J. Mol. Biol. (1995) **252**, 709–720

JMB

Recognizing Native Folds by the Arrangement of Hydrophobic and Polar Residues

Enoch S. Huang*, S. Subbiah and Michael Levitt

Beckman Laboratories for Structural Biology Department of Structural Biology, Stanford University School of Medicine, Stanford CA 94305-5400, USA

**Corresponding author*

Central to the *ab initio* protein folding problem is the development of an energy function for which the correct native structure has a lower energy than all other conformations. Existing potentials of mean force typically rely extensively on database-derived contact frequencies or knowledge of three-dimensional structural information in order to be successful in the problem of recognizing the native fold for a given sequence from a set of decoy backbone conformations. Is the detailed statistical information or sophisticated analysis used by these knowledge-based potentials needed to achieve the observed degree of success in fold recognition? Here we introduce a novel pairwise energy function that enumerates contacts between hydrophobic residues while weighting their sum by the total number of residues surrounding these hydrophobic residues. Thus it effectively selects compact folds with the desired structural feature of a buried, intact core. This approach represents an advance over using pairwise terms whose energies of interaction that are independent of the position in the protein and greatly improves the discrimination capability of an energy function. Our results show that 85% of a set of 195 representative native folds were recognized correctly. The 29 exceptions were lipophilic proteins, small proteins with prosthetic groups or disulfide bonds, and oligomeric proteins. Overall, our method separates the native fold from incorrect folds by a larger margin (measured in standard deviation units) than has been previously demonstrated by more sophisticated methods. The arrangement of hydrophobic and polar residues alone as evaluated by our novel scoring scheme, is unexpectedly effective at recognizing native folds in general. It is surprising that a simple binary pattern of hydrophobic and polar residues apparently selects a given unique fold topology.

Keywords: protein folding; hydrophobic interaction; fold recognition; contact potential; threading

Introduction

Since the classic work by Kauzmann (1959), it has been hypothesized that hydrophobic interactions play a major role in organizing and stabilizing the architecture of proteins. This phenomenon, loosely defined, is related to the relative insolubility of non-polar substances in water (Tanford, 1980). It has long been observed that residues with hydrophobic side-chains tend to segregate into the interior of a globular protein, thus constituting a core in which they interact with each other rather than with water, whereas residues with charged and polar side-chains remain exposed to the solvent (Perutz *et al*., 1965; Rose *et al*., 1985). It is now widely accepted that hydrophobicity is a dominant force of protein folding (Dill, 1990).

The patterning of hydrophobic and polar residues (a ''binary code'') in a sequence appears to be a strong determinant of the fold of a protein. Recent experiments from Hecht and co-workers have demonstrated that a binary code is sufficient for folding a polypeptide into the topology of a four-helix bundle, even though the side-chain packing for these molecules is still uncharacterized (Kamtekar *et al*., 1993). There is also an array of evidence indicating that although the identities and sizes of residues aligned on a given fold vary greatly, the polarities are strongly preserved (Lesk & Chothia, 1980; Bashford *et al*., 1987; Sweet & Eisenberg, 1983). Consequently, one may determine the compatibility of a sequence for a given fold by simply aligning hydrophobic residues to buried positions in the structural motif (Bowie *et al*., 1990a).

Since the polarities, but not necessarily the identities, of residues are conserved on a fold, it appears that hydrophobic interactions are non-

0022–2836/95/400709–12 $12.00/0

specific. Inter-residue contact frequencies, garnered from the structural database, fail to show bias towards particular pairs of hydrophobic residues other than cysteine-cysteine in the context of a disulfide bond (Miyazawa & Jernigan, 1985; Bryant & Amzel, 1987). Dill and co-workers, in their theoretical work, have mounted sequences composed of only two types of residues, hydrophobic (H) and polar (P), upon square and cubic lattices (Lau & Dill, 1989; Yue & Dill, 1992; Dill *et al*., 1993). They concluded that for these crude models, favorable attraction between the hydrophobic monomers alone was sufficient to drive cooperative folding of these copolymers into unique compact shapes with well-defined hydrophobic cores. Moreover, the relative composition of H and P monomers in the sequence appeared to be important for maintaining a unique global minimum for this model. Experimental data for real polypeptides have been corroborative. Residues in the hydrophobic core are generally tolerant of substitution by other non-polar residues (Bowie *et al*., 1990b; Lim & Sauer, 1989). Although individual residues exposed to the solvent are more robust to replacement, the surface as a whole must remain polar to a large extent, otherwise competing conformations may predominate (Bowie *et al*., 1990b).

The interaction and placement of hydrophobic residues seems to be a more critical determinant of protein structure than the role of polar residues and local sequence-dependent interactions. Studies using site-directed mutagenesis have indicated that the stability and activity of proteins are greatly compromised by replacement of hydrophobic residues by polar and charged residues (Lim & Sauer, 1989; Shortle & Meeker, 1986; Hecht *et al*., 1984). Moreover, the contribution that a hydrophobic residue makes to the stability of a protein varies roughly with its extent of burial (Shortle *et al*., 1990; Serrano *et al*., 1992). In contrast, polar residues at the surface may play a less important role in dictating the structure of the protein because they generally have low information content (Reidhaar-Olson & Sauer, 1988) and do not contribute appreciably to the stability of some proteins (Sali *et al*., 1991). Hydrophobic interactions can also overwhelm other factors, such as sequence-dependent propensities towards formation of certain substructures (Kabsch & Sander, 1984; Argos, 1987; Cohen *et al*., 1993).

In sum, it appears that the sequestering of hydrophobic residues into the interior of the protein may be the most important link between sequence and structure. Whether or not a fold actually forms a stable core is also dependent on how well the side-chain shapes can be packed together (Ponder & Richards, 1987; Lim & Sauer, 1989; Lee & Subbiah, 1991; Lee & Levitt, 1991; Eriksson *et al*., 1992). This notwithstanding, it would seem that the partitioning of hydrophobic residues is a necessary, if incomplete, hallmark of the compatibility of a given sequence with a selected fold.

