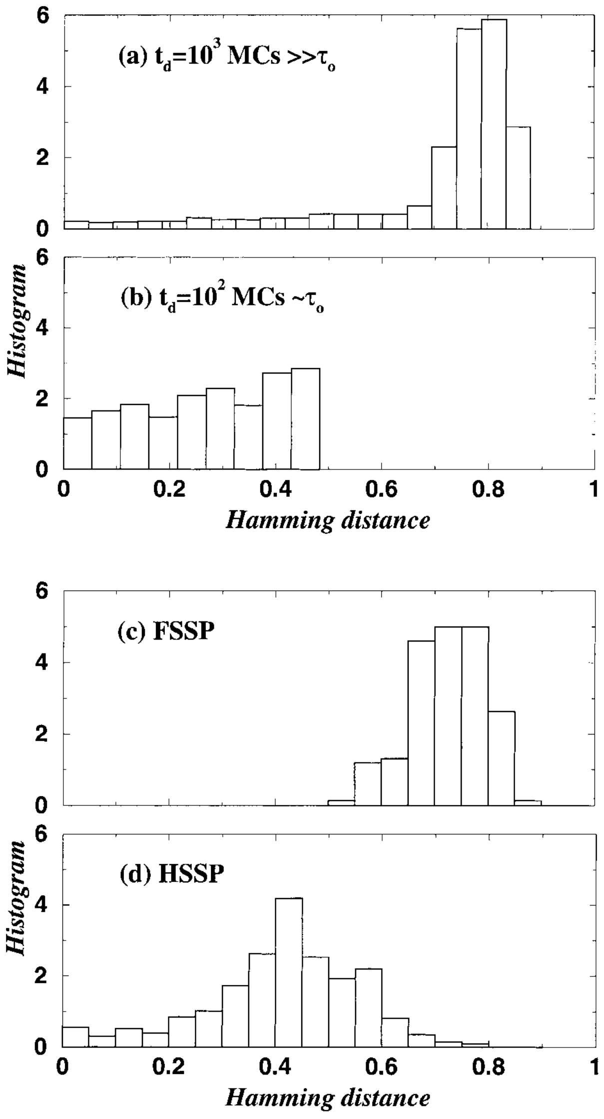

Figure 5. The histograms of hamming distances (equation (14)) between all designed sequences for the OB fold family for two design times (a) $t_d = 10^3 \gg \tau_o$ and (b) $t_d = 10^2 \sim \tau_o$ Monte Carlo steps. The histograms of hamming distances for 1MJC family of actual protein sequences: (c) analogs taken from the FSSP database; and (d) homologs taken from the HSSP database. In the computation of the histograms in case (b) we omit all sequences with sequence similarity less than $ID = 55\,\%$ to mimic sequence collection in the HSSP database. The threshold sequence similarity, *ID*, is chosen so that the hamming distance histogram is derived from the actual sequences in HSSP database (d). All histograms are normalized to unit area.

including similar sequences in our histogram. In the FSSP database, on the other hand, only distant sequences are present so that the tail corresponding to the close sequences in Figure 5(c) is absent.

The distributions of hamming distances in protein families are in qualitative agreement with those observed in simulations and with our picture of hierarchical protein evolution. At short time-scales sequences are not strongly separated from each other forming families of homologs, while at long time-scales, a family of analogs is formed, comprised of strongly separated sequences but structurally similar proteins.

Determination of the family formation time-scale

To quantify our observation of evolutionary time-scales separation, we compute the relaxation times of the correlation function (at each protein position, *k*) in the course of *Z*-score design defined as:

$$C_k(\tau) = \frac{1}{t_d N_s}\sum_{\alpha=1}^{N_s}\int_0^{t_d}\chi_k^{(\alpha)}(t,\tau)dt = \langle\langle\chi_k(\tau)\rangle\rangle_{t_d,N_s} \qquad (15)$$

where $\langle\langle\cdots\rangle\rangle_{t_d,\,N_s}$ denotes average over simulation design time, t_d, and the number, N_s, of initial sequences. $\chi_k^{(\alpha)}(t,\tau)$ is a Boolean indicator of whether an amino acid $\sigma_k(t+\tau)$ at position *k* at time $t+\tau$ is the same as the amino acid at time $\sigma_k(t)$ at the same position at time *t*:

$$\chi_k^{(\alpha)}(t,\tau) = \begin{cases} 1, & \sigma_k(t) = \sigma_k(t+\tau) \\ 0, & \sigma_k(t) \neq \sigma_k(t+\tau) \end{cases} \qquad (16)$$

$C_k(\tau)$ measures the probability that a mutation does not occur at the position *k* in time τ. This function for most equilibrium systems decays exponentially:

$$C(\tau) \sim \exp(-\tau/\tau_o) \qquad (17)$$

where τ_o is the relaxation time that is the average mutation time between subsequent mutations. The quantity inverse to the relaxation time $(1/\tau_o)$ is proportional to the average substitution rate of a site.[53]

We also find that the correlation function computed for $N_s = 10^3$ and for $t_d = 10^3$ decays exponentially (see Figure 6) and relaxation times τ_o depend strongly on the positions of the amino acids under consideration. For example, the relaxation of the correlation functions for positions 1 (Ser in 1MJC) and 31 (Val) in 1MJC design vary by almost a factor of 2: τ_o(Ser1) = 143 and τ_o(Val31) = 387 Monte Carlo steps. The fact that τ_o(Ser1) is more than twice as large as τ_o(Val31) indicates that Ser1 is likely to mutate more than twice in the time-span of a single Val31 mutation.

In addition, the distribution of the relaxation times (see Figure 5) exhibits a pronounced peak at $\tau_o = 170$ Monte Carlo steps, indicating that for most protein positions relaxation occurs with this typical relaxation time. The relaxation times found from the correlation function analysis are in agreement with our observations. The long non-Gaussian tail in the histogram of the relaxation times also suggests the presence of the conserved positions. In fact, this tail, composed of the conserved positions, strongly deviates from the rest of the

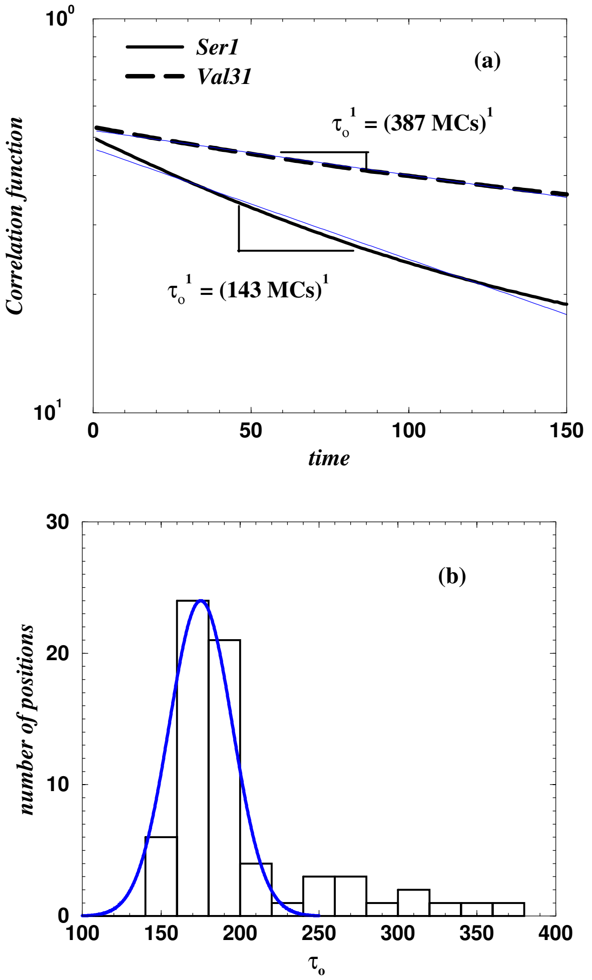

Figure 6. (a) Plot of the correlation functions *versus* time for two positions in 1MJC, Ser1 and Val31, obtained in the course of Z-score design of 10^3 sequences for 10^3 Monte Carlo steps. In semilogarithmic scale $C_{k=1}(\tau)$ and $C_{k=31}(\tau)$ are straight lines with slopes $\tau_o = 143$ and $\tau_o = 387$ Monte Carlo steps, correspondingly. (b) The histogram of the relaxation times τ_o for all positions in 1MJC obtained in the course of Z-score design of 10^3 sequences for 10^3 Monte Carlo steps. The histogram is well fit by a Gaussian function in the region $100 < \tau_o < 250$ (continuous line). (The regression coefficient for the Gaussian fit is $r \approx 0.98$.) The long tail that strongly deviates from the Gaussian distribution (over seven standard deviations) indicates the presence of the conserved positions in the course of design.

distribution, which is well approximated by a Gaussian distribution.

