J. Mol. Biol. (1996) **257**, 716–725

JMB

Using a Hydrophobic Contact Potential to Evaluate Native and Near-native Folds Generated by Molecular Dynamics Simulations

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

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

**Corresponding author*

There are several knowledge-based energy functions that can distinguish the native fold from a pool of grossly misfolded decoys for a given sequence of amino acids. These decoys, which are typically generated by mounting, or ''threading'', the sequence onto the backbones of unrelated protein structures, tend to be non-compact and quite different from the native structure: the root-mean-squared (RMS) deviations from the native are commonly in the range of 15 to 20 Å. Effective energy functions should also demonstrate a similar recognition capability when presented with compact decoys that depart only slightly in conformation from the correct structure (i.e. those with RMS deviations of $\sim$5 Å or less). Recently, we developed a simple yet powerful method for native fold recognition based on the tendency for native folds to form hydrophobic cores. Our energy measure, which we call the hydrophobic fitness score, is challenged to recognize the native fold from 2000 near-native structures generated for each of five small monomeric proteins. First, 1000 conformations for each protein were generated by molecular dynamics simulation at room temperature. The average RMS deviation of this set of 5000 was 1.5 Å. A total of 323 decoys had energies lower than native; however, none of these had RMS deviations greater than 2 Å. Another 1000 structures were generated for each at high temperature, in which a greater range of conformational space was explored (4.3 Å average RMS deviation). Out of this set, only seven decoys were misrecognized. The hydrophobic fitness energy of a conformation is strongly dependent upon the RMS deviation. On average our potential yields energy values which are lowest for the population of structures generated at room temperature, intermediate for those produced at high temperature and highest for those constructed by threading methods. In general, the lowest energy decoy conformations have backbones very close to native structure. The possible utility of our method for screening backbone candidates for the purpose of modelling by side-chain packing optimization is discussed.

Keywords: protein folding; hydrophobic interaction; fold recognition; molecular dynamics simulation; side-chain packing

Introduction

A successful approach to the *ab initio* protein folding problem must surmount two key obstacles. The first is the satisfactory presentation of candidate folds either by enumeration or by minimization in a suitably continuous conformational space. The second is the development of an energy function for which the correct native structure has a lower energy than all other conformations. Progress towards the latter has been made using knowledge-based potentials, most of which are derived by applying the principles of statistical mechanics to the observed relationships between sequence and structure in the database (Kocher *et al*., 1994; Sippl, 1995; see also references therein). Other potentials, such as those first developed by Eisenberg & McLachlan (1986), exploit the tendency for non-polar atoms (or groups of atoms) to be inaccessible

Abbreviations used: RMS, root-mean-squared; PDB, Protein Data Bank; MD, molecular dynamics; HF, hydrophobic fitness; cRMS, coordinate root-mean-squared; sd, standard deviation.

0022–2836/96/130716–10 $18.00/0

to solvent in the native conformations of globular proteins (Vila *et al*., 1991; Koehl & Delarue, 1994a; Wang *et al*., 1995a,b). Energy functions of both types can be used to discriminate native folds from incorrect decoys built by mounting, or ''threading'', the sequence of a polypeptide on the backbone fragments taken from known three-dimensional structures. This technique was pioneered by Sippl and co-workers (Hendlich *et al*., 1990) and has since been used extensively by others (Maiorov & Crippen, 1992; Bryant & Lawrence, 1993; Kocher *et al*., 1994; Bauer & Beyer, 1994; Huang *et al*., 1995).

Although the structures generated by threading upon existing backbones automatically satisfy excluded volume constraints, these alternative folds are unsuitable for the sequence, as they are generally non-compact and quite different from the native structure (Wang *et al*., 1995a,b; Monge *et al*., 1995). Thus, most energy functions can easily discriminate the native fold from quite dissimilar folds. High quality energy functions should also be able to discriminate the correct conformation from near-native conformations. It is reasonable to expect that such potentials will assign energies to near-native conformations that are between the relatively high energies of unrelated structures (i.e. those built from threading) and the energy minimum corresponding to the crystal structure. It would also be useful if the ''energies'' of these backbones were to vary monotonically with the corresponding root-mean-squared (RMS) deviation from the native structure. The ability to judge among many near-native backbones has broadly reaching implications for structure prediction. For instance, methods which model all-atom structures by side-chain packing optimization require fixed backbones that are very close to the crystal structure on which to arrange the mobile side-chains (Lee & Subbiah, 1991; Holm & Sander, 1991; Tuffery *et al*., 1991; Desmet *et al*., 1992; Dunbrack & Karplus, 1993; Eisenmenger *et al*., 1993; Wilson *et al*., 1993; Koehl & Delarue, 1994a; Lee, 1994; Tanimura *et al*., 1994). For the prediction of an unknown structure using side-chain packing methods, one must typically borrow backbones from homologous structures after assigning equivalent residues, normally by sequence alignment or threading techniques (Chung & Subbiah, 1995a). This approach would benefit from the use of a method which could triage poor backbones from suitable ones in the large pool of candidates that are typically available. It is not necessarily true that the backbones of proteins with sequences closer to the sequence to be modelled make better templates than others that are somewhat less similar in sequence. An energy function capable of evaluating the quality of candidate backbones prior to any detailed side-chain prediction is helpful, since the energy functions used to pack side-chains cannot be used reliably to generate accurate models if the backbones deviate from the native form by more than approximately 2 Å (Chung & Subbiah, 1995a, 1996). Another area in which a discrimination function could prove useful is related to *ab initio* folding. By holding the units of secondary structure fixed (i.e. assuming a perfect secondary structure prediction), it is now possible to explore conformational space by enumeration (Park & Levitt, 1995) or minimization (Monge *et al*., 1995). Potentials used to screen the millions of candidates generated from such methods need to be more discriminating than those used to screen the thousands of decoys generated by threading trials.

