Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins

SUSANNE MOELBERT,[1,2] ELDON EMBERLY,[1,3] AND CHAO TANG[1]

[1]NEC Laboratories America, Princeton, New Jersey 08540, USA
[2]Institut de Physique Théorique, Université de Lausanne, 1015 Lausanne, Switzerland
[3]Center for Studies in Physics and Biology, Rockefeller University, New York, New York 10021, USA

(RECEIVED September 10, 2003; FINAL REVISION November 26, 2003; ACCEPTED November 28, 2003)

Abstract

Hydrophobicity is thought to be one of the primary forces driving the folding of proteins. On average, hydrophobic residues occur preferentially in the core, whereas polar residues tend to occur at the surface of a folded protein. By analyzing the known protein structures, we quantify the degree to which the hydrophobicity sequence of a protein correlates with its pattern of surface exposure. We have assessed the statistical significance of this correlation for several hydrophobicity scales in the literature, and find that the computed correlations are significant but far from optimal. We show that this less than optimal correlation arises primarily from the large degree of mutations that naturally occurring proteins can tolerate. Lesser effects are due in part to forces other than hydrophobicity, and we quantify this by analyzing the surface-exposure distributions of all amino acids. Lastly, we show that our database findings are consistent with those found from an off-lattice hydrophobic–polar model of protein folding.

Keywords: hydrophobicity; protein folding; surface exposure; secondary structure; designability

One of the most persistent challenges in modern molecular biology is to understand how proteins fold into their unique conformations (Anfinsen 1973). The challenge lies in the fact that there are a variety of forces that contribute to the folding process and that these act over a range of length scales. Despite the many interactions, it is known that a wide variety of different protein sequences can adopt very similar folds. Analysis of the >20,000 known structures in the Protein Data Bank (PDB) resulted in only a few hundred different folds (Murzin et al. 1995). Although the number of determined sequences and structures increases rapidly, the number of "new folds" increases only slowly, which indicates that the total number of possible structures is extremely small (Chothia 1992). What leads to this many-to-one mapping of sequence to structure?

Of the many forces involved, it is argued that the hydrophobic interaction plays a central role in determining the overall fold of a protein (Kauzmann 1959; Tanford 1978). Each of the 20 amino acids has a characteristic hydrophobicity—a measure of the nonpolarity (insolubility in water) of a molecule. On average, hydrophobic residues tend to be in the core of a protein, where solvent accessibility is low, whereas polar residues tend to reside on the surface, where solvent accessibility is high (Rose et al. 1985; Miller et al. 1987; Lesser and Rose 1990; Lins et al. 2003). Many attempts based on different approaches have been made to determine the hydrophobicity of the amino acids (Nozaki and Tanford 1971; Kyte and Doolittle 1982; Engelman et al. 1986; Nauchitel and Somorjai 1994; Miyazawa and Jernigan 1996, 1999; DeVido et al. 1998; Branden and Tooze 1999). However, the various scales in the literature sometimes disagree as to these hydrophobicity rankings (Nauchitel and Somorjai 1994), which has been attributed to the fact that hydrophobicity is a relative quantity that depends on the environment and reference molecules used in the measurement (DeVido et al. 1998). Empirical hydrophobicity measurements may not truly reflect the energetics of solvation in protein folding (Lee 1993). Statistical scales may better reflect the role of solvation in folding.

Reprint requests to: Chao Tang, NEC Laboratories America, Princeton, NJ 08540, USA; e-mail: tang@nec-labs.com; fax: (609) 951-2483.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.03431704.

Although on average there is a correlation between hydrophobicity and surface exposure (Chothia 1974; Rose et al. 1985; Miller et al. 1987), the extent to which a fold of a protein, and hence its specific surface-exposure pattern, correlates with the hydrophobic pattern dictated by its amino acid sequence remains unclear. If the average hydrophobic behavior of amino acids is generally true, one might expect that there should be a statistically significant correlation between the hydrophobicity sequence and the corresponding surface-exposure pattern. However, theoretical studies of protein folding using only hydrophobicity models (Dill 1985; Lau and Dill 1989) have shown that there can be significant variations among hydrophobic–polar sequences that adopt a given structure (Li et al. 1998). This translates into the theoretical structures having a large degree of mutational stability (Li et al. 1996). Do real proteins also display this behavior? Quantifying the degree of variation between sequence and structure will be relevant to protein design based purely on hydrophobic–polar (HP) patterning, in which the hydrophobicity sequence is assumed to dictate the final fold (Kamtekar et al. 1993).

In this article, we analyze on a structure-to-structure basis the correlation between hydrophobicity sequence and surface-exposure pattern for several commonly used hydrophobicity scales. We find that all the scales yield similar distributions of correlation coefficients, and that these distributions are statistically significant when compared with a null model in which the amino acid sequences are randomized. However the distributions are broad, and the means are far from the fully correlated limit. We explore various factors that influence this less-than-optimal correlation between sequence and surface-exposure pattern. This encompasses looking at how the degree of mutational stability (i.e., sequence entropy/designability) affects the correlation, along with other lesser effects such as the actual surface-exposure propensities of the amino acids and secondary-structural influences. We show that the less-than-optimal correlation between sequence and structure for naturally occurring proteins is a manifestation of designability, and may also be selected for to "design out" competing folds.

Results

Testing hydrophobicity scales

In this section we compute the correlation coefficient between the hydrophobicity sequence and surface-exposure pattern of 3242 representative protein folds (see Materials and Methods), where the hydrophobicities of the amino acids are taken from several widely used hydrophobicity scales. The scales that we have chosen to analyze are based on different approaches: measurements of water-vapor transfer free energies and analysis of side-chain distributions (Kyte and Doolittle 1982), semitheoretical approaches determining transfer free energies for α-helical amino acid side chains from water to a nonaqueous environment (Engelman et al. 1986), determination of transfer free energies by measuring solubilities in water and ethanol relative to the reference amino acid glycine (Nozaki and Tanford 1971), calculating residue–residue potentials with pairwise contact energies (Miyazawa and Jernigan 1996), and a refined study of the latter using the Bethe approximation for determination of relative contact energies with respect to the native state (Miyazawa and Jernigan 1999). These scales cover a broad range of methods used to characterize hydrophobicity, ranging from empirical to statistical approaches.

