PROTEINS: Structure, Function, and Genetics 52:561–572 (2003)

Spatial Profiling of Protein Hydrophobicity: Native Vs. Decoy Structures

Ruhong Zhou,* B. David Silverman, Ajay K. Royyuru, and Prasanna Athma
IBM Thomas J. Watson Research Center, Yorktown Heights, New York

ABSTRACT **A recent study of 30 soluble globular protein structures revealed a quasi-invariant called the hydrophobic ratio. This invariant, which is the ratio of the distance at which the second order hydrophobic moment vanished to the distance at which the zero order moment vanished, was found to be 0.75 ± 0.05 for 30 protein structures. This report first describes the results of the hydrophobic profiling of 5,387 non-redundant globular protein domains of the Protein Data Bank, which yields a hydrophobic ratio of 0.71 ± 0.08. Then, a new hydrophobic score is defined based on the hydrophobic profiling to discriminate native-like proteins from decoy structures. This is tested on three widely used decoy sets, namely the Holm and Sander decoys, Park and Levitt decoys, and Baker decoys. Since the hydrophobic moment profiling characterizes a global feature and requires reasonably good statistics, this imposes a constraint upon the size of the protein structures in order to yield relatively smooth moment profiles. We show that even subject to the limitations of protein size (both Park & Levitt and Baker sets are small protein decoys), the hydrophobic moment profiling and hydrophobic score can provide useful information that should be complementary to the information provided by force field calculations. Proteins 2003;52:561–572.**

Key words: protein decoys; hydrophobic profiling; hydrophobic ratio; hydrophobic score; second-order moment; globular proteins

INTRODUCTION

One essential requirement of protein structure prediction methods is the ability to discriminate native and native-like conformations from significantly misfolded ones or so-called protein decoys. Present methods can be roughly catalogued into three categories: knowledge-based, physics-based, or a combination of the two.[1–3] Several varieties of knowledge-based empirical scoring functions have been proposed for ranking protein conformations.[4–9] One recent interesting observation made by Silverman[10] is that 30 diverse globular native proteins exhibit some common features of their hydrophobic moment profiles. A relatively universal constant of 0.75 called the hydrophobic ratio was found, which is defined as the ratio of radii from the protein centroid at which the second order hydrophobic moment and the zero order moment vanished (a detailed definition is given in Molecular Moments and Hydrophobicity Profiling). It is of interest to see if (1) this remains true for a large number of globular soluble proteins in the Protein Data Bank (PDB), and (2) this observation can be used to discriminate decoys from native-like structures.

As described previously,[10] the universal spatial transition from the hydrophobic core to the hydrophilic exterior of globular proteins motivated the detailed spatial profiling. With an ellipsoidal characterization of protein shape, an appropriate scaling of residue hydrophobicity and a second-order ellipsoidal moment, it was shown that 30 diverse globular soluble proteins shared detailed spatial features of this transition, with a quasi-invariant hydrophobic ratio of 0.75 ± 0.05 for the protein structures examined. Furthermore, the profiling clearly distinguished some decoys from their native structures.[10,11] In this report, we will examine all the nonredundant soluble globular proteins in PDB, as well as the three widely used decoy sets, namely the Holm and Sander decoys,[8] Park and Levitt decoys,[7,12] and Baker decoys.[5,13] Particular attention will be paid to decoys with small sizes, e.g., the Park and Levitt and Baker decoys.