A final application relates to the increasingly common use of structures derived from nuclear magnetic resonance (NMR) spectroscopy as main-chain search models in a molecular replacement approach to solving the higher resolution crystal structures of the same or a related protein. Given the ensemble of some 20 NMR structures that are deposited into the Brookhaven Protein Data Bank (PDB; Bernstein *et al*., 1977), it would be useful to select and use the backbone from the model that is likely to be most accurate. Such selection ability is also useful when multiple backbone models are available from related proteins obtained by previous X-ray crystal studies.

Early studies by Scheraga and co-workers produced a set of 39 near-native models using a Monte Carlo algorithm followed by local energy minimization (Vila *et al*., 1991). These exhibited RMS deviations from the native ranging from 0.68 to 1.33 Å from all C^{α} atoms. A series of different empirical solvation energy functions was then developed and applied to this set, yielding lowest energies for the native fold in many cases. Moreover, most of these functions showed a positive monotonic relationship with the RMS deviation from the native conformation. However, their test set of 39 conformations is quite modest in size and for that reason may not have posed a sufficient challenge for the energy functions they used.

Alternatively, molecular dynamics (MD) simulations of proteins have been used to generate pools of near-native conformations for evaluation by energy functions. Scott and co-workers have shown convincingly that their atomic solvation model nearly always distinguishes the native conformation from hundreds of decoys generated by MD at various temperatures (Wang *et al*., 1995b). Furthermore, their particular atomic solvation model was found to be more effective at selecting the native structure from such alternate conformations than several assorted empirical energy functions, including contact models, mean force models, local backbone models and combinations thereof (Wang *et al*., 1995a), as well as other atomic solvation models (Wang *et al*., 1995b).

Recently, we have designed an extremely simple yet powerful discrimination function. This novel method is not based on a statistical mechanical framework, nor is it dependent upon empirical solvation parameters; rather, it seeks the structural feature of a hydrophobic core (Huang *et al*., 1995). We treat the polypeptide as a ''binary

code'' of hydrophobic and polar residues. Our application of this model towards fold recognition complements the theoretical work by Dill on protein folding and stability (Dill *et al*., 1995, and references therein) and the experimental approach of Hecht towards *de novo* protein design (Kamtekar *et al*., 1993). In our earlier work we showed that our method was capable of recognizing the native fold from threaded decoys by a larger average margin (measured in standard deviation units) than other methods. It is only natural that our potential, having tested well for discriminating native from quite unrelated folds, should now be tested against a more challenging pool of near-native conformations.

In this work, we evaluate the ability of our energy function to discriminate the native structures in a test set of proteins from a thousand decoys generated by MD at 298 K. We then repeat this trial for the same set of proteins at 498 K. The simulations are a nanosecond in length and have explicit hydrogen atoms and water molecules present. We are unaware of another study that generates as many near-native folds over an MD simulation of similar length in water for the purpose of testing a discrimination function. The possible utility of our potential for structure prediction, especially in the area of homology modelling by side-chain packing, is discussed.

Results

Table 1 summarizes all the results for the five proteins tested in this study. For each population of backbones recorded during simulation, we list the hydrophobic fitness (HF) score of the native structure only, the mean and standard deviation of the HF scores ($\langle$HF$\rangle$), the mean and standard deviation of the co-ordinate RMS deviations ($\langle$RMS$\rangle$), the mean and standard deviation of the radii of gyration ($\langle R_G \rangle$), the number of false positives (i.e. those structures that have HF scores more favorable than the native), and the range of RMS deviations for the false positives.

We have suggested that two desirable qualities for a potential are (1) the ability to indicate near-native folds as those that have mean energies higher than the respective native structure but lower than misfolded structures and (2) a reasonable dependence of the energy on the extent to which a structure differs from the native. To test if the former criterion is met, we performed MD simulations at 298 K to produce a large number of structures that deviate slightly from the native crystal structure. Simulations at 498 K were used to collect an ensemble of structures that deviate somewhat further from the native. A large separation in energy between the set of structures generated at 298 K and a set of 1000 structures generated by threading is easily seen when plotted as histograms. As expected, the energies of the 1000 structures at 498 K fell in between these extremes (Figure 1a to e). Out of the 5000 structures generated at room temperature, 4677 (94%) were higher in energy than the corresponding native structure. These data are very encouraging, because the false positives were all very close to the crystal structure in conformation. Only seven false positives were present in the five high temperature set, reflecting the greater average departure from the native structure. To test if the latter criterion was satisfied, we plotted the HF score as a function of RMS error for the 298 K and 498 K simulations and 1000 decoys built by threading (Figure 2a to e). As expected, the average HF score of a 498 K simulation is less favorable than that of the corresponding 298 K simulation, but more favorable than that of the threaded structures (Table 1). The function does not monotonically increase with RMS deviation; i.e. between a pair of structures close in RMS deviation, the one with the more favorable HF score is not necessarily the one with lower RMS deviation. Nevertheless, on aggregate, the HF score for all 2000 simulated structures very clearly increases with RMS deviation from the native structure, satisfactorily meeting the second criterion.