Figure 1 shows the distributions of computed correlations between the hydrophobicity sequences and surface-exposure patterns of the 3242 structures in our data set using the above scales. The black histograms were computed using all the amino acids. None of the means exceed 0.5, with the highest being $\mu_{data} = \langle c^S \rangle_{database} = 0.454$ for the scale in Miyazawa and Jernigan 1999. Nevertheless, the computed distributions are significantly different from the null model, which considers the same set of structures but uses randomized versions of their amino acid sequences. (For each representative structure, we computed the correlation coefficient between its surface-exposure pattern and 25 random versions of its hydrophobicity sequence.) The distribution of correlation coefficients computed for the null model is shown in blue for each scale. Despite several discrepancies in classification between the scales, it can be seen that all yield similar distributions of correlation coefficients and that all have similar scores $Z = (\mu_{data} - \mu_{null})/\sigma_{null}$ when compared with their null models, with values between $Z = 2.46$ and $Z = 2.91$ (see Table 1).

The above results show that a protein fold's hydrophobicity sequence and its surface-exposure pattern are far from being completely correlated. We now explore potential reasons for this finding. In Figure 2, we show that the correlation between hydrophobicity sequence and surface-area pattern can be improved if limits are placed on either the sequence or the structure. For each representative structure, we have a set of aligned structures whose sequences also adopt the same/similar fold (see Materials and Methods). From these sequences and structures, we are able to compute an average hydrophobicity sequence and surface-exposure pattern. We find a significant improvement in the computed correlation coefficients if the average sequences and surface patterns are used (Fig. 2D; Table 1). Using averaging over sequences to help improve structural predictions was suggested by Finkelstein (1998) and later shown theoretically for an HP model (Cui and Wong 2000). In both those papers, it was argued that averaging was helping to reduce the noise in the energy parameter set. With respect to sequence–structure correlation, by averaging, one is reducing the noise contributed by sites that are not essential to dictating the final fold. The poor correlation seen at the

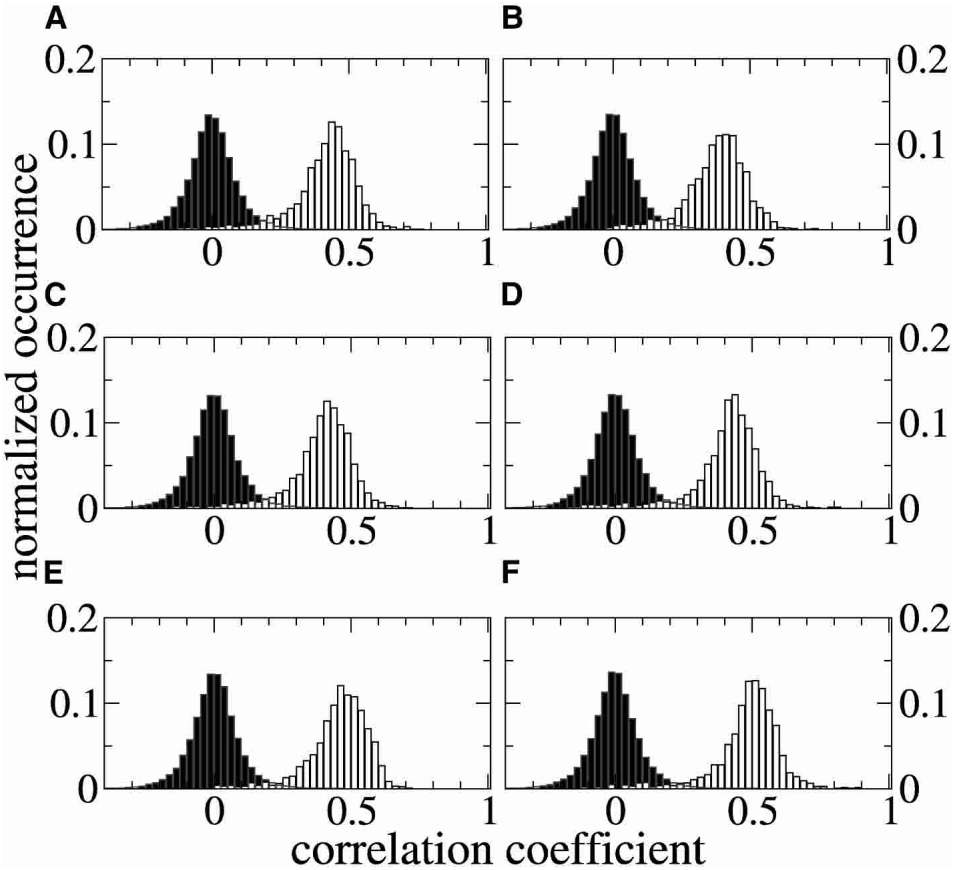

Figure 1. Histograms of correlation coefficients between single surface-exposure sequences and hydrophobicity sequences (white bars) for the 3242 representative structures obtained using the following hydrophobicity scales: (*A*) Kyte and Doolittle (1982), (*B*) Engelman et al. (1986); (*C*) Nozaki and Tanford (1971); (*D*) Miyazawa and Jernigan 1996; (*E*) Miyazawa and Jernigan 1999; and (*F*) ASA. Also shown are the histograms for the correlation coefficients of random amino acid sequences (black bars). The average correlation coefficients and the *Z* scores are (*a*) $\mu_{data} = 0.421$, $Z = 2.7$; (*B*) $\mu_{data} = 0.384$, $Z = 2.46$; (*C*) $\mu_{data} = 0.384$, $Z = 2.46$; (*D*) $\mu_{data} = 0.397$, $Z = 2.55$; (*E*) $\mu_{data} = 0.454$, $Z = 2.91$; and (*F*) $\mu_{data} = 0.492$, $Z = 3.15$.

single sequence level is evidence of naturally occurring proteins having significant mutational stability or designability. We discuss this further in the context of a model below.

A second contributing factor is that there are amino acids for which hydrophobicity is not the prime factor in determining exposure: As examples, amino acids such as glycine can appear either on the surface or in the core, and charged amino acids can form salt bridges. Including such amino acids can only lessen the correlation between hydrophobicity and surface exposure. We find that further statistical significance can be achieved if only a subset of the most hydrophobic and polar amino acids is chosen. We have found that taking the set of amino acids [ILFVRENQ] results in an appreciable improvement in the *Z* score (Fig. 2B; Table 1). The four hydrophobic residues were chosen because they are the largest, and adding others reduced *Z*. The four polar residues were selected because they have the largest ratio of polar surface area to hydrophobic surface area. Hence, including those amino acids for which hydrophobicity is most likely to be the dominant force in determining their surface exposure within a protein fold indeed improves the correlation. In the next section we explore in much more detail the propensities for surface exposure of each of the amino acids.

Lastly, we consider the improvements to the correlation between sequence and structure if only residues that form secondary-structural elements are used. Many helices and strands have one side that is hydrophobic and hence tends to be in the core, whereas the other side is polar and tends to be exposed on the surface. Turns tend to be flexible and irregular. Including turns may increase the noise in the data. Figure 2C shows that a slight improvement is gained by only considering helices and strands. We further break down the connection to secondary-structural elements and surface exposure for the various amino acids below.