Rates of amino acid substitutions and conservatism

A number of authors suggested[56–58] (and see Dokhdyan *et al.*†) that study of the conserved amino acids in families of structurally similar proteins can shed light on the functionally, kinetically and thermodynamically important amino acids in proteins. The basic belief behind the majority of such studies is that evolution optimizes, to a certain extent, the properties of proteins so that they become more stable and have better folding and functional properties. Here, we use the "optimization" hypothesis of molecular evolution to understand the universe of protein sequences by implication of molecular evolution. The link between conserved amino acids and their role in proteins has been widely studied.[58,59–63]

A recent study by Mirny & Shakhnovich[35] identified the presence of universally conserved amino acids across the families of proteins sharing the same fold. These conserved residues have been linked to protein stability, kinetic properties or function. Various experiments[59,64–73] have identified some of the conserved residues to have predicted specific roles.

Direct evidence of the relationship between conservatism and the physical properties of amino acids can be accessed by calculating the rates of amino acid substitutions in the course of the Z-score design. By comparing mutational rates at various positions of the proteins, we attempt to reconstruct the conservatism of these positions across the family of analogous proteins. Starting with the sequence of a representative protein of a given fold we perform Z-score design for $t_d = 10^8$ Monte Carlo steps. The substitution rates are defined as:

$$R(k) \equiv \frac{N_m(k)}{\hat{t}_d} = \frac{N}{t_d}\sum_{t=1}^{t_d}[1 - \delta(\sigma_k(t) - \sigma_k(t-1))] \quad (18)$$

where $N_m(k)$ is the number of mutations that occurred at the position k, $\delta(x)$ is a Kronecker function, equal to 1 if $x = 0$ and 0 otherwise, $\sigma_k(t)$ is an amino acid σ at the position k at time t, and $\hat{t}_d = t_d/N$ is the average number of attempted mutations per position in a protein. Thus, $R(k)$ from equation (18) is inversely proportional to the average time between subsequent substitutions of amino acids at the position k; the lower the $R(k)$ value the longer the amino acid at the position k remains unchanged and, therefore, the more conservative is this position in the course of design.

We find that the rates of substitutions, $R(k)$, correlate with the conservatism patterns, S_{acr} (see Figures 7-11). Since there is no obvious relation between $R(k)$ and S_{acr}, and, moreover, there is no reason to assume linear relation between these quantities, the linear regression has only an illustrative meaning of the correlations observed between $R(k)$ and S_{acr} (see Figures 7-1(b) and Table 1). Despite the likely lack of linear relation between the rates and the entropy, the correlation observed based on the assumption of linear dependence between $R(k)$ and S_{acr} is feasible.

The computation of mutational rates, $R(k)$, does not involve the tuning of any parameters. We can

---

† http://arxiv.org/abs/cond-mat/0007084

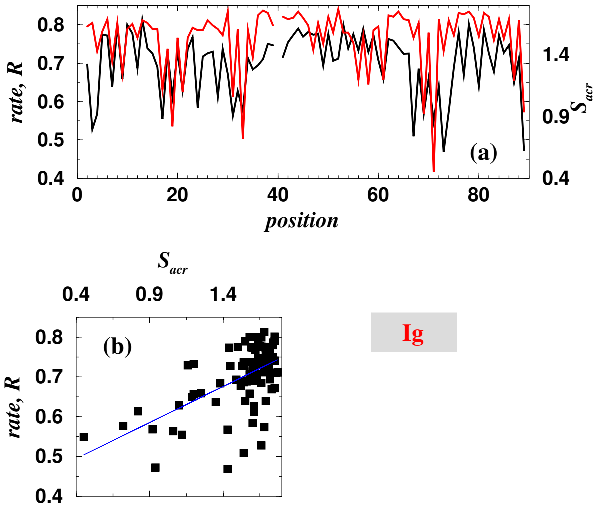

Figure 7. (a) The values $R(k)$ (black line) and $S_{\mathrm{acr}}(k)$ (red line) for all positions, k, for the Ig-fold. The lower the values of $R(k)$ the more conservative amino acids are at these positions. (b) The scatter plot of $R(k)$ *versus* observed $S_{\mathrm{acr}}(k)$. The linear regression correlation coefficients are shown in Table 1. The blue line is the linear regression approximation. In both (a) and (b) rates are multiplied by the length of the representative protein.

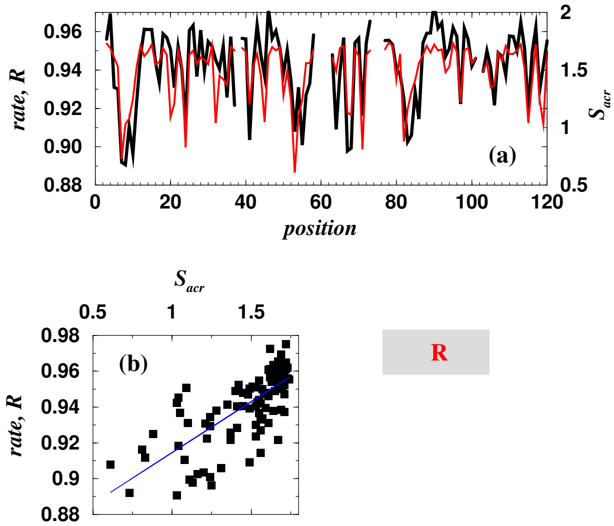

Figure 9. (a)-(c) The same as Figure 7 but for the R-fold.

choose any non-zero temperature, given that the total number of Monte Carlo steps, t_{d}, in the course of design is large enough to obtain statistically significant values of $R(k)$. We also find that at $T_{\mathrm{des}} = 0.25$ the data for $R(k)$ are identical after 10^7 Monte Carlo steps to that after 10^8 Monte Carlo steps, so the values of $R(k)$ are statistically significant.

Interestingly, the fastest rates are at most twice as fast as the slowest rates. Such variability of rates might be due to the variability in physical properties of amino acids. It has been shown[74] that there are only two principal eigenvalues of Miyazawa-Jernigan energy matrix[75,76] and the remaining eigenvalues are close to each other. Such a "degeneracy" in eigenvalues accounts for the similarities in physical properties of amino acids. In addition, the limitations in the derivation of the knowledge-based potential of amino acid interactions may be partially responsible for the observed range of rate variability.

Another possible reason for such a limited range of rate variability is the absence of the side-chains in our model. The side-chains are an additional fac-

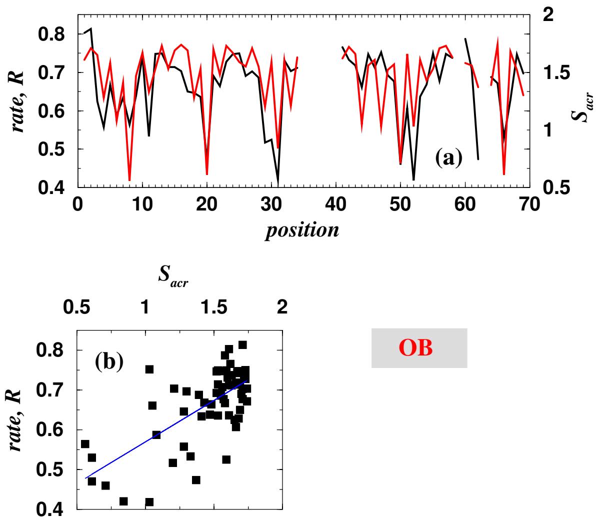

Figure 8. (a)-(c) The same as Figure 7 but for the OB-fold.

