J. Mol. Biol. (1996) **258**, 367–392

JMB

Energy Functions that Discriminate X-ray and Near-native Folds from Well-constructed Decoys

Britt Park* and Michael Levitt

Beckman Laboratory for Structural Biology Department of Structural Biology, Stanford School of Medicine, Stanford CA 94305, USA

**Corresponding author*

This study generates ensembles of decoy or test structures for eight small proteins with a variety of different folds. Between 35,000 and 200,000 decoys were generated for each protein using our four-state off-lattice model together with a novel relaxation method. These give compact self-avoiding conformations each constrained to have native secondary structure. Ensembles of these decoy conformations were used to test the ability of several types of empirical contact, surface area and distance-dependent energy functions to distinguish between correct and incorrect conformations. These tests have shown that none of the functions is able to distinguish consistently either the X-ray conformation or the near-native conformations from others which are incorrect. Certain combinations of two of these energy functions were able, however, consistently to identify X-ray structures from amongst the decoy conformations. These same combinations are better also at identifying near-native conformations, consistently finding them with a hundred-fold higher frequency than chance. The fact that these combination energy functions perform better than generally accepted energy functions suggests their future use in folding simulations and perhaps threading predictions.

Keywords: reduced representation model protein energy discrimination

Introduction

The key roadblock in the way of the successful prediction of the conformations of proteins *ab initio* is the lack of reliable ways to distinguish correct from incorrect folds. The literature abounds in empirical energy functions (reviewed by Sippl, 1995; Wodak & Rooman, 1993). Many of these have been generated for ''threading'' applications, where the goal is to distinguish the best possible conformation of a protein from among a collection of known protein conformations (Bauer & Beyer, 1994; Bowie *et al*., 1991; Bryant & Lawrence, 1993; Casari & Sippl, 1992; Godzik *et al*., 1992; Hendlich *et al*., 1990; Huang *et al*., 1995; Jones *et al*., 1992; Maiorov & Crippen, 1992; Ouzounis *et al*., 1993; Sippl & Weitckus, 1992; Sippl *et al*., 1994). These have been partially successful at achieving their goals. Others have been designed for *ab initio* folding and have sometimes been used successfully to fold small helical proteins and peptides (Bowie & Eisenberg, 1994; Covell, 1992; Dandekar & Argos, 1994; Gunn *et al*., 1994; Kolinski & Skolnick, 1994; Levitt, 1976, 1983; Levitt & Warshel, 1975; Skolnick *et al*., 1993; Sun, 1993; Vieth *et al*., 1994, 1995; Wallqvist & Ullner, 1994; Wilson & Doniach, 1989). There is no guarantee, however, that energy functions which are good at distinguishing correct from incorrect conformations in a threading context will be useful for the *ab initio* problem or that energy functions devised for *ab initio* studies will do well in a threading context.

The most immediate question is whether an energy function exists that is able to distinguish the correct native fold from all other possible alternatives. It is often assumed that the native structure is at a global free-energy minimum (Anfinsen, 1973). Even if this is so, it is not very helpful in devising a discrimination function; the free energy is not a simple function of atomic positions, depending, as it does, on the extent of motion and degree of order. The potential energy is a function of atomic position and is an important component of the free energy, and it is often assumed that correct folds should have low values of the potential energy.

The efficiency with which the potential energy can be calculated depends strongly on the degree of detail used in studying a particular system. At the most detailed level are quantum mechanical

Abbreviation used: cRMS, coordinate root-mean-square deviation.

0022–2836/96/170367–26 $18.00/0

treatments, which, although they are physically sound, are computationally infeasible for systems the size of proteins. At the next level are all atom empirical energy functions which, except for small peptides and systems in which the structure of a target protein is already known to a close approximation, are too computationally complex to be useful for structure prediction. At the level of least detail are models which treat proteins as simple chains of interacting centers, each of which represents anywhere from two residues to parts of a single residue. (Covell, 1992; Covell & Jernigan, 1990; Dandekar & Argos, 1994; Hinds & Levitt, 1992, 1994; Kolinski & Skolnick, 1994; Levitt, 1976; Park & Levitt, 1995; Rooman *et al*., 1991; Skolnick *et al*., 1993; Sun, 1993; Wallqvist & Ullner, 1994; reviewed by Wodak & Rooman, 1993).

Energy parameters for these kinds of low-resolution models, with which this paper is concerned, have been derived in several ways. Levitt (1976) generated potentials of mean force by averaging energies over all relative orientations of pairs of side-chains. More recently these kinds of energy functions have been derived as potentials of mean force from the ever-growing database of known protein structures (see the references in Sippl, 1995). Huang *et al*. (1995) have devised a potential which does not explicitly use the database of known structures; they use only a simple classification of different residues as hydrophobic or hydrophilic, reminiscent of the theoretical energy models of Dill *et al*. (reviewed by Dill *et al*., 1995; Yue & Dill, 1995). Maiorov & Crippen (1992) generated a potential function by an optimization procedure which sought to maximize the difference in energy between correct and incorrect protein conformations.

These simplified potentials have generally been tested by trying to predict the observed native structure of a protein from its amino acid residue sequence. This can be done in two ways. (1) In folding simulations, the conformation of the polypeptide chain is changed to get arrangements with the lowest possible potential energy values using energy minimization (Levitt & Warshel, 1975), molecular dynamics (Levitt, 1983), Monte Carlo techniques (Covell, 1992; Gunn *et al*., 1994; Kolinski & Skolnick, 1994; Skolnick *et al*., 1993; Vieth *et al*., 1994,1995; Wilson & Doniach, 1989) or genetic algorithms (Bowie & Eisenberg, 1994; Dandekar & Argos, 1994; Unger & Moult, 1993; Sun, 1993). For any of these methods to work, it is clear that the desired native conformation must have a lower value of the potential energy than all other accessible states. (2) In inverse folding experiments, a known three-dimensional structure is used to pick out the sequence that would favor the particular fold. This method is becoming increasingly important as the number of different three-dimensional structures continues to increase rapidly and there are many sequences that seem to prefer very similar conformations (Chothia, 1992; Orengo *et al*., 1994). Rather than change the conformation, it is necessary to thread the sequence through the conformation using templates (Bowie *et al*., 1991), profiles (Ouzounis *et al*., 1993), or pairwise potentials (Hendlich *et al*., 1990; Sippl & Weitckus, 1992; Jones *et al*., 1992; Godzik *et al*., 1993; Bryant & Lawrence, 1993). Again, using this technique, the potential energy function must have the lowest value for the correct matching of structure and sequence.

In most of the *ab initio* studies mentioned above the success or failure of particular methodologies was not generally attributable to either the energy functions, search strategies or models used. Inverse folding studies have looked at energy functions more explicitly but as Kocher *et al*. (1994) have pointed out, the database of known protein conformations is a poor challenge for many empirical energy functions.

In response to this inadequacy of the database approach to energy function testing, work is increasingly turning to alternative methods. In general these other approaches involve the generation of sets of decoy protein conformations. The severity of these tests depends critically on the quality and quantity of the decoys. Decoy structures must: (1) include structures that are close to the native X-ray structure; (2) be native-like in all properties of the real polypeptide chain except the overall folded conformation, otherwise they could easily be distinguished by trivial tests; (3) be diverse so as to sample all possible arrangements and (4) be numerous for more sensitive testing. The sets of incorrect folds generated by Levitt (1983, 1992) were small in number and not sufficiently diverse. Covell & Jernigan (1990) generated all-lattice conformations within a volume shaped like the correct fold of several small proteins (several thousand conformations), and tested a statistically derived contact energy function, finding that the conformation nearest the native was always within the top l% of conformations. At low resolution (5 to 7 Å root-mean-square deviation) Hinds & Levitt (1992, 1994) have generated exhaustive sets of decoys on a diamond lattice. Williams *et al*. (1992) have used small sets of Monte Carlo generated near-native conformations to examine the effectiveness of different solvation models. Monge *et al*. (1995) have used a Monte Carlo algorithm to generate a small number of conformations 4 to 10 Å from X-ray structures to examine the performance of all-atom and reduced-representation energy functions. Wang *et al*. (1995) have used molecular dynamics to generate small ensembles of alternative protein conformations 2.8 Å to 7.8 Å from the X-ray structure, which they used to test various empirical energy functions both all-atom and reduced representation.

Here we generate large ensembles (hundreds of thousands) of test structures for eight small proteins with between 54 and 76 residues. These proteins have a variety of different folds and include all the well-refined structures with less than 80 residues available to us at the start of the work. Our decoys are native-like in that they are all constrained to

have native secondary structure. This technique has been used before to simplify and explore conformational space (Cohen *et al*., 1979; Gunn *et al*., 1994; Monge *et al*., 1994). By using a simple four-state model and exhaustive enumeration of all conformations of a carefully chosen set of loop residues we ensure diversity and proper sampling of conformations. We generate about one million conformations for each protein and then exclude shapes that are not compact or have an excessive number of interpenetrating residues. By repeating the conformational generation with eight different four-state models, we end up with samples of from 35,000 to 200,000 decoys that include several near-native structures within 2.5 Å coordinate root-mean-square deviation (cRMS) and dozens within 4.0 Å of the native structure. With this sample of decoys, we test a variety of simple energy functions that depend individually on residue contacts, solvent-accessible surface area and smoothed inter-residue distance histograms.

Our results show that while none of the individual energy functions can distinguish the native X-ray structure from amongst the decoys, certain simple additive combinations of these functions can. These better-discriminating combinations are also able to find near-native structures, albeit less well than X-ray structures, suggesting that energy functions like those we have used will find applications in folding simulations and, with suitable modifications, to threading attempts.

Methods

Database and test set of well-refined protein structures

In order to parameterize various of the empirical energy functions used in this study we used a database of 232 well-refined protein structures. The proteins are identical to those used by Hinds & Levitt (1994) except that 15 structures were excluded because they contained chain discontinuities, which would have made the compilation of some energy parameters unnecessarily complicated.

In addition we used a set of eight small proteins, not included in the above database, to generate ensembles of secondary structure constrained conformations as described below. Their Protein Data Bank (PDB) designations are 4rxn, 4pti, 1r69, 2cro, 1sn3, 1ctf, 3icb and 1ubq (Bernstein *et al*., 1977). With the exception of 2cro and 1r69 which are homologous but distant in sequence, all of these are structurally distinct. They range in size from 54 to 76 residues.

Discrete-state models

We have used a highly simplified model of protein structure in which the amino acid residue backbone is represented geometrically by a chain of connected C^α atoms. The distance between adjacent C^α atoms is fixed at the average observed value of 3.8 Å. The conformation of the model is specified by (ϕ, ψ) angles for each residue. For the purpose of calculating C^α coordinates these backbone internal coordinates are converted into α and τ angles, the C^α backbone pseudo-dihedral and pseudo-torsion angles (Park & Levitt, 1995).

In addition our model is further simplified by allowing each residue to assume one of only four different conformations, each characterized by different (ϕ, ψ) angles. In this study we use eight different sets of four (ϕ, ψ) states. Each of these state sets was generated using an optimization procedure described previously by Park & Levitt (1995), where the actual (ϕ, ψ) values can be found.

For several of the energy functions we evaluate in this study an additional interacting center located at the C^β atom or at the side-chain centroid is needed for each residue. These we construct geometrically from the C^α coordinates (r_i values) in the following way. We first calculate two unit vectors, **x** and **y** by the relations:

$$\mathbf{x} = \frac{(r_i - r_{i-1}) + (r_i - r_{i+1})}{|(r_i - r_{i-1}) + (r_i - r_{i+1})|}$$

and:

$$\mathbf{y} = \frac{(r_i - r_{i-1}) \times (r_i - r_{i+1})}{|(r_i - r_{i-1}) \times (r_i - r_{i+1})|} \quad (1)$$

We then calculate the position of the side-chain centroid or C^β atom from the relation:

$$r_\beta = l \cos\theta \mathbf{x} + l \sin\theta \mathbf{y} \quad (2)$$

where l is the distance of the C^β atom or side-chain centroid from the C^α atom, and θ is the out-of-plane angle (we used $\theta = 37.6°$). For C^β atoms we used $l = 3.0$ Å, which is almost twice the length of the actual C^α to C^β distance but gives better results for the energy functions used here. For side-chain centroids, l depended on the residue type and was set to the average distance between the side-chain centroid and the C^α atom for that residue type (see l_{sc} in Table 1). Unit vectors **x** and **y** are indeterminate for the first and last residues of a protein and the C^β atom or side-chain centroid coordinates for these were simply set to the C^α coordinates.

Native-like structures with the four-state model

Using four-state models to define the conformation of each residue is efficient in that the number of possible conformations is much reduced compared to models with more states. It is crucially important that the model be of high enough resolution to be able to represent the native structure sufficiently well. Previously we have shown that by using a build-up procedure we can find a near best fit of any discrete state model to an X-ray structure (Park & Levitt, 1995). For the eight different optimized models used here, the mean cRMS value over a sample of 149 well-refined proteins was 2.38 Å. For the eight small test proteins used here the best fits have a lower average cRMS of 1.92 Å. We use these best fit conformations as the starting points for the generation of our ensembles. Thus, in spite of their low complexity, the set of decoys generated with our four-state models will contain some structures that are clearly native-like.