Surface-exposure distributions of the amino acids

As shown above, the known hydrophobicity scales yield statistically significant correlations between a protein's pat-

Table 1. *Summary of correlation analysis*

Scale	a	b	c	d	e	f
No average:						
μ_{data}	0.421	0.384	0.384	0.397	0.454	0.492
Z	2.7	2.46	2.46	2.55	2.91	3.15
ILVFRENQ:						
μ_{data}	0.52	0.494	0.428	0.486	0.499	0.516
Z	3.3	3.16	2.75	3.11	3.19	3.3
Helices + strands:						
μ_{data}	0.47	0.467	0.443	0.417	0.458	0.499
Z	3.01	2.99	2.84	2.67	2.93	3.19
Averages:						
μ_{data}	0.572	0.535	0.555	0.579	0.591	0.613
Z	3.65	3.42	3.55	3.69	3.77	3.91

The scales used are (a) Kyte and Doolittle 1982; (b) Engelman et al. 1986; (c) Nozaki and Tanford 1971; (d) Miyazawa and Jernigan 1996; (e) Miyazawa and Jernigan 1999; and (f) ASA. The mean correlation coefficient (μ_{data}) of each distribution is given along with the $Z = (\mu_{data} - \mu_{random})/\sigma_{random}$ for several different conditions. No average corresponds to using just individual sequences and structures. ILVFRENQ considered only those positions with the given amino acids. Helices + strands used only those residues that formed secondary structural elements. Finally, averages computed the correlation coefficient using an average sequence computed from the set of aligned sequences for a given representative structure.

tern of surface exposure and the hydrophobicities of its amino acid sequence. However, despite this statistical significance, the correlations are far from the case in which hydrophobicity and exposure patterns are completely correlated. In this section, we show that this departure from optimal correlation can be partly attributed to the broad distribution of surface exposures that some amino acids tolerate. In the spirit of the work by Rose et al. (1985), for each amino acid we have computed its surface-exposure distribution within the representative set of structures. From the distributions we derive a surface-exposure propensity that reflects the tendency of each amino acid to be either exposed or buried in the core, and show that this scale leads to a better correlation between sequence and surface pattern.

Before considering the surface-exposure distributions of each amino acid, we examine the probability distributions for surface exposure and amino acid occurrence within the database of structures. Folded proteins are dense, three-dimensional (3D) clusters of amino acids. The core thus represents a considerable portion of the whole protein, whereas only a relatively small number of amino acids are to some extent exposed to the aqueous solvent. In Figure 3A we show the probability $p(A)$ for a given surface exposure A using all of the side-chain exposures from the 3242 representative structures. It is clear that a large fraction of residues reside in the core, where surface exposure is low. The probability of occurrence for the individual amino acids, $p(a.a.)$, is also nonuniform. Figure 3B shows the occurrence frequencies of the amino acids within the sequences used in the data set. These distributions will be used to examine whether the occurrence of an amino acid with a given surface exposure is correlated or independent.

For each amino acid, we compute the joint probability of observing a given surface exposure, $p(a.a.\ \&\ A)$. To extend the analysis of Rose et al. (1985) and to better characterize the propensity of a given amino acid to appear with a given surface exposure, we compare the joint probability with the null model in which the occurrence of an amino acid and the surface exposure are independent. This is expressed by the ratio,

$$P = \frac{p(a.a.\&A)}{p(a.a.)p(A)}, \quad (1)$$

where values >1 indicate favored for the given surface exposure, whereas those <1 are less favored.

Figures 4–6 show the distributions of P for the 20 amino acids. The distributions are rather broad. Tests using only a half of the database, and others using only a half of the length of the sequences, led to very similar results. As was found by Rose et al. (1985), our distributions are also suggestive of three classes of amino acids: core amino acids (C) with a peak at low surface exposure, surface amino acids (S) with a peak at high surface exposure, and intermediate amino acids (M) with relatively flat distributions. We are in agreement with Rose et al. regarding core amino acids; however, there are discrepancies between our classification of intermediate and surface amino acids. Nominally some of our intermediate amino acids show preferences for being on the surface when only secondary structure is considered—this is discussed below.

For each amino acid, the mean of the P distribution gives a weighted average surface accessibility (ASA) for each amino acid. Table 2 shows the computed ASAs of the 20 amino acids. Although the surface-exposure scale ranges from 0 (completely hidden in the core) to 1 (100% exposed to water), the averages do not take extreme values. Eleven amino acids have rather moderate tendencies to prefer the core of proteins, whereas nine are more polar. Tyrosine occurs mostly in the core, and thus shows quite hydrophobic properties in a protein environment. Charged amino acids including aspartic acid, glutamic acid, lysine, and arginine, not surprisingly, tend to occur on the surface. Cysteine is the monomer most frequently found in the core, and thus represents the most markedly hydrophobic amino acid. Thus, despite cysteine having a polar group, it has a strong tendency to be buried in the core, which can be attributed to its ability to form disulfide bonds within the cores of protein structures.

Comparison to the hydrophobicity scales shows that the ASA scale agrees in large part with the method of Miyazawa and Jernigan (1999) as regards the broad distinction between hydrophobic and polar amino acids. However, the