Discussion

Our results demonstrate that our hydrophobic contact potential has responded well to the

Table 1. Summary of the results for the five proteins tested in this work

Protein (residues)	HF (native)	Protocol	$\langle$RMS$\rangle \pm$sd (Å)	$\langle$Rg$\rangle \pm$sd (Å)	$\langle$HF$\rangle \pm$sd	No. decoys	No. false positives (fp)	RMS range fp (Å)
1ctf	–31.48	MD@298 K	1.14($\pm$0.12)	10.54($\pm$0.11)	–29.45($\pm$3.27)	1000	271	0.82 to 1.43
(68)		MD@498 K	4.17($\pm$1.27)	10.84($\pm$0.23)	–12.20($\pm$7.67)	1000	7	0.78 to 2.63
		Threading	13.76($\pm$2.44)	19.97($\pm$4.93)	0.38($\pm$1.92)	19,589	0	—
1hdd:D	–36.46	MD@298 K	1.92($\pm$0.53)	11.08($\pm$0.11)	–23.28($\pm$3.24)	1000	0	—
(57)		MD@498 K	4.58($\pm$1.05)	11.46($\pm$0.34)	–10.71($\pm$5.25)	1000	0	—
		Threading	12.37($\pm$2.49)	18.60($\pm$5.20)	0.25($\pm$2.06)	20,707	0	—
1r69	–32.26	MD@298 K	0.94($\pm$0.13)	10.38($\pm$0.08)	–23.7($\pm$2.98)	1000	2	0.77 to 0.87
(63)		MD@498 K	3.81($\pm$0.89)	11.26($\pm$0.36)	–13.82($\pm$24.74)	1000	0	—
		Threading	12.50($\pm$2.21)	18.79($\pm$5.02)	–0.38($\pm$1.94)	20,083	0	—
1ubq	–43.82	MD@298 K	1.65($\pm$0.26)	11.77($\pm$0.16)	–38.36($\pm$3.21)	1000	50	1.46 to 1.98
(79)		MD@498 K	4.30($\pm$1.30)	12.51($\pm$0.37)	–21.50($\pm$6.63)	1000	0	—
		Threading	14.77($\pm$2.41)	19.25($\pm$4.80)	0.25($\pm$1.90)	18,820	0	—
4icb	–40.56	MD@298 K	1.96($\pm$0.27)	11.73($\pm$0.10)	–28.84($\pm$3.12)	1000	0	—
(76)		MD@498 K	4.80($\pm$0.95)	12.73($\pm$0.28)	–14.80($\pm$4.10)	1000	0	—
		Threading	13.59($\pm$2.23)	19.25($\pm$4.80)	–0.36($\pm$2.25)	18,820	0	—