Decoy structures of small globular soluble proteins have provided test sets for the evaluation of energy functions used in the ab-initio prediction of native protein structures. While an ideal objective would be the determination of a free energy function that selects structures that are either minimally displaced spatially from the native structure or a function that selects the native structure itself, success has not been forthcoming. One suspects that a difficulty in the determination of an appropriate free energy function is related to the approximate manner in which the calculations treat the entropic character of solvation. One global structural feature arising from solvation is the ubiquitous hydrophobic core and hydrophilic exterior of soluble globular proteins. This feature has been used to identify protein structures that might be candidates that approximate the native structure or used to eliminate candidate structures that might not.[5,7,8] Considerations of hydrophobicity together with free energy ap-

*Correspondence to: Ruhong Zhou, IBM Thomas J. Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598. E-mail: ruhongz@us.ibm.com

Received 11 October 2002; Accepted 10 January 2003

TABLE I. Eisenburg Hydrophobicity Consensus Values for Each Amino Acid†

Residue	Consensus
Arginine	−1.76
Lysine	−1.10
Aspartic acid	−0.72
Glutamine	−0.69
Asparagine	−0.64
Glutamic acid	−0.62
Histidine	−0.40
Serine	−0.26
Threonine	−0.18
Proline	−0.07
Tyrosine	0.02
Cysteine	0.04
Glycine	0.16
Alanine	0.25
Methionine	0.26
Tryptophan	0.37
Leucine	0.53
Valine	0.54
Phenylalanine	0.61
Isoleucine	0.73

†See references[15,16] for details of these consensus values.

proaches[2,6,12,14] can provide a more selective procedure than the use of either alone.

Small native protein structures had not been selected[10] to avoid statistical irregularities in the moment profiles. The smallest protein among the set of the thirty previously examined consisted of 64 residues. The next largest consisted of 96 residues. The protein decoys that had been examined were restricted to have a residue number of no less than 100. Since the small protein decoys of Park and Levitt, and those of the Baker group, have been central to ab-initio procedures in discriminating decoys from native structures, it is of interest to see if moment profiling could yield useful supplemental information, even in the regime of profile irregularities due to the discrete spatial distribution of the residues. The intent of the present report is to, therefore, first validate the hydrophobic ratio for all nonredundant soluble globular proteins in PDB and second to develop a new scoring function based on the hydrophobic moment profiling, which can provide useful discrimination between native and decoy structures.

MOLECULAR MOMENTS AND HYDROPHOBICITY PROFILING

Hydrophobicity is widely used to describe the solvation of small organic molecules, proteins, or other molecules in a water solvent. For proteins, each residue exhibits a different degree of hydrophobicity or hydrophilicity, based upon its solubility in water. A value of hydrophobicity, h_i, can then be assigned to each residue of type, i. Table I lists the Eisenberg hydrophobicity consensus values for each amino acid.[15,16]

Since the distribution of hydrophobicity is profiled from the protein interior to the exterior of globular proteins, an ellipsoidal profiling shape had been chosen with axes determined by the inertial tensor $\bar{\mathrm{I}}$, which has components

$$I_{jk} = \int_V \rho\ (\vec{r})\ (r^2\delta_{jk} - x_j x_k)dV, \tag{1}$$

where $\rho(\vec{r})$ is the density of the residue centroids of unit mass, δ_{jk} is the Kronecker delta function with value of 1 if $j = k$ and 0 otherwise. Diagonizing the inertial tensor, one obtains the three principal axes as well as the moments of geometry. The x, y, and z axes are then aligned with the principal axes. The moments of geometry are designated as g_1, g_2 and g_3, with $g_1 < g_2 < g_3$. The ellipsoidal representation generated by these moments is,

$$x^2 + g'_2 y^2 + g'_3 z^2 = d^2, \tag{2}$$

where $g'_2 = g_2/g_1$, $g'_3 = g_3/g_1$. The value d is the major principal axis of the ellipsoid and can be considered as a generalized ellipsoidal radius.

Whatever the initial distribution of residue hydrophobicity, h_i, chosen, the distribution is shifted such that the net hydrophobicity of each protein vanishes. The distribution is then normalized to yield a standard deviation of one. Shifting the residue hydrophobicity distribution for each protein selects a common structural reference and thus enables the quantitative comparison of protein profile shapes and profile features such as the hydrophobic ratio. After scaling, residues with positive hydrophobicity values are referred to as "hydrophobic residues" and those with negative values as "hydrophilic residues" in the following.