If hydrophobic interactions are essential for the stability of the native states of proteins, then they somehow must be included in any theoretical treatment of the conformational energy of a polypeptide (see the review by Wodak & Rooman, 1993). Novotny *et al*. (1984, 1988) were the first to investigate the consequences of draping sequences onto non-native backbone conformations. By including terms in their potential energy function that account for the effect of solvent, the misfolded structures had higher free energy than the native structures, since the incorrect models were characterized by the exposure of non-polar side-chains and the burial of charged groups. Later work used similar polarity criteria to distinguish correct from incorrect models (Baumann *et al*., 1989). Alternatively, solvent effects may be included implicitly through statistically derived contact energies; these contain information reflecting the partitioning of hydrophobic residues to the interior of proteins (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985; Sippl, 1990; Hinds & Levitt, 1992, 1994; Bryant & Lawrence, 1993; Sun, 1993). Other approaches, such as using atomic solvation parameters or three-dimensional profile methods, judge the quality of models in part by explicitly considering the degree to which residues are buried or solvated (Eisenberg & McLachlan, 1986; Chiche *et al*., 1990; Lüthy *et al*., 1992; Ouzounis *et al*., 1993). Finally, hydrophobic forces modeled from statistical data can be used alone to select native folds from non-native decoys (Casari & Sippl, 1992). Such potentials are commonly applied as a means of recognizing the native conformation from all other folds.

The problem of fold recognition has been approached primarily by ''threading'' a specific polypeptide sequence on backbones of equal or greater length taken from the database, thereby generating a set of discrete decoys. By rank-ordering the energies of the various conformations, one expects the native conformation to receive the most favorable value. One method adjusted parameters in a contact potential such that the pseudo-energies of native structures in a training set were global minima (Maiorov & Crippen, 1992). Although this function merely mimics only free energy, it successfully recognizes the native fold for nearly all compact protein structures. Statistically derived potentials can competently select the native structure from a large pool of alternatives generated from known structural motifs (Sippl, 1990, 1993; Casari & Sippl, 1992; Bryant & Lawrence, 1993; Bauer & Beyer, 1994).

Here we describe a new contact potential that does not depend on any information from known protein structures other than the rule of ''hydrophobic residues inside, polar residues outside''. Although not all proteins apparently depend on hydrophobic interactions for stability to the same extent, we discover that our function can clearly distinguish the native fold from decoys for those proteins for which this generalization holds true. Our method regards a given sequence as a chain of binary-encoded residues; the actual residue types are important only to the extent that they are either polar or hydrophobic. Although free energies *per se* are not

computed, we believe that this function captures the essence of all energy scoring methods while circumventing the problem of scarce data (Sippl, 1990) inherent to statistically derived potentials. Moreover, the specifics of the potential select for a specific structural feature: an intact, buried core of hydrophobic residues. Finally, a potential with high discrimination power (i.e. the score of the native fold is separated by a large margin from the scores of the non-native folds) may be more useful for evaluating the fitness of folds that are exhaustively enumerated by lattice (Hinds & Levitt, 1992, 1994; Covell & Jernigan, 1990) or off-lattice (Park & Levitt, 1995) models than those with lower discrimination power. Our results clearly suggest that for most proteins, the simple arrangement of non-specific hydrophobic and polar residues upon a folding motif is in itself sufficient to select the native fold from discrete alternatives with a high degree of discrimination.

Method Development

In accordance with our goal of using only the most basic of premises, our method relies on a simplified representation of actual protein structures (Levitt, 1976). We therefore begin by describing how one may reduce a detailed structure of a protein into an adequate model. Next, we present the computation of our contact potential, which considers only the placement of hydrophobic residues on a structural motif. Finally, we discuss the particular manner by which we explicitly analyze the discrimination power of our method.

Simplified representation of proteins

Each residue is defined by its C^{α} coordinate and a representative coordinate approximating the centroid of its side-chain (glycine has only the former). This virtual side-chain is a point 3.0 Å from C^{α} along the C^{α}-C^{β} vector. This representation is like that used by Bryant & Lawrence (1993), who computed the mean projection of all side-chain centroids on the C^{α}-C^{β} vector in a set of 161 proteins. The C^{β} coordinates, when unavailable, are computed using the backbone coordinates and standard bond geometry. We consider two residues to be in contact if their effective side-chains are within 7.3 Å of each other and if they are not adjacent to each other in the sequence. This is similar to the contact distance of 6.5 Å, determined by Miyazawa & Jernigan (1985) to be the radial distribution of residues excluding nearest-neighbors along the polypeptide chain. We tested these parameters on the structure of barnase, whose Brookhaven Protein Data Bank, or PDB, entry is 1rnb (Bernstein *et al*., 1977). Fersht and co-workers (Serrano *et al*., 1992) have defined the residues that comprise the three separate hydrophobic cores in this enzyme. Using these parameters, our method produced a list of contacts. As would be expected for a reasonable model, only residues within each of the three hydrophobic cores were in contact with each other; no inter-core contact was present in the list. Furthermore, all the core residues described by Serrano *et al*. (1992) were included. It was encouraging that the residues that make the most hydrophobic contacts within each of the three cores, according to our representation, correspond to the central core residues designated by Fersht and colleagues. To summarize, our choice of parameters seems reasonable for our desired task of both enumerating and subsequently defining the composition of hydrophobic cores in a macromolecule.

Computing the hydrophobic fitness score

In order to test the fitness of a fold for a given sequence, the environment surrounding each hydrophobic residue is probed. The contact function is designed to yield favorable values for the conformations that form hydrophobic cores. We use two criteria in computing our score: (1) the polarity of the environment surrounding each hydrophobic residue; and (2) the extent to which that residue is buried.

Seven amino acid residues are designated as hydrophobic, Cys, Ile, Leu, Met, Phe, Trp, and Val; the others are considered to be polar. Since tyrosine is often not buried in known protein structures (Wertz & Scheraga, 1978; Miyazawa & Jernigan, 1985), we do not regard it to be a hydrophobic residue. For each of these residues we evaluate the polarity of its environment by enumerating the contacts it makes with other non-polar residues. While tyrosine is not itself defined as one of the seven hydrophobic residues, contacts with tyrosine are included in this summation. Approximately twice as many hydrophobic contacts are made in native proteins as would be expected on a random basis (Bryant & Amzel, 1987); hence, structures that make such contacts should be rewarded with higher scores. First, we compute H_i, the total number of contacts (as defined in the previous section) a given hydrophobic residue i makes with the non-polar residues. To penalize conformations that do not place the hydrophobic residue in a non-polar environment, we subtract from H_i the number of hydrophobic contacts expected by chance. This expected value, which we designate as H_i°, is based on the sequence composition and spatial position of the hydrophobic residue. Thus, a positive value of $H_i - H_i^{\circ}$ contributes favorably towards our ''energy'' score, since hydrophobic residue i is contacted by more non-polar residues than would be expected on a random basis. We sum this quantity $(H_i - H_i^{\circ})$ over all hydrophobic residues i and divide by the number of hydrophobic residues n to yield the ''hydrophobic component'' of our score:

$$\text{Hydrophobic term} = \frac{\sum_i (H_i - H_i^{\circ})}{n}$$

Since a hydrophobic residue is typically buried in the native state, it also tends to have many neighboring residues surrounding it. Moreover, it has been discovered that the average number of neighbors within 10 Å of a given residue correlates optimally with its hydrophobicity, based on solvent-accessibility data (Viswanadhan, 1987), and with its expected effect on stability (Shortle *et al*., 1990). We define the burial of a given residue i, B_i, very simply as the number of residues within 10 Å. As with the first term, this quantity is summed over all hydrophobic residues i and divided by n to yield the ''burial'' term of our score:

$$\text{Burial term} = \frac{\sum_i B_i}{n}$$

Our definition of burial is simpler and more approximate than those measures based on accessible surface area (Lee & Richards, 1971).

Finally, we combine the two terms by weighting the hydrophobic term by the burial term to obtain the overall score. This value, which we call the hydrophobic fitness (*HF*) score, indicates how well the conformation assembles a hydrophobic core (or cores) using the non-polar residues in its sequence. A given fold with random sequences threaded upon it will, on average, receive an *HF* score of zero, as will a given sequence in random conformations. We multiply *HF* by −1 so that favorable conformations receive lower values, as is customary for analogous free energy calculations. A schematic of our method is shown in Figure 1. Computation of *HF* is summarized by the following equation:

$$\text{Hydrophobic Fitness score } HF = -\frac{\left(\sum_i B_i\right)\left(\sum_i (H_i - H_i^\circ)\right)}{n^2}$$

where the summation is over all hydrophobic (C, F, I, L, M, V, W) residues i; for each hydrophobic residue i, B_i is the burial, or the number of virtual side-chain centroids within 10 Å; H_i is the number of contacts made with non-polar side-chains (the seven hydrophobic residues plus Y), i.e. the number of non-polar centroids within 7.3 Å; n is the number of hydrophobic residues (excluding tyrosine) in the sequence. H_i° is the expected number of hydrophobic contacts, defined as:

$$H_i^\circ = C_i\left(\frac{h_i}{N_i}\right)$$

where C_i is the number of all side-chain centroids that contact residue i, h_i is the number of hydrophobic residues (including tyrosine) in the sequence, except for residue i and any that immediately precede or follow residue i; N_i is the total number of residues, excluding residue i and the residues immediately flanking it.

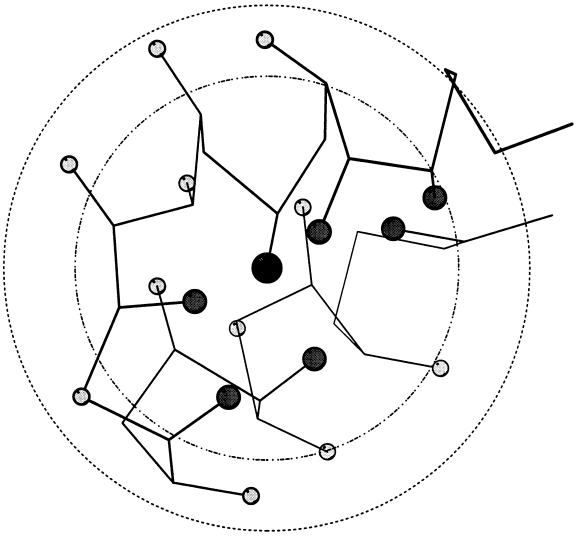

Figure 1. Schematic of the scoring method. The function probes the environment surrounding every hydrophobic (C, F, I, L, M, V, W) residue i. For such a residue, the polarity of its surroundings is evaluated by enumerating the number of contacts the virtual side-chain, shown in black, makes with other hydrophobic residues plus Tyr. This value is equal to H_i (see the text). The virtual side-chains of such non-polar residues within 7.3 Å of this central residue are depicted in dark gray; the inner concentric ring represents the 7.3 Å sphere. The burial of the residue (B_i) is assessed by counting the number of virtual residues within 10 Å (represented by the outer concentric ring). All other residues within this radius are drawn in light gray. This Figure was prepared using the software MOLSCRIPT (Kraulis, 1991).

Evaluating the discrimination power of the potential

Each sequence is threaded upon its native fold and the hydrophobic fitness score is computed. Then for that sequence, we challenge the potential to recognize its correct structure from a set of unrelated backbone segments of equal or greater length obtained from the structural database. Each sequence is threaded upon the backbone scaffold using every possible ungapped alignment and the corresponding scores are computed. For example, a sequence of length 100 residues mounted on a backbone of 150 residues would assume (150 − 100 + 1) = 51 conformations. It is hoped that the native fold would score more favorably than all other conformations generated in this fashion.

Moreover, it is also desirable that the discrimination of the contact function be high. To evaluate the discrimination, we use a Z-score, Z_t, defined as:

$$Z_t = \frac{HF_{\text{native}} - \langle HF \rangle}{\sigma}$$

where HF_{native} is the score of the native fold; $\langle HF \rangle$ is the mean score of the entire distribution of *HF* scores; and σ is the standard deviation of the distribution.

The contact potential should also be able to recognize the correct patterning of hydrophobic and polar residues upon a given motif. To this end, the

score of the native sequence is compared with the scores of many random permutations of that sequence mounted on the same motif. The *HF* score of the native sequence is transformed into another *Z*-score, Z_p, for permuted sequence decoys, defined as the separation of the native score from the average score of the sequence-randomized set in standard deviation units. The quantity Z_p represents the level to which the native sequence is compatible with its correct fold according to our function. It has been suggested that 1000 permutations of the sequence is sufficient to approximate the standard deviation of the distribution (Bryant & Lawrence, 1993); we find that 2500 provides a more converged estimate (data not shown). Taken together, Z_t and Z_p measure the compatibility of a native sequence with its three-dimensional structure.

Classifying the test set of proteins

The main premise on which our method is based is that proteins generally have hydrophobic cores in which the non-polar side-chains interact, shielded from solvent. However, at least three main classes of polypeptides are notable exceptions to this ''hydrophobic in, polar out'' generalization. For these proteins, our potential might be expected to yield less favorable *Z*-scores or fail to assign the native sequence-structure pair as being the most favorable in ''energy''.