Decoys preserve native secondary structure

The sets of decoys used here are generated by complete enumeration of the four (ϕ, ψ) states for each of a limited set of flexible residues. The small proteins that we deal with here have between 54 and 76 residues; even with our four-state model, there are $4^{54-3} = 5 \times 10^{30}$ possible

Table 1. Contact distances for interacting centers, R_{ij}^{a}, and other geometric data

α	Ala	Val	Leu	Ile	Pro	Asp	Glu	Asn	Gln	Lys	Arg	Ser	Thr	Met	Cys	Tyr	Trp	His	Phe	
3.5	2.7	3.8	4.2	4.2	3.6	3.3	3.8	3.5	3.9	4.3	4.3	2.8	3.4	4.2	3.3	4.6	4.9	4.1	4.5	R_i^s[a]
—	1.5	2.0	2.6	2.3	1.9	2.5	3.1	2.5	3.1	3.5	4.1	1.9	1.9	3.0	2.0	3.8	3.9	3.1	3.4	l_{sc}[b]
5.6	3.7	4.1	4.3	4.3	4.2	3.9	4.1	4.0	4.2	4.4	4.6	3.8	4.0	4.2	3.8	4.5	4.7	4.4	4.4	α
	3.8	4.5	4.7	4.6	4.3	4.2	4.4	4.3	4.5	4.7	4.8	3.9	4.4	4.6	4.2	4.8	5.0	4.5	4.7	Ala
		5.2	5.3	5.3	5.1	4.9	5.0	5.1	5.0	5.3	5.3	4.7	5.0	5.2	5.0	5.4	5.5	5.1	5.3	Val
			5.5	5.5	5.2	5.1	5.2	5.1	5.2	5.4	5.4	4.8	5.2	5.4	5.0	5.5	5.7	5.2	5.5	Leu
				5.6	5.3	5.2	5.2	5.3	5.3	5.4	5.6	4.9	5.3	5.4	5.1	5.6	5.8	5.3	5.5	Ile
					4.9	4.8	4.9	4.9	4.9	5.1	5.3	4.6	4.8	5.0	4.7	5.1	5.2	4.9	5.0	Pro
						4.8	5.0	4.8	5.0	4.9	5.1	4.2	4.7	5.0	4.7	5.6	5.6	5.0	5.4	Asp
							5.4	5.0	5.2	5.1	5.2	4.5	4.8	5.2	4.8	5.5	5.8	5.2	5.4	Glu
								4.9	5.0	5.1	5.4	4.5	4.8	5.2	4.8	5.4	5.4	5.1	5.3	Asn
									5.3	5.2	5.3	4.7	4.9	5.1	4.9	5.4	5.6	5.0	5.3	Gln
										5.6	5.9	4.8	5.1	5.5	5.1	5.3	5.2	4.9	5.3	Lys
											6.0	5.2	5.3	5.5	5.3	5.6	5.3	5.4	5.4	Arg
												4.2	4.5	4.9	4.6	5.2	5.3	4.7	5.1	Ser
													4.7	5.2	4.8	5.5	5.7	4.9	5.4	Thr
														5.1	4.9	5.6	5.7	5.2	5.5	Met
															3.2	5.3	5.2	4.9	5.1	Cys
																5.9	6.2	5.6	5.9	Tyr
																	6.3	5.5	6.0	Trp
																		5.2	5.4	His
																			5.7	Phe

[a] R_i^s values are the radii used for surface area calculations.
[b] l_{sc} values are the distances between α-carbon centers and side-chain centroids.

Table 2. Secondary structure and movable positions for test set proteins

Residue Position[1]

```
Protein    1         11        21        31        41        51        61        71
           |         |         |         |         |         |         |         |
4rxn       -SSSSS---SSSSS-SSSS----SSSS--------------------SSSS--
                     JJ        JJ        JJ        JJ    JJ

4pti       ---------------SSSSSSSSSS--SSSSSSSSS----------HHHHHHHHHH--
                  JJ     JJ          JJ     JJ        JJ

1r69       HHHHHHHHHHHH---HHHHHHH-----HHHHHHHHH--------HHHHHHH----HHHHHH
                       JJJ         JJ            JJJ           JJ

2cro       --HHHHHHHHHHHH---HHHHHHH-----HHHHHHHHH--------HHHHHHH----HHHHHH--
                        JJ          JJ          JJJ           JJ

1sn3       SSSS-----------------HHHHHHHHHH----SSSSS----SSSSS---------------
                     JJ        JJ          JJ        JJ        JJ

1ctf       -SSSSSSS---HHHHHHHHHHHHHH-HHHHHHHHHH---SSSSSSSHHHHHHHHHHHHHHH--SSSSS
                    JJJ           JJ            JJ      JJJ

3icb       -HHHHHHHHHHHHHHH------HHHHHHHHHHHHHHHHH---HHHHHHHHHH-------HHHHHHHHHHHHHH
                          JJJ          JJ[2]      JJ            JJJ

1ubq       SSSSSSS--SSSSSSSS-----HHHHHHHHHHHH-----SSSSSS--SSS-----HHHH----SSSSSSSSS---
                    JJ        JJ              JJ        JJ          JJ
```

[a] For each protein the secondary structures for each residue are indicated by S = β-strand, H = α-helix or - = neither. The hinge residues are marked with J below the secondary structure.
[b] This position is actually between two adjacent α-helices.

conformations for our smallest protein, 4rxn with 54 residues. Clearly, we can only generate decoy structures by changing the conformational states of a subset of the residues. Here we chose to allow only ten flexible residues for each protein so that we have to enumerate 4^{10} = 1,048,576 conformations for each protein. Choice of such a small number of residues must be made with care and satisfy a number of different criteria. (1) At least two successive residues must be flexible allowing at least $4 \times 4 = 16$ conformations for each hinge. (2) The hinge residues must be relatively exposed to solvent so that they can change conformation without disturbing the local structure (a hinge in the middle of an α-helix would cause bad clashes as it moved). (3) The rigid segments between hinges should be as straight as possible so that each hinge moves as large a lever arm as possible. (4) The rigid segments should correspond as much as possible to the regions of α-helix and β-strand secondary structure.

Satisfying all criteria at once is not easy and we have chosen the following scheme; (1) parse the structure into a small number of the most linear segments, (2) merge adjacent segments to have no more than six segments (five hinges), (3) ensure that the hinges correspond to regions between segments of secondary structure.

Parsing a protein into a small number of linear segments can be done rigorously by using the same dynamic programming algorithm commonly used for aligning sequences (Needleman & Wunsch, 1970). Consider a line joining C^{α} positions $i-n$ and i. Each intervening C^{α} position, k from $i-n$ to i, will deviate from this line by the perpendicular distance:

$$d_k = |(r_k - r_i) - (r_k - r_i)[(r_k - r_i)\cdot\mathbf{u}]| \tag{3}$$

where $\mathbf{u}$ is the unit vector along the line:

$$\mathbf{u} = \frac{r_{i-n} - r_i}{|r_{i-n} - r_i|} \tag{4}$$

The total squared deviation of C_{α} positions from this line is $D(i-n, i) = \Sigma(d_k)^2$. If there are several line segments, $D(i-n, i)$ must be summed over them to give the total squared deviation for the entire chain. The problem is to find the set of hinge points $i-n$ and i such that there are the required number of line segments. This is done using an inductive scheme. Define $D_{\min}(i)$ as the smallest possible value of total deviation for residues from 1 to i. The initial value of $D_{\min}(i) = 0$ for $i = 0$ and 1. New values of $D_{\min}(i)$ are calculated iteratively from previous values, $D_{\min}(i-n-1)$, which have been tabulated, as follows:

$$D_{\min}(i) = \min[D_{\min}(i-n-1), D(i-n, i) + \gamma] \tag{5}$$

for $n = 1, i-1$. The penalty value, γ, controls the number of segments. For $\gamma = 0$, every C^{α} position becomes a hinge to give $D_{\min}(i) = 0$ for all i. As γ increases, the cost of making an additional segment increases and will only occur if the fit of the C^{α} chain to the new segment is good enough. For each chain position i, it is also necessary to record the value of j that gives the minimum, $D_{\min}(i)$, making it possible to find all the hinge points by tracing back from $D_{\min}$ at the end of the chain.

Table 3. Movable residues of test proteins

Structure	Length	C_b[a]	Best RMS (Å)	Ensemble Size	Number conformations <[b] 3.5 Å	4.0 Å	4.5 Å
4rxn	54	10	1.63	187,298	48	105	196
4pti	58	13	1.83	179,339	36	62	124
1r69	63	5	1.44	199,943	72	136	301
2cro	63	7	1.46	197,572	187	437	830
1sn3	65	27	1.92	134,456	18	50	97
1ctf	68	25	1.67	159,340	49	93	171
3icb	75	10	1.67	188,767	67	132	224
1ubq	76	59	1.73	35,650	20	48	74

[a] C_b is the average ''bad contact'' cutoff distance used for ensemble generation.
[b] Shown are the number of conformations with RMS deviations less than 3.5, 4.0 and 4.5 Å from the X-ray.

We have used this scheme to divide each of the eight small proteins considered into five or six linear segments separated by four or five hinges that contain a total of ten flexible residues. Table 2 shows how the hinges correspond to the loop regions between secondary structure. All of the flexible hinge residues are located in the regions between or at the ends of secondary structure elements, as was desired. Some loop regions involve many residues and it would not have been obvious how to select hinge points without objective parsing of the chain into linear segments. In fact, early in the study random hinge choices were used. It turned out to be all too easy to make poor choices which yielded ensembles which contained too many conformations with large numbers of steric conflicts, or which had only small numbers of conformations of sufficient compactness.

Enumeration of ensembles

With these sets of flexible or movable residues we are able to generate ensembles of secondary structure constrained conformations. We do not do this by simply specifying idealized secondary structure for each of our test proteins and then enumerating the possible conformations of the loop residues. The secondary structure in actual proteins mostly differs significantly from idealized forms. This is particularly true for small proteins like the ones in our test set. Therefore using idealized secondary structure would yield ensembles that are unlikely to contain significant numbers of conformations sufficiently near the correct structure. Our ensembles for each protein are therefore based on the best fit conformations of the four-state models to the X-ray structure (Park & Levitt, 1995) and take all but ten residue states from these conformations. Exhaustive enumeration of the conformations associated with the ten flexible residues, however, yields ensembles which are still too large ($4^{10} \approx 1{,}000{,}000$ conformations) for convenient analysis, and moreover contain a preponderance of non-compact and therefore uninteresting conformations.

To cut down the ensemble size we apply two filters. The first is a radius of gyration (R_g) cutoff set to be $3n^{1/3}$ Å, where n is the number of residues in a protein. Any conformation whose R_g is greater than this value is discarded. The second is a filter which discards conformations which have greater than C_b bad contacts, where C_b is chosen for each ensemble to generate about 20,000 final conformations. For this purpose bad contacts are those inter-residue contacts nearer than 3.5 Å. Finally we combine all eight ensembles for the eight different four-state models into single large ensembles. Table 3 shows the final number of conformations for each of the eight test proteins.

Calculating the potential energy

Conformational relaxation. As can be deduced from Table 3, the enumerated conformations in our ensembles often have ''bad contacts'' which real protein conformations plainly do not have. In addition to making many enumerated conformations un-protein-like, these ''bad contacts'' degrade the performance of many empirical energy functions. Before evaluating energy functions, therefore, we attempt to ''relax'' the enumerated conformations by a rapid minimization, designed to remove steric conflicts while minimally changing conformation. We use conjugate gradient minimization (Press *et al*., 1988) and a target function that is the sum of two terms:

$$E_{\text{steric}} = \sum_{(1 \leqslant i \leqslant N)} \sum_{(i+2 \leqslant j \leqslant N)} \begin{cases} (r_{ij} - 3{,}5)^2 & \text{if } r_{ij} \leqslant 3.5 \\ 0 & \text{if } r_{ij} > 3.5 \end{cases}$$

and:

$$E_{\text{constraint}} = \sum_{1 \leqslant i \leqslant N} |r_i - r_{0i}|^2 \qquad (6)$$

where r_{ij} is the distance between the C^{α} atoms of residues i and j, r_i is the position of residue i, and r_{0i} is the original unrelaxed position of residue i. Relaxation, using this formulation, takes less than one second per conformation for proteins of 60 to 80 residues on a Silicon Graphics Indigo workstation (MIPS R3000 at 33 MHz), and, except for topologically knotted conformations, removes essentially all bad contacts.

Contact energy functions

The simplest and best known type of energy function that we have tested is a contact potential derived from the database of X-ray conformations. We use two different formulations of this kind of function. Both of our versions of the contact potential use the same model of interacting centers. Each residue consists of two centers, the α-carbon atom and the side-chain centroid. All α-carbon atoms are considered to be energetically equivalent. Different side-chain types are considered distinct. There are therefore 20 different types of interacting centers, 19 amino acid side-chains (glycine has no side-chain) and the

α-carbon. The energy of a conformation for both formulations is:

$$E_{\text{con}} = \sum_{(1 \leqslant i \leqslant N)} \sum_{(i-4 \leqslant j \leqslant N)} \epsilon_{ij}[\text{if } R_b < r_{ij} < R_{ij}^c] + \epsilon_b[\text{if } r_{ij} < R_b] + \epsilon_{\alpha\alpha}[\text{if } R_b < r_{\alpha_i\alpha_j} < R_{i\alpha}^c] + \epsilon_b[\text{if } r_{\alpha_i\alpha_j} < R_b] + \sum_{(1 \leqslant i \leqslant N)} \sum_{((1 \leqslant j \leqslant i-3) \cup (i+3 \leqslant j \leqslant N))} \epsilon_{i\alpha}[\text{if } R_b < r_{i\alpha_j} < R_{i\alpha}^c] + \epsilon_b[\text{if } r_{i\alpha_j} < R_b] \quad (7)$$

where N is the number of residues in the protein, ϵ_{ij} is the contact energy between residue side-chains of the types of residues i and j, $\epsilon_{\alpha\alpha}$ is the contact energy for two α-carbons, $\epsilon_{\alpha j}$ is the contact energy for an α-carbon and a side-chain of the type of residue j, r_{ij} is the distance between the side-chains of residues i and j, R_{ij}^c is the contact cutoff distance for residue side-chains of the types of residues i and j, $r_{\alpha_i\alpha_j}$ is the distance between the α-carbons of residues i and j, $r_{j\alpha_i}$ is the distance between the α-carbon of residue i and the side-chain of residue j, ϵ_b is a bad contact penalty (3.0 in kT units) and R_b is a bad contact cutoff distance (set to 3 Å).