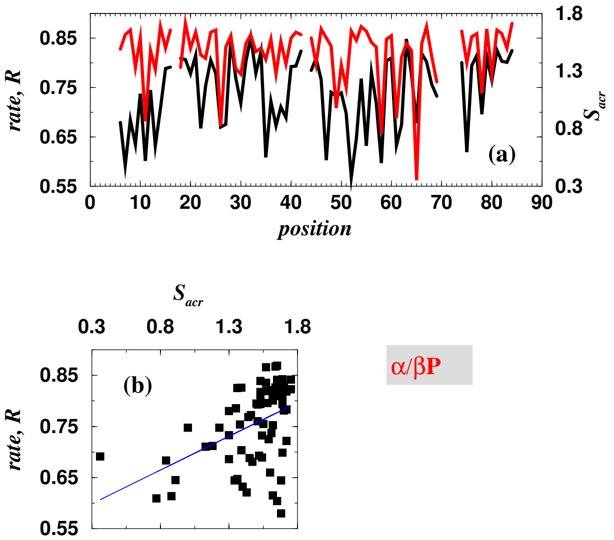

Figure 10. (a)-(c) The same as Figure 7 but for the α/β-P-fold.

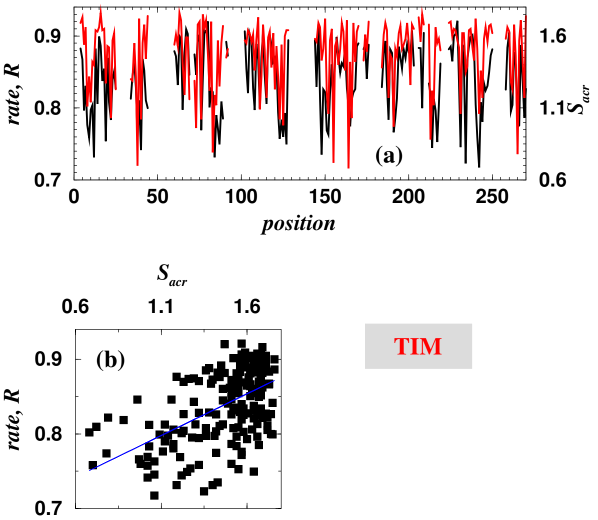

Figure 11. (a)-(c) The same as Figure 7 but for the TIM-fold.

tor that slow the rates because of the frustrations caused by the multiple side-chain conformations.[77] Despite the limitations of our model, the correlation between $R(k)$ and S_{acr} is significant, which indicates that the model does qualitatively capture the evolutionary selection of proteins.

Profile solution

The correlation between $S_{\mathrm{P}}(k)$ and $S_{\mathrm{acr}}(k)$ is remarkable for all five folds and indicates that our profile approximation is able to select conserved amino acids in protein fold families and properly describe the formation of families on the short time-scales (Table 1). It is fully expected that the correlation coefficient is smaller than 1. The reason for this is that computation of $S_{\mathrm{P}}(k)$ takes into account evolutionary selection for stability only and it does not take into account possible additional pressure to optimize kinetic or functional properties.

The additional evolutionary pressure due to the kinetic or functional importance of amino acids results in pronounced deviations of S_{P} from S_{acr} for a few amino acids that may be kinetically or functionally important. A number of amino acids whose conservatism is much greater than predicted by our model form a group of "outliers" from otherwise very close correspondence between S_{P} and S_{acr}. To demonstrate that some of those amino acids are important for folding kinetics and, as such, they can be under additional evolutionary pressure, we color data points on the S_{P} *versus* S_{acr} scatter plot according to the range of ϕ-values[78] that the corresponding amino acids fall into. The thermodynamic and kinetic roles of individual amino acids were studied extensively (i) by Hamill *et al.*[80] for the TNfn3 (1TEN) protein, (ii) by López-Hernández & Serrano[66] for the chemotactic protein (CheY, pdb:3CHY), and (iii) by Chiti *et al.*[78] for muscle acylphosphatase (AcP, pdb:2ACY).

We use the ϕ-values for individual amino acids obtained by López-Hernández & Serrano and by Hamill *et al.*[66,79] We observe that: (i) for TNfn3 protein most of the points on Figure 12(b) that belong to the outlier group have ϕ-values ranging from 0.2 to 1; (ii) for CheY protein most of the points (for which ϕ-values are known) on Figure 14(b) that belong to the outlier have ϕ-values ranging from 0.3 to 1; and (iii) for AcP protein, one nucleic amino acid, Tyr11, is strongly conserved, more than predicted by the profile solution, while the second amino acid, Pro54, belonging to the nucleus[80] does not appear to be conserved. The third nucleus amino acid, Phe94, in AcP protein is excluded from our analysis due to the lack of data at position 94. The discrepancy of the Pro54 conservatism and its kinetic role may be attributed to the poor statistical significance of $S_{\mathrm{acr}}(k)$ calculation at this position. Figures 12(b), 14(b), and 15(b) demonstrate that the presence of additional evolutionary pressure due to the kinetic importance of amino acids results in stronger conservatism of specific positions than predicted by profile solution.

It has been conjectured (e.g. see Rost[11]) that on average only 3-4% of residues are "anchor residues", i.e. those that are more significantly conserved than the rest of the residues. In fact, this observation is supported by the $S_{\mathrm{acr}}(k)$ profile of the sequences and their profile estimates $S_{\mathrm{P}}(k)$ (see Figures 12(a)-16(a)). These 3-4% of anchor residues are the principal "gates" to the structure/kinetics of a given family of proteins. For example, it has been shown[77] that the number of residues that belong to the nucleus of a model protein is about 5%; we expect the same low percentage of residues that determine the kinetics of real proteins. The number of key residues that form a functional site is also a small fraction of the total number of residues in proteins.

In order to demonstrate the statistical significance of the outliers' kinetic importance, we show that the number of sites with high values of ϕ found among the outliers is larger than that expected if such sites were randomly distributed across all values of S_{P}. For tenascin, the total number of residues is $N_{\mathrm{tot}} = 89$, the number of sites with $\phi > 0.2$ is $N_{\mathrm{tot}}(\phi > 0.2) = 17$, the number of outliers is $N_{\mathrm{out}} = 13$, and the expected number of sites with $\phi > 0.2$ among outliers is $N_{\mathrm{out}}^{\mathrm{exp}}(\phi > 0.2) = N_{\mathrm{tot}}(\phi > 0.2) \quad N_{\mathrm{out}}/N_{\mathrm{tot}} \approx 2.5$. The observed number of sites with $\phi > 0.2$ among outliers is $N_{\mathrm{out}}(\phi > 0.2) = 8$, which is over three times more than expected. A similar estimate for $\phi > 0.5$ gives $N_{\mathrm{out}}^{\mathrm{exp}}(\phi > 0.5) = 0.75$ and $N_{\mathrm{out}}(\phi > 0.5) = 2$, which is nearly three times more than expected. For CheY, the total number of residues is $N_{\mathrm{tot}} = 128$, the number of sites with $\phi > 0.3$ is $N_{\mathrm{tot}}(\phi > 0.3) = 11$, the number of outliers: $N_{\mathrm{out}} = 22$, and the expected number of sites with

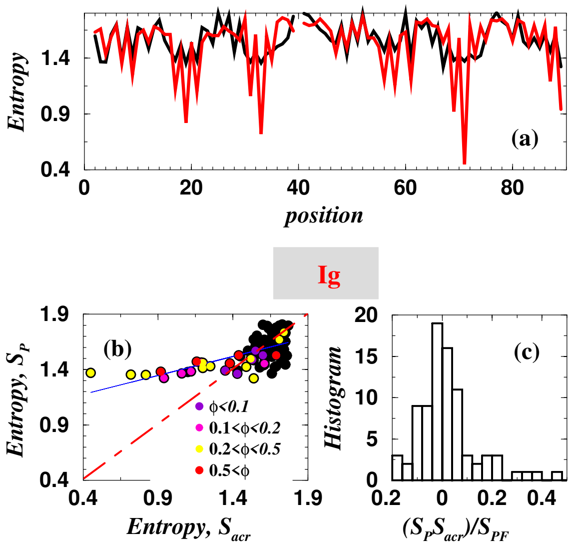

Figure 12. (a) The values $S_P(k)$ (black line) and $S_{\mathrm{acr}}(k)$ (red line) for all positions, k, for the Ig-fold. The lower the values of $S_P(k)$, the more conservative amino acids are at these positions. (b) The scatter plot of predicted $S_P(k)$ *versus* observed $S_{\mathrm{acr}}(k)$. The linear regression correlation coefficients are shown in Table 1. The blue line is the linear regression that has a slope different from 1 (red line), corresponding to the $S_P(k) = S_{\mathrm{acr}}(k)$ relation. (c) The histogram of the relative differences between $S_P(k)$ and $S_{\mathrm{acr}}(k)$. In (b) we assign colors to data points corresponding to amino acids with the specific range of ϕ-values:[79] red, if $0.5 < \phi < 1$; yellow, if $0.2 < \phi < 0.5$; magenta, if $0.1 < \phi < 0.2$; violet if $\phi < 0.1$; and black if ϕ-values are not determined.