The first group is composed of small proteins. Their native conformations are likely to depend on additional sources of stability other than simple hydrophobic interactions among residues alone. Small proteins commonly have disulfide bonds conferring stability (for a review, see Betz, 1993). In the absence of prosthetic groups, small proteins such as cytochrome b_{562} will often fail to assume the conformation of the holoprotein (Feng *et al*., 1994). Since our potential is based solely on a simple-minded counting of only hydrophobic residues, it does not address the thermodynamic effects of these additional factors. Besides disregarding whatever stabilization the prosthetic groups may exert in the native fold, our function is further handicapped by the disruption such groups cause to the integrity of the hydrophobic cores themselves. This effect is proportional to the size of the core relative to the size of the intervening group. When the hydrophobic cores of large proteins are disrupted, they are left more intact than those of small proteins. Thus, for cases with prosthetic groups one could anticipate success for large proteins but not for small proteins. Our function likewise may struggle when challenged by small proteins rich in disulfide bonds.

The second group is primarily composed of oligomeric proteins. While a number of these form independently stable monomers, the majority are, to varying degrees, integral parts of large multi-component assemblies. Consistent with the hydrophobic nature of the surfaces involved in such oligomeric interaction, these proteins also disobey the ''hydrophobic in, polar out'' generalization (Miller, 1989; Janin & Chothia, 1990). While not necessarily oligomeric, complexes between protease inhibitors and their targets also depend on hydrophobic interactions over relatively large interfaces (Young *et al*., 1994). As with oligomeric assemblies, our potential should not score these particularly well without prior information concerning the existence and location of such hydrophobic interfaces at the surface of the monomers. Small proteins in this class are also expected to score less well than larger proteins because of their higher surface area to volume ratio. It is also worth noting that a few members of this second group have extended or non-compact structures. Such folds are particularly problematic (Maiorov & Crippen, 1992).

The third group is lipophilic, e.g. transmembrane proteins, lipid-binding proteins and phospholipases. Because of their extensive interaction with lipids, they too disregard the ''hydrophobic in, polar out'' rule. Hence, our potential is not expected to apply to these cases.

Results

Our discrimination function was applied to the complete set of representative structures from the Protein Data Bank compiled by Orengo *et al*. (1993). This compilation includes 208 chains with less than 35% sequence identity with one another. We excluded 12 structures determined by nuclear magnetic resonance spectroscopy and also the longest chain (8acn) as it could not be threaded onto any of the other shorter chains (Table 1). All the sequences of the remaining 195 chains were used to test the ability of our method to recognize native folds by threading onto a set of decoy chains. This set of decoys, which comprises the 107 structurally most dissimilar folds in this set of 195 as identified by Orengo *et al*. (1993), is indicated by asterisks in Table 1.

Because only backbones equal to or greater in length than the test sequence are eligible for serving as decoy folds, the number of available alternative conformations decreases with chain length. Hence, the distribution of scores for long polypeptides may be insufficiently large for accurate determination of the variances. We have found that in these cases, the variances are larger than they would be in a normal distribution so that our computed Z_t-scores would err on the conservative side. Because the longest sequences do not present our potential with a sufficiently challenging number of decoys, we also insist that for these sequences the Z_t score be acceptably high in addition to the native fold being the lowest in energy. We designate $Z_t = -5.0$ as a threshold for effectiveness for sequences that do not have at least 2500 admissible decoy conformations. The corresponding odds of spurious recognition are 1 in about 3,000,000, assuming a normal distribution.

We have separated the entire set of 195 chains into two categories, those that are expected to score well (set A) and those that could cause our function difficulty (set B). The classification for each chain is

Table 1. The two sets of proteins used in this work

Set A				Set B		
*1ace	*1phh	*2pmg:A	4icd	1bbp:A	*1utg	*3ebx
1acx	1pii	2psg	4mdh:A	1bmv:1	*1ycc	3hla:B
1ald	*1pyp	*2reb	*4pfk	1bmv:2	*256b:A	3sdh:A
1col:A	*1r69	2scp:A	4ptp	*1bov:A	2ccy:A	*3sdp:A
*1csc	*1rhd	2sga	5cpa	1cc5	*2cdv	3sgb:I
1cse:E	1rnb:A	*2taa:A	*5p21	1cd8	2ci2:I	451c
1ctf	*1rnh	2trx:A	5rub:A	1cob:A	*2cy3	*4bp2
1ecd	*1rve:A	*2ts1	*6ldh	*1crn	2fxb	4sbv:A
1etu	*1sgt	*2tsc:A	*6tmn:E	*1cse:I	*2hip:A	5fd1
*1f3g	1snc	*2yhx	6xia	1fdx	2hmz:A	5hvp:A
*1fbp:A	1trb	*3adk	*7aat:A	1fia:A	2lhb	*5pti
1fcb:A	*1ubq	*3bcl	7api:A	1fxd	*2ltn:A	*5rxn
*1fkf	*1vsg:A	*3blm	*7rsa	1fxi:A	2mev:1	*8atc:B
*1fnr	1wsy:A	3cd4	8abp	1gpl:A	2mev:2	
*1gd1:O	*1wsy:B	3chy	*8adh	*1hge:B	2mev:3	
1gky	2alp	*3cox	*8atc:A	*1hoe	*2ovo	
*1gmf:A	2aza:A	3dfr	*8cat:A	1hrh:A	*2pab:A	
1gox	*2ca2	*3gap:A	8dfr	1ith:A	2plv:1	
*1gst:A	*2cna	*3gly	8ilb	1le2	2plv:2	
1hdd:D	*2cpk:E	*3grs	*9pap	1lmb:A	2plv:3	
*1hge:A	*2cpp	*3hla:A	*9rnt	*1msb:A	*2por	
1ifc	*2cyp	*3lzm	*9wga:A	*1pi2	*2rhe	
*1ipd	*2er7:E	*3pgk		*1prc:C	2rsp:A	
*1lap	2fb4:H	*3pgm		*1prc:H	2sar:A	
*1lfi	2fcr	*3tim:A		*1prc:L	*2sic:I	
1lz1	2fx2	*4cla		1prc:M	*2sn3	
1mba	*2gbp	*4cpv		*1rbp	2snv	
*1mbc	*2gcr	*4dfr:A		1rop:A	*2stv	
*1nsb:A	*2gls:A	*4enl		1tgs:I	2tbv:A	
*1ova:A	2lh3	*4fgf		1thb:A	*2tmv:P	
*1paz	2liv	4fxn		1tnf:A	2wrp:R	
*1pgd	2pcy	4icb		*1tpk:A	*3b5c	

Notation: A four-letter Protein Data Bank (Bernstein *et al*., 1977) code followed by a chain identifier (if applicable), separated by a colon.