The two versions of the contact potential differ in the way that the ϵ_{ij} values are derived. For our first formulation, referred to as the Contact(HL) function, the energy parameters are derived similarly to those used by Hinds & Levitt (1992, 1994) by the relation:

$$\epsilon_{ij} = -\ln\left(\frac{n_{ij}}{n_{ij\text{exp}}}\right)$$

where:

$$n_{ij} = \sum_p n_{pij} \quad (8)$$

The n_{pij} are the numbers of contacts between interacting centers of type i and j in protein p and $n_{ij\text{exp}}$ is the number of contacts expected from a random distribution of contacts. A particular pair of interacting centers are considered to be in contact if any pair of their constituent atoms are within 4.0 Å of each other (the constituent atoms of the C^α center are the backbone atoms of the residue, C, N, C^α and O). At the same time that contacts are counted, we also calculate the average distances between interacting centers in contact, R_{ij}^a, and use these values, which are given in Table 1, to calculate the contact cutoff distance R_{ij}^c as $1.2R_{ij}^a$. For the backbone the C^α atom and for side-chains the centroid are used for calculating distances. The $n_{ij\text{exp}}$ values are calculated explicitly for each protein p as the expected frequency of type-specific contacts for a randomly mixed set of interacting centers with the same composition and number of contacts as protein p. More specifically, for a given protein and interacting center, i, the expected number of contacts made with other residue types is calculated based on the frequency of different residue types in the protein but excluding residues $i-3$ to $i+3$ for each residue i. The full set of ϵ_{ij} parameters is given in Table 4; the mean value of ϵ_{ij} is –0.522 (kT units).

This first formulation of the contact function takes no account of solvent effects so that a zero total energy value corresponds to a hypothetical randomly mixed compact state. Our second version of the contact function, referred to as the Contact(MJ) function, does, and is modeled after that of Miyazawa & Jernigan (1985). Here, in addition to the 20 types of interacting centers from the proteins, pseudo-solvent molecules are introduced to approximate solvent effects. To calculate the parameters for this energy formulation an effective coordination number q_i is estimated as the average number of contacts, excluding those made by near neighbors, made for each residue type when buried. The q_i values are shown in Table 5. Then the n_{pij} are tabulated the same as for the Hinds & Levitt formulation except that contacts with pseudo-solvents are added; if an interacting center in a database protein makes k contacts with other centers and k is less than q_i for that center, then it also is assumed to form $q_i - k$ contacts with pseudo-solvents. The remaining free parameter for this formulation is an estimate of the effective number of pseudo-solvent molecules for each protein, or equivalently the number of solvent-solvent contacts. In this study we make the simple assumption that there are two effective solvent molecules for each residue of a protein, which is probably an overestimate for small proteins but gives results very similar to those of Miyazawa & Jernigan (1985).

With the set of ϵ_{ij} values, including solvent-residue, ϵ_{wi}, and solvent-solvent, ϵ_{ww} terms, calculated the same way as those for the Contact(HL) function, we then calculated the net energies for each interacting center pair by:

$$\epsilon_{\text{net}ij} = \epsilon_{ij} - \epsilon_{wi} - \epsilon_{wj} + \epsilon_{ww} \quad (9)$$

The full set of $\epsilon_{\text{net}ij}$ values for this formulation are given in Table 5; the mean value of $\epsilon_{\text{net}ij}$ is –3.18, which is much more negative than for the Hinds & Levitt version. Both versions of the contact potential use the same set of R_{ij}^a values (Table 1).

van der Waals energy function

As will be seen in the Results, ''on-off'' contact potential functions have distinct disadvantages for the kind of model used here. Therefore we also developed distance-dependent versions of the contact potentials presented above. These functions, referred to as the VdW(HL) and VdW(MJ) functions are similar in form to the van der Waals energy functions used for interatomic interactions:

$$E = \sum_{(1 \leqslant i \leqslant N)} \sum_{(i+4 \leqslant j \leqslant N)} \frac{A_{ij}}{r_{ij}^8} - \frac{B_{ij}}{r_{ij}^4} + \sum_{(1 \leqslant i \leqslant N)} \sum_{(i+4 \leqslant j \leqslant N)} \frac{A_{\alpha\alpha}}{r_{\alpha_i\alpha_j}^8} - \frac{B_{\alpha\alpha}}{r_{\alpha_i\alpha_j}^4} + \sum_{(1 \leqslant i \leqslant N)} \sum_{((1 \leqslant j \leqslant i-3) \cup (i+3 \leqslant j \leqslant N))} \frac{A_{i\alpha}}{r_{i\alpha_j}^8} - \frac{B_{i\alpha}}{r_{i\alpha_j}^4} \quad (10)$$

The A and B energy parameters are calculated from the ϵ_{ij} and R_{ij}^a parameters generated for the simple contact functions above as:

$$A_{ij} = -\epsilon_{ij}(R_{ij}^a)^8$$

$$B_{ij} = -2\epsilon_{ij}(R_{ij}^a)^4 \quad (11)$$

A functional form similar to this has been used by Wallqvist & Ullner (1994) for their simplified protein potential.

Table 4. Contact energies: Hinds & Levitt (HL) variant (*kT* units)

α	Ala	Val	Leu	Ile	Pro	Asp	Glu	Asn	Gln	Lys	Arg	Ser	Thr	Met	Cys	Tyr	Trp	His	Phe	
0.1	0.7	0.1	0.1	0.0	0.5	0.4	0.6	0.1	0.0	0.4	–0.1	0.4	0.2	–0.1	–0.1	–0.4	–0.7	–0.0	–0.3	α
	0.5	–0.3	–0.4	–0.4	0.6	0.3	0.6	0.3	0.0	1.0	0.2	0.5	0.0	–0.5	0.3	–0.7	–0.8	0.0	–0.8	Ala
		–1.1	–1.2	–1.2	–0.0	0.4	0.0	–0.0	–0.4	0.1	–0.5	0.0	–0.3	–1.0	–0.5	–1.2	–1.6	–0.5	–1.5	Val
			–1.4	–1.4	–0.1	0.0	–0.1	–0.1	–0.6	0.1	–0.6	0.0	–0.3	–1.3	–0.8	–1.4	–1.7	–0.7	–1.6	Leu
				–1.5	–0.1	0.0	–0.2	–0.1	–0.4	–0.0	–0.7	–0.1	–0.6	–1.4	–0.8	–1.4	–1.8	–0.8	–1.7	Ile
					0.1	0.1	0.1	–0.1	–0.3	0.6	–0.2	0.2	–0.0	–0.5	0.0	–1.0	–1.3	–0.4	–0.7	Pro
						–0.0	0.0	–0.6	–0.3	–1.0	–1.4	–0.3	–0.3	0.1	0.0	–1.0	–0.6	–1.1	–0.3	Osp
							0.1	–0.6	–0.4	–1.1	–1.5	–0.2	–0.3	–0.3	0.1	–1.0	–0.8	–1.0	–0.5	Glu
								–0.7	–0.7	–0.3	–0.8	–0.1	–0.4	–0.3	–0.0	–0.8	–0.8	–0.8	–0.6	Osn
									–0.5	–0.4	–0.9	–0.0	–0.5	–0.6	–0.2	–1.1	–1.0	–0.5	–0.8	Gln
										0.7	0.1	0.1	–0.0	–0.1	0.5	–1.0	–0.8	–0.0	–0.4	Lys
											–0.9	–0.4	–0.6	–0.5	0.0	–1.4	–1.3	–1.0	–0.9	Arg
												0.0	–0.2	–0.1	–0.1	–0.6	–0.6	–0.6	–0.4	Ser
													–0.5	–0.6	–0.3	–0.8	–0.9	–0.7	–0.7	Thr
														–1.5	–0.8	–1.5	–2.0	–0.9	–1.9	Met
															–2.7	–0.8	–1.3	–0.6	–1.2	Cys
																–1.6	–1.8	–1.5	–1.7	Tyr
																	–2.2	–1.5	–2.0	Trp
																		–1.6	–1.2	His
																			–2.0	Phe

Table 5. Contact energies: Miyazawa & Jernigan variant (*kT* units)

α	Ala	Val	Leu	Ile	Pro	Asp	Glu	Asn	Gln	Lys	Arg	Ser	Thr	Met	Cys	Tyr	Trp	His	Phe	
3	2	3	4	4	3	3	3	3	4	4	4	3	3	4	4	6	7	5	6	q[a]
−2.1	−2.2	−3.0	−2.5	−2.7	−1.8	−1.9	−1.3	−2.4	−2.0	−1.9	−2.2	−1.9	−2.4	−2.8	−3.0	−2.8	−3.1	−2.1	−2.6	α
	−2.9	−4.1	−3.7	−3.9	−2.3	−2.6	−1.9	−2.8	−2.7	−1.9	−2.4	−2.4	−3.2	−3.8	−3.1	−3.7	−3.8	−2.6	−3.7	Ala
		−5.1	−4.7	−5.0	−3.2	−2.8	−2.8	−3.5	−3.4	−3.1	−3.5	−3.1	−3.9	−4.7	−4.3	−4.5	−4.8	−3.5	−4.6	Val
			−4.3	−4.6	−2.8	−2.7	−2.4	−3.0	−3.1	−2.5	−3.0	−2.6	−3.3	−4.4	−4.0	−4.1	−4.4	−3.1	−4.2	Leu
				−4.9	−3.0	−2.9	−2.7	−3.2	−3.2	−2.9	−3.3	−3.0	−3.8	−4.7	−4.2	−4.4	−4.7	−3.4	−4.5	Ile
					−2.2	−2.3	−1.8	−2.7	−2.5	−1.8	−2.4	−2.2	−2.8	−3.4	−2.8	−3.5	−3.7	−2.6	−3.1	Pro
						−2.6	−2.0	−3.3	−2.7	−3.5	−3.7	−2.8	−3.2	−2.7	−2.9	−3.5	−3.2	−3.3	−2.7	Asp
							−1.5	−2.8	−2.2	−3.2	−3.2	−2.3	−2.7	−2.7	−2.4	−3.0	−2.8	−2.8	−2.4	Glu
								−3.6	−3.2	−3.0	−3.2	−2.8	−3.4	−3.3	−3.2	−3.5	−3.5	−3.2	−3.2	Asn
									−2.6	−2.7	−2.9	−2.3	−3.1	−3.3	−2.9	−3.4	−3.3	−2.5	−3.0	Gln
										−1.7	−2.0	−2.3	−2.8	−2.9	−2.4	−3.5	−3.3	−2.2	−2.8	Lys
											−2.9	−2.6	−3.1	−3.1	−2.6	−3.6	−3.5	−2.9	−3.0	Arg
												−2.4	−2.9	−3.0	−3.0	−3.1	−3.1	−2.7	−2.8	Ser
													−3.5	−3.8	−3.5	−3.6	−3.7	−3.2	−3.4	Thr
														−4.8	−4.2	−4.4	−4.8	−3.5	−4.6	Met
															−6.1	−3.8	−4.3	−3.3	−4.1	Cys
																−4.1	−4.3	−3.7	−4.1	Tyr
																	−4.7	−3.6	−4.4	Trp
																		−3.5	−3.3	His
																			−4.3	Phe

[a] q is the estimated coordination number for each interacting center type.

Surface area energy function

The third type of function that we have used, referred to as the surface area energy function, is a measure of exposed hydrophobic surface area. Here the energy of a conformation is calculated as:

$$E = \sum_{1 \leqslant i \leqslant N} s_i[\text{if residue } i \text{ is Met, Val, Leu, Ile, Phe, Trp, Tyr or Cys}] \quad (12)$$

where s_i is the exposed solvent-accessible surface area of the side-chain of residue i. The s_i values are calculated using the approximate method of Wodak & Janin (1980). Each side-chain is treated as a sphere whose radius is proportional to the square root of the side-chain volume. The proportionality constant was adjusted manually to maximize the difference in surface area between buried and un-buried side-chains in X-ray structures. The values of the radii used, R_i^s, are shown in Table 1. The backbone atoms were treated similarly as single centers located at the α-carbon atom, but only for the purpose of determining the burial of side-chain spheres. Burial of the backbone centers does not affect the conformational energy. This function is very similar in form to one used by Kocher *et al*. (1994), and similar in intent to many energy functions which try to model hydrophobic forces (Huang *et al*., 1995; Chiche *et al*., 1990; Eisenberg & McLachlan, 1986).

Histogram energy function

The last energy function we consider also uses a statistical analysis of known protein structures. Here instead of recovering a single residue-residue interaction energy for each residue pair, a histogram of energies is calculated for different residue separations and for ten different topological levels, which correspond to different separations in sequence between interacting residues. Eight of the ten topological levels correspond to short-range interactions (one to three through one to ten). The ninth corresponds to medium range interactions (one to 11 through one to 50) and the last corresponds to long-range interactions (all others.) The basic form of this function was devised by Sippl *et al*. (1989) and is described therein. We refer to this function as the histogram function, and calculate it as:

$$E = \sum_{(3 \leqslant l \leqslant 10)} \sum_{(1 \leqslant j \leqslant N - l + 1)} e_{l,j,j+l-1}(r_{j,j+l-1}) + \sum_{(1 \leqslant j \leqslant N - 11)} \sum_{((j + 10 \leqslant k \leqslant j + 50) \cap (k \leqslant N))} e_{\text{med},j,k}(r_{jk}) + \sum_{(1 \leqslant j \leqslant N - 51)} \sum_{(j + 50 \leqslant k \leqslant N)} e_{\text{long},j,k}(r_{jk}) \quad (13)$$

where the r_{jk} values are the distances between the β-carbons of residues j and k, and the $e(r)$ are energies tabulated over 20 equally sized distance ranges. The 20 values for each topological level are determined statistically, like the pure contact and continuous contact functions, to give potentials of mean force. In order to calculate these $e(r)$ values, contact distance histograms are tabulated both for individual residue type pairs and for all residues in aggregate over each topological level. In the following, n_{lijr} is the number of occurrences of residues of type i and j on topological level l separated by distance range r. The symbol n_{lr} is the same as n_{lijr} except that it is for pair distances for all residue types in aggregate. Given these definitions the $e(r)$ are calculated as:

$$e_{l,i,j}(r) = -\ln\left(\frac{n_{lijr} + \sigma n_{lr} / n_l}{(n_{lij} + \sigma) n_{lr} / n_l}\right) \quad (14)$$

The energy of interaction of a particular residue pair (i, j) separated by a particular distance r is calculated as a potential of mean force relative to the behavior of all residues in aggregate. The σ in the equation is a weighting factor whose purpose is to compensate for sparse data. Each residue-specific frequency is combined with the corresponding residue non-specific frequency weighted by a factor of σ (50 in this case).