$\phi > 0.3$ among outliers is $N_{\mathrm{out}}^{\mathrm{exp}}(\phi > 0.3) = N_{\mathrm{tot}}(\phi > 0.3)\ N_{\mathrm{out}}/N_{\mathrm{tot}} \approx 1.9$. The observed number of sites with $\phi > 0.3$ among outliers is $N_{\mathrm{out}}(\phi > 0.3) = 4$, which is over twice that expected. These crude estimates demonstrate that outliers have, in fact, a higher than expected number of residues with pronounced kinetic role, hinting towards an additional evolutionary pressure exerted on kinetically important amino acids.

Convergent or divergent evolution?

It has been a long-standing question[11,13,82–84] whether the presently known proteins have evolved from a smaller family of prebiotic proteins (“divergent” evolution scenario) or whether they evolved from ancestors with distant homology and due to thermodynamic, kinetic, and functional pressure exerted by evolution they converged to structurally similar proteins (“convergent” evolution scenario). The model of evolution proposed in this work does not rule out any of these scenarios. However, the similarity in distribution of hamming distances in the family of homologous proteins produced by our model to that taken from nature is striking (Figure 5), serving as a hint in favor of divergent evolution.

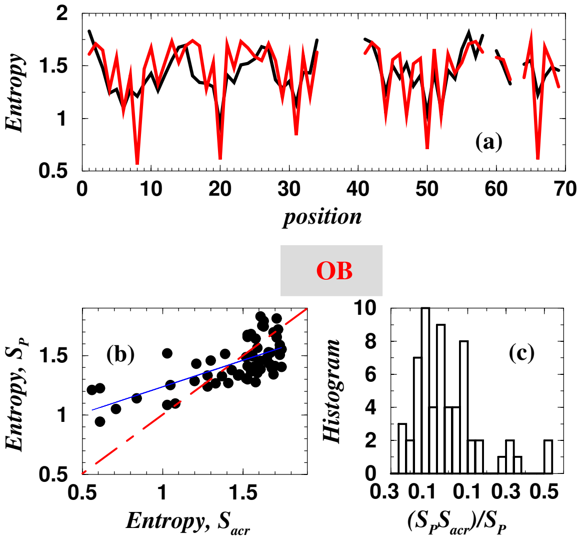

Figure 13. (a)-(c) The same as Figure 12 but for the OB-fold.

Proteins’ functions are related to the their structures in several cases,[14,17] so there is an evolutionary pressure to preserve structures. If a protein were to change the structure in the course of evolution, it would affect its functionality (there are, of course, possible exceptions). There is a growing evidence[78,85,86] that a small subset of all amino

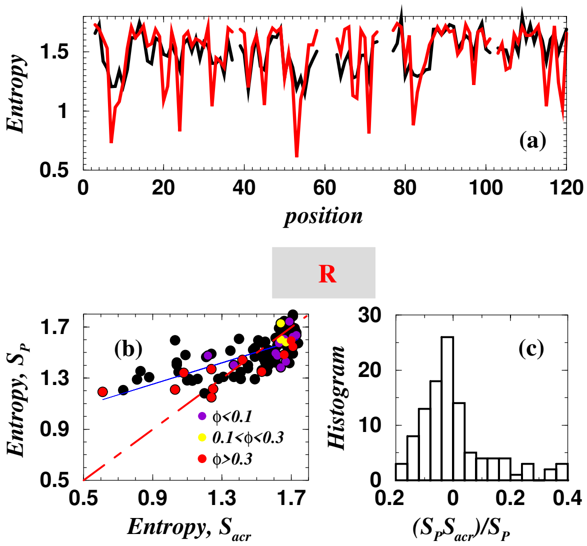

Figure 14. (a)-(c) The same as Figure 12 but for the R-fold. In (b) we assign colors to data points corresponding to amino acids with the specific range of ϕ-values:[64] red, if $0.3 < \phi < 1$; yellow, if $0.1 < \phi < 0.3$; violet if $\phi < 0.1$, and black if ϕ-values are not determined.

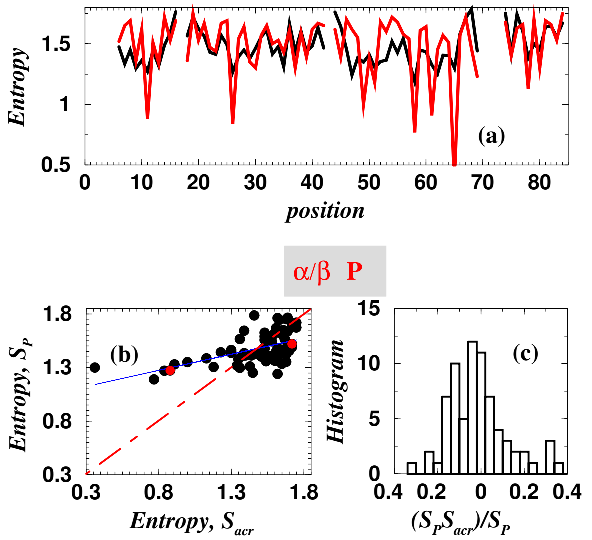

Figure 15. (a)-(c) The same as Figure 12 but for the α/β-P-fold. In (b) we color red (two out three) nucleic amino acids, Tyr11 and Pro54.[80]

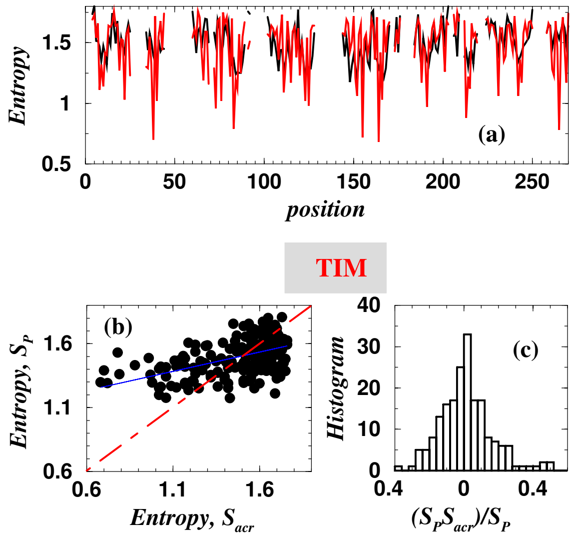

Figure 16. (a)-(c) The same as Figure 12 but for the TIM-fold.

acids of proteins is responsible for protein thermodynamic stability of the native structure and the folding kinetics. Thus, an important argument favoring divergent evolution is that nature has to preserve only a small subset of amino acids that contribute the most to the protein stability and folding kinetics. However, Murzin[14] proposed a way for functionless protein to survive by fusion with another functional protein and evolving already as a unit to a multi-functional protein. One of the most prominent examples are the DNA polymerases that are composed of similar domains with different sequence composition.[87–89]

If we set aside multi-domain proteins, the fact that there is a limited number of folds (<1000 according to Chothia[4] or <7920 according to Orengo[6]) has been extensively used to favor convergent evolution.[5,8,10,12,14] Li *et al.*[10] used the designability principle to show from full enumeration of lattice protein models that the number of members of a fold family depends on the stability gap, Δ. This dependence means that many unrelated sequences search in the course of evolution for the stable conformations and as soon as they reach a basin of a certain fold with large enough energy gap they stay within that basin. The scenario proposed[5,8,10] also explains why various fold families are unequally populated,[6] the number of family members depends on the energy gap. The more pronounced the energy gap is, the more mutations such a fold can tolerate. Buchler & Goldstein[16] argued that the energy gap depends on the number of non-local contacts of a given fold.