The zero-order hydrophobic moment H_0 of the accumulated residue distribution within the ellipsoidal surface specified by d is then written,

$$H_0(d) = \sum_{r<d} h'_i = \sum_{r<d} (h_i - \bar{h})/ < (h_j - \bar{h})^2 >^{1/2}, \tag{3}$$

where the prime designates the value of hydrophobicity of each residue after shifting and normalizing the distribution, $\bar{\mathrm{h}}$ is the mean of the h_i, and $< (h_j - \bar{\mathrm{h}})^2 >^{1/2}$ represents the standard deviation. Therefore, when the value of d is just sufficiently large enough to collect all of the residues, the net hydrophobicity of the protein vanishes. This value of d_0, for which $H_0(d)$ vanishes, assigns a surface as common structural reference for each protein.

Second-order moments amplify the differences between hydrophobic and hydrophilic residues that contribute to the spatial profile of the hydrophobicity distribution. The second-order hydrophobic moment H_2 is defined as,

$$H_2(d) = \sum_{r<d} h'_i\ (x_i^2 + g'_2 y_i^2 + g'_3 z_i^2), \tag{4}$$

where the (x_i, y_i, z_i) denote the position of the ith residue centroid. For globular soluble native protein structures, the zero and second-order moments are positive when d is small. Both increase with distance, d, within the region of the hydrophobic core. At greater values of d, the ratio of hydrophilic to hydrophobic residues increases. The increase of both the zero- and second-order moments with distance then slows and turns around, decreasing with increasing d. Since the second-order moment amplifies

differences in the distribution, this moment will cross zero, becoming negative at a distance below the value of, *d,* at which the zero-order moment vanishes. The location at which the second-order moment vanishes is defined as d_2. As mentioned earlier, the location at which the zero-order moment vanishes is denoted as d_0. The hydrophobic-ratio is then defined as,

$$R_H = d_2/d_0. \quad (5)$$

The study by Silverman[10] showed the hydrophobic-ratio to be a quasi-invariant for 30 globular proteins. The origin of this invariance has been recently identified.[17] In Protein Selection, the hydrophobic ratio will be shown to characterize native and near-native structures. Such a ratio, however, cannot always be defined for arbitrary protein structures. This is particularly true if the second-order moment profile does not exhibit the smooth generic native behavior expected. The hydrophobic ratio would then be unable to provide a continuous score with respect to how deviant a decoy profile is with respect to its native profile. To provide such continuous ranking of each decoy profile with respect to its native profile, a new scoring function will be defined.

PROTEIN SELECTION

The extensive number of globular proteins extracted from the PDB is obtained by the following procedure. All proteins in PDB were downloaded as of February 2002. Conflicts in residue sequences in SEQRES and ATOM records of the PDB files are resolved for each protein chain, resulting in total 30,856 PDB chains (some proteins have multiple chains). SCOP (version 1.53)[18] is then used to identify soluble globular protein domains (class a–e). The domain definition in SCOP is mapped onto the residue ranges in the PDB chains. A nonredundant subset of domain length protein sequences is obtained through a pairwise sequence alignment process that retains domains that have sequence identities below 95%. This gives us a total of 5,786 soluble globular protein domains. Then, 77 multi-chain domains in class e are removed to avoid complexity. As mentioned above, there is a limit in protein size in order to get smooth hydrophobic moment profiles with meaningful statistics. We limit our selection to proteins having more than 70 residues in this study, which gives us a total of 5,387 protein domains.[†]