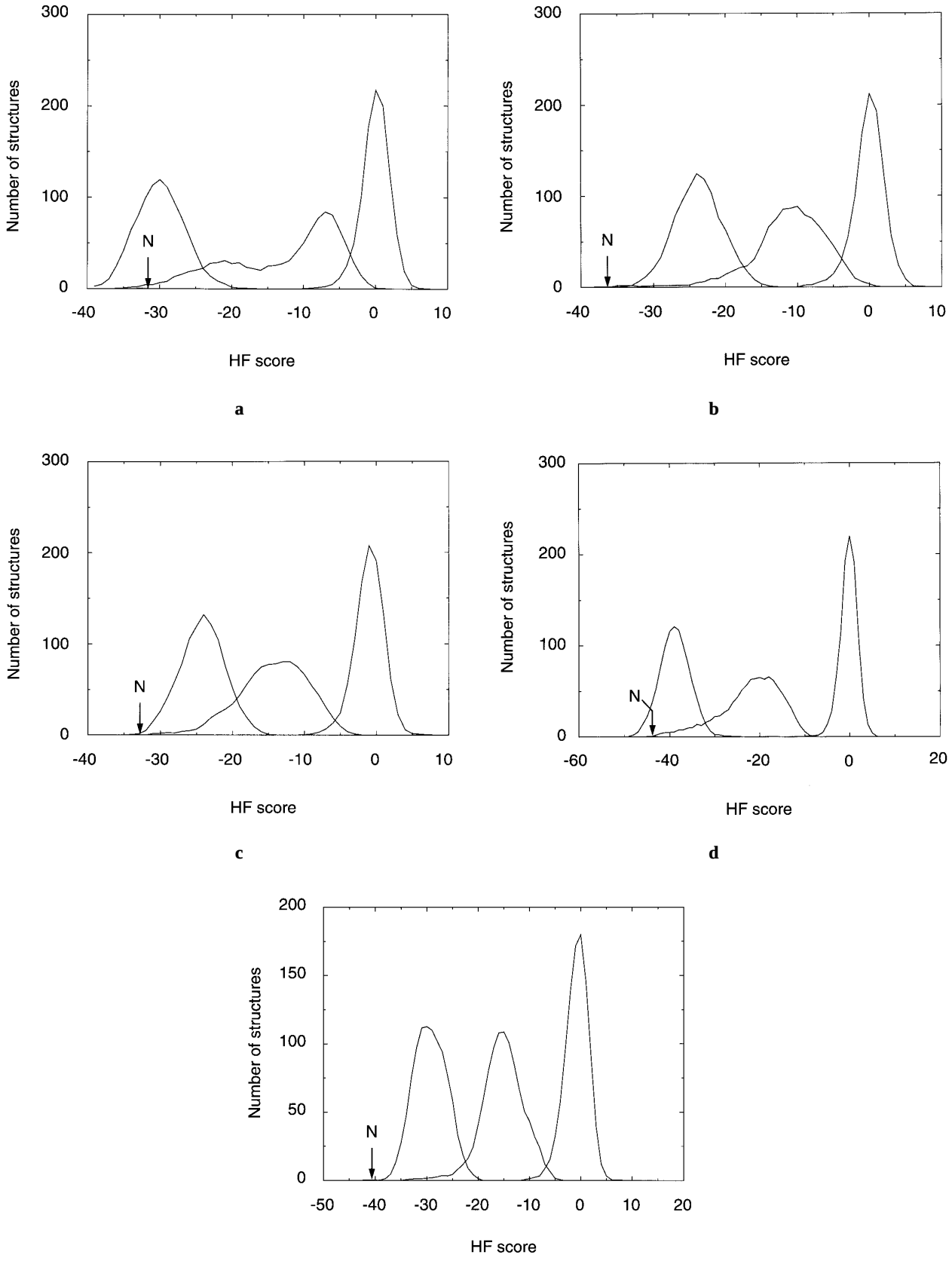
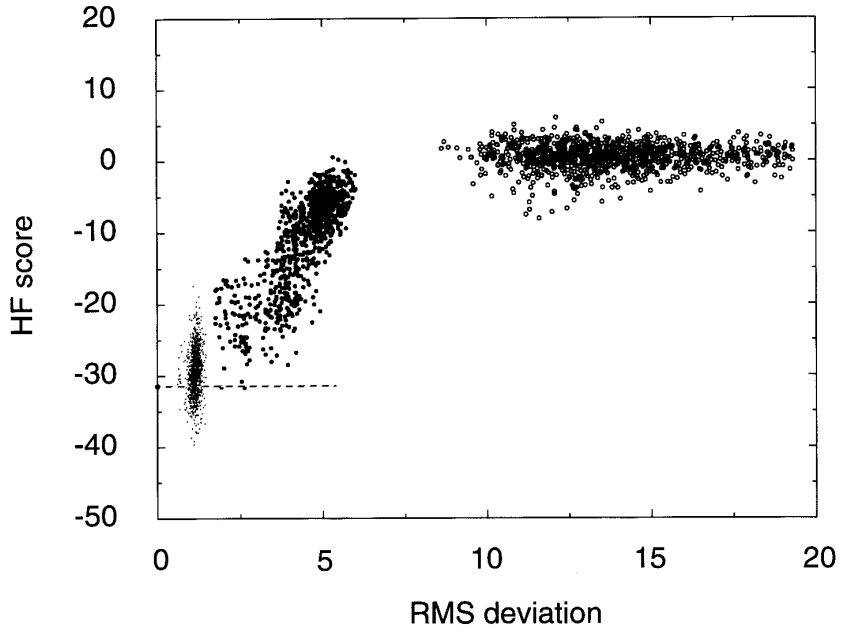


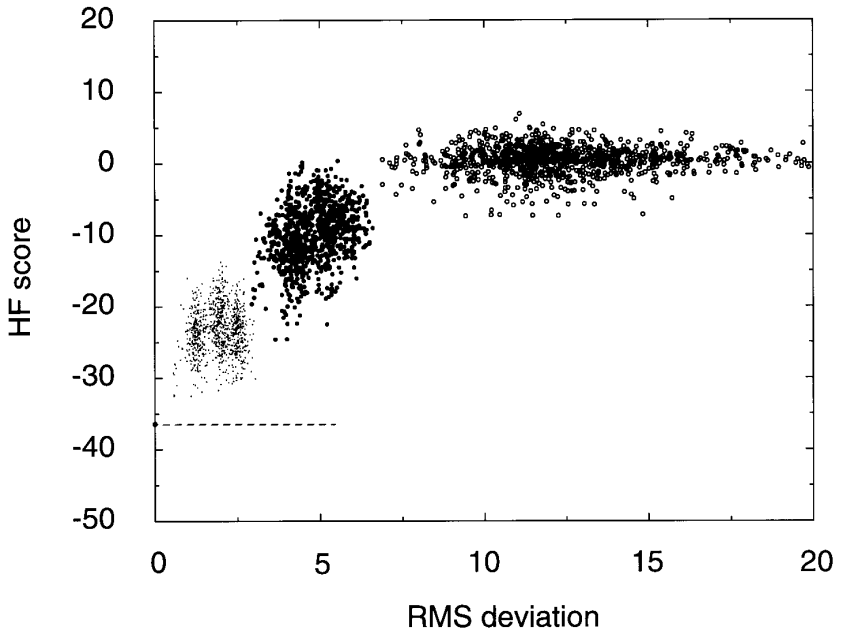
Figure 1. Histograms of HF scores for five proteins. a, 1ctf; b, 1hdd:D; c, 1r69; d, 1ubq; e, 4icb. For each, the curve on the left, center, and right represent the energies of 1000 conformations generated by molecular dynamics (MD) simulation at 298 K, by MD simulation at 498 K, and by threading. The native score for each is indicated by an arrow labeled with an N.

challenge of native fold recognition. Given that the correct backbone is present amongst a large pool of decoys, our method will either recognize the native as the one lowest in energy or place the native amongst the very lowest in energy. Overall, 97% of the decoys had HF scores less favorable than their respective native structures. In nearly all the cases where a near-native structure was lower in energy than the native structure, it was very close in conformation to the respective native structure. For example, only one decoy, one of the 498 K structures of 1ctf, had a lower energy score than the native structure and an RMS deviation greater than 2 Å. In their review, Maiorov & Crippen (1994) have suggested that deviations in conformation on the order of 1 to 2 Å are small and typically yield similar values from energy functions. We also highlight this 2 Å cutoff because of its apparent importance in modelling by side-chain packing. Recent data suggest that modelling by this method requires templates within ~2 Å RMS deviation of the ideal for any degree of accuracy (Chung & Subbiah, 1996). Thus, we regard energy functions that can consistently identify the native fold, or an alternative within 2 Å, as useful.