There are several questions about that scenario. First, it is not clear if nature exploits all possible folds:[6] even though there are only 1000 folds,[4] it is possible that nature simply does not need more of them. Second, Chothia & Gerstein[12] argued that the restriction on the divergence of proteins from one another does not come from a stability requirement (which is, of course, important) but from the separation of the mutated residue from the active site. Thus, the extent of sequence divergence is inversely proportional to that of protein function(s). The experiments by Gassner *et al.*[90] and Axe *et al.*[91] support the arguments of Chothia & Gerstein.[12] In these experiments substitution of the several amino acids in the hydrophobic core of the T4 lysozyme conserved, to a certain extent, the function and the structure of the protein. Third, we note that there is a limited amount of types of chemical elements that are part of the ligand structures that are bound by the active site. Thus, we expect that there are groups of evolutionarily unrelated proteins with similar binding sites and structure. In fact, there are examples of proteins sharing the same site, also called a "super-site". For example, both transforming protein p21H-RAS-1 fragment (pdb:1CTQA) and chemotactic protein 3CHY have similar binding sites, the root-mean-square deviation of one protein from another is 3.2 Å while there are only 13 identical residues (i.e. $ID < 10\,\%$).[92,93] Interestingly, the active site of 1CTQA is centered around Mg^{2+}, while the active site of 3CHY is built by residues only (Asp12, Asp13, Met17, and Asp57).

It is possible that evolution follows several paths at the same time and the question of whether evolution is divergent or convergent is just ill-posed. To understand which pathways are favored by nature we need more evidence to make statistically valid conclusions.

Roles of stability, folding kinetics and function

There are three essential ingredients of proteins that nature exerts pressure on in the course of evolution, their thermodynamic stability, folding kinetics, and function. These ingredients are not mutually independent. For example, disruption of the folding kinetics or function of a protein may discard this protein from the cell's life. Out of these three ingredients, the evolutionary selection of proteins by their function may result in the strongest conservatism of amino acids. However, in many proteins this pressure may be localized to a small number of amino acids (e.g. binding site) that are responsible for its function. Therefore, there may be only few functionally important amino acids that are under functional evolutionary pressure. Of course there are exceptions, such as histones, where function is the structure of these proteins and, thus, most amino acids are conserved in the course of evolution.[17] In addition, the loss of a protein function may not result in elimination of this protein from the cell's life, because there are mechanisms by which functionless proteins can be fused with another functional protein[14] and evolve as a unit of a multi-functional protein.

The evolutionary pressure to preserve the folding kinetics of proteins may be crucial and necessary for proteins to survive. It has been pointed out[57,81,94] that there are few amino acids, the protein folding nucleus, that are responsible for the folding kinetics of the proteins. Thus, to preserve folding properties, nature may exert evolutionary pressure on the few amino acids that are part of the folding nucleus.

The evolutionary selection of proteins by their thermodynamic stability may be most robust, because if the stability of a protein is lost, the function and folding kinetics of this protein may become irrelevant. It seems that there is a larger amount of amino acids that are important for protein stabilization than for the folding kinetics.[81,95,96] Thus, evolutionary pressure to preserve thermodynamic stability of proteins may be less specific and affect more amino acids.

It should be noted that our model, while not accounting explicitly for protein function does so implicitly. Conservation of function (not considered explicitly in our model) may be a primary reason for conservation of stable structure, which is the major premise of our analysis of superfamily formation. There are several examples of protein superfamilies (according to SCOP classification) that have similar or identical function, similar structures but vastly different sequences whose similarity is undetectable by most sensitive sequence alignment methods such as PSI-BLAST (e.g. Zn-dependent exopeptidases, glutathione synthetase ATP-binding domain-like, serin/threonine and tyrosine kinases). In terms of our model this situation fits the hierarchy of time-scales whereby the time-scale for functional divergence τ_F is much greater than the longest time-scale τ_o on which a superfamily is born. This reflects the well-known fact of considerable plasticity of sequences preserving function (and structure).

As we already pointed out, conservation of stable structure has the most wide-ranging impact on protein sequences and the main goal of our analysis here is to understand the implications of preservation of protein structure and stability for sequence evolution.

Homology in proteins: a rigorous way out of a terminological muddle

In 1987, 11 leading evolutionary biologists[27] made a statement asking the scientific community for the appropriate usage of the term "homology". Two proteins are said to be homologous if they posses a common evolutionary origin (e.g. Fitch[97]). Because many proteins that have high sequence similarity are homologous, this term has been used loosely in the discussion of any proteins with high sequence identity. Proteins that have no common ancestor, but posses structural similarity, are called analogs.

If the sequence identity of two structurally similar proteins is high ($ID > 25\text{-}30\,\%$), there is a high probability that these proteins share a common ancestor, and thus, statistically, one would rarely be mistaken when calling these two proteins homologs. If the sequence similarity of two structurally similar proteins is low ($ID < 25\,\%$), it is difficult to establish whether these proteins are homologs or analogs. In fact, despite clever efforts,[97] it is still questionable whether there is a unique solution to the problem of determining whether two proteins with low sequence identity are homologs or analogs, i.e. whether they evolved by divergent or convergent evolution.

Two proteins are likely to be homologs that diverged from the same root if they still carry the same function (i.e. if the evolutionary time elapsed from their common divergence point is smaller than functional relaxation time τ_F). However, if two structurally similar proteins with low sequence identity have significantly different functions, then there is little information with which to identify them as homologs or analogs. These two proteins might be homologs, although one of them has evolved to possess a new function.[14] However, these two proteins can also be analogs and their similarity in structure is purely accidental or, for example, is due to a potential similarity of the structure of the binding site. The question then becomes, how can we retrace the history of these two proteins?

Our results suggest that it may be impossible to retrace the history of two structurally similar proteins with low sequence identity. In this case, the ancestral relation classification terms, homologs and analogs, become meaningless. There are two reasons we believe this to be so. Firstly, the correlation function (equation (15)) decays exponentially, so that beyond the correlation function relaxation time one cannot relate the sequences. Secondly, it does not make a difference if we start our design procedure from one sequence or from two unrelated sequences. After $\tau \gg \tau_o$, sequences diverge so much from each other that one cannot say what initial sequence we used in the design procedure. Furthermore, results suggest that some degree of homology may occur even between sequences that converged from an unrelated root to the same structure, i.e. in clear analogs. The reason for that is that, as we show here, some positions may feature conserved residues due to the physical requirement of stability of a common fold. Physical conservation of certain classes of amino acids at some positions in protein folds may be reflected on the genetic level due to the specifics of the genetic code. Such conservation in some cases may be confused with homology due to origin of sequences in divergent evolution.

A rigorous definition of analogs and homologs can therefore come only from the understanding of the correlation times τ between consecutive mutations. If the time-scale is smaller than τ_o then the homology is well defined: the homologous sequences in this case have high sequence similarity, while the analogous sequences have low sequence similarity. At a longer time-scale $\tau \gg \tau_o$, unless there is a high sequence similarity between sequences, the notion of homology and analogy becomes meaningless. Thus, we just use the terms analogs or homologs to refer to their sequence identity. In our model we know exactly the ancestral information during the design procedure. Since this is not the case for real fold families, we use similar terminology applied to real protein folds to avoid confusion.

Conclusion

We present a hierarchical model that attempts to explain sequence conservation caused by the most basic and universal evolutionary pressure in proteins to maintain stability. Using this model, we show that separation of basic time-scales (that constitute a broad distribution with long tails) in evolution is a plausible scenario for the sequence heterogeneity of analogous proteins. The two basic time-scales are τ_o and $\tau \gg \tau_o$; (i) at time-scale, τ_o, most mutations that occur in protein sequences do not alter the protein's thermodynamically and/or kinetically important sites and form families of homologous proteins; (ii) at time-scales $\tau \gg \tau_o$, mutations occur that would alter several amino acids at the important sites of the proteins in such a way that the properties of the proteins are not compromised. At time-scale τ, the family of analogs is formed. Mutational rates, directly computed during Z-score design, show agreement with the conservatism profiles of the fold families.