The Holm and Sander,[8] Park and Levitt[12] and Baker decoy sets[13] examined in this study have been downloaded from the web (http://dd.standford.edu for the Holm and Sander and Park and Levitt set, and http://depts.washington.edu/bakerpg for the David Baker set). Since the hydrophobic moments and ratios involve the spatial profiling of the residue distribution, and this distribution is discretely distributed in space, a typical window of 1 Å in generalized ellipsoidal radius, *d,* had been used to generate the nested ellipsoidal surfaces. This provided reasonable resolution in obtaining the generally smooth moment profiles over the range of protein sizes previously investigated. Protein size imposes a constraint upon the ability to generate relatively smooth profiles. It is found that a relatively smooth profile can be obtained for proteins with a residue number greater than 100. Since Holm and Sander decoys have reasonably large sizes, we selected those with more than 100 residues. This resulted in a total of 14 decoy sets out of total 26, with a protein size ranging from 107 residues to 317 residues. The Park and Levitt and Baker decoys range in size well below this limit so proteins chosen for the present study are limited to a residue number of no less than 60. This is a smaller cutoff than that used for the entire PDB database or the Holm and Sander decoys.

For the Baker decoy sets, we have also applied two other criteria to eliminate decoy sets from the total of 92. The objective is to examine decoys with a broad range of RMSD's and hence a broad range of "similarity" to their native structures:

1. those decoy sets where 10% or less of the decoys have RMSD's from the native structure that are less than 8 Å were eliminated.
2. those decoy sets having the smallest RMSD larger than 4 Å were eliminated.

Thus, decoys significantly displaced in RMSD from their native structures have not been included. This selects the decoys that should be more difficult to distinguish from their native structure. This decoy set elimination together with the residue number limitation reduces the number of Baker sets studied to 11 from the total of 92. The residue number restriction imposed on the Park and Levitt decoy sets reduces the number of sets examined to 4 from a total of 7 (one decoy set has outdated native PDB structures, which has also been eliminated). The PDB entries and number of residues for the proteins finally selected for this study are summarized in Table II. The numbers of residues of these proteins range from 60 to 75. These protein sizes are insufficient, in most cases, to yield smooth hydrophobic moment profiles. It will, however, be shown that even subject to this limitation, the moment profiling can provide useful complementary information to that obtained from energy minimization procedures. The RMSD values for the Park and Levitt decoy sets are supplied by the authors on their web site. These are RMSDs for the C_α atoms. The RMSD values for the Baker decoy sets are not available from the web site and are, therefore, recomputed with the IMPACT program[19,20] for all backbone atoms. The RMSD values based on the C_α atoms, backbone atoms, or all of the atoms will be slightly different, but for the case at hand, they should be equally instructive.

RESULTS

As demonstrated in a previous report on 30 native proteins,[10] the hydrophobic ratio R_H is a "quasi-invariant," which provides a feature based on a second-order moment profile that enables comparison between different native structures. We have presently examined the PDB data

[†]Details of selection process and the final list of protein domains are available upon request.

TABLE II. Native PDB Entries of the Decoy Sets Selected From Both the Park and Levitt Set and Baker Set, and Their Number of Residues and Hydrophobic Ratio

Decoy set	PDB entry	Residues	R_H
ParkLevitt	1ctf	68	0.722
	1r69	63	0.762
	2cro	65	0.722
	3icb	75	0.750
Baker	1c5a	62	0.727
	1ctf	67	0.722
	1hsn	62	0.679
	1leb	63	0.684
	1mzm	67	0.773
	1nkl	70	0.737
	1r69	61	0.762
	1sro	66	0.640
	2ezh	65	0.667
	2fow	66	0.750
	2ptl	60	0.682

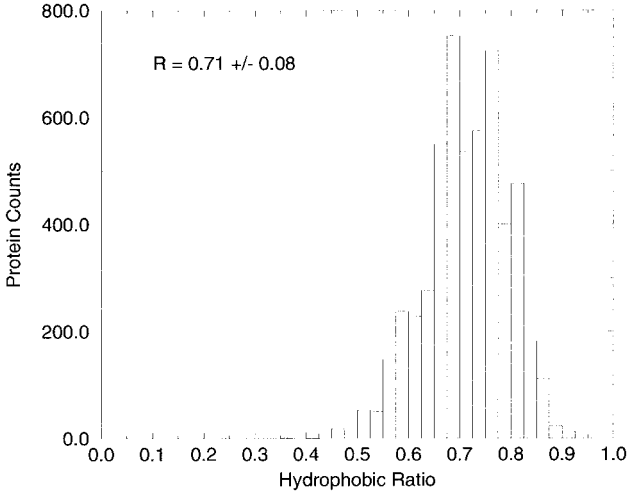

Fig. 1. Hydrophobic ratio R_H for the 5,387 protein domains from the entire Protein Database Bank. It shows a R_H value of 0.71 ± 0.08.