Measures of significance

RMS deviation

To quantify the similarity of different conformations we use the coordinate root mean square deviation (RMS):

$$\text{RMS} = \left(\frac{\sum_{1 \leqslant i \leqslant N} |r_{ai} - r_{bi}|^2}{N}\right)^{1/2} \quad (15)$$

where r_{ai} and r_{bi} are the positions of atom i of structure a and structure b, respectively, and where structures a and b have been optimally superimposed (Kabsch, 1978).

Z-scores

The literature of energy functions used either for the prediction or the identification of protein conformations, commonly measures success using the Z-score (Bowie *et al*., 1991; Bryant & Lawrence, 1993; Godzik *et al*., 1992; Huang *et al*., 1995; Kocher *et al*., 1994; reviewed by Bryant & Altschul, 1995), which expresses the deviation from the mean in units of the standard deviation. The Z-score, Z_i, of a particular conformation with energy E_i is:

$$Z_i = \frac{E_i - \bar{E}}{s} \quad (16)$$

where s is the standard deviation of the energy distribution and $\bar{E}$ is the average energy of that distribution.

Ranking-scores

In any discrimination task, it is also useful to rank the energy of the conformation one is looking for relative to the energies of the entire distribution. Throughout this paper we list the rank scores for different energy function and protein conformation combinations. The rank score is simply the position of a target conformation in the sorted list of all energies in a particular ensemble. Others have used similar ranking scores (Hendlich *et al*., 1990; Ouzounis *et al*., 1993). We also need to compare the discrimination power of different energy functions and need, for example, to decide whether it is better to find one out of 100 conformations or eight out of 1000. We therefore derive a measure which allows these comparisons.

A conformation selected at random has equal probability of having a rank between one and M, where

Table 6. X-ray ranks

Protein	Surface	Contact(MJ)	VdW(MJ)	Contact(HL)	VdW(HL)	Histo[a]	$\langle Q_p \rangle$[b]
4rxn	285	29,184	49	10,154	324	12,277	2.07
4pti	10,963	126,900	286	129,987	5074	1488	1.32
1r69	2939	19,423	77	78	222	2703	2.42
2cro	19,092	57,484	160	478	8	8	2.67
1sn3	31,000	55,387	2	16,067	8	6188	2.06
1ctf	2592	54,658	2	46	16	1	3.32
3icb	6965	141,032	1327	234	32	1	2.61
1ubq	719	22,527	1	63	1	7	2.91
$\langle Q_f \rangle$[c]	1.55	0.46	3.66	2.19	3.53	3.13	2.42

[a] Histo = Histogram.
[b] $\langle Q_p \rangle$ is the average quality factor for all energy functions for individual proteins.
[c] $\langle Q_f \rangle$ is the average quality factor for all proteins for each individual energy function.

the ensemble contains a total of M conformations. Consider the probability, $P(r)$, that a randomly chosen state has a rank of r or less (i.e., that the state is amongst the best r entries of ensemble M). If there is one conformation that we are trying to find, $P(r)$ is r/M (the chance of picking one of the first r entries out of a total of M entries). If there are n different conformations that we trying to find, the probability that any one of these rank in the top r is higher. If the ensemble size, M, is big compared to r and n, $P(r)$ is approximately nr/M (this formula is almost exact for $m > 1000$ and $nr < 0.1$ M). This makes intuitive sense as it is more likely to find a conformation ranking in the top r if (1) the ensemble size, M, is smaller and (2) the number of equally acceptable conformations, n, is greater. We define a quality score as

$$Q(r) = -\log_{10} P(r) = \log_{10}(M/nr) \qquad (17)$$

A $Q(r)$ score of 2 indicates that the particular ranking had a chance occurrence of 0.01 (one in 100). The maximum $Q(r)$ score occurs for $r = 1$ and is $\log_{10}(M/n)$. In comparing the performance of different energy functions on a series of different proteins, we add $Q(r)$ values, which is equivalent to multiplying $P(r)$ values to give the joint probability.

These two scoring methods are complementary in that the Z-score uses the deviation from the mean to get the significance of an extreme value whereas the Q-score uses the probability of extrema directly.

Results

Suitable ensembles

In outline, our procedure in this study was to generate large ensembles of decoy structures for eight test proteins using a set of eight simple four-state models of protein structure (see Methods). Starting with near best fits of each four-state model to X-ray structures, we enumerated all possible conformations for ten carefully chosen flexible residues per protein, yielding 1,048,576 conformations for each four-state model or 8,388,608 total conformations for each test protein. We reduced the size of these ensembles by excluding conformations whose radius of gyration was too large or which had too many steric conflicts. These filters left ensembles of from 35,000 to 200,000 conformations. After applying a rapid relaxation procedure we evaluated six empirical energy functions for each conformation.

Table 2 shows the flexible residues chosen for each protein superimposed over the corresponding secondary structure. In all cases movable residues are found in regions between or at the ends of secondary structure elements. Table 3 shows the relevant enumeration parameters and results for each of the test proteins, as well as the number of conformations in each of the ensembles which have RMS deviations from their corresponding X-ray structures of less than 3.5, 4.0 and 4.5 Å. Unless otherwise noted all results which refer to native-like conformations refer to conformations <4.0 Å from the X-ray structure. Using this figure every ensemble contains at least 48 native-like conformations, and 133 on average.

The rest of this paper is an examination of the effectiveness of the different energy functions at identifying X-ray and native-like structures from among the large number of decoy conformations in our ensembles.

Comparing individual energy functions

Finding the X-ray fold

We present the discrimination results for the different potential functions in several alternative ways. The first, which we consider to be most telling, is the order rank of the X-ray or best scoring native-like conformation relative to the rest of the ensemble for a particular potential function. We prefer this measure because it directly reflects what one is trying to do with potential functions, namely trying to identify the correct conformation from many other conformations of a given sequence. A rank score is a direct indication of the number of best-scoring conformations one will have to keep for further analysis and be sure that one has found the target (or targets).

Table 6 shows the rank scores for X-ray conformations relative to enumerated conformations for all the test protein-potential function combinations. Our first reaction to these results was

Table 7. X-ray Z-scores

Protein	Surface	Contact(MJ)	VdW(MJ)	Contact(HL)	VdW(HL)	Histo
4rxn	–2.65	–1.02	–3.37	–1.47	–2.39	–1.51
4pti	–1.57	0.58	–3.27	0.55	–1.67	–2.69
1r69	–1.99	–1.35	–3.79	–2.63	–2.53	–2.17
2cro	–1.28	–0.52	–3.52	–2.42	–3.05	–3.53
1sn3	–0.75	–0.13	–5.12	–1.12	–2.75	–1.67
1ctf	–2.29	–0.38	–5.31	–2.49	–2.38	–4.24
3icb	–1.76	0.69	–2.67	–2.71	–2.76	–4.06
1ubq	–1.98	0.43	–4.58	–2.54	–2.73	–3.38
Mean	–1.78	–0.21	–3.95	–1.85	–2.53	–2.91

bewilderment at the inconsistency of individual functions over different proteins. Three of the functions are able to rank the X-ray structure first for individual proteins but no energy function can consistently rank X-ray structures highly for all proteins.

However, some trends in the data of Table 6 are discernible. To see these more easily we calculated the quality factor, $Q(r)$ for each rank (see Methods), and then averaged these values $Q(r)$ over different proteins for the same function ($\langle Q_f \rangle$) and over different functions for the same protein ($\langle Q_p \rangle$). The potential functions can be divided into two classes by their overall performance. The VdW(MJ), VdW(HL) and Histogram functions, which have $\langle Q_f \rangle$ scores above 3, are plainly better than the Surface, Contact(MJ) and Contact(HL) functions whose $\langle Q_f \rangle$ scores range from 0.5 to 2.2. Of the different proteins, 1ctf is most easy to find with a high $\langle Q_p \rangle$ of 3.3, and 4pti is hardest with $\langle Q_p \rangle = 1.3$.

Table 7 shows an alternative scoring scheme, in which Z-scores are calculated for each protein-energy function combination. We see that the Contact(MJ) function does no better than chance (its average Z-score is near 0), and that the best discrimination by any function is only 5.31 standard deviations (1sn3 and the VdW(MJ) function). Kocher *et al*. (1994), for instance, using a database threading approach report Z-scores considerably better than ours for many similar types of energy functions. This discrepancy is an indication that our ensembles are a more challenging test of certain energy functions. The Z-scores averaged over different proteins also show that the energy functions, VdW(MJ), VdW(HL) and Histogram, do better than the three others, paralleling what is seen in Table 6 using Q-scores.

Finding near-native folds

As important a characteristic of a potential function as being able to identify the X-ray conformation from a group of incorrect conformations is, it is perhaps more important for a potential function to be able to identify native-like conformations. Any real attempt to predict protein structure *ab initio* is likely to generate conformations some distance from the correct conformation. Even when inverse folding methods are used and the goal is to identify the nearest structural homolog to a given protein sequence in a database of protein folds, that nearest homolog is likely to be significantly different from the correct conformation, with α-carbon RMS deviations greater than 1 Å and sometimes greater that 2 Å.

We have therefore also examined how well the different energy functions discriminate native-like from non native-like conformations. Table 8 shows the rank scores, $\langle Q_f \rangle$ scores and $\langle Q_p \rangle$ scores for the best-scoring native-like conformations. Here we find that of the different proteins in the test set, the native-like conformations of 4rxn are most easily found, and those of 1sn3 least easily.

In comparing the different energy functions we find that the overall average scores for native-like structure identification are all lower than for discrimination of the X-ray structure (0.9 compared to 2.4). In general it seems that it is harder to find a near-native structure than the actual X-ray structure. Also, in contrast to the results for X-ray structure identification, we do not see here an obvious dichotomy between good functions and bad. The $\langle Q_f \rangle$ scores vary continuously from –0.4 for Contact(MJ) to 1.5 for Contact(HL). For the most part they do parallel the results for X-ray structure identification.

Table 8. Native-like ranks

Protein	Surface	Contact(MJ)	VdW(MJ)	Contact(HL)	VdW(HL)	Histo	$\langle Q_p \rangle$
4rxn	330	2116	15	1	5	1840	1.42
4pti	768	3949	101	364	35	416	0.93
1r69	496	1270	71	115	135	34	0.94
2cro	153	195	27	20	129	15	0.91
1sn3	1141	16,239	149	122	3525	7530	0.27
1ctf	2454	24,156	98	28	28	40	0.86
3icb	373	3538	533	86	2	6	1.18
1ubq	22	9895	18	46	22	182	0.89
$\langle Q_f \rangle$	0.57	–0.40	1.33	1.52	1.51	1.01	0.92

Table 9. Average native-like Z-scores

Protein	Surface	Contact(MJ)	VdW(MJ)	Contact(HL)	VdW(HL)	Histo
4rxn	–0.78	–0.11	–0.63	–1.08	–0.61	–0.94
4pti	–1.35	–0.30	–0.47	–0.34	–0.11	–0.89
1r69	–0.66	0.22	–1.05	–0.44	–1.13	–1.06
2cro	–0.67	–0.18	–1.33	–0.93	–0.83	–1.48
1sn3	–0.90	–0.02	–1.59	–0.63	–0.67	–0.61
1ctf	–0.59	–0.82	–0.82	–0.73	–1.04	–1.76
3icb	–0.99	0.53	–1.06	–0.79	–1.26	–1.65
1ubq	–1.03	0.89	–0.90	–0.97	–1.43	–1.75
Mean	–0.87	0.03	–0.98	–0.74	–0.89	–1.27

To calculate Z-scores for the best native-like conformations of the different ensembles would be misleading, because by chance alone individual native-like conformations may score significantly better than the average of the entire ensemble and therefore have good Z-scores. In general the Z-score will increase if more of the decoys are native-like thus making the direct comparison of Z-scores uninformative. Table 9 shows a more useful Z-score measure for native-like conformations. Here the Z-scores are calculated as averages over all native-like conformations for each protein-energy function combination. The results for this measure are much more in line with the Q-score results, although the spread in average Z-scores is narrow making the differentiation of energy functions difficult.

Variation of results with different ''native-like'' cutoffs

To reassure ourselves about our choice of 4 Å as the dividing line between native-like and non-native-like conformations we calculated the $\langle Q_f \rangle$ scores for native-like structure identification using cutoffs of 3.5 Å and 4.5 Å (Table 10). The trends are similar for both alternate cutoffs. We chose 4 Å as the standard cutoff because for proteins the size of those in our test set we found 4 Å to represent the largest RMS deviation for which two structures seem consistently to be subjectively the ''same'' structure. Structures 4.5 Å apart sometimes crossed that line. A 3.5 Å cutoff would have probably served as well as the 4 Å, but we believed, all other things being equal, that it would be better to have more rather than fewer native-like structures in each of our ensembles since it is the native-like conformations we are finally interested in.