Set A, 118 Protein chains expected to score well according to our method.

Set B, 77 Protein chains with disulfide bonds ($\leqslant$100 residues), protein chains with prosthetic groups ($\leqslant$120 residues), oligomeric proteins ($\leqslant$200 residues), viral coat proteins, protease inhibitors and lipophilic proteins.

The backbones of the proteins marked with asterisks form a structurally diverse subset and were used as threading decoys. An additional entry, 8acn, was used in the pool of threading decoys but not in the threading or sequence permutation trials.

also listed in Table 1. The latter category is composed of the three classes of proteins described in the previous section. Specifically, we designate the potentially problematic proteins to be those with disulfide bonds of length 100 or shorter, those with prosthetic groups of length 120 or shorter, and those that form oligomeric complexes of length 200 or shorter. In addition, we include viral coat proteins, protease inhibitors, and lipophilic proteins of all sizes.

Figure 2(a) depicts the Z_t values for the sequences ordered by size. The average Z_t-score over all 195 sequences was –11.9. The difference in Z_t-score between the two categories of proteins is marked. Histograms of the Z_t-scores for set A and set B are found in Figure 2(b) and (c), respectively. The average Z_t for the 118 proteins in set A was –13.8. In contrast, the 77 proteins in set B had a mean Z_t-score of –8.8.

Our function was very successful in distinguishing the native fold for the proteins in set A: 117 out of 118 native folds (99%) were recognized correctly. The only exception was leghemoglobin, a heme-binding protein of 153 amino acid residues, whose native conformation placed second out of 12,421 alternative folds. It was classified as a set A protein because of our stringent size restrictions for placement in set B. As expected, our potential was less successful for set B, but nearly 64% (49 out of 77) of the native folds were still correctly recognized.

Taking sets A and B together, 166 of the 195 native conformations (85%) had the lowest ''energy'' relative to their respective pools of non-native conformations. While the native fold was not the lowest in energy for 29 cases, the correct conformation always placed near the best of the score-sorted list: 23 were better than 99% of the scores and five of the remaining six were better than 90%. A representative histogram of *HF* scores generated by threading a sequence on a set of alternative folds is shown in Figure 3 (turkey egg-white lysozyme, PDB entry 1lz3).

Each of the 195 sequences was tested against a set of at least 2500 decoy conformations obtained from our library of backbones, except for 30, which were too long to be threaded on this number of folds (the smallest pool of alternative folds was 64). Our function did recognize the respective native fold for these 30 sequences, albeit from a smaller set of

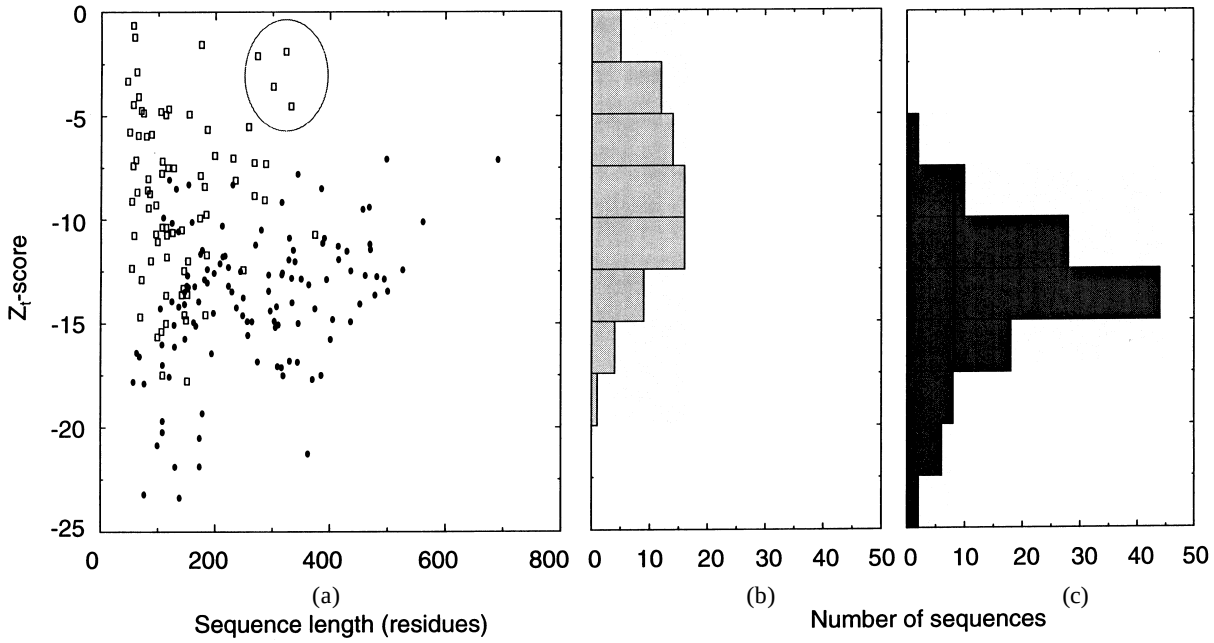

Figure 2. Plot of Z_t-scores *versus* sequence length. The Z_t-score is the number of standard deviations the hydrophobic fitness (*HF*) score of the native fold departs from the mean *HF* score for the entire pool of conformations for a given sequence. (a) The Z_t-scores of the native folds of all 195 sequences, each sharing less than 35% identity with the others, plotted as a function of the sequence length. The set of sequences is divided into two classes: those expected to adhere closely to the ''hydrophobic in, polar out'' generalization (set A), and those that are not (set B). The latter set, depicted as open squares, is composed of all lipophilic proteins, protease inhibitors, viral coat proteins, proteins with disulfide bonds (of length $\leqslant$100 residues), proteins with prosthetic groups (of length $\leqslant$120 residues) and oligomeric proteins (of length $\leqslant$200 residues). Three components of the photosynthetic reaction center and porin (lipophilic membrane proteins) are encircled. Set A is drawn as filled circles. (b) Histogram of Z_t-scores for the 77 chains from set B. The average Z_t-score is −8.8. (c) Histogram of Z_t-scores for the 118 chains in set A, which are expected to perform well by our method. The average Z_t-score is −13.8.

admissible conformations. These were regarded as successful cases, however, since all had Z_t-scores lower than −5.0.