Although the discrimination capability for our simple energy function was impressive overall, it was not perfect. For two proteins (1hdd:D and 4icb),

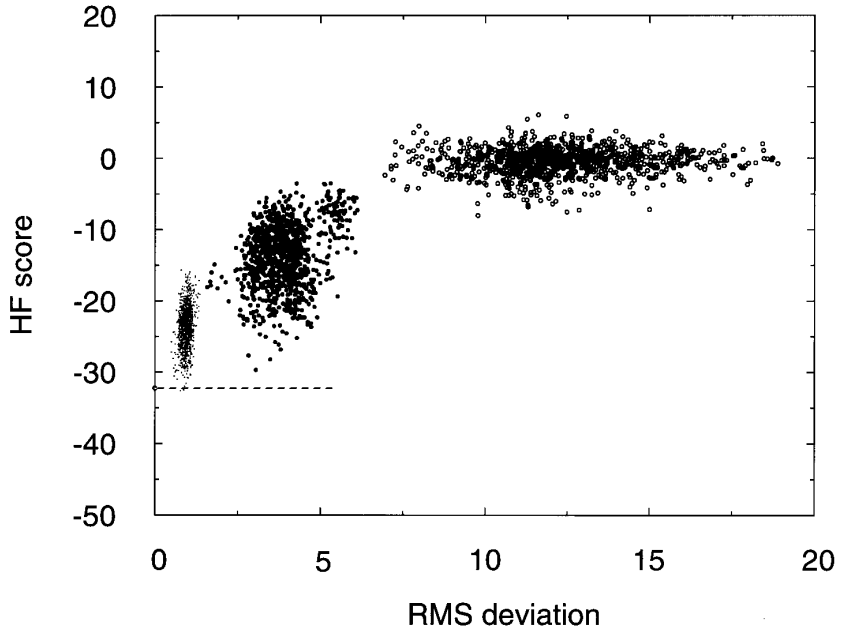


a

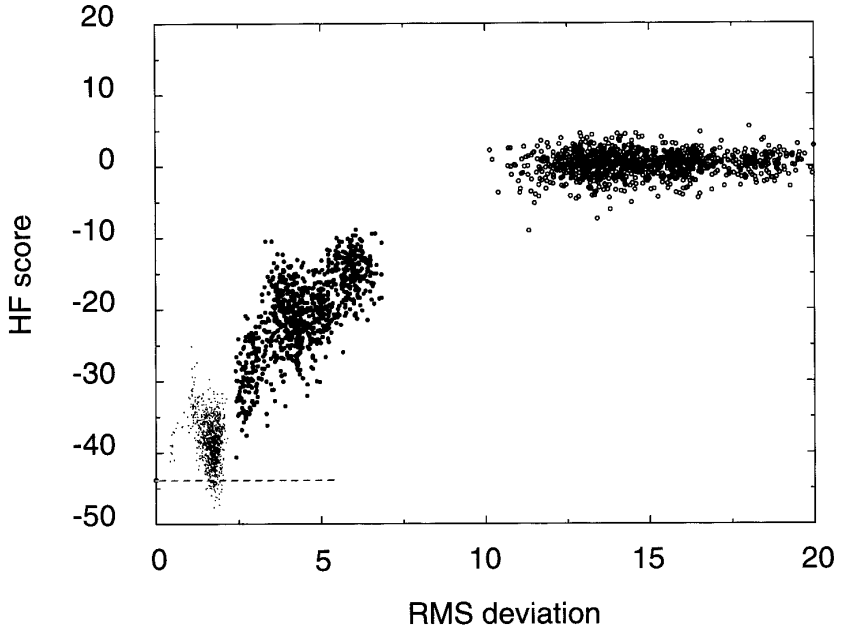


b

Figure 2a–b *(legend on page 722)*



c



d

Figure 2c–d *(legend on page 722)*

all 1000 decoys generated at 298 K were higher in energy than the native conformation, but for another two proteins (1ctf and 1ubq), several decoy structures were misrecognized as native. The function does not incorrectly identify decoy structures as native for 1hdd:D and 4icb, presumably because the structures generated are sufficiently different from the crystal structure (as demonstrated by the relatively high ⟨RMS⟩ for these simulations) and therefore fail to preserve the ''native'' hydrophobic contacts. False positives arise when additional contacts not present in the crystal structure are made by the hydrophobic residues of a slightly deformed structure. The ''hard cutoff'' we impose in our criterion for contact at 7.3 Å, coupled with many small changes in conformation at room temperature, can lead to these additional contacts (see Method Development). The simulation of 1ctf generated the most false positives (271). Close inspection of the inter-residue contact map of the crystal structure of 1ctf, i.e. considering the distances from one virtual centroid to another, reveals two pairs of hydrophobic residues just outside the 7.3 Å range we defined for contact. Specifically, Phe54·Val56 and Leu106·Val117 were both within 0.1 Å of making ''contact'' in the crystal