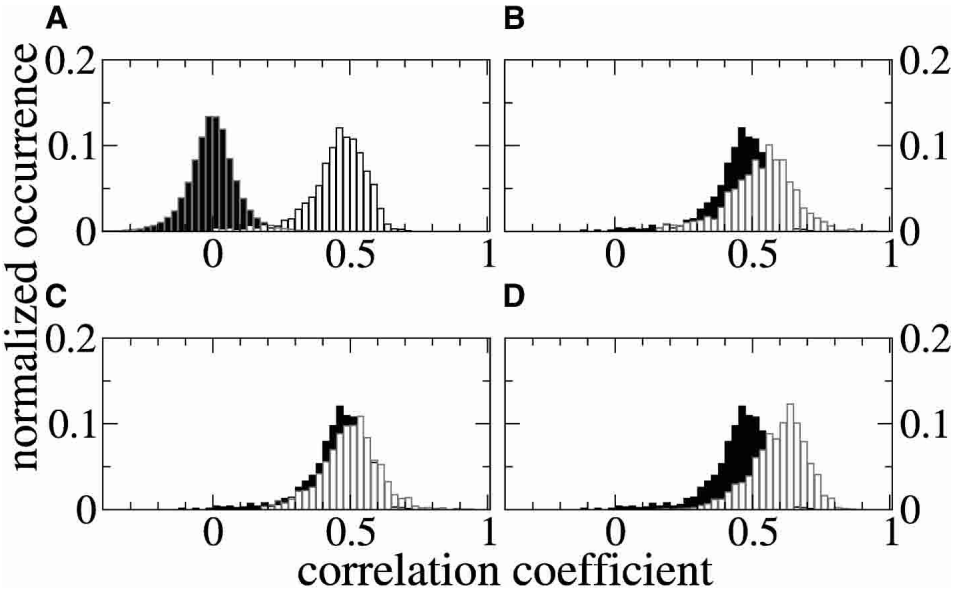

Figure 2. Correlation between hydrophobicity sequence and surface exposure for the 3242 representative structures using the scale of Miyazawa and Jernigan 1999 as a function of different factors. (*A*) No sequence averaging (white); randomized data (black). (*B*) Subset of amino acids (ILVFRENQ; white); all amino acids (black). (*C*) Only secondary structure (white); whole proteins (black). (*D*) Average over sequences that adopt the same fold (white); no averaging (black).

specific rankings are rather different. Correlations between the ASA values for the 20 amino acids and their hydrophobicity values determined using the scales under consideration are shown in Figure 7. The three scales based on the transfer free energies of amino acid side chains from water into either vapor or nonaqueous solvents have the lowest correlation with the ASA scale. An improvement is observed for the scales obtained by determination of the pairwise interaction between amino acids. Thus, the database-derived hydrophobicity scales correlate best with our statistically derived surface-exposure propensities. The lesser correlation to empirical scales highlights the context dependence of hydrophobicity, and that there are departures between how an amino acid behaves in liquid solution versus the environment of densely packed protein. This highlights how energetics depends on the reference state whose effects on the correlation between a similar set of parameter sets was discussed by Godzik et al. (1995).

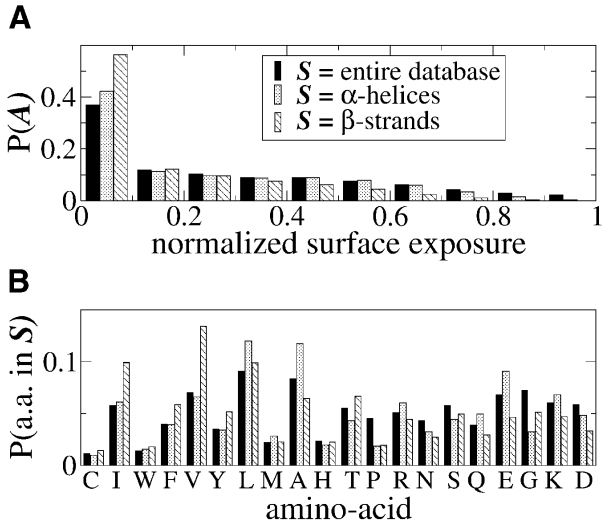

Figure 3. (*A*) Probability of finding a residue at a given degree of surface exposure *A* ($A = 0$, core; $A = 1$, surface) compared with the probability of finding an α-helix residue and a β-strand residue at a given degree of surface exposure *A*. (*B*) Probability of finding a residue in an α-helix and in a β-strand compared with the probability of finding it at any position in a protein. The total number of residues in proteins is 352,707, in α-helices 129,643, and in β-strands 74,543.

We conclude this section by re-examining the correlation between the amino acid sequence and surface-exposure pattern of a protein. Using the ASAs in Table 2, we assign to each amino acid sequence a most probable surface-exposure pattern. Table 1 shows the results of the correlation analysis using this scale. These database-derived mean surface exposures for each amino acid consistently yield better correlation coefficients than the hydrophobicity scales. Thus, using the above surface-exposure distributions to derive statistical surface propensities may offer a better alternative to the hydrophobicity scales that we have examined.

Secondary-structure analysis

The native configuration of a folded protein is characterized by secondary-structure elements, α-helices and β-strands,

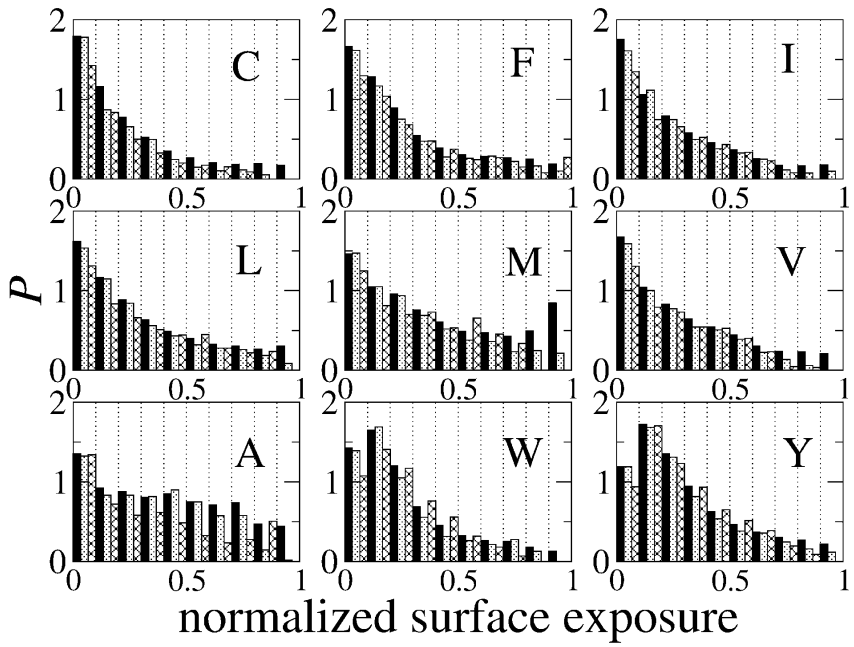

Figure 4. Histograms of degree of surface exposure of the core amino acids (*C*) in the complete database, only in α-helices, and only in β-strands. Legend as in Figure 5.

which are connected by turns (Levitt and Chothia 1976). It was shown above that considering only the sequence and surface patterns of secondary structural elements led to a slight improvement in the correlation between hydrophobicity and exposure. In this section, we break down the occurrence of the 20 amino acids in these structural elements and their corresponding surface-exposure patterns. We first consider the distribution of surface exposures within secondary elements irrespective of amino acid: Figure 3A shows that most of the residues in α-helices and β-strands occur in the interior of native protein configurations. However, this effect is much stronger for β-strands indicating that residues making up β-strands have a higher tendency to be in the core than those making up helices.