Comparing combinations of energy functions

The results so far have shown that none of the functions perform as well as one could wish. Our hope then was that some combination of energy functions might improve significantly the discrimination both of X-ray and native-like structures. We therefore looked at the performance of all the possible pairwise combinations of energy functions.

We used a straightforward scheme to combine pairs of energy functions; with no *a priori* reason to believe that one energy function of any pair should be given more weight than another, we simply defined the combined score for two energy functions as the sum of the two functions' energies individually scaled by their respective standard deviations.

Ranking scores

Table 11 shows the X-ray rank results for five of the most illustrative of these combinations of energy functions; Surface + VdW(MJ), Surface + VdW(HL), VdW(MJ) + VdW(HL), Surface + Histogram and VdW(MJ) + Histogram. A quick glance tells us that three of the combinations, Surface + VdW(HL), VdW(MJ) + VdW(HL) and VdW(MJ) + Histogram, are significant improvements over their constituent functions (none of the combinations not shown showed any significant improvement). Each of these combinations can usually identify the X-ray structure unequivocally (rank = 1) the exceptional proteins being 4pti, 2cro, 1sn3 and 4rxn. The $\langle Q_f \rangle$ values for each of these combined functions is around 5, compared to $\langle Q_f \rangle$ values of the component functions from 3 to 4. Remembering that a quality factor increase of 1 corresponds to an order of magnitude better discrimination, these results show that certain

Table 10. $\langle Q_f \rangle$ scores for best native-like rankings with 3.5 Å and 4.5 Å cutoffs

Cutoff (Å)	Surface	Contact(MJ)	VdW(MJ)	Contact(HL)	VdW(HL)	Histo	Average
3.5	0.86	–0.19	1.23	1.36	1.09	0.95	0.88
4.0	0.57	–0.40	1.33	1.52	1.51	1.01	0.92
4.5	0.92	–0.19	1.28	1.04	1.44	1.04	0.92

Table 11. X-ray rank for combined energy functions

Protein	Surf + V(MJ)[a]	Surf + V(HL)[a]	V(MJ) + V(HL)[a]	Surf + Histo[a]	V(MJ) + Histo[a]	$\langle Q_p \rangle$
4rxn	18	1	1	373	2	4.45
4pti	370	4	3	456	9	3.80
1r69	72	1	1	128	1	4.51
2cro	754	2	1	100	1	4.26
1sn3	52	9	1	9283	1	3.80
1ctf	10	1	1	1	1	5.00
3icb	1273	1	1	1	1	4.65
1ubq	1	1	1	3	1	4.46
$\langle Q_f \rangle$	3.36	4.93	5.10	3.44	5.00	4.37

[a] Surf = Surface, V(MJ) = VdW(MJ), V(HL) = VdW(HL), Histo = Histogram.

combinations have significantly greater discriminating power.

The combined potential functions are also improved in their ability to pick out native-like conformations (Table 12). Parallel with the improvements in X-ray conformation identification we see that the Surface + VdW(HL), VdW(MJ) + VdW(HL) and VdW(MJ) + Histogram combinations are better than their constituent functions, but by less than for X-ray conformation identification.

Z-scores

Table 13 shows the Z-score results for combined function identification of X-ray structures. They mostly parallel the results in Table 11 for Q-scores; we provide them for comparison with other studies. The average Z-scores for all native-like conformations, shown in Table 14, parallel those from X-ray Z-scores and X-ray and native-like Q-scores. Once again we see that the Z-scores do not differentiate one combination from another well.

Energy/RMS plots

Q-scores and Z-scores are ways of extracting single numbers from what are in fact very complicated sets of data. Another interesting way to see the behaviors of different energy functions is to look at the relationship between energy values and RMS deviation for whole ensembles; doing so allows one to see all the data at once. But what does one expect to see for ''good'' energy functions? To the idealist, a good energy function will have a simple linear relationship between energy and RMS. A little reflection will tell us, however, that this sort of relationship is impossible. For any but a strangely biased ensemble of conformations, structures which are distant from the X-ray structure (high RMS deviation) will be distant from each other and therefore unlikely to have similar energies. The best that one can expect is a funnel-like distribution that has a wide dispersion of energies for conformations far from the correct structure and approaches a linear relationship between energy and RMS as energies approach that of the correct structure. This kind of relationship has indeed been seen in some other studies, but only for limited sets of decoy conformations. (Bowie & Eisenberg, 1994; Levitt, 1983, 1992; Monge *et al*., 1995; Williams *et al*., 1992).

The reality of energy-RMS distributions for our ensembles falls far short of this ideal. Figure 1, showing the energy-RMS distribution of the Surface energy function for 1ctf, illustrates the worst case. There is very little funnel like about this distribution. The native-like conformations do have a lower average energy but only by a small amount. In fact, the Surface energy function performs quite poorly as measured by both X-ray rank and best native-like rank scores. Perhaps the distributions for functions which discriminate more effectively will be better.

Figure 2, unfortunately, squashes this hope. The VdW(MJ) function is much more effective at finding both the X-ray structure and native-like structures than the Surface function, yet the energy-RMS distribution looks as bad as that of the previous example. We therefore cannot conclude that the global distribution of energies is necessarily a good

Table 12. Combination native-like ranks

Protein	Surf + V(MJ)	Surf + V(HL)	V(MJ) + V(HL)	Surf + Histo	V(MJ) + Histo	$\langle Q_p \rangle$
4rxn	6	5	32	662	29	1.798
4pti	99	40	229	217	188	1.347
1r69	123	3	75	114	68	1.501
2cro	10	10	9	70	24	1.419
1sn3	176	213	17	1592	16	1.387
1ctf	403	43	4	22	1	1.997
3icb	186	13	42	34	3	1.752
1ubq	16	1	6	14	2	2.184
$\langle Q_f \rangle$	1.379	2.046	1.787	1.117	2.036	1.673

Table 13. Combined function X-ray Z-scores

Protein	Surf + V(MJ)	Surf + V(HL)	V(MJ) + V(HL)	Surf + Histo	V(MJ) + Histo
4rxn	–3.48	–3.94	–3.57	–2.60	–3.43
4pti	–3.72	–3.03	–2.44	–2.68	–4.20
1r69	–3.42	–3.88	–4.37	–2.81	–4.34
2cro	–5.13	–2.80	–5.51	–3.13	–5.19
1sn3	–3.58	–3.36	–5.32	–1.51	–4.50
1ctf	–4.73	–4.47	–5.83	–4.75	–7.19
3icb	–4.92	–2.63	–5.78	–4.04	–5.25
1ubq	–4.09	–4.39	–5.15	–3.75	–5.98
Mean	–4.13	–3.56	–4.75	–3.16	–5.01

indicator of a function's discriminatory power. To see the relative success of the VdW(MJ) function we are forced to examine two important details. The first is obvious: The X-ray structure has a lower energy than all but one of the ensemble conformations. The second detail concerns the energies of the native-like conformations. Although many are only slightly distinguished energetically from the mass of ensemble conformations, several are quite well resolved. When one is actually trying to predict protein conformations what one needs is some function which assuredly will find *some* native-like conformations among a small number of best-scoring conformations. In the case of the VdW(MJ) function the number of low energy conformations one has to examine to find one which is native-like is small. The overall distribution of energies of an energy function is not necessarily related to its ability to discriminate.

Not all energy-RMS distributions are as far from ideal as those in Figures 1 and 2. The energy-RMS distribution for the VdW(HL) energy over the 1CTF ensemble (Figure 3) looks considerably better. With a little imagination, especially if one takes into account the low-energy points (they are in fact the most important) the distribution is funnel like. The funnel, however, is a peculiar one. Its bottom boundary is essentially horizontal (or even slopes up to higher energies as the RMS deviation decreases). This means that when one selects some fraction of the best-scoring conformations for this function, one gets an assortment that contains a significant proportion of high RMS deviation structures. The reason that this function does better than others is because, once again, the X-ray structure is lower in energy than most of the ensemble and there are several native-like conformations which are well separated in energy from the bulk.

What we have presented so far tells us that the energy functions we have examined are less than ideal, at least for unequivocally identifying native-like structures 2 to 4 Å from the correct structure. Several combinations of functions do consistently identify X-ray structures and can find native-like conformations 100 times better than chance. The question remaining now is why these functions fail when they do fail. Why do wrong structures sometimes have lower energies than right?

Viewing low-energy folds

In order to answer at least partially this question we carefully examined with molecular graphics the 30 best-scoring conformations for each energy function and energy function combination. The most surprising thing we found is that our ensemble generation procedure produces conformations which score well for some functions, are wrong (i.e. unrelated to the X-ray structure) and yet are architecturally reasonable looking. The following is an account of what we found function by function.

Contact(MJ) and Contact(HL)

The problems with these functions are fundamental. Almost all good scoring but wrong conformations for these functions are those which, because of their topologies, were not fully relaxed by our minimization procedure. That is, these conformations contained knotted or threaded regions. The result of this phenomenon for an on/off contact energy function is that certain residues will

Table 14. Combined average native-like Z-scores

Protein	Surf + V(MJ)	Surf + V(HL)	V(MJ) + V(HL)	Surf + Histo	V(MJ) + Histo
4rxn	–0.82	–1.09	–0.77	–1.08	–1.10
4pti	–1.13	–1.51	–0.40	–1.40	–0.95
1r69	–1.01	–1.54	–1.51	–1.16	–1.54
2cro	–1.17	–1.41	–1.62	–1.40	–2.07
1sn3	–1.52	–1.50	–1.53	–0.94	–1.99
1ctf	–0.88	–1.56	–1.41	–1.71	–1.95
3icb	–1.22	–2.16	–1.59	–1.84	–2.12
1ubq	–1.21	–2.29	–1.63	–1.95	–1.79
Mean	–1.12	–1.63	–1.31	–1.44	–1.69

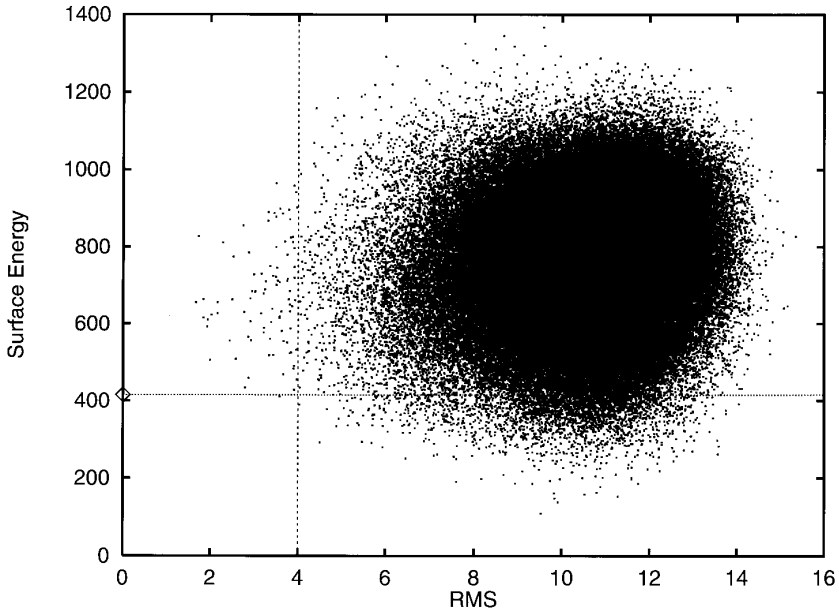

Figure 1. Energy plotted as a function of RMS deviation from the X-ray structure for the Surface energy function and lctf. The energy of the X-ray structure is shown as a diamond on the *y*-axis. The dashed horizontal line corresponds to this energy, and the dashed vertical line is placed at 4 Å, corresponding to the native-like cutoff. There is no perceptible trend towards the lower left corner where low energy and low RMS deviation meet. Admittedly this function discriminates native from non-native poorly, but the energy-RMS distributions for other functions which discriminate better are only somewhat improved.

form a much larger number of contacts than they could in a real protein. If most of these extra contacts are favorable a falsely low energy results. This is perhaps why the Contact(HL) functions performs better than the Contact(MJ) function. Since the (MJ) formulation is calculated relative to the unfolded polypeptide, almost all residue-residue interactions are negative. The (HL) formulation, on the other hand, is calculated relative to the compact state and its residue-residue energies cluster around zero. It is therefore less likely to be fooled by an over-abundance of contacts.

Surface

The surface energy function has much the same problem as the contact energy functions. It too is fooled by conformations which are over-crowded or knotted, (Figure 4). If hydrophobic residues in a particular conformation are crowded together because of the failure of the relaxation procedure, they will have small exposed surface areas and therefore low energies. Another way in which this energy function fails is unrelated to incomplete relaxation, however. In some wrong conformations the hydrophobic residues are all buried (i.e. have small exposed surface areas) without forming a true hydrophobic core (see Figure 4B). In these cases the hydrophobic residues form two or more hydrophobic clusters. This phenomenon perhaps reflects the fact that while the hydrophobic effect is cooperative and favors a single hydrophobic core, the surface energy is non-cooperative

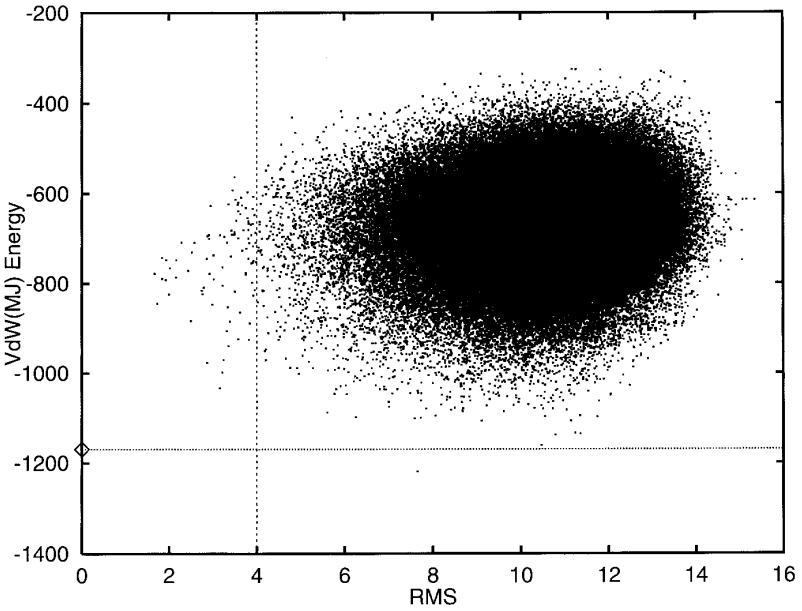

Figure 2. Energy plotted as a function of RMS deviation from the X-ray structure for the VdW(MJ) function and 1ctf. This distribution looks much like that of Figure 1. But note two things. The X-ray energy is better than all but one of the ensemble conformations. There are several very low energy native-like conformations (RMS < 4 Å) which score better than all but a handful of non-native conformations.