To confirm that our method is capable of matching a native sequence to a given fold, we calculated

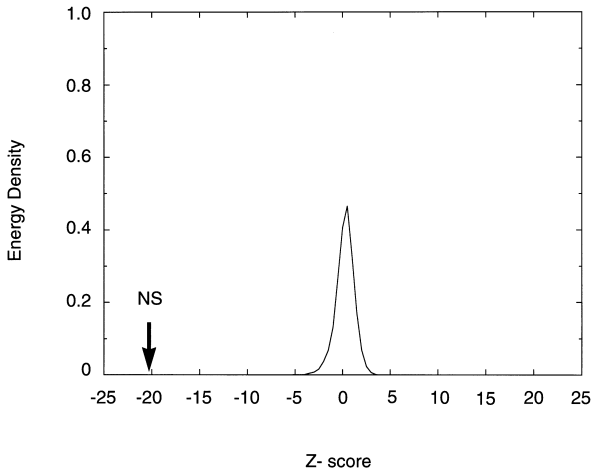

Figure 3. Histogram of hydrophobic fitness (*HF*) scores, expressed in units of standard deviation. The native sequence of turkey egg-white lysozyme (PDB entry 1lz3) was threaded onto a set of 14,210 different folds. The departure of the native score (NS) from the mean score of the misfolded structures was over 20 standard deviations.

Z_p-scores by permuting the sequences for each of the 195 folds (Table 1). The average Z_p-score for the 195 motifs was −9.0. Again, the folds in set A outperformed those in set B, with mean Z_p values of −10.6 and −6.6, respectively. In 172 cases (88%), the native sequence scored higher than its respective set of 2500 scrambled sequences. Out of the 23 native sequences that did not yield the best score, 17 placed in the 99th centile and four placed in the 90th centile. A plot of Z_p-scores against motifs (ordered by size) is shown in Figure 4; data from set A and set B are depicted with different symbols.

Discussion

Discrimination from a simple hydrophobic potential

Our results clearly demonstrate that proper use of information regarding the favorable arrangement of hydrophobic residues is sufficient to design a contact energy function that recognizes a sequence for a given fold and the fold for a given sequence. We show this for a collection of 195 proteins representing a diversity of sequences and folds. For a typical protein in set A, the score of the native ordering of polar and non-polar residues is over ten standard deviations more favorable than the mean score for its set of 2500 randomly ordered sequences. Moreover, it was

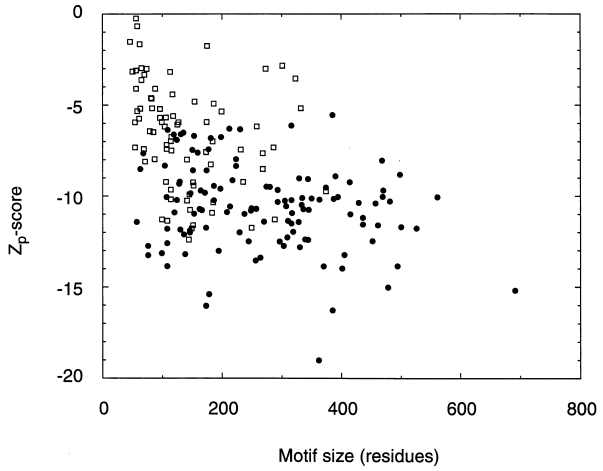

Figure 4. Plot of Z_p-scores *versus* chain size. The Z_p-score is the departure, in standard deviation units, of the hydrophobic fitness (*HF*) score of the native sequence from the mean *HF* score for 2500 randomly permuted sequences for a given fold. The Z_p-scores of the native sequences of the 195 fold motifs are plotted against the motif size (in residues). For 172 folds the native sequence scored better than any other in its respective set of 2500 randomly permuted sequences. The Z_p-scores of the 77 cases in set B are shown as open squares, those for the other 118 (set A) are depicted as filled circles.

similarly successful in fold recognition: the native structure was clearly identified in 117 out of 118 trial sequences (99%) in set A. Larger proteins tended to have better Z-scores, except the longest chains, which did not have sufficient threading decoys to estimate accurate variances.

More than half of the 29 polypeptides for which our method failed were shorter than 100 residues. The function does not perform as well for the shorter polypeptides because they often fall into the class of proteins for which the ''hydrophobic in, polar out'' model does not suffice. The difference in Z_t-score between set A (proteins expected to score well) and set B (the exceptional proteins) is striking. When none of these complicating factors is present, the function does extremely well, even with small proteins such as the homeodomain (PDB entry 1hdd:D, 57 residues; $Z_t = -17.8$). One may counter the problems associated with chains in oligomeric assemblies by repeating the threading or the sequence permutation tests for the entire multi-chain complexes. After having done so, we found that our potential then gave excellent discrimination for both tests (data not shown). Of the 29 sequences for which the native fold was not recognized, 28 were classified as set B. The lone exception was a protein with complicating factors but larger than the size cutoff for set B. In short, our method finds the native conformation for those proteins that are expected to have prominent and intact hydrophobic cores.

We conclude from our threading trials that our quasi-energy function can select the native fold with excellent confidence from a large set of decoys for the sequences found in set A. More significantly, when the potential succeeded for these proteins, it did so convincingly: on average the native *HF* score was nearly 14 standard deviations better than the mean score of the misfolded structures. Even though the results for set B are less compelling, nearly two-thirds of the native folds for these sequences were recognized.

Like many other discrimination functions, the contribution of a given residue to the overall energy is computed by considering the interactions that it makes with neighboring residues. These traditional pairwise potentials simply sum up the energies of residues in contact (Hinds & Levitt, 1992, 1994; Bryant & Lawrence, 1993). While developing our potential, we had evaluated such a function: the contacts between hydrophobic residues were summed without regard to burial. The average *Z*-score for the 195 test cases was −8.5. Other potentials sample the overall environment (e.g. burial) of every residue without explicitly considering pairwise interactions (Lüthy *et al*., 1992). To estimate the effectiveness of solvation terms alone, we scored the 195 folds by our measure of burial. The average *Z*-score was −3.6. In contrast, our hydrophobic fitness function considers these two factors concomitantly and yields a significantly higher average *Z*-score of −13.8. While existing potentials typically select structures with hydrophobic residues merely contacting each other or those that simply partition non-polar residues away from solvent, our method seeks a distinct structural feature that is characteristic of native folds, namely, a compact, buried and intact core of non-polar residues. It is this special aspect of our method that confers the higher degree of discrimination power in assessing sequence-structure complementarity.