It is well known that the various amino acids have different propensities to form either α-helices or β-strands

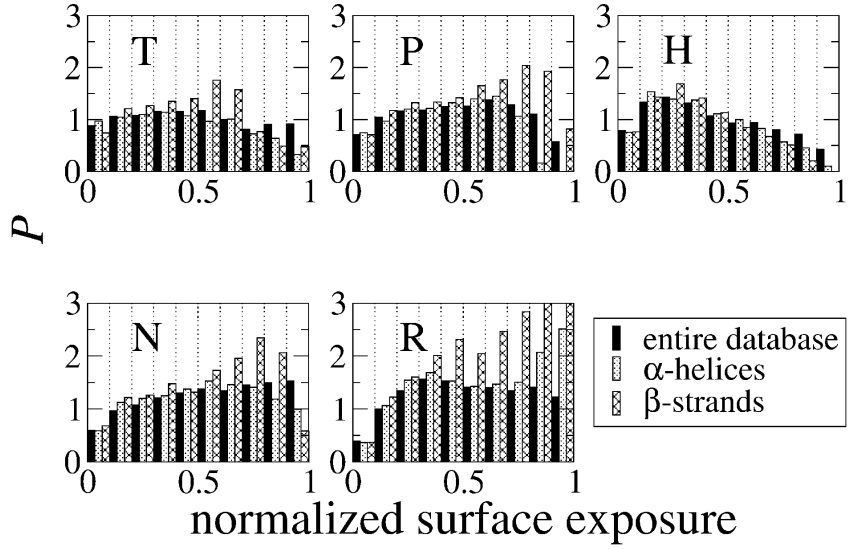

Figure 5. Histograms of degree of surface exposure of the intermediate amino acids (*M*) in the entire database, only in α-helices, and only in β-strands.

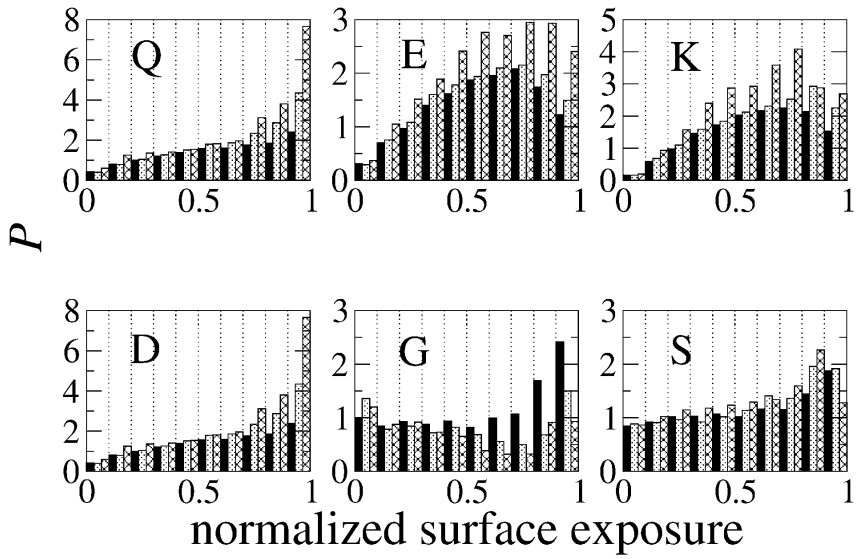

Figure 6. Histograms of degree of surface exposure of the surface amino acids (*S*) in the complete database, only in α-helices, and only in β-strands. Legend as in Figure 5.

(Munoz and Serrano 1994). Figure 3B shows the frequency of occurrence of each amino acid in α-helices and β-strands compared with the frequency of occurrence over the whole database. The amino acids are arranged according to their ASA values in increasing order. Compared with the total database, β-strands tend to be composed of a high portion of amino acids with low ASA and rather large side chains, such as V, I, and T, or with an aromatic ring as in F, Y, and W, whereas charged amino acids occur less frequently than expected. For α-helices, strong helix-formers such as alanine are particularly prominent, and the residues that are found more frequently in other parts of the proteins are divided into comparable numbers of amino acids with low and high ASA.

Figures 4–6 show the surface-exposure distributions *P* of the 20 amino acids in α-helices and in β-strands, juxtaposed with the distributions for the entire database. For the core

Table 2. *ASAs of amino acids obtained by analysis of the complete structure and sequence database, and their classifcation based on surface-accessibility distribution (Figs. 4–6)*

Amino acid		ASA	σ	Class
Cystein	C	0.268	0.248	C
Isoleucine	I	0.273	0.247	C
Tryptophan	W	0.279	0.236	C
Phenylalanine	F	0.290	0.261	C
Valine	V	0.306	0.252	C
Tyrosine	Y	0.319	0.250	C
Leucine	L	0.321	0.266	C
Methionine	M	0.364	0.288	C
Alanine	A	0.405	0.288	C
Histidine	H	0.425	0.274	M
Threonine	T	0.480	0.274	M
Proline	P	0.502	0.268	M
Arginine	R	0.539	0.255	M
Asparagine	N	0.568	0.275	M
Serine	S	0.568	0.288	S
Glutamine	Q	0.573	0.254	S
Glutamic Acid	E	0.586	0.247	S
Glycine	G	0.588	0.295	S
Lysine	K	0.607	0.231	S
Aspartic Acid	D	0.615	0.265	S

The variances, σ, of each distribution are also given.

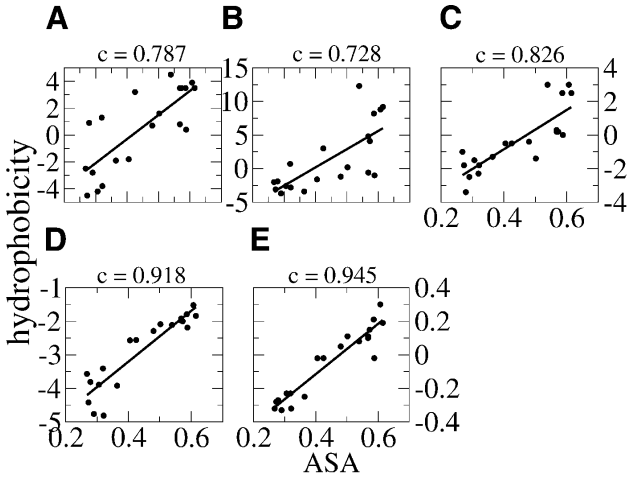

Figure 7. Correlation between ASA values of the 20 amino acids (Table 2) and their hydrophobicity values deduced from the scales of (*A*) Kyte and Doolittle (1982); (*B*) Engelman et al. (1986); (*C*) Nozaki and Tanford (1971); (*D*) Miyazawa and Jernigan 1996; and (*E*) Miyazawa and Jernigan 1999.