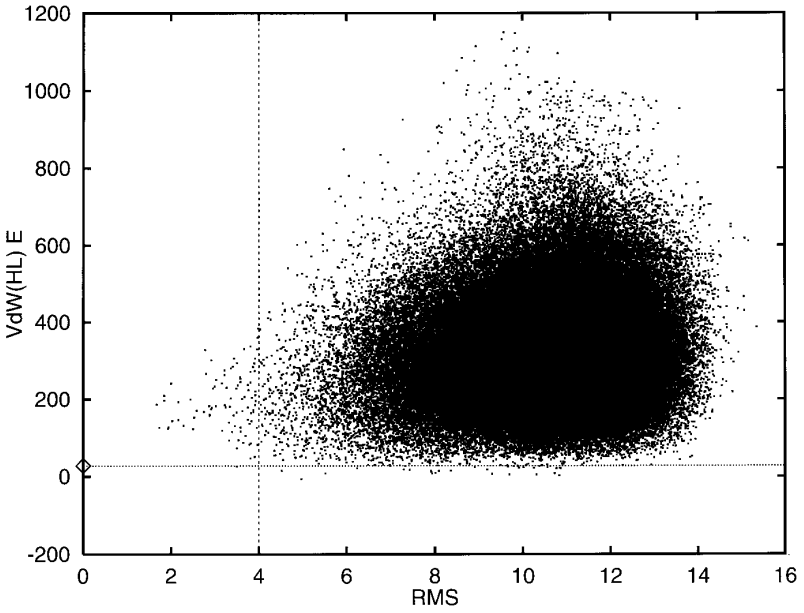

Figure 3. Energy plotted as a function of RMS deviation from the X-ray structure for the VdW(HL) function and 1ctf. This distribution is considerably more satisfactory than those shown in Figures 1 and 2. The vast majority of conformations are clustered in an ovoid blob, but the outlying points do form a kind of funnel pointing towards low energy and low RMS. Note, however, that the bottom boundary of the funnel is horizontal (perhaps even tilted up towards higher energies and lower RMS deviations.)

and favors hydrophobic burial by hydrophilic residues as much as by other hydrophobic residues.

VdW(MJ)

The VdW(MJ) function also suffers from this predilection for conformations which do not form single hydrophobic cores (Figure 5). The gratifying (or horrifying, depending on one's perspective) thing about the wrong conformations which score well with this function is that they commonly look reasonable and protein like. Many of the wrong conformations which score well with the Surface or Contact functions are trivially wrong in that simple tests, (topological or steric) can eliminate them. The distance dependence of the VdW(MJ) function, however, assures that trivial tests are not applicable to its mistakes.

VdW(HL)

Like the (MJ) formulation the (HL) formulation of the VdW function finds some low-energy, wrong conformations which are architecturally reasonable and protein like. More characteristically, however, its wrong conformations are inadequately packed. This inadequacy may be manifested by one or two secondary structure elements which are arranged separately from the rest of the protein, but more often takes the form of a general structural looseness. Figure 6 shows examples. This tendency towards inadequate packing is not entirely surprising, since the inter-residue energies for this function are calculated relative to the compact state. In contrast to the (MJ) formulation the (HL) formulation has less of a hydrophobic driving component and therefore is less likely to favor properly packed conformations.

Histogram

The Histogram energy function fails, at least superficially, in a way similar to the VdW(HL) function. Its parameters too are calculated relative to the compact state. Very few of its wrong conformations, however, are protein like. The salient characteristic of most wrong low-energy conformations for the Histogram function is a favoring of local at the expense of long-range interactions. Short segments of these wrong conformations look correct but the global architecture is clearly wrong (Figure 7).

Combinations of functions

In general the combinations of energy functions show the failings of their constituent functions. For the good combinations, however, the constituent functions often compensate for each other's deficiencies. The result is that for the VdW(MJ) + VdW(HL) combination, for example, a large number of low-energy but wrong structures are architecturally reasonable, like those shown in Figure 7 for the VdW(MJ) function.

High energy native-like conformations

We also examined another set of conformations, namely those of native-like structures. These, for most energy functions and combinations, had higher energies than the corresponding X-ray structures. In Figure 8 we show three native-like conformations superimposed over the X-ray structures of three different proteins. The differences between X-ray structures and native-like conformations were, in terms of distances, slight (the RMS deviations of all native-like conformations shown in Figure 8 are less than 2 Å). However, one should

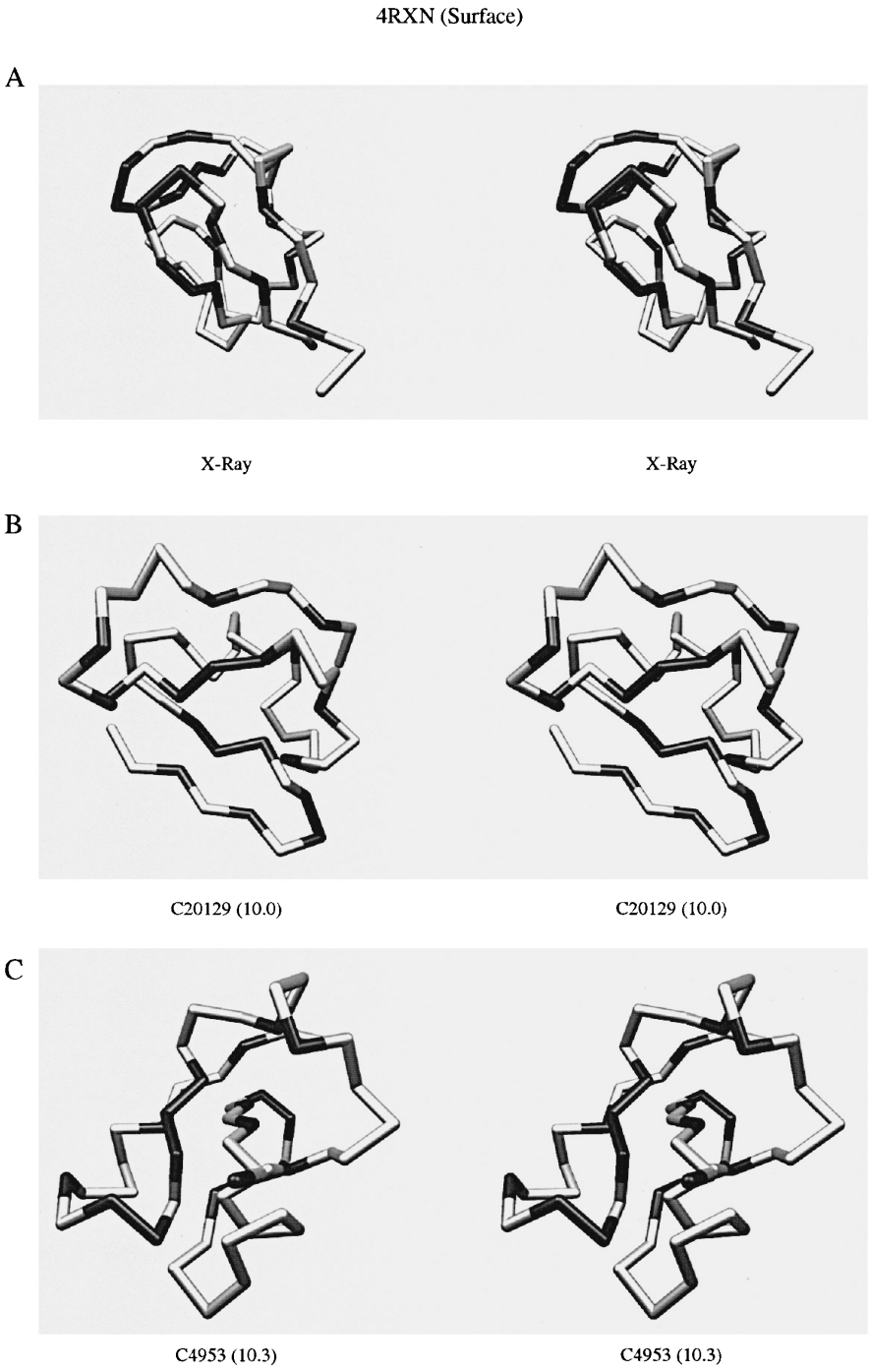

Figure 4. Stereo diagrams (cross-eyed) of low-energy conformations of 4rxn for the Surface area energy function. Hydrophobic residues are shown in dark gray while all others are light gray. The numbers in parentheses below the structures are RMS deviations from the X-ray structures. A is the X-ray conformation for 4rxn. The conformation in B illustrates a common occurrence for all E-functions that do well. This structure (number C20129) looks reasonable. The β-strands are paired and hydrophobic residues are making contacts with each other. One notices, however, that there is no true hydrophobic core. The structure in C is illustrative of a common failure mode for this function and for others. The N-terminal β-strand is threaded through the core of the protein. This effectively buries the two hydrophobic residues of this strand.

note that the local orientations of α-carbons in native-like conformations are often quite different from those of the corresponding X-ray structures. In particular, β-strands in native-like conformations have orientations that commonly place their side-chains in positions very different from those of the X-ray structure, i.e., facing the exterior rather than the interior of the protein. The same can be said of the loops between secondary structure elements. In 1r69 (Figure 8B) and 3icb (Figure 8C),

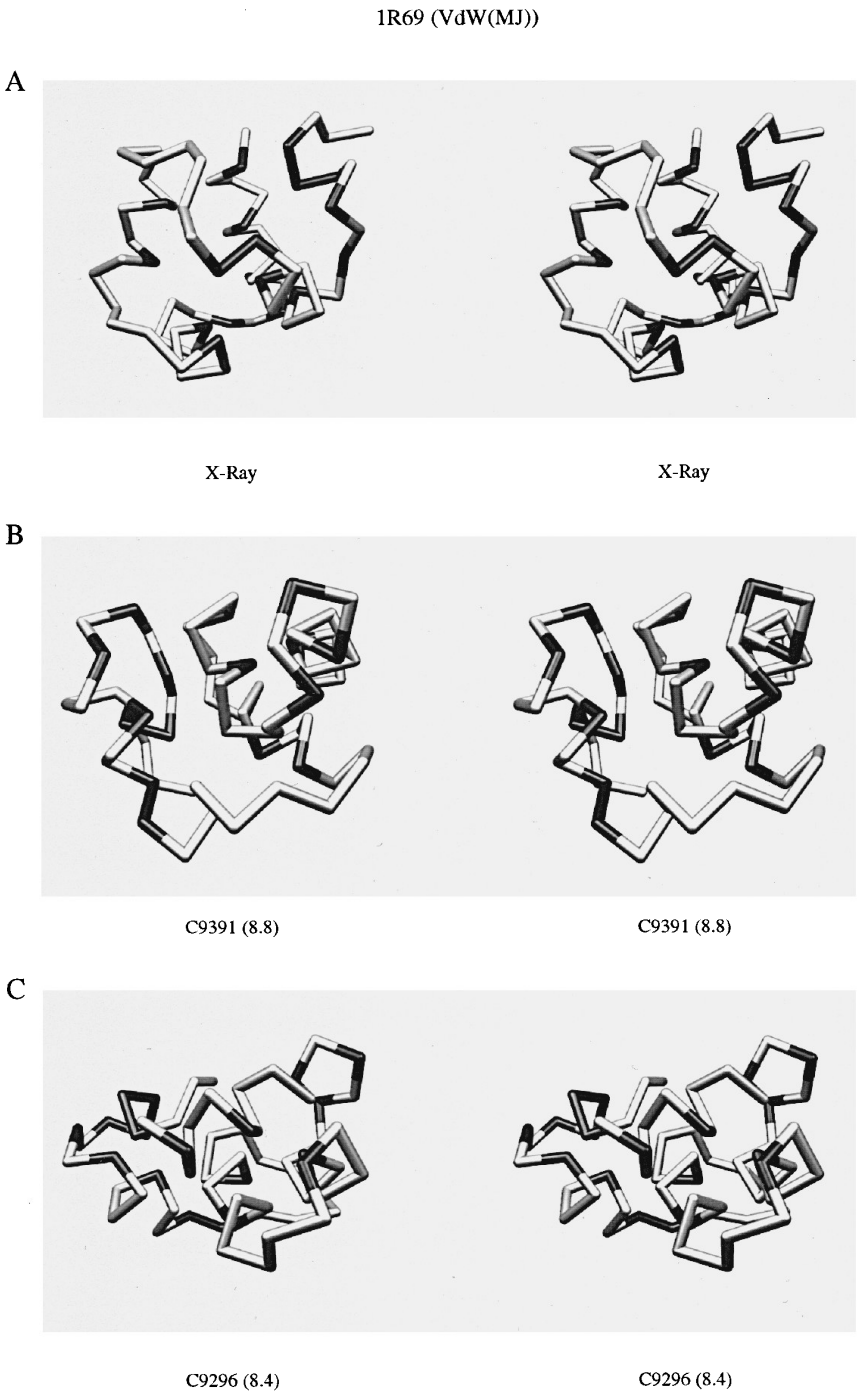

Figure 5. Stereo diagrams for low-energy conformations of 1r69 for the VdW(MJ) functions. A is the X-ray conformation. Structure C9391, B, is a good example of a distressingly reasonable-looking conformation which is wrong. This is an alternate packing of the helices in lr69 which for the most part succeeds in following the rules of good protein structure. The overall shape is compact. Hydrophobic residues form clusters. Structure C9296, C, represents the majority of those conformations which score well with this function but are incorrect. The structure is manifestly wrong. The two most C-terminal helices are not packed into the structure. The other three helices do, however, form a reasonable hydrophobic core.

both all-α proteins, the orientation of α-carbons in the helices are usually quite similar to those of the X-ray structure. Their loop regions however are quite different.