The profile solution predicts sequence entropy reasonably well for the majority of, but not all, amino acids. The amino acids that exhibit considerably higher conservatism than predicted from stability pressure alone are likely to be important for function and/or folding. Comparison of the "base-level" stability conservatism $S_P(k)$ with $S_{acr}(k)$, actual conservatism profile of a protein fold, allows one to identify functionally and kinetically important amino acid residues and potentially gain specific insights into folding and function of a protein.

Analysis of the correlation function confirms (i) the presence of an intrinsic time-scale, τ_o, at which designed sequences are similar and beyond which they differ strongly, (ii) the presence of the conserved positions in the course of Z-score design. The distributions of hamming distances between sequences reveal "clustering" of similar sequences (with low Hd) at short time-scales $\tau \sim \tau_o$ and disappearance of similarity at larger time-scales $\tau \gg \tau_o$. The above distributions are in accord with the distribution of hamming distances in the families of homologs, taken from the HSSP database, and with that of analogs, taken from the FSSP database, correspondingly.

The proposed study offers a plausible explanation of the clustering of structurally similar protein into families of homologs and analogs. From the perspective of the proposed view of evolution, the conserved amino acids appear as thermodynamically and kinetically important centers, mutations of which result in other (possibly strong) sequence modifications to preserve the physical properties of the parental proteins. Such modifications result in a new family of homologous proteins. In addition, the proposed model can be utilized to search for the thermodynamically and kinetically important amino acids in silica.

Evolution is an extremely complex phenomenon, driven by numerous factors, such as history, preservation of function, folding kinetics and stability of proteins in response to change in cell/body environment. It is remarkable, however, that our simple model was able to qualitatively capture certain aspects of protein evolution without any adjustable parameters (except for the contact definition threshold and the empirical matrix of amino acid pairwise interactions).

In addition, our model provides a possible scenario of divergent evolution. Possibly both divergent and convergent paths exist in nature. However, the question of whether evolution prefers a divergent or a convergent path is yet to be resolved. Extensive theoretical, phenomenological and experimental effort may bring insight to this puzzle.

Materials and Methods

Protein model

We use the C^{β} representation of proteins in which each pair of amino acids is in contact if their C^{β}s (C^{α} in the case of Gly) are within the distance 7.5 Å.[98] We use Miyazawa-Jernigan (MJ)[76] matrix of pair potentials to

Table 2. A six-letter potential derived for MJ 20-letter potential

	l	r	p	+	−	s
l	−0.31	−0.39	−0.22	0.01	−0.41	−0.12
r	−0.39	−0.27	−0.32	−0.02	−0.28	−0.12
p	−0.22	−0.32	−0.41	−0.25	0.07	−0.29
+	0.01	−0.02	−0.25	−0.10	−0.18	−0.18
−	−0.41	−0.28	0.07	−0.18	0.01	−0.05
s	−0.12	−0.12	−0.29	−0.18	−0.05	0.04

The symbols l, r, p, +, − and s denote six distinct corresponding groups of amino acids: aliphatic (A, V, L, I, M and C), aromatic (F, W, Y and H), polar (S, T, N and Q), positively charged (K and R), negatively charged (D and E), and special (reflecting their special conformational properties) (G and P).

represent the interaction between each pair of 20 amino acids. The total potential energy of the protein can be written as follows:

$$E = \frac{1}{2}\sum_{i \neq j}^{N} U(\sigma_i, \sigma_j)\Delta_{ij} \tag{19}$$

where N is the length of the protein, σ_i is an amino acid at the position $i = 1, \ldots, N$. $U(\sigma_i, \sigma_j)$ is the corresponding element of the MJ matrix of pairwise interactions between amino acids σ_i and σ_j. Δ_{ij} is the element of the contact matrix, that is defined to be 1 if contact between amino acids i and j exists (i.e. the distance between these amino acids in the native (ground) state is smaller than 7.5 Å), and 0, if the above contact does not exist:

$$\Delta_{ij} \equiv \begin{cases} 1, & |r_i^{\mathrm{NS}} - r_j^{\mathrm{NS}}| \leqslant 7.5\mathrm{A} \\ 0, & |r_i^{\mathrm{NS}} - r_j^{\mathrm{NS}}| > 7.5\mathrm{A} \end{cases} \tag{20}$$

where r_i^{NS} is the position of the ith residue when the protein is in the native conformation.

The six-letter potential

Due to the similarities in properties of the 20 types of amino acid, one can classify these amino acids into six distinct groups: aliphatic (A, V, L, I, M and C), aromatic (F, W, Y and H), polar (S, T, N and Q), positive (K and R), negative (D and E), and special (reflecting their special conformational properties) (G and P). We construct the potential of interaction, $U_6(\hat{\sigma}_i, \hat{\sigma}_j)$, between the six groups of amino acids, $\hat{\sigma}$, by computing the average interaction between these groups, i.e.:

$$U_6(\hat{\sigma}_i, \hat{\sigma}_j) = \frac{1}{N_{\hat{\sigma}_i} N_{\hat{\sigma}_j}} \sum_{\sigma_k \in \hat{\sigma}_i, \sigma_l \in \hat{\sigma}_j} U_{20}(\sigma_k, \sigma_l) \tag{21}$$

where σ denotes amino acids in 20-letter representation and $U_{20}(\sigma_k, \sigma_l)$ is the 20-letter matrix of interaction MJ; $\hat{\sigma}$ denotes amino acids in six-letter representation. $N_{\hat{\sigma}}$ is the number of actual amino acids of type $\hat{\sigma}$, e.g. for the aliphatic group $N_{\hat{\sigma}} = 6$. The six-letter interaction potential for MJ 20-letter potential is given in Table 2.

The measure of the information context of the sequences

In both the energy gap model and the profile solution, to study the information context of the sequences, we compute the sequence entropy, $S_X(k)$, at each position, k, of the sequence:

$$S_X(k) = -\sum_{\sigma} P_X(\sigma_k) \ln P_X(\sigma_k) \tag{22}$$

where $P_X(\sigma_k)$ is the probability that we observe an amino acid σ_k at the kth position. Subscript $X = Z$ or P denotes the energy gap model or profile solution correspondingly. The summation is taken over all possible values of σ_k.

The effect of switching to a six-letter representation of amino acids from the 20-letter representation on the sequence entropy, $S_6(k)$, is that all values $S_6(k)$ are typically smaller than that of $S_{20}(k)$. For an M-letter alphabet with all letters equally represented, i.e. $P_X(\sigma_k) = 1/M$, the entropy is equal to ln M. Thus, we expect that the difference between the typical values of $S_{20}(k)$ and $S_6(k)$ is approximately $\ln(20/6) \approx 1.2$. The case when all letters of an M-letter alphabet are equally presented corresponds to the maximal value of the entropy, i.e.:

$$S_M(k) \leqslant \ln M \tag{23}$$

The entropy of the protein fold families

Theoretical predictions from statistical-mechanical analysis can be compared with data on real proteins. In order to determine conservatism in real proteins we assume that the space of sequences that fold into the same protein structure presents a two-tier system, where homologous sequences are grouped into families and there is no recognizable sequence homology between families despite the fact that they fold into closely related structures.[11,35,99]

Using the database of protein families with close sequence similarity (HSSP database[22]), we compute frequencies of amino acids at each position, k, of aligned sequences, $P_m(\sigma_k)$, for a given, mth, family of proteins. We average these frequencies across all N_s families sharing the same fold that are present in FSSP database:[25]

$$P_{\mathrm{acr}}(\sigma_k) = \frac{1}{N_s}\sum_{m=1}^{N_s} P_m(\sigma_k) \tag{24}$$

Next, we determine the sequence entropy, $S_{\mathrm{acr}}(k)$, at each position, k, of structurally aligned protein analogs:

$$S_{\mathrm{acr}}(k) = -\sum_{\sigma} P_{\mathrm{acr}}(\sigma_k) \ln P_{\mathrm{acr}}(\sigma_k) \tag{25}$$

Acknowledgments

We thank R. S. Dokholyan for careful reading of the manuscript and S. V. Buldyrev, A. V. Finkelstein, N. V. Grishin, A. Yu. Grosberg, and L. A. Mirny for helpful discussions. The profile solution was developed with L. A. Mirny. N.V.D. is supported by NIH postdoctoral fellowship GM20251-01. E.I.S. is supported by NIH grant RO1-52126.