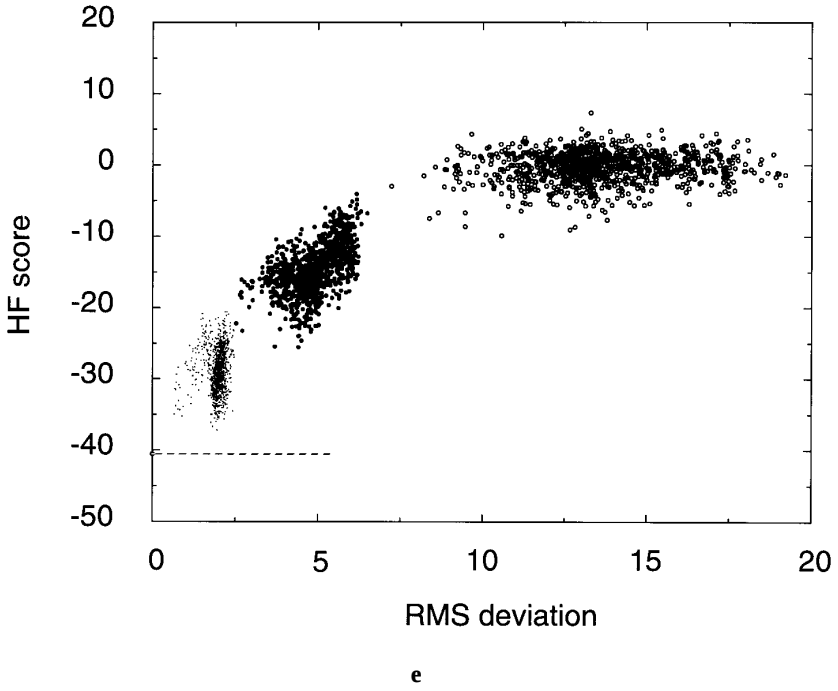


Figure 2. Plots of HF scores *versus* root-mean-squared (RMS) deviation from the native structure. a, 1ctf; b, 1hdd:D; c, 1r69; d, 1ubq; e, 4icb. The population of structures generated by MD simulation at 298 K is depicted as dots, the population at 498 K as filled circles and the collection of threaded structures as open circles. Each set contains 1000 structures; for clarity, data from the 498 K set, that overlapped with the 298 K set, are not shown. The HF score of the native fold is indicated by a broken line; hence, all structures with values below the line are regarded as false positives.

structure according to our simplified model. Thus, the crystal structure fails to make two contacts that are easily accessible to the structures generated by MD simulation. Since the decoys of 1ctf exhibited a low ⟨RMS⟩ (Table 1), many structures were generated that made these ''excess'' contacts while preserving the other hydrophobic contacts, leading to the relatively high number of false positives. Analysis of the room temperature structures revealed that the contacts Phe54:Val56 and Leu106:Val117 were present in 163 and 242 out of the 271 false positives, respectively. In the case of 1ubq, two pairs of hydrophobic residues in the crystal structure were within 0.1 Å of contact, but since the ⟨RMS⟩ of the simulation was higher than that of 1ctf, only 50 false positives resulted (Table 1). The simulation of 1r69 had the lowest ⟨RMS⟩ of all, yet only yielded two false positives. As might be expected, there were no hydrophobic pairs within 0.1 Å of contact in the crystal structure. It therefore appears that false positives are artefacts caused by simplifying side-chains to single coordinates and by defining inter-residue contacts at a hard cutoff at 7.3 Å. However, since these false positives must preserve native contacts in addition to making the excess contacts, no false positive (in this local sampling of conformational space) can depart significantly from the crystal structure overall. Our results are consistent with this expectation.

It is important to compare our method with established potentials applied towards this same problem, such as the recent method of Wang *et al*. (1995a,b) which has demonstrated impressive discrimination. This method subjects each all-atom model to 1000 cycles of energy minimization, after which it calculates accessible surface areas and solvation energies. It is a computationally intensive procedure that considers complete all-atom models, the final energies of which are used to classify those backbones as likely to be near-native. As such, it is not designed explicitly for rapid screening of candidate backbones for further modelling. On the other hand, the detailed information required for this potential might very well result in higher effectiveness: only 7 out of 8200 non-native structures (generated by Monte Carlo and molecular dynamics simulation) were assigned energies lower than the native.

Superficially, it might appear that our HF method is a less discriminating measure than that of Wang *et al*. (1995a,b), since 330 of the total 10,000 near-native structures generated were misrecognized. However, the MD protocols used to generate the decoys closest to the native structures in our study and that of Wang *et al*. (1995b) differ in at least two respects. First, it appears that the incorrect structures that we generated at room temperature were, on average, closer to the native than the corresponding structures used by Wang *et al*. (1995b). For example, ubiquitin was selected as one of the test cases in both studies. The ⟨RMS⟩ from our 298 K simulation was 1.65 Å (0.3 Å standard deviation); for the corresponding simulation by Wang *et al*. (1995b), it was 3.12 Å (0.2 Å standard deviation). The overall ⟨RMS⟩ for all room temperature simulations (1.5 Å) was also lower than the overall ⟨RMS⟩ for those of Wang *et al*. (1995b) (4.1 Å), but one must be mindful that the test sets in these two studies also differed. Moreover, none of the other protocols (e.g. Monte Carlo or threading) used by Wang *et al*. (1995b) produced decoy conformations with lower ⟨RMS⟩ than those from our 298 K MD simulations. Second, we presented our function with 1000 decoys gathered