From these comparisons we can make some conjectures about what structural features allow certain energy functions to distinguish X-ray structures but not native-like structures from the rest of our ensembles. First in β-strand-containing proteins the rather large errors in the α-carbon orientation of native-like conformations are bound to affect native-like energies adversely. The poor

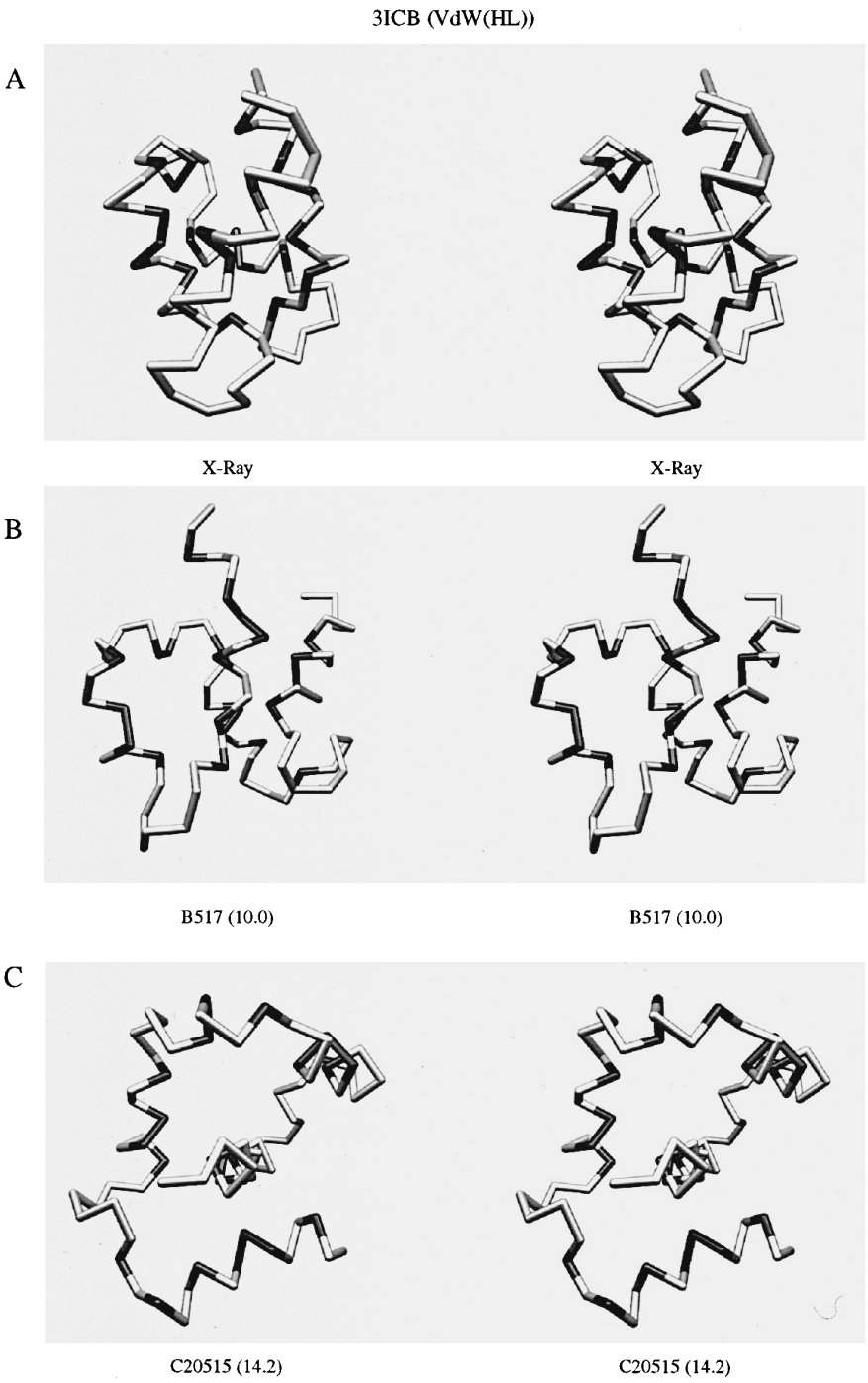

Figure 6. Stereo diagrams of low-energy conformations of 3icb for the VdW(HL) function. A is the X-ray conformation. In B we again see in structure B517 another example of a wrong structure which looks reasonable. It appears to be an alternate arrangement of helices. A close examination shows that there is a certain lack of coherence to the hydrophobic residues. Those of helices 1 and 2 do not quite point inward. Structure C20515, C, is more representative of the way in which this function fails. The conformation is obviously not compact, and a cursory examination leaves one wondering what is energetically satisfactory about it. A closer look shows that the hydrophobic sides of helices 2 through 5 are laid on helix 1. Local hydrophobic tendencies are satisfied without a true hydrophobic core being formed.

reproduction of loop residue orientations may also have similar effects. Finally there may be subtle differences in the precise way that packed secondary structure elements mesh that native-like conformations, at least those generated by our four-state models, cannot capture.

Discussion

Our goal has been to generate ensembles of decoy conformations that offer a different and in some ways more rigorous test of energy functions than the usual collections of X-ray structures. There are

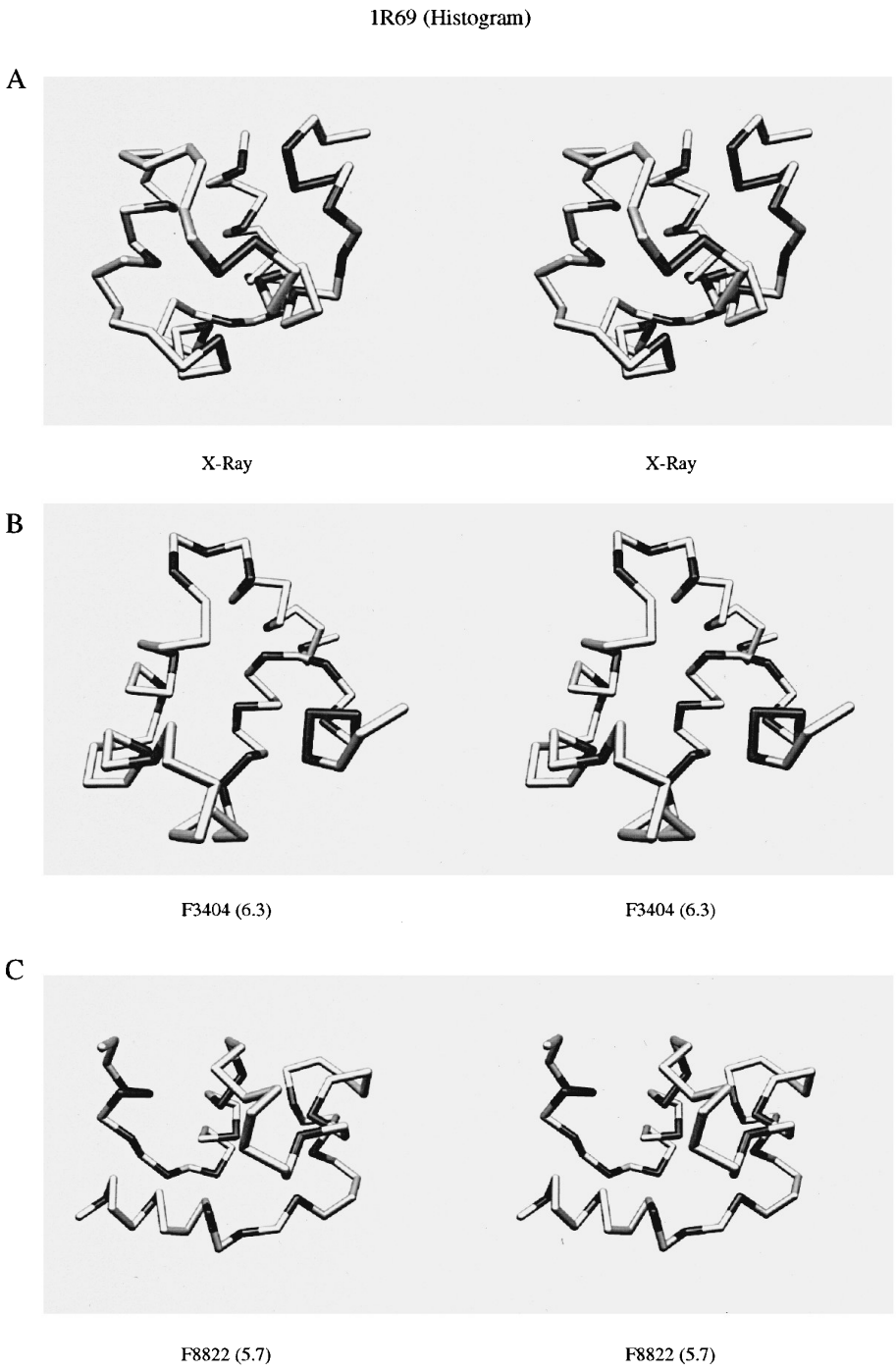

Figure 7. Stereo diagrams of low-energy conformations of 1r69 for the Histogram energy function. A is the X-ray conformation. The two wrong structures in this illustration are unsatisfying after even casual examination. The placement of the N-terminal helix of F3404, in B, seems unlikely. The segment of the conformation composed of it and helix 4 are unnaturally flat and immediately suggest that this is not a real protein conformation. In C, the N and C-terminal helices of F8822 form an annex to the overall conformation. The central helices 2, 3 and 4 form a reasonable hydrophobic core, and the other two helices are placed almost as an afterthought. Together these two conformations illustrate the tendency of the histogram energy function to produce reasonable local conformations at the expense of global architecture.

many energy functions which find the correct X-ray structure from among other incorrect X-ray structures (Casari & Sippl, 1992; Huang *et al*., 1995; Kocher *et al*., 1994). It is extremely difficult, on the other hand, to generate energy functions which can successfully find correct conformations using an *ab initio* approach. For energy functions directed towards *ab initio* prediction our ensembles are

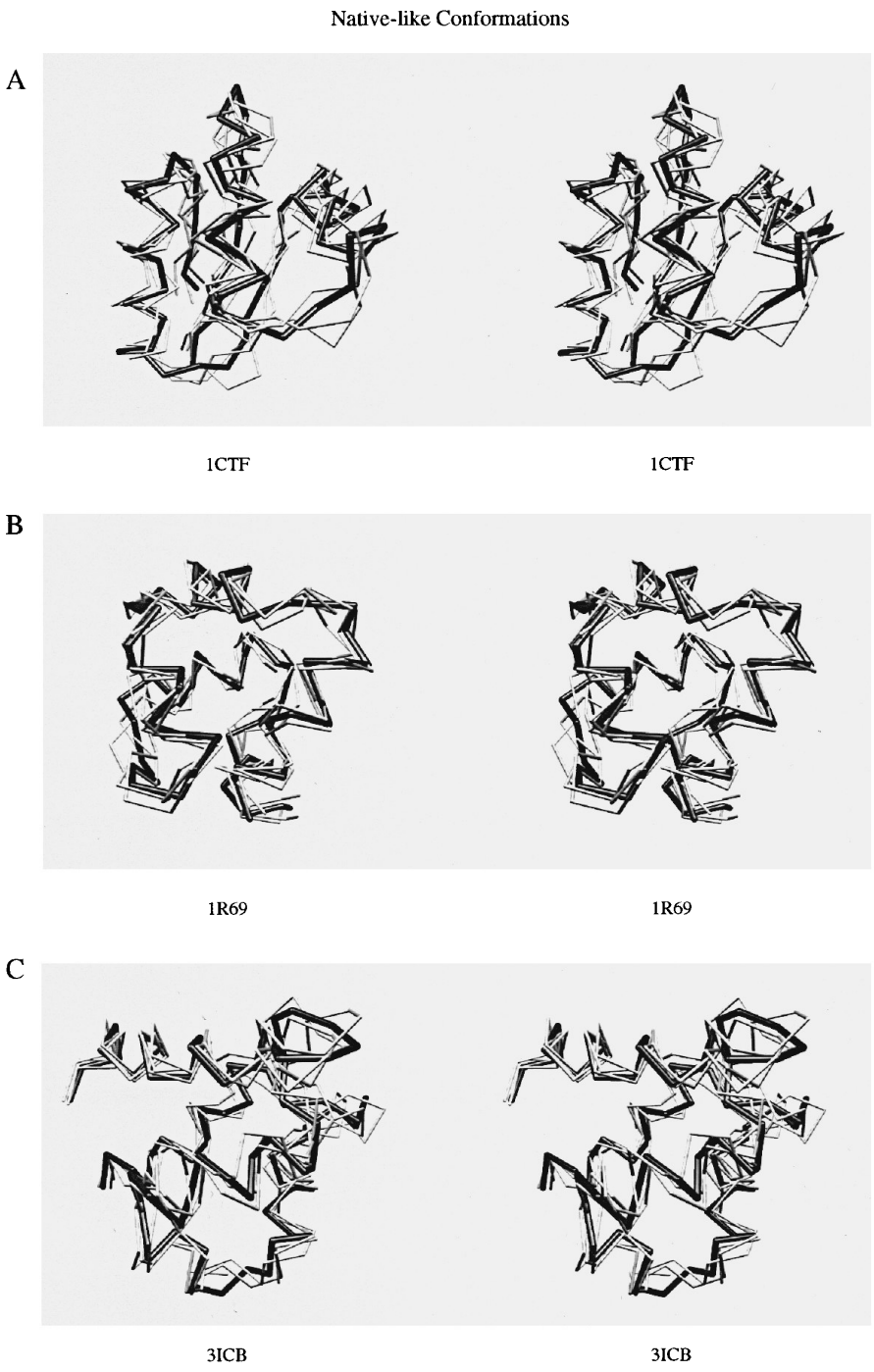

Figure 8. Stereo pairs showing native-like conformations superimposed over X-ray structures. The X-ray structures are the thick, dark gray skeletons. Three native-like conformations are shown for each protein as thinner light-colored skeletons. A, 1ctf; B, 1r69; C, 3icb.

indeed a more rigorous test of energy functions than a set of X-ray structures. The Z-scores for our ensembles are considerably lower than Z-scores reported for database threading using almost identical energy functions (Kocher *et al*., 1994), and our ensembles contain wrong conformations which score better than the X-ray structure for energy functions which have been used (sometimes successfully) in *ab initio* studies (Bowie & Eisenberg, 1994; Covell, 1992; Dandekar & Argos, 1994; Monge *et al*., 1994; Sun, 1993; Wallqvist & Ullner, 1994). Our ensembles have the additional strength of being able to realistically test energy functions for *ab initio* structure prediction while being independent of any particular folding algorithm.