(*C*) amino acids, the differences are rather small. However, for the intermediate (*M*) amino acids, both arginine and asparagine (which are nominally polar) appear prominently as being exposed in β-strands. Arginine is also seen to have a tendency to appear on the exposed surfaces of helices. For those nominally polar amino acids (*S*) classified as residing on the surface, the propensity to be exposed is further increased within secondary structures when compared with the results obtained from the whole database. These slight enhancements in surface-exposure propensity for certain amino acids while in secondary-structural elements leads to the marginal improvement in correlation between sequence and surface exposure seen above when only secondary elements were included.

Model

Theoretically, hydrophobic–polar (HP) models have been studied for some time to help clarify the nature of the hydrophobic force in the folding process. Correlations have been studied in the context of sequence (White and Jacobs 1990), and nonrandomness has been detected both in real protein sequences and theoretical models (Irbäck et al. 1996; Irbäck and Sandelin 2000). Here, we consider the correlations between hydrophobicity sequences and surface-exposure patterns that emerge in a protein-folding model based solely on hydrophobicity. Does the less than perfect correlation between hydrophobicity sequence and surface pattern still remain when only solvation energy is considered? If so, is it caused by the large variation of sequences that can be tolerated by highly designable structures (Li et al. 1996)? How does averaging improve the correlation in the model results?

We study the folding of random amino acid sequences using an HP model (see Materials and Methods), in which the single energy entering the analysis is a solvation energy dependent only on the hydrophobicities of the side chains and their corresponding surface exposures in a fold. Because it is not computationally feasible to consider the continuum of possible structures that a large set of random sequences could adopt, we choose to use only a finite number of compact representative folds, formed in this case by a statistically complete set of four-helix bundles. The designability of this set of structures has been studied previously, and many of the top designable helix structures in this set correspond to naturally occurring four-helix bundles (Emberly et al. 2002). The set has the following advantages: (1) The folds are 60-mers and hence are much longer than structures generated by enumerating all possible structures using a finite set of dihedral angles (Miller et al. 2002); (2) it is more diverse than decoy sets generated from a specific native fold. A set of random amino acid sequences was folded onto the above set of structures using the HP model (see Materials and Methods). We chose the top 250 designable structures and their corresponding sequences to form the database on which to perform the correlation analysis. These structures represent plausibly thermodynamically stable folds and their corresponding sequences, although just a mere sample of the sequences that actually fold into these structures are assumed to be good folders. Lattice studies have shown that removing the compactness constraint can lead to a different set of designable structures (Chan and Bornberg-Bauer 2002), but the correlation findings below undoubtedly would not change.

Figure 8 shows the distribution of correlation coefficients between the hydrophobicity sequences and surface-exposure patterns of the model. The green histogram was computed using only a single sequence, randomly selected from the pool that fold to the corresponding structure, for each structure. This is nearly identical to what was found from the database, namely, that the correlation between a hydrophobicity sequence and its structure is less than optimal. The red histogram is for a randomized version of the data. Thus, as before the correlation between sequence and structure is not random and has some statistical significance. Because for each of the 250 designable structures we have several hundred sequences that fold into them, we can assess the effects of sampling. As in the analysis for the real protein structures, the mean hydrophobicity sequence was computed for each set of sequences that adopt the same fold. Although the mean is somewhat greater than those of the database distributions, the model distribution remains similar to the results computed from the database structures and sequences. Reducing the number of sequences used to compute the average (10%) still leads to an improvement in the correlation and is more in line with the improvement seen in

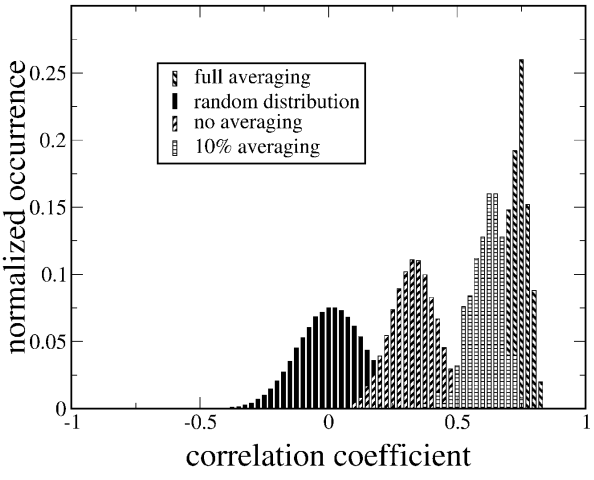

Figure 8. Histograms of correlation coefficients computed for the average hydrophobicity sequences and surface-exposure patterns of the top 250 designable model four-helix bundles. The distribution of correlation coefficients for the null model where the sequences were randomized is also shown. The scale of Nozaki and Tanford (1971) was used.

the database analysis. We discuss the implications of the theoretical findings in light of the database results below.

Discussion

Hydrophobicity has long been considered as one of the primary driving forces in the folding of proteins. It has been shown, and reconfirmed by our results, that the hydrophobicity of an amino acid is, indeed, correlated with its average surface exposure. However, the degree to which this correlation extends to the relationship between specific amino acid sequences and surface patterns has received little investigation. We have now quantified this correlation for several widely used hydrophobicity scales, and have shown that amino acid hydrophobicity does play a statistically significant role in shaping the surface-exposure pattern of a structure. However the distributions of correlation coefficients are broad, and remain far from the optimal case in which the surface-exposure pattern would show a perfect correlation with the hydrophobicity pattern.

The origin of this suboptimal correlation may lie in the fact that there are factors other than hydrophobicity that contribute to the determination of a protein's final fold. There are clearly other forces at work in determining a protein's ultimate fold, for example, a recent study suggested that hydrophobicity alone cannot account for the observed thermodynamics of protein folding (Chan 2000). Thus, some residues' behavior may not be solely dictated by hydrophobicity. Using updated data, we carried out an analysis similar to Rose et al. (1985) to determine the surface-exposure distributions of each of the amino acids, and found that many were rather broad. Indeed, several amino acids have essentially flat distributions, and hence their exposure seems to be uncorrelated with their hydrophobicity. Such broad distributions are in part responsible for the less than optimal correlation, and we showed that using only a subset of amino acids that have more peaked distributions led to improved correlations. The exposure distributions reflect all of the forces that are involved in the folding process, and we have found several discrepancies between the most likely exposure of an individual amino acid and its hydrophobicity. An example is provided by cysteine, for which the ability to form disulfide bonds with other cysteine residues constitutes a factor independent of hydrophobicity that influences surface exposure. From the distributions we computed a scale that reflects the surface-exposure propensities of the amino acids. This goes beyond just hydrophobicity and leads to an improvement in the correlation between sequence and the surface-exposure pattern of a fold. Hence, for folding studies that use energy models that are based solely on side-chain solvation, using these database-derived distributions (or the ASAs computed from them) over the empirical hydrophobicity scales should lead to a better performance.