Comparison with other methods

There is already established work on native fold recognition using contact potentials such as those of the groups of Sippl (Hendlich *et al*., 1990) and Crippen (Maiorov & Crippen, 1992). In their earliest work, these methods were tested on a common set of 64 proteins with less than 200 residues; each sequence was threaded upon a large number of decoy backbone segments. For each sequence, the energy of the native fold was compared directly with the lowest energy in a set of non-native folds. Recognition was deemed successful if the native fold had the lower energy of the two. Accordingly, the effectiveness of the method was judged by the difference between the two values. However, there is a drawback in simply rating the potentials by the ability to find the native fold amongst an arbitary pool of decoys. Whether or not the function successfully recognizes the native fold depends on the number of alternative backbone conformations: the more decoys, the lesser the margin in energy between native and best non-native folds (and the greater the chance of failure). Consequently, long sequences cannot be compared on the same basis as shorter sequences; hence, they restricted the length

of the test proteins. More recently, the capability of potentials to recognize native folds has not been judged on this basis but by *Z*-scores (see the review by Sippl, 1995). Later work by Sippl (1993) has ingeniously threaded test sequences on a ''polyprotein'', which is an extremely long continuous backbone constructed by covalently joining many known backbone motifs together. The use of a ''polyprotein'' ensures that even longer sequences may be subjected to $\sim$50,000 alternative conformations and an accurate estimate of variance may thereby be obtained. It should be stressed that while using *Z*-scores is robust with respect to the number of alternative structures available for a sequence, a good *Z*-score does not guarantee that the native fold scores better than all other folds in the pool.

A high level of discrimination is a feature that should not be understated. For the majority of proteins, the presence of an intact hydrophobic core is expected. For all these cases, the separation in energy between the native and non-native folds is quite large. This is perhaps a more useful measure of the effectiveness of a function. Mere recognition of the native fold amongst a pool size of several thousand is impressive, but when challenged with millions of possible alternatives, greater discrimination power is required. When a potential faces the formidable task of selecting a native fold from an exhaustively generated set of discrete alternatives, the number of decoys can be of the order of tens of millions. It remains to be seen whether our potential can indeed render such a problem more tractable by virtue of its consistently high level of discrimination.

Because every method employs different approaches and computations in deriving scores, we find that the most objective means of comparing the effectiveness of different energy functions is also by *Z*-scores. The knowledge-based potential of mean force of Sippl has been tested in this manner, but the absence of *Z*-score evaluations for other threading trials precludes reasonable comparisons with these methods (Maiorov & Crippen, 1992; Bryant & Lawrence, 1993; Bauer & Beyer, 1994). The most recent work by Sippl indicates that the average *Z*-score obtained by that method is –9.66 (Sippl, 1995; see also references therein). One example was cited: the native conformation of turkey egg-white lysozyme (PDB code 1lz3) received a Z-score slightly better than –8.0. Scored by our function, the same fold received a Z-score of –20.2 (Figure 3). It is interesting to note that the recent work by Sippl reports that both residue-residue contacts and protein-solvent contacts were used in conjunction to maximize the average *Z*-score; either component alone does not achieve a similarly high level of discrimination. This is consistent with our present findings, where both pairwise contacts among hydrophobic residues and a burial term are considered together for our most favorable results. Given that our simple hydrophobic potential achieves overall better *Z*-scores (–9.7 *versus* –11.9) than that of Sippl, two implications are suggested.

First, the relative complexity inherent in existing knowledge-based methods with their dependence on large sets of statistical information collected from native protein structures, might not be necessary to achieve the observed high degree of fold recognition for most proteins. Simple consideration of the arrangement of the hydrophobic residues seems to be sufficient for most proteins. From the perspective of information content, one may think of our potential as requiring only 23 parameters. The first two parameters are the radii we define, one for enumerating the contacts between residues (7.3 Å), and the other to compute burial (10 Å). In addition, we set the hydrophobic or polar nature of 19 amino acid residues (seven hydrophobic, 12 polar), and recognize the special dual identity of tyrosine (an additional two pieces of information), for a total of 23 parameters, most of which are single bits of information and obvious from elementary physical chemistry. Importantly, it also has an inherent ''cooperative mechanism'' that rewards compact structures by promoting clustering and burial of hydrophobic residues. In contrast, other potentials may use far more parameters in order to achieve a degree of success in fold recognition no better than ours. For example, the method of Sippl utilizes interaction energies for all 20 amino acid residues (210 parameters) at 15 sequence separations, for a total of 3150 histograms as a function of C^{β}-C^{β} distance. Obviously, since these parameters are derived from the hundreds of known crystal structures of polypeptide chains, whereas our 23 parameters are not, a direct comparison would not be equitable. Nevertheless, it seems that the vast information harnessed by such statistical energy functions is largely summarized by the 23 parameters we show to be sufficient. Furthermore, the exhaustive nature of these potentials exploits all or nearly all of the available information regarding protein structure (e.g. knowledge of the chemical properties of residues or statistics derived from the database), leaving no information to harness for improving these potentials. By reducing the information to 23 key parameters, our work allows all the other information to be explored independently. It is hoped that such studies will identify other key parameters that can be combined with our present potential for increased selectivity. Finally, it is worth noting that, unlike other methods, ours is neither susceptible to the bias inherent in choosing a set of chains from which to compile the potentials nor to the problem of data scarcity.

Second, the significance of our straightforward energy function is related to the recent results of Hecht and co-workers. They showed that after placing randomly chosen hydrophobic residues at their designed core positions and polar residues at the intended surface locations, approximately half of the astronomically large sequence permutations could be expected to fold into some native-like compact structure, albeit without taking into account the internal packing of side-chains. While it is unclear at this stage whether or not these polypeptides

have well-packed cores, it is likely that a number of these may be molten globules that have topologies and secondary structure approximating a native fold. In any case, their work suggests that many different sequences that have the same ''binary code'' of hydrophobic and polar residues could indeed be predisposed to assume the same overall motif in solution. From this perspective, our successful native fold recognition using a similar, but computational, assay for a binary pattern of residues, while not necessarily implying the conclusions of the experimental work, seems consistent with it. Our work shows that a binary pattern of hydrophobic and polar residues alone can recognize the actual fold from the rather small library of currently known topologies. Unlike the experimental work, it cannot make any claims about the fraction of such similarly encoded sequences that will actually assume the expected fold. Even without considering the whole issue of the internal packing of side-chains, our equal treatment of the seven hydrophobic residues, each in reality unique in its shape, propensities, and properties, prevents our potential from predicting either the stability of an individual sequence or its ability to fold, even when the desired binary pattern of residue polarity is obeyed. Moreover, our work cannot and does not suggest that (1) simple hydrophobic considerations alone are sufficient to fold a protein or (2) that other considerations such as electrostatic interactions, hydrogen bonds, or even the presence of chaperones are unnecessary. Our results merely suggest that native folds can be selectively recognized by exploiting the structural feature of hydrophobic cores characteristic of globular proteins.