Table 2. RMS deviation of ten lowest energy non-native structures for each protein

Protein	⟨RMSD⟩ ± sd	No. RMSD ⩾ 2.0 Å
1ctf	1.29(±0.61)	1
1hdd:D	0.83(±0.16)	0
1r69	1.72(±1.37)	4
1ubq	1.36(±0.90)	2
4icb	0.83(±0.19)	0

from the 298 K simulation for each protein, compared with only 100 decoys per protein in the investigation by Wang *et al*. (1995b). Given that we have apparently challenged our potential with more structures closer to the native than Wang *et al*. (1995b), a direct comparison between the two methods must be made cautiously. Since the ⟨RMS⟩ for all our high temperature simulations was 4.3 Å, these structures may be more representative of the decoy sets used by Wang *et al*. (1995b). There were only seven false positives amongst the 5000 structures generated at high temperature; of these seven, the ⟨RMS⟩ was 1.4 Å. These results are comparable to those of Wang *et al*. (1995b); hence, the extent to which our method is less discriminating (if at all) than their more sophisticated and computationally expensive measure is unclear.

A related challenge we pose to our energy function is the selection of a suitable backbone for further modelling (e.g. by side-chain packing) from a pool of candidates lacking the ideal backbone. We have already discussed the utility of backbones that deviate slightly (⩽2 Å) from ideal for this purpose. Yet, since our function does not show a perfectly monotonic increase in the HF score with respect to RMS, we cannot rule out the possibility that the lowest energy conformation is not the one closest to the native structure. Nor is it guaranteed that all low-energy structures are reasonably close to the native at all. Thus, we took the ten non-native structures with lowest energy from each 498 K simulation and examined the RMS deviation of each. We chose the high temperature simulations because they generate backbones with more structural diversity and higher ⟨RMS⟩. The results in Table 2 demonstrate that choosing the lowest energy structures generally yields backbones with low RMS deviations. Since the 498 K simulation used as its starting structure the X-ray crystal structure, these low-energy backbones were those generated early in the simulation. Only seven out of the 50 low-energy structures exhibited RMS deviations greater than 2 Å, and the ⟨RMS⟩ over the entire low-energy set was 1.2 Å. For comparison, the ⟨RMS⟩ over our five 498 K simulations was 4.3 Å. Thus, using the HF score as a means of ''purifying'' the collection of candidate backbones for those with the minimal deviation from native is effective.

We draw three conclusions from our data. First, our energy function, which performed well at recognizing native folds from grossly misfolded structures, may also be applied towards discriminating native from near-native structures. It certainly demonstrates a potential for effective screening against structures that deviate by ∼4 Å (i.e. those akin to a high temperature ensemble). Even for structures that deviate between 1 and 2 Å from the native, the discrimination capability is respectable. Second, since our function generally assigns low energies to those structures that are close in conformation to the native, it can also serve as a means of evaluating backbones for homology modelling by side-chain packing or perhaps as a tool for selecting amongst available models for solving X-ray crystal structures by molecular replacement. Finally, since the discrimination capability is apparently not compromised by the simplicity of our method, we regard its speed and ease of implementation as advantageous.

Method Development

Selection of a test set of proteins

We selected a set of five small monomeric proteins from the Brookhaven PDB on which to test our energy function. This set is composed of the D monomer of the engrailed homeodomain from fruit fly (PDB entry 1hdd:D), the amino-terminal domain of 434 repressor (1r69), the carboxy-terminal domain of L7/L12 50 S ribosomal protein from *Escherichia coli* (1ctf), ubiquitin from human erythrocytes (1ubq), and bovine calbindin D9K (4icb). These are well suited for evaluation by our potential because they lack additional stabilizing interactions such as disulfide bonds, prosthetic groups and metal ions, all of which our method disregards for the sake of simplicity. Other studies have also chosen test sets based on these criteria (Wang *et al*., 1995b).

Generation of incorrect backbones by molecular dynamics simulation

Molecular dynamics simulations of each protein in water were used to produce large pools of near-native backbone conformations. We ran two simulations for each protein, one at room temperature (298 K) and another at high temperature (498 K). The program ENCAD (Energy Calculation and Dynamics) was used for all simulations. The program and force field are described elsewhere (Levitt, 1983; Levitt *et al*., 1995). Each protein was placed in an appropriately sized box of water. All water molecules closer than 1.67 Å to the non-hydrogen atoms of the solute were removed. The box was scaled to the appropriate volume using the experimental density of water (0.997 g/ml at 298 K and 0.829 g/ml at 498 K; Kell, 1967). The system was run through a cycle of minimizations and dynamics to allow the protein to relax in solution. Then the system was equilibrated to the desired temperature using increments of 0.2 deg.K. The simulations used a two femtosecond time step, a periodic box and a smooth force-shifting truncation applied previously by ENCAD for protein and pure water simulations (Levitt *et al*., 1995). We recorded the coordinates at every picosecond over a total simulation time of one nanosecond, thereby generating 1000 backbones per simulation. Each simulation required approximately seven CPU days on a Digital Equipment Corporation Alphastation 250 4/266.