References

1. Govindarajan, S. & Goldstein, R. A. (1997). The foldability landscape of model protein. *Biopolymers*, **42**, 427-438.
2. Govindarajan, S. & Goldstein, R. A. (1997). Evolution of model proteins on a foldability landscape. *Proteins: Struct. Funct. Genet.* **29**, 461-466.
3. Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, **261**, 552-558.
4. Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, **357**, 543-544.
5. Finkelstein, A. V., Gutin, A. & Badretdinov, A. (1993). Why are some protein structures so common?. *FEBS Letters*, **325**, 23-28.
6. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631-634.
7. Davidson, A. R. & Sauer, R. T. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl Acad. Sci. USA*, **91**, 2146-2150.
8. Finkelstein, A. V., Gutin, A. & Badretdinov, A. (1995). Why are the same protein folds used to perform different functions? *Proteins: Struct. Funct. Genet.* **23**, 142-149.
9. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
10. Li, H., Helling, R., Tang, C. & Wingreen, N. S. (1996). Emergence of preferred structures in a simple model of protein folding. *Science*, **273**, 666-669.
11. Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold. Des.* **2**, S19-S24.
12. Chothia, C. & Gerstein, M. (1997). Protein evolution - how far can sequences diverge?. *Nature*, **385**, 579.
13. Grishin, N. V. (1997). Estimation of evolutionary distances from protein spatial structures. *J. Mol. Evol.* **45**, 359-369.
14. Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380-387.
15. Holm, L. (1998). Unification of protein families. *Curr. Opin. Struct. Biol.* **8**, 372-379.
16. Buchler, N. E. G. & Goldstein, R. A. (2000). Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: a consensus. *J. Chem. Phys.* **112**, 2533-2547.
17. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1994). *Molecular Biology of the Cell*, Garland Publishing, New York.
18. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56-68.
19. Flahetry, K. M., McKay, D. B., Kabsch, W. & Holmes, K. C. (1991). Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70k heat-shock cognate protein. *Proc. Natl Acad. Sci. USA*, **88**, 5041-5045.
20. Holmes, K. C., Sander, C. & Valencia, A. (1993). A new ATP-binding fold in actin, hexokinase and Hsc70. *Trends Cell Biol.* **3**, 53-59.
21. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). Cath - a hierarchic classification of protein domain structures. *Structure*, **5**, 1093-1108.
22. Dodge, C., Schneider, R. & Sander, C. (1998). The HSSP database of protein structure-sequence alignments and family profiles. *Nucl. Acids Res.* **26**, 313-315.
23. Sánchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S. *et al.* (2000). Protein structure modeling for structural genomics. *Nature Struct. Biol.* **7**, 986-990.
24. Perl, F. M. G., Lee, D., Bray, J. E., Sillitoe, I., Todd, A. E., Harrison, A. P. *et al.* (2000). Assigning genomic sequences to CATH. *Nucl. Acids Res.* **28**, 277-282.
25. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.
26. Holm, L. & Sander, C. (1997). An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins: Struct. Funct. Genet.* **28**, 72-82.
27. Reeck, G. R., de Haen, C., Teller, D. C., Doolittle, R. F., Fitch, W. M., Dickerson, R. E., *et al.* (1987). ''Homology'' in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, **50**, 667.
28. Goldstein, R., Luthey-Schulten, Z. A. & Wolynes, P. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl Acad. Sci. USA*, **89**, 4918-4922.
29. Shakhnovich, E. I. & Gutin, A. M. (1993). Engineering of stable and fast folding sequences of model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 7195-7199.
30. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1996). Improved design of stable and fast-folding model proteins. *Fold. Des.* **1**, 221-230.
31. Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.
32. Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524-7528.
33. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1998). Theory of kinetic partitioning in protein folding with possible applications to prions. *Proteins: Struct. Funct. Genet.* **31**, 335-344.
34. Shakhnovich, E. I. (1998). Protein design: a perspective from simple tractable models. *Fold. Des.* **3**, R45-R58.
35. Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177-196.
36. Altschuh, D., Vernet, T., Berti, P., Moras, D. & Nagai, K. (1988). Coordinated amino acid changes in homologous protein families. *Protein Eng.* **2**, 193-199.
37. Thomas, D., Casari, G. & Sander, C. (1996). The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* **9**, 941-948.

38. Gutin, A. M., Sali, A., Abkevich, V. I., Karplus, M. & Shakhnovich, E. I. (1998). Temperature dependence of the folding rate in a simple protein model: search for a "glass" transition. *J. Chem. Phys.* **108**, 6466-6483.
39. Buchler, N. E. G. & Goldstein, R. A. (1999). Universal correlation between energy gap and foldability for the random energy model and lattice proteins. *J. Chem. Phys.* **111**, 6599-6609.
40. Mirny, L. A., Finkelstein, A. V. & Shakhnovich, E. I. (2000). Statistical significance of protein structure prediction by threading. *Proc. Natl Acad. Sci. USA,* **97**, 9978-9983.
41. Gutin, A. M. & Shakhnovich, E. I. (1993). Ground state of random copolymers and the discrete random energy model. *J. Chem. Phys.* **98**, 8174-8177.
42. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Jr, Mayer, E. F., Brice, M. D., Rodgers, J. R. *et al.* (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
43. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein data bank. In *Crystallographic Databases-Information Content, Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G. & Sievers, R., eds), pp. 107-132, Data Commission of the International Union of Crystallography, Cambridge.
44. Ramanathan, S. & Shakhnovich, E. I. (1994). Statistical mechanics of proteins with evolutionary "selected" sequences. *Phys. Rev. ser. E,* **50**, 1303-1312.
45. Sali, A., Shakhnovich, E. I. & Karplus, M. (1994). Kinetics of protein folding. a lattice model study for the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614-1636.
46. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1995). Freezing transition of random heteropolymers consisting of arbitrary sets of monomers. *Phys. Rev. ser. E,* **51**, 3381-3393.
47. Saven, J. & Wolynes, P. (1997). Statistical mechanics of the combinatorial synthesis and analysis of folding molecules. *J. Phys. Chem. ser. B,* **101**, 8375-8389.
48. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known 3-dimensional structure. *Science,* **253**, 164-170.
49. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1989). *Numerical Recipes*, Cambridge University Press, Cambridge.
50. Shakhnovich, E. I. & Gutin, A. M. (1989). Formation of unique structure in polypeptide chains: theoretical investigation with the aid of a replica approach. *Biophys. Chem.* **34**, 187-199.
51. Shakhnovich, E. I. & Gutin, A. M. (1990). Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature,* **346**, 773-775.
52. Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997). Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73**, 3192-3210.
53. Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Tech. J.* **29**, 147-160.
54. Rzhetsky, A. & Sitnikova, T. (1996). When is it safe to use an oversimplified substitution model in tree-making?. *Mol. Biol. Evol.* **13**, 1255-1265.
55. Feng, D.-F. & Doolittle, R. F. (1997). Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships. *J. Mol. Evol.* **44**, 361-370.
56. Ptitsyn, O. B. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* **278**, 655-666.
57. Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA,* **95**, 4976-4981.
58. Ptitsyn, O. B. & Ting, K.-L. H. (1999). Non-functional conserved residues in globins and their possible role as a folding nucleus. *J. Mol. Biol.* **291**, 671-682.
59. Krebs, H., Schmid, F. X. & Jaenicke, R. (1983). Folding of homologous proteins. *J. Mol. Biol.* **169**, 619-635.
60. Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225-270.
61. Hollecker, M. & Creighton, T. E. (1983). Evolutionary conservation and variation of protein folding pathways. *J. Mol. Biol.* **168**, 409-437.
62. Plaxco, K. W., Spitzfaden, C., Campbell, I. D. & Dobson, C. M. (1997). A comparison of the folding kinetics and thermodynamics of two homologous fibronectin type III modules. *J. Mol. Biol.* **270**, 763-770.
63. Nishimura, C., Prytulla, S., Dyson, H. J. & Wright, P. E. (2000). Conservation of folding pathways in evolutionarily distant globin sequences. *Nature Struct. Biol.* **7**, 679-686.
64. Lorch, M., Mason, J., Clarke, A. & Parker, M. (1999). Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the i-state. *Biochemistry,* **38**, 1377-1385.
65. Schindler, T., Perl, D., Graumann, P., Sieber, V., Marahiel, M. & Schmid, F. (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.
66. López-Hernández, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein chey resembles that of a smaller protein, CI-2. *Fold. Des.* **1**, 43-55.
67. Russell, R., Sasieni, P. & Sternberg, M. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.
68. Welch, M., Oosawa, K., Aizawa, S. & Eisenbach, M. (1994). Effects of phosphorylation, Mg^{2+}, and conformation of the chemotaxis protein chey on its binding to the flagellar switch protein film. *Biochemistry,* **33**, 10470-10476.
69. Bellsolell, L., Cronet, P., Majolero, M., Serrano, L. & Coll, M. (1996). The three-dimensional structure of two mutants of the signal transduction protein chey suggests its molecular activation mechanism. *J. Mol. Biol.* **257**, 116-128.
70. Wilcock, D., Pissabarro, M. T., López-Hernández, E., Serrano, L. & Coll, M. (1998). Structure analysis of two chey mutants: importance of the hydrogen bond contribution to protein stability. *Acta Crystallog. sect. D,* **54**, 378-385.
71. Villegas, V., Martínez, J. C., Avilés, F. X. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027-1036.
72. van Nuland, N. A. J., Chiti, H., Taddei, N., Raugei, G., Ramponi, G. & Dobson, C. (1998a). Slow folding