By far the greatest improvement was achieved when we computed the correlation coefficients between average hydrophobicity sequences and structures. The average hydrophobicity sequence gives a better measure of the sequence that best matches the structure (Finkelstein 1998). The low correlation observed at the single-sequence level shows that there can be a broad variation from that of the "best match" sequence. From theoretical models, it is predicted that thermodynamically stable folds are those that are also highly designable; that is, they have a large number of sequences that fold into them (Li et al. 1996; Emberly et al. 2002; Miller et al. 2002). This large degree of mutational stability for designable folds means that there can be significant departures from the lowest energy sequence. In fact, if sequences were selected at random from a large pool of sequences that fold into a designable structure, it would be more likely to select a sequence far from the central "best match" sequence than not. Even if a sequence started near the "center" (best match sequence), its "neutral" evolution would lead it to somewhere farther away from the center in the sequence space owing to the sequence entropy (Li et al. 1998; Taverna and Goldstein 2002a). Hence, the lack of strong correlation between sequence and structure found in the database could be a signature of designability in nature. It has also been postulated that it may even be advantageous for sequences to select against being near the "best match," as such selection helps to improve plasticity in sequence space (Taverna and Goldstein 2002b).

We have shown that the correlation improves when one uses the average hydrophobicity sequence; however, we have also found that even the average sequence is not perfectly correlated with the surface-exposure pattern. This could simply be because of insufficient sampling of sequence space or could be evidence of something more fundamental. It has been argued that having a suboptimal correlation between a protein's amino acid sequence and surface-exposure pattern may help to improve the thermodynamic stability of the fold and "design out" competing folds (DeGrado 1997). All of the average hydrophobicity patterns for the most designable model helix structures have "misspellings" at various locations, where a misspelling involves the placement of a hydrophobic residue on an exposed site or a polar residue in the core. These departures from the optimal pairing of hydrophobicity with exposure have been shown in other theoretical studies (Emberly et al. 2002) to help increase the energy gap between the ground state and competing structures. If the hypothesis of designing out competing structures through suboptimal correlation is valid, this has important consequences for structural design based on binary patterning (Kamtekar et al. 1993). The surface pattern of the structure may act as a starting point for the selection of an amino acid sequence, but it may then prove advantageous to depart from this blueprint to improve thermodynamic stability. Database analysis of the type per-

formed here may form the basis for advanced techniques to detect further correlations between sequence and structure that would help to better design sequences in protein design.

Materials and methods

Representative set of database structures

To have a nonredundant set of protein structures for analysis, we have chosen to use the 3242 representative structures from the FSSP database (Holm and Sander 1996). The FSSP database is the result of an all-against-all structure analysis that groups protein structures into a hierarchical tree based on their level of structural similarity. All residues of the known protein structures are compared in three dimensions, and the results are reported in the form of alignments of equivalent residues. Redundancy is eliminated by removing proteins with mutual sequence identity >25%, because they result in almost complete structural overlap. There are 30,624 known protein chains grouped to one of the representative structures in the FSSP. Each representative structure has a set of aligned structures. Each structure, in turn, has a corresponding amino acid sequence. Thus, for each representative structure in the FSSP, we have a list of aligned structures along with a corresponding set of amino acid sequences, all of which are assumed to fold into a similar fold as the representative structure in the aligned regions.

Correlation analysis

A hydrophobicity scale s assigns a hydrophobicity value $h_{\text{a.a.}}{}^s$ to each amino acid (a.a.). $h_{i,j}{}^s$ is the hydrophobicity of the i-th aligned residue of sequence j that is aligned with a representative structure, based on the hydrophobicity scale s. For the set of amino acid sequences that fold into a given structure, we wish to consider what the average hydrophobicity sequence for the set is. We consider the average sequence because it gives a good characterization of the hydrophobicity sequence that adopts the given representative structure (Finkelstein 1998; Cui and Wong 2000). The average hydrophobicity value $\overline{h_i^s}$ at position i within this representative structure using scale s is:

$$\overline{h_i^s} = \frac{1}{M}\sum_{j=1}^{M} h_{i,j}^s, \tag{2}$$

where M is the number of sequences in the alignment at residue i. Calculating this average for all residues of the representative structure with length N gives the average hydrophobicity sequence of this structure: $(\overline{h_i^s})_{i=1..N} = (\overline{h_1^s}, \overline{h_2^s}, \cdots, \overline{h_N^s})$.

The surface exposure a_i of residue i in a structure is quantified as the amount of surface area of the side chain atoms (represented as spheres) that is accessible to water (represented by a sphere of radius 1.4 Å). For each structure, we obtain the surface exposures of each of its residues from the FSSP file. We normalize each surface exposure by the total surface area of the side-chain atoms making up the given residue (Creighton 1993). This yields a fractional exposure for each residue in a structure. We compute an average surface-exposure pattern for a structure using its FSSP alignment:

$$\overline{a_i} = \frac{1}{L}\sum_{j=1}^{L} a_{i,j}^{\gamma}, \tag{3}$$

where L is the number of known structures that have a residue aligned with residue i and $a_{i,j}$ denotes the surface-accessible area of residue i in structure j of the alignment. Performing this procedure for each residue i of the representative structure leads to a sequence of surface accessibilities $(\overline{a_i})_{i=1..N} = (\overline{a_1}, \overline{a_2}, \cdots, \overline{a_N})$.

The correlation coefficient c^s between the hydrophobicity sequence $(\overline{h_i^s})_{i=1..N}$ and the accessible surface-area sequence $(\overline{a_i})_{i=1..N}$ of a structure is given by:

$$c^s = \frac{\sum_i^N (\overline{a_i} - \overline{a})(\overline{h_i^s} - \overline{h^s})}{\sqrt{\sum_i^N (\overline{a_i} - \overline{a})^2 \sum_i^N (\overline{h_i^s} - \overline{h^s})^2}}. \tag{4}$$

Hydrophobic–polar model

In hydrophobic–polar (HP) models, hydrophobicity is the sole force driving the folding process (Dill 1985; Lau and Dill 1989). For an amino acid sequence that corresponds to a sequence of hydrophobicities $\{h_i\}$, the solvation energy of the sequence on a given structure γ is

$$E^{\gamma} = \sum_{i=1}^{N} h_i(1 - a_i^{\gamma}) \tag{5}$$

where $a_i{}^{\gamma}$ is the surface exposure of residue i in structure γ. The native fold of a sequence is the one that minimizes this energy.