Evaluation of backbones by hydrophobic fitness score

The criterion by which we measure the compatibility of the backbone for the given sequence of amino acids is a measure called the HF score (Huang *et al*., 1995). This is a simple and fast method that uses a reduced representation for both sequence and structure. Each amino acid residue is summarized by a virtual side-chain centroid and is classified as either hydrophobic or polar. We reduce the 1000 all-atom structures generated by molecular dynamics to this simplified form and apply our energy potential to them, yielding a HF score for each. The computation of the HF score is summarized below:

Hydrophobic Fitness (HF) score =

$$-\frac{\left(\sum_i B_i\right)\left(\sum_i (H_i - H_i^\circ)\right)}{n^2}$$

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

Generation of incorrect structures by threading

We compare the HF scores of near-native structures generated by MD with those of grossly incorrect structures produced from threading. We construct these models by mounting the binary-encoded sequence of each test protein onto the backbones of a set of 107 unrelated structures using every available ungapped alignment, as in our earlier study (Huang *et al*., 1995).

General computational methods

To assess the extent to which each of the backbones differs from the crystal structure, we compute the coordinate root-mean-squared (cRMS) deviation of corresponding C^α atoms between the crystal structure and each non-native structure with the following equation:

$$\text{cRMS deviation} = \left(\sum_{i=1}^{n} |\mathbf{r}_{ai} - \mathbf{r}_{bi}|^2 \Big/ n\right)^{1/2}$$

where $\mathbf{r}_{ai}$ and $\mathbf{r}_{bi}$ are the positions of the ith C^α atom of the crystal structure and the incorrect structure, respectively, after optimal superimposition of corresponding C^α atoms (Kabsch, 1978) and n is the number of residues. In this study all RMS deviations are computed as cRMS deviations.

For each model, we also calculate the radius of gyration (R_G) using the C^α atoms. R_G is a simple measure of compactness for a given fold, and is computed as:

$$R_G = \left[(1/n)\sum_{i=1}^{n} R_i^2\right]^{1/2}$$

where R_i is the distance of the ith C^α atom from the center of mass of the protein and n is the number of residues.

Acknowledgements

This work was supported by the Department of Energy (grant number DE-FG03-95ER62135) and the National Institutes of Health (grant number GM 41455). E.S.H. is a National Science Foundation Pre-doctoral Fellow. The authors thank S. Chung, D. Hinds, B. Park and M. Gerstein for their helpful discussion.

References

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

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

Chung, S. Y. & Subbia, S. (1995). The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. *Protein Sci.* **4**, 2300–2309.

Chung, S. Y. & Subbiah, S. (1996). How similar must a template protein be for homology modeling by side-chain packing methods? In *Proceedings of the Pacific Symposium on Biocomputing, Hawaii, USA* (Hunter, L. & Klein, T. E., eds), pp. 126–141, World Scientific, New Jersey.

Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, **356**, 539–542.

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.

Dunbrack, R. L., Jr & Karplus, M. (1993). Backbone-dependent rotamer library for proteins: application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.

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

Eisenmenger, F., Argos, P. & Abagyan, R. (1993). A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* **231**, 849–860.

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.

Holm, L. & Sander, C. (1991). Database algorithm for generating protein backbone and side-chain coordinates from a C^α trace: application to model building and detection of coordinate errors. *J. Mol. Biol.* **218**, 183–194.

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.

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.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary

patterning of polar and nonpolar amino acids. *Science*, **262**, 1680–1685.
Kell, G. S. (1967). Precise representation of volume properties of water at one atmosphere. *J. Chem. Eng. Data*, **12**, 66–69.
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.
Koehl, P. & Delarue, M. (1994a). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249–275.
Koehl, P. & Delarue, M. (1994b). Polar and nonpolar atomic environments in the protein core: implications for folding and binding. *Proteins: Struct. Funct. Genet.* **20**, 264–278.
Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* **236**, 918–939.
Lee, C. & Subbiah, S. (1991). Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* **217**, 373–388.
Levitt, M. (1983). Molecular dynamics of native proteins. I. Computer simulation of trajectories. *J. Mol. Biol.* **168**, 595–620.
Levitt, M., Hirschberg, M., Sharon, R. & Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Commun.* **91**, 215–231.
Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.
Maiorov, V. N. & Crippen, G. M. (1994). Learning about protein folding via potential functions. *Proteins: Struct. Funct. Genet.* **20**, 167–173.
Monge, A., Lathrop, E. J. P., Gunn, J. R., Shenkin, P. S. & Friesner, R. A. (1995). Computer modeling of protein folding: conformational and energy analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995–1012.
Park, B. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493–507.
Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.
Tanimura, R., Kidera, A. & Nakamura, H. (1994). Determinants of protein side-chain packing. *Protein Sci.* **3**, 2358–2365.
Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). A new approach to the rapid determination of protein side-chain conformations. *J. Biomol. Struct. Dynam.* **8**, 1267–1289.
Vila, J., Williams, R. L., Vásquez, M. & Scheraga, H. A. (1991). Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins: Struct. Funct. Genet.* **10**, 199–218.
Wang, Y., Zhang, H., Li, W. & Scott, R. A. (1995a). Discriminating compact nonnative structures from the native structure of globular proteins. *Proc. Natl Acad. Sci. USA*, **92**, 709–713.
Wang, Y., Zhang, H. & Scott, R. A. (1995b). A new computational model for protein folding based on atomic solvation. *Protein Sci.* **4**, 1402–1411.
Wilson, C., Gregoret, L. M. & Agard, D. A. (1993). Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.* **229**, 996–1006.

Edited by F. E. Cohen

(Received 12 October 1995; accepted 9 January 1996)