The test that our ensembles provide has told us much about what kinds of energy functions are likely to do well under the special circumstances

of *ab initio* structure prediction. It is plain that functions which are distance dependent (in more than an on/off sense) are more effective than those which are not. The VdW and Histogram functions, are on average more discriminating than the Contact functions. There are two possible reasons for this. The first is simply that the forces that drive actual protein folding are distance dependent. One expects approximate functions which have this property to work better *a priori*. The second is that our discrete-state models require smooth or distance-dependent functions to compensate for the models' own shortcomings. Studies of protein folding which use lattice models of protein structure often do well with simple contact functions (Hinds & Levitt, 1992, 1994; Kolinski & Skolnick, 1994). For those lattice models residue-residue contact distances are always correct, at least to within the models' accuracies and a simple on/off function will work adequately. Our off-lattice models, while being considerably more representationally accurate than low-resolution lattice models, allow a large distribution of possible residue-residue contact distances some of which, even after relaxation, are less than sterically ideal. A function which is distance dependent can therefore weigh the contributions of particular residue-residue pairs by their distances. A simple on/off cutoff would give a considerably less accurate relative accounting of different interactions within a conformation, sterically ideal or not.

The different energy functions we have examined can be distinguished from each other also by the extent to which they incorporate hydrophobic energies. The Surface energy function does so explicitly. Its design goal was to quantify the extent to which hydrophobic surface is buried. The (MJ) formulations of the Contact and VdW functions also incorporate hydrophobic energies by explicitly referring to the unfolded solvated state (see Methods). The other energy functions, based as they are on the hypothetical randomly mixed compact state of proteins, incorporate hydrophobic information to a lesser extent. The differences between these two sets of functions are not seen when they are used individually, but are clearly evident when they are combined. The two sets of functions form a complementary pair; combinations of two functions both from the same group yields no improvement in discriminating power; combinations of two energy functions one from each group, produce considerably improved results. Witness the combinations VdW(MJ) + VdW(HL), VdW(HL) + Surface, and Surface + Histogram, which are the most successful.

Why do combination energy functions work well?

Why do combinations of functions which incorporate hydrophobic information and those which do not work so much better than either alone? The obvious answer is that proteins do not fold by hydrophobic interactions alone, nor by specific residue-residue interactions alone. This certainly seems most likely, but we are still puzzled by the (relatively) poor performance of the VdW(MJ) energy function. It should encompass both hydrophobic and specific residue-residue energies. Our best answer is that the VdW(MJ) function over-emphasizes hydrophobic forces. Therefore its combination with an energy function which under emphasizes-hydrophobic contributions yields a happy mean. The lesson we must take from this is that the balance of hydrophobic and residue-specific energies is likely to be a critical parameter in any successful energy function.

What's wrong with the energy functions?

Many readers will find the energy-RMS deviation distributions shown in Figures 1 to 3 disappointing. They are for the most part unsatisfactory looking. There is little or no obvious correlation between energy and RMS deviation. Although it is pleasing that various combination functions almost always find that the X-ray structure has the lowest energy, they are not as good at distinguishing native-like from non-native-like conformations. This latter fact suggests a difficulty that those trying to predict protein conformations may run into, and indeed have run into. One group, for example, had initial success reconstructing the tertiary structure of four-helical bundles (Monge *et al*., 1994) and myoglobin (Gunn *et al*., 1994), given fixed correct secondary structure, and an energy function based on the work of Casari & Sippl (1992). In later work (Monge *et al*., 1995), however, they found that their energy function and methodology were incapable of reproducing the tertiary structure of the 434 repressor protein, for example, and in general that RMS deviations and energies are not well correlated. Their results and ours may mean that the best energy functions will only start to discriminate effectively when structures closer to the correct conformation than those in our ensembles can be examined. In fact, our discrimination results for the Contact(HL) and VdW(HL) functions lend some further weight to this supposition. For identifying X-ray structures the VdW(HL) function is clearly superior, with a $\langle Q_f \rangle$ 3.5 *versus* 2.2 for the Contact(HL) function. For identifying native-like conformations, however, the functions are indistinguishable, with $\langle Q_f \rangle$ values of 1.5 each. We conjecture that the noise level for conformations 2 to 4 Å from the X-ray structure is high enough that the disadvantages of the on/off nature of the Contact(HL) function are masked.

Until we do look at conformations considerably closer to the X-ray structure we cannot tell if the discrimination of energy functions will improve. The fact that X-ray structures are almost always found to be extremely low in energy by our best function combinations suggests that they will. It is

still an open question, however, where the actual minima of different energy functions lie. It is clear that the goal of energy function development must be to move these minima towards the X-ray structure as well as to extend the region around the minima in which discrimination occurs.

None of this is to say, of course, that the kinds of functions we have examined in this study are not useful for structure prediction at the 2 to 4 Å range or perhaps even higher resolution. Indeed our combined energy functions can usually find native-like conformations within the top 0.01% of our ensembles. This is heartening, in that the functions we have examined here are only a first pass, an initial attempt to identify some of the characteristics that energy functions will need. It is still an open question how much improvement can be expected from energy functions, aimed at low to medium resolution structure prediction. Our ensembles provide a good test bed for future study.

Future work

Our results suggest several new lines of inquiry. It is plain that there is a need for test sets which contain conformations nearer to correct structures than those contained in the present ensemble in order to understand where in conformational space the true minima of empirical energy functions lie. Several methods of generating these come to mind. Molecular dynamics can easily produce conformations which are from 0 to 2.0 Å distant from X-ray structures (Troyer & Cohen, 1995; Wang *et al*., 1995). At raised temperatures structures more highly divergent should be accessible (Huang *et al*., in press). Monte Carlo techniques have also been used, but only to generate fairly small ensembles (Monge *et al*., 1995, Williams *et al*., 1992). We are currently working on extending the methods of this paper to generate new ensembles which contain conformations nearer the X-ray structures, either by using more accurate models of protein structure, or alternatively by modifying the conformations in our existing ensembles.

It is also plain from our results that the kinds of energy functions we have looked at need considerable improvement in their ability to discriminate native-like from non-native-like conformations, the ability most relevant to *ab initio* structure prediction. We believe that this inadequacy is a result of the high sensitivity of most energy functions to small geometric perturbations. Therefore, one approach to devising improved energy functions for *ab initio* prediction is by mathematical smoothing. Future studies will look at the effects various smoothing techniques have on the ability of different energy functions to distinguish native-like from non-native-like, and the effects such smoothing has on the identification of the X-ray structure.

Finally the current work can be extended by using the basic techniques presented in this paper and Park & Levitt (1995) to look at more realistic approaches to the actual prediction of protein structure. In particular we hope to ask how well a build-up procedure (Vasquez & Scheraga, 1988), constrained by known secondary structure (from NMR or multiple sequence alignment) or by a small number of nuclear Overhauser enhancement distance constraints, can predict protein structure, using one of our combination energy functions.

Acknowledgements

This work was supported by National Institutes of Health Award GM41455 and Department of Energy Award DE-FG03-95ER62135. B.H.P. holds a National Science Foundation fellowship from the Program in Mathematics and Molecular Biology.

References

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

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

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

Bowie, J. U. & Eisenberg, D. (1994). An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl Acad. Sci. USA*, **91**, 4436–4440.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236–244.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the 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.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Cohen, F. E., Richmond, T. J. & Richards, F. M. (1979). Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J. Mol. Biol.* **132**, 275–288.

Covell, D. G. (1992). Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Funct. Genet.* **14**, 409–420.

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

Dandekar, T. & Argos, P. (1994). Folding the main-chain

of small proteins with the genetic algorithm. *J. Mol. Biol.* **236**, 844–861.

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). Principles of protein folding-a perspective from simple exact models. *Protein Sci.* **4**, 561–602.

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

Godzik, A., Kolinski, A. & Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227–238.

Gunn, J. R., Monge, A., Friesner, R. A. & Marshall, C. H. (1994). Hierarchical algorithm for computer modeling of protein tertiary structure: folding of myoglobin to 6.2 Å resolution. *J. Phys. Chem.* **98**, 702–711.

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.

Huang, E. S., Subbiah, S. & Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* **252**, 709–720.

Huang, E. S., Subbiah, S., Tsai, J. & Levitt, M. (1996). Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J. Mol. Biol*. **257**, 716–725.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A*, **34**, 827–828.

Kocher, J.-P. A., Rooman, M. J. & Wodak, S. J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598–1613.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. 1. Lattice model and interaction scheme. *Proteins: Struct. Funct. Genet.* **18**, 338–352.

Levitt, M. (1976). A simplified representation of protein conformation for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

Levitt, M. (1983). Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.* **170**, 723–764.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507–533.

Levitt, M. & Warshel, A. (1975). Computer simulation of protein folding. *Nature*, **253**, 694–698.

Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Monge, A, Friesner, R. A. & Honig, B. (1994). An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl Acad. Sci. USA*, **91**, 5027–5029.

Monge, A., Lathrop, E. J. P., Gunn, J. R., Shenkin, P. S. & Friesner, R. A. (1995). Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995–1012.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.

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. H. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493–507.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK.

Rooman, M. J., Kocher, J. A. & Wodak, S. J. (1991). Prediction of protein backbone conformation based on seven structure assignments. *J. Mol. Biol.* **221**, 961–979.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.

Sippl, M. J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins: Struct. Funct. Genet.* **13**, 258–271.

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., Jaritz, M., Hendlich, M., Ortner, M. & Lackner, P. (1994). Application of knowledge-based mean fields in the determination of protein structures. In *Nato Asi Series Series B Physics* (Doniach, S., ed), vol. 325, *Statistical Mechanistics, Protein Structure, and Protein Substrate Interactions*. Plenum Press, New York.

Skolnick, J., Kolinski, A., Brooks, C. L., Godzik, A. & Rey, A. (1993). A method for predicting protein structure from sequence. *Curr. Biol.* **3**, 414–423.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

Troyer, J. M. & Cohen, F. E. (1995). Protein conformational landscapes: energy minimization and clustering of a long molecular dynamics trajectory. *Proteins: Struct. Funct. Genet.* **23**, 97–110.

Unger R. & Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.* **231**, 75–81.

Vasquez, M. & Scheraga, H. A. (1988). Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data. *J. Biomol. Struct. Dyn.* **5**, 705–755.

Vieth, M., Kolinski, A., Brooks, C. L., III & Skolnick, J. (1994). Prediction of the folding pathways and structure of the GCN4 leucine zipper. *J. Mol. Biol.* **237**, 361–367.

Vieth, M., Kolinski, A., Brooks, C. L., III & Skolnick, J. (1995). Prediction of quaternary structure of coiled coils. Applications to mutants of the GCN4 leucine zipper. *J. Mol. Biol.* **251**, 448–467.

Wallqvist, A. & Ullner, M. (1994). A simplified amino acid potential for use in structure prediction of proteins. *Proteins: Struct. Funct. Genet.* **18**, 267–280.

Wang, Y., Zhang, H., Li, W. & Scott, R. A. (1995). Discriminating compact nonnative structures from the native structure of globular proteins. *Proc. Natl Acad. Sci. USA*, **92**, 709–713.

Williams, R. L., Vila, J., Perrot, G. & Scheraga, H. A. (1992). Empirical solvation models in the context of conformational energy searches: application to bovine pancreatic trypsin inhibitor. *Proteins: Struct. Funct. Genet.* **14**, 110–119.

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins: Struct. Funct. Genet.* **6**, 193–209.

Wodak, S. J. & Janin, J. (1980). Analytical approximation to the accessible surface area of proteins. *Proc. Natl Acad. Sci. USA*, **77**, 1736–1740.

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

Yue, K. & Dill, K. A. (1995). Forces of tertiary structural organization of globular proteins. *Proc. Natl Acad. Sci. USA*, **92**, 146–150.

Edited by F. E. Cohen

(Received 31 October 1995; received in revised form 1 February 1996; accepted 8 February 1996)