We use a representative set of structures to act as the space of potential folds. For a given amino acid sequence, we then use the above energy equation to determine the structure that has the lowest energy within the set of competing structures. We deem this to be the native fold of the sequence. Studies have shown that folding numerous random amino acid sequences in this way results in a nonuniform mapping of sequences to structures: Some structures turn out to be native folds far more often than others, and have been designated "designable" structures (Li et al. 1996).

Here we consider a representative set of 203,282 four-helix bundles for the competing set of structures (Emberly et al. 2002). This set was shown to cover the space of all possible four-helix folds at the 95% confidence level, and hence represents a relatively complete set of compact folds on which an HP sequence can compete. Then 10^6 random amino acid sequences (the hydrophobicity scale based on transfer free energy between water and ethanol was used; Nozaki and Tanford 1971) were folded by selecting the ground-state structure for each sequence. The top 250 designable structures (each with several hundred sequences that fold into it) and their corresponding hydrophobicity sequences formed the model database on which the correlation analysis was performed.

Acknowledgments

We thank Jonathan Miller and Bruce Normand for many helpful comments and discussions. This research was partially supported by the Swiss Study Foundation and by the Swiss National Science Foundation through grants FNRS 21-61397.00 and 2000-67886.02, and the National Natural Science Foundation of China (no. 20228306).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby

marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181:** 223–230.

Branden, C. and Tooze, J. 1999. *Introduction to protein structure*, pp. 6–7. Garland Publishing, New York.

Chan, H.S. 2000. Modeling protein density of states: Additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. *Proteins* **40:** 543–571.

Chan, H.S. and Bornberg-Bauer, E. 2002. Perspectives on protein evolution from simple exact models. *App. Bioinformatics* **1:** 121–144.

Chothia, C. 1974. Hydrophobic bonding and accessible surface area in proteins. *Nature* **248:** 338–339.

———. 1992. One thousand families for the molecular biologist. *Nature* **357:** 543–544.

Creighton, T.E. 1993. Hydrophobicity scale (p. 154). Surface accessibilities of amino acids (p. 142). In *Proteins: Structures and molecular principles*, 2nd ed. W.H. Freeman, New York.

Cui, Y. and Wong, W.H. 2000. Multiple-sequence information provides protection against mis-specified potential energy functions in the lattice model of proteins. *Phys. Rev. Lett.* **85:** 5242–5245.

DeGrado, W.F. 1997. Proteins from scratch. *Science* **278:** 80–81.

DeVido, D.R., Dorsey, J.D., Chan, H.S., and Dill, K.A. 1998. Oil/water portioning has a different thermodynamic signature when the oil solvent chains are aligned than when they are amorphous. *J. Phys. Chem.* **102:** 7272–7279.

Dill, K.A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* **24:** 1501–1509.

Emberly, E.G., Wingreen, N.S., and Tang, C. 2002. Designability of α-helical proteins. *Proc. Natl. Acad. Sci.* **99:** 11163–11168.

Engelman, D.M., Steitz, T.A., and Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biomol. Struct.* **15:** 321–353.

Finkelstein, A.V. 1998. 3D protein folds: Homologs against errors—A simple estimate based on the random energy model. *Phys. Rev. Lett.* **80:** 4823–4825.

Godzik, A., Kolinski, A., and Skolnick, J. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* **4:** 2107–2117.

Holm, L. and Sander, C. 1996. Mapping the protein universe. *Science* **273:** 595–602.

Irbäck, A. and Sandelin, E. 2000. On hydrophobicity correlations in protein chains. *Biophys. J.* **79:** 2252–2258.

Irbäck, A., Peterson, C., and Potthast, F. 1996. Evidence for nonrandom hydrophobicity structures in protein chains. *Proc. Natl. Acad. Sci.* **93:** 9533–9538.

Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., and Hecht, M.H. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262:** 1680–1685.

Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14:** 1–63.

Kyte, J., and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157:** 105–132.

Lau, K.F. and Dill, K.A. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22:** 3986–3997.

Lee, B. 1993. Estimation of the maximum change in stability of globular proteins upon mutation of a hydrophobic residue to another of smaller size. *Protein Sci.* **2:** 733–738.

Lesser, G.J. and Rose, G.D. 1990. Hydrophobicity of amino acid subgroups in proteins. *Proteins* **8:** 6–13.

Levitt, M. and Chothia, C. 1976. Structural patterns in globular proteins. *Nature* **261:** 552–558.

Li, H., Helling, R., Tang, C., and Wingreen, N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* **273:** 666–669.

Li, H., Tang, C., and Wingreen, N. 1998. Are protein folds atypical? *Proc. Natl. Acad. Sci.* **95:** 4987–4990.

Lins, L., Thomas, A., and Brasseur, R. 2003. Analysis of accessible surface of residues in proteins. *Protein Sci.* **12:** 1406–1417.

Miller, S., Janin, J., Lesk, A.M., and Chothia, C. 1987. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196:** 641–656.

Miller, J., Zeng, C., Wingreen, N., and Tang, C. 2002. Emergence of highly designable protein-backbone conformations in an off-lattice model. *Proteins* **47:** 506–512.

Miyazawa, S. and Jernigan, R.L. 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing term, for simulation and threading. *J. Mol. Biol.* **256:** 623–644.

———. 1999. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins: Struct. Mol. Principles* **34:** 49–68.

Munoz, V. and Serrano, L. 1994. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ–ψ matrices: Comparison with experimental scales. *Proteins* **20:** 301–311.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Nauchitel, V.V. and Somorjai, R.L. 1994. Spatial and free energy distribution patterns of amino acid residues in water soluble proteins. *Biophys. Chem.* **51:** 327–336.

Nozaki, Y. and Tanford, C. 1971. Solubility of amino acids and 2 glycine peptides in aqueous ethanol and dioxane solutions—Establishment of a hydrophobicity scale. *J. Biol. Chem.* **246:** 2211.

Rose, G., Geselowitz, A., Lesser, G., Lee, R., and Zehfus, M. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* **289:** 834–839.

Tanford, C. 1978. Hydrophobic effect and organization of living matter. *Science* **200:** 1012–1018.

Taverna, D. and Goldstein, R.A. 2002a. Why are proteins marginally stable? *Proteins* **46:** 105–109.

———. 2002b. Why are proteins so robust to site mutations? *J. Mol. Biol.* **315:** 479–484.

White, S.H. and Jacobs, R.E. 1990. Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution. *Biophys. J.* **57:** 911–921.