of muscle acylphosphatase in the absence of intermediates. *J. Mol. Biol.* **283**, 883-891.
73. van Nuland, N. A. J., Meijberg, W., Warner, J., Forge, V., Scheek, R., Robbilard, G. & Dobson, C. (1998b). Slow cooperative folding of a small globular protein hpr. *Biochemistry,* **37**, 622-637.
74. Li, H., Tang, C. & Wingreen, N. S. (1997). Nature of driving force for protein folding: a result from analyzing the statistical potential. *Phys. Rev. Letters,* **79**, 765-768.
75. Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules,* **18**, 534-552.
76. Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623-644.
77. Kussell, E., Shimada, J. & Shakhnovich, E. I. (2001). Excluded volume in protein sidechain packing. *J. Mol. Biol.* **311**, 183-193.
78. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods - evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
79. Hamill, S. J., Steward, A. & Clarke, J. (2000). The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165-178.
80. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005-1009.
81. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2000). Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296**, 1183-1188.
82. Zhang, G., Liu, Y., Ruoho, A. E. & Hurley, J. H. (1987). Structure of the adenylylcyclase catalytic core. *Nature,* **386**, 247-253.
83. Artymiuk, P. J., Poirrette, A. R., Rice, D. W. & Willett, P. (1997). A polymerase I palm in adenylyl cyclase. *Nature,* **388**, 33-34.
84. Bryant, S. H., Madej, T., Liu, Y., Ruoho, A. E., Zhang, G. & Hurley, J. H. (1997). A polymerase I palm in adenylyl cyclase-reply. *Nature,* **388**, 34.
85. Kim, D. E., Gu, H. & Baker, D. (1998). The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl Acad. Sci. USA,* **95**, 4982-4986.
86. Gegoret, L. M. & Sauer, R. T. (1998). Tolerance of a protein helix to multiple alanine and valine substitutions. *Fold. Des.* **3**, 119-126.
87. Wang, J., Sattar, A. K. M. A., Wang, C. C., Karam, J. D., Konigsberg, W. H. & Steitz, T. A. (1997). Crystal structure of a pol α family replication DNA polymerase from bacteriophage RB69. *Cell,* **89**, 1087-1099.
88. Doublie, S., Tabor, S., Long, A. M., Richardson, C. C. & Ellenberg, T. (1997). Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature,* **391**, 251-258.
89. Kiefer, J. R., Mao, C., Braman, J. C. & Beese, L. S. (1997). Visualizing DNA replication in a catalytically active *Bacillus* DNA polymerase crystals. *Nature,* **391**, 304-307.
90. 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.
91. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996). Active barnase variants with completely random hydrophobic core. *Proc. Natl Acad. Sci. USA,* **93**, 5590-5594.
92. Bellsolell, L., Prieto, J., Serrano, L. & Coll, M. (1994). Magnesium binding to the bacterial chemotaxis protein CheY results in large conformational changes involving its functional surface. *J. Mol. Biol.* **238**, 489-495.
93. Berger, J. M., Fass, D., Wang, J. C. & Harrison, S. C. (1998). Structural similarities between topoisomerases that cleave one or both DNA strands. *Proc. Natl Acad. Sci. USA,* **95**, 7876-7881.
94. Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.
95. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (1998). Molecular dynamics studies of folding of a protein-like model. *Fold. Des.* **3**, 577-587.
96. Scala, A., Dokholyan, N. V., Buldyrev, S. V. & Stanley, H. E. (2001). Thermodynamically important contacts in folding of model proteins. *Phys. Rev. ser. E,* **63**, 032901.
97. Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99-113.
98. Jernigan, R. L. & Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **6**, 195-209.
99. Tiana, G., Broglia, R. & Shakhnovich, E. I. (2000). Hiking in the energy landscape in sequence space: a bumpy road to good folders. *Proteins: Struct. Funct. Genet.* **39**, 244-251.
100. Berezovsky, I. N., Grosberg, A. Y. & Trifonov, E. N. (2000). Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters,* **466**, 283-286.

Appendix

Determination of Contact Frequencies from Homopolymer Conformations

The estimation of frequencies is one of the key ingredients in protein design. An alternative approach to that proposed above is to assume that the set of conformational decoys is the set of all possible random coil states of a homopolymer collapsed at the temperatures below theta-point temperature, $T < T_\theta$, these are the states that decoy random heteropolymers explore at the folding transition temperature. Thus, we can determine the frequencies of contacts in an ensemble of random heteropolymers by taking the time average of a contact matrix element Δ_{ij} in the possible conformations of a homopolymer at $T < T_\theta$.

To compute the frequencies of contacts for a homopolymer of length N, we use discrete molecular dynamics simulations.[1-3] We model a homopolymer by N beads on a string with the interaction distances scaled to 7.5 Å. (see Dokholyan *et al.*[2] for a detailed description of the model and the algorithm). We run the simulation at the temperature T_θ (ε parameter[2] is set to -1) for 10^7 time units. (In the discrete molecular dynamics algorithm, the time unit is the average time between subsequent collisions.). After 10^7 time units of simulations we compute the frequency f_{ij} of each of the $N(N-1)/2$ contacts in our homopolymer.

There are two principal drawbacks of the second method: (1) the probability of occurrence of stable elements of the structure in homopolymers resembling secondary structure in proteins is so low that the distribution of contact lengths, $P(\ell)$, in homopolymer (not shown) drastically differs from that shown for real proteins in Figure 2 of the main text. (2) The model of a homopolymer used in the simulations strongly differs (e.g. in flexibility) from real proteins. In fact, the problem of building an appropriate model for chain flexibility is so important that small variations in it result in drastically different kinetics from a realistic one (e.g. appearance of the intermediate states) (Borreguero *et al.*, unpublished results; Ding *et al.*, unpublished results). We find that both of these drawbacks make the ''homopolymer'' approach of estimating the frequencies very inefficient for existing protein models, so we omit it in our studies.

There are two strong advantages of this approach, which make it worthwhile to explore it in the future, upon the availability of realistic protein models: (i) the possibility to generate a large amount of decoy conformations and, thus, achieve statistically highly significant contact frequency spectra; (ii) the independence of the produced decoys from various database biases.

References

1. Dokholyhan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2000). Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296**, 1183-1188.
2. Dokholyhan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (1998). Molecular dynamics studies of folding of a protein-like model. *Fold. Des.* **3**, 577-587.
3. Zhou, Y. & Karplus, M. (1997). Folding thermodynamics of a three-helix-bundle protein. *Proc Natl Acad. Sci. USA*, **94**, 14429-14432.

Edited by J. Thornton

(Received 9 April 2001; received in revised form 17 July 2001; accepted 19 July 2001)