In conclusion, both the computational approach presented here and the experimental results from Hecht *et al*. addressed the following question: ''How much of the overall fold topology is defined in terms of the classic arrangement of hydrophobic and polar residues?'' Both suggest that this consideration alone achieves much more than might have been expected by theorists or experimentalists alike: the selection of a target fold topology from a large number of alternatives.

Acknowledgements

The authors thank Peter David, Mark Gerstein, David Hinds, Doug Laurents, Kenneth Ng, Britt Park and Jack Yu for their helpful discussion and Judy Chen for her careful reading of the manuscript. E.S.H. is a National Science Foundation Pre-doctoral Fellow; M.L. acknowledges support from grant GM 41455 from the National Institutes of Health. S.S. acknowledges partial support from the Department of Energy through the Stanford Synchrotron Radiation Laboratory.

References

Argos, P. (1987). Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis. *J. Mol. Biol.* **197**, 331–348.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.

Bauer, A. & Beyer, A. (1994). An improved pair potential to recognize native protein folds. *Proteins: Struct. Funct. Genet.* **18**, 254–261.

Baumann, G., Frömmel, C. & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng.* **2**, 329–334.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Jr, Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Betz, S. F. (1993). Disulfide bonds and the stability of globular proteins. *Protein Sci.* **2**, 1551–1558.

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990a). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins: Struct. Funct. Genet.* **7**, 257–264.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990b). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**, 1306–1310.

Bryant, S. H. & Amzel, L. M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. *Int. J. Pept. Protein Res.* **29**, 46–52.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92–112.

Casari, G. & Sippl, M. J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725–732.

Chiche, L., Gregoret, L. M., Cohen, F. E. & Kollman, P. A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl Acad. Sci. USA*, **87**, 3240–3243.

Cohen, B. I., Presnell, S. R. & Cohen, F. E. (1993). Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci.* **2**, 2134–2145.

Covell, D. G. & Jernigan, R. L. (1990). Conformations of folded proteins in restricted spaces. *Biochemistry*, **29**, 3287–3294.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.

Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993). Cooperativity in protein-folding kinetics. *Proc. Natl Acad. Sci. USA*, **90**, 1942–1946.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.

Eriksson, A. E., Basse, W. A., Zhang, X.-J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**, 178–183.

Feng, Y., Sligar, S. G. & Wand, A. J. (1994). Solution structure of apocytochrome b_{562}. *Nature Struct. Biol.* **1**, 30–34.

Hecht, M. H., Sturtevant, J. M. & Sauer, R. T. (1984). Effect of single amino acid replacements on the thermal stability of the NH_2-terminal domain of phage λ repressor. *Proc. Natl Acad. Sci. USA*, **81**, 5685–5689.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The

calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167–180.
Hinds, D. A. & Levitt, M. (1992). A lattice model for protein structure prediction at low resolution. *Proc. Natl Acad. Sci. USA*, **89**, 2536–2540.
Hinds, D. A. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* **243**, 668–682.
Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 16027–16030.
Kabsch, W. & Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Natl Acad. Sci. USA*, **81**, 1075–1078.
Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & 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. *Advan. Protein Chem.* **14**, 1–63.
Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
Lau, K. F. & Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, **22**, 3986–3997.
Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
Lee, C. & Levitt, M. (1991). Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature*, **352**, 448–451.
Lee, C. & Subbiah, S. (1991). Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* **217**, 373–388.
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.
Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.
Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of λ repressor. *Nature*, **339**, 31–36.
Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.
Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.
Miller, S. (1989). The structures of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng.* **3**, 77–83.
Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
Novotny, J., Bruccoleri, R. & Karplus, M. (1984). An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol.* **177**, 787–818.
Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins: Struct. Funct. Genet.* **4**, 19–30.
Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.
Ouzounis, C., Sander, C., Scharf, M. & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.* **232**, 805–825.
Park, B. & Levitt, M. (1995) The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493–507.
Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of haemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13**, 669–678.
Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.
Reidhaar-Olson, J. F. & Sauer, R. T. (1988). Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science*, **241**, 53–57.
Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**, 834–838.
Sali, D., Bycroft, M. & Fersht, A. R. (1991). Surface electrostatic interactions contribute little to stability of barnase. *J. Mol. Biol.* **220**, 779–788.
Serrano, L., Kellis, J. T., Jr, Cann, P., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **224**, 783–804.
Shortle, D. & Meeker, A. K. (1986). Mutant forms of staphylococcal nuclease with altered patterns of guanidine hydrochloride and urea denaturation. *Proteins: Struct. Funct. Genet.* **1**, 81–89.
Shortle, D., Stites, W. E. & Meeker, A. K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **29**, 8033–8041.
Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.
Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct. Funct. Genet.* **17**, 355–362.
Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.
Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.
Sweet, R. M. & Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* **171**, 479–488.
Tanaka, S. & Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**, 945–950.
Tanford, C. (1980). *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd edit. Wiley, New York.
Viswanadhan, V. N. (1987). Hydrophobicity and residue-residue contacts in globular proteins. *Int. J. Biol. Macromol.* **9**, 39–48.

Wertz, D. H. & Scheraga, H. A. (1978). Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules*, **11**, 9–15.

Wodak, S. J. & Rooman, M. J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.* **3**, 247–259.

Yue, K. & Dill, K. A. (1992). Inverse protein folding problem: designing polymer sequences. *Proc. Natl Acad. Sci. USA*, **89**, 4163–4167.

Young, L., Jernigan, R. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717–729.

Edited by F. E. Cohen

(Received 13 March 1995; accepted 18 July 1995)

Note added in proof: We would like to bring to attention the noteworthy results of Wodak and co-workers, who have approached the problem of fold recognition using statistically derived potentials, including those that are based on backbone dihedral angle preferences (Kocher *et al.* (1994) *J. Mol. Biol.* **235**, 1598–1613). They have also found that combining such potentials with a term reflecting solvent effects improved their recognition performance. Their best performance with such a combined function was an average *Z*-score of –9.7, similar to that of Sippl (1995).