bank of all the protein structures deposited as of February 2002. A total of 5,786 globular protein domains (SCOP classes a–e) have been extracted and examined. As previously mentioned, there is a limitation in protein size in order to get smooth hydrophobic moment profiles with good statistics. We consequently selected proteins having more than 70 residues in this study, which resulted in a total of 5,387 protein domains. The results for this large set of nonredundant soluble globular proteins are shown in Figure 1. The hydrophobic ratio R_H is found to have a mean value of 0.71 with a standard deviation of 0.08. Given that this covers all the soluble globular proteins in PDB, there is indeed a relatively constant of 0.71 ± 0.08 for the hydrophobic ratio.

The origin of the quasi-invariance of the hydrophobic ratio may be of interest. Scaling the values of residue hydrophobicity such that the total hydrophobicity of the protein vanishes sets a length scale for each protein. For the present calculations, it is just the principal major axis of the ellipsoid that encloses all residues. All protein lengths normalized to this distance enable comparison between different proteins. Two other features contribute to the invariance. First, the accumulation of hydrophobic residues (0th order) is found to be greater than the accumulation of hydrophilic residues over the entire range of accumulation with distance. The hydrophilic residues are distributed more towards the exterior and hydrophobic ones more towards the interior, thus the second order moment will favor the hydrophilic residues over the hydrophobic ones at large distances, which result in a crossover in the second-order moment away from the surface, about 7/10th from the center.[17] The crossover distance over the total distance or the hydrophobic ratio is fairly independent of the differential accumulation of hydrophobic and hydrophilic residues. These calculated values of the hydrophobic ratios correspond to the predicted values from a simple two-component nucleation model of hydrophobicity. The decrease in residue density over the length scales as the protein exterior is approached is found to be comparable for different proteins, which is necessary for the hydrophobic ratio to fall within the observed range. These features, contributing to the invariance, are simply revealed by performing calculations on an idealized two-component model of protein hydrophobicity.[17]

In the following, much of the attention will focus on the protein decoys. Holm and Sander decoys had been generated to test their solvation preference method[8] designed to distinguish native from decoy structures. Figure 2 shows the second-order hydrophobic moment profiles for 14 such decoys (one decoy for each protein). All native structures exhibit a second-order profile shape that had been previously found for native proteins. All of the decoy structures, on the other hand, do not show the significant separation between the hydrophobic residues forming the native core and hydrophilic exterior. Their second-order moments fluctuate around zero on the abscissa axis. The hydrophobic ratio cannot be defined for these decoy structures.

The second-order moment profiles of the thousands of Park and Levitt and Baker decoy structures do not, however, always exhibit easy patterns to be discriminated against as in the Holm and Sander single decoy sets. It is also not feasible to visually or manually inspect thousands of profiles. Therefore, a new scoring function is needed to quantitatively rank each decoy profile with respect to an expected native profile. Before such a scoring function is defined, it is of interest to examine the hydrophobic ratios and profiles of these very small-sized native proteins of the Park and Levitt and Baker decoy sets. Interestingly, even subject to this small size limit, all native second-order moment profiles still show a hydrophobic core and a sharp plunge to negative values in the transition from hydrophobic core to hydrophilic exterior. Similar to previous results, the native decoy structures have R_H values that range from 0.64 to 0.77, with a mean of 0.72. The values of R_H for each of the native structures are listed in Table II.

Examination of a few of the decoy profiles reveals several interesting features involved in defining the new scoring function. Figure 3 shows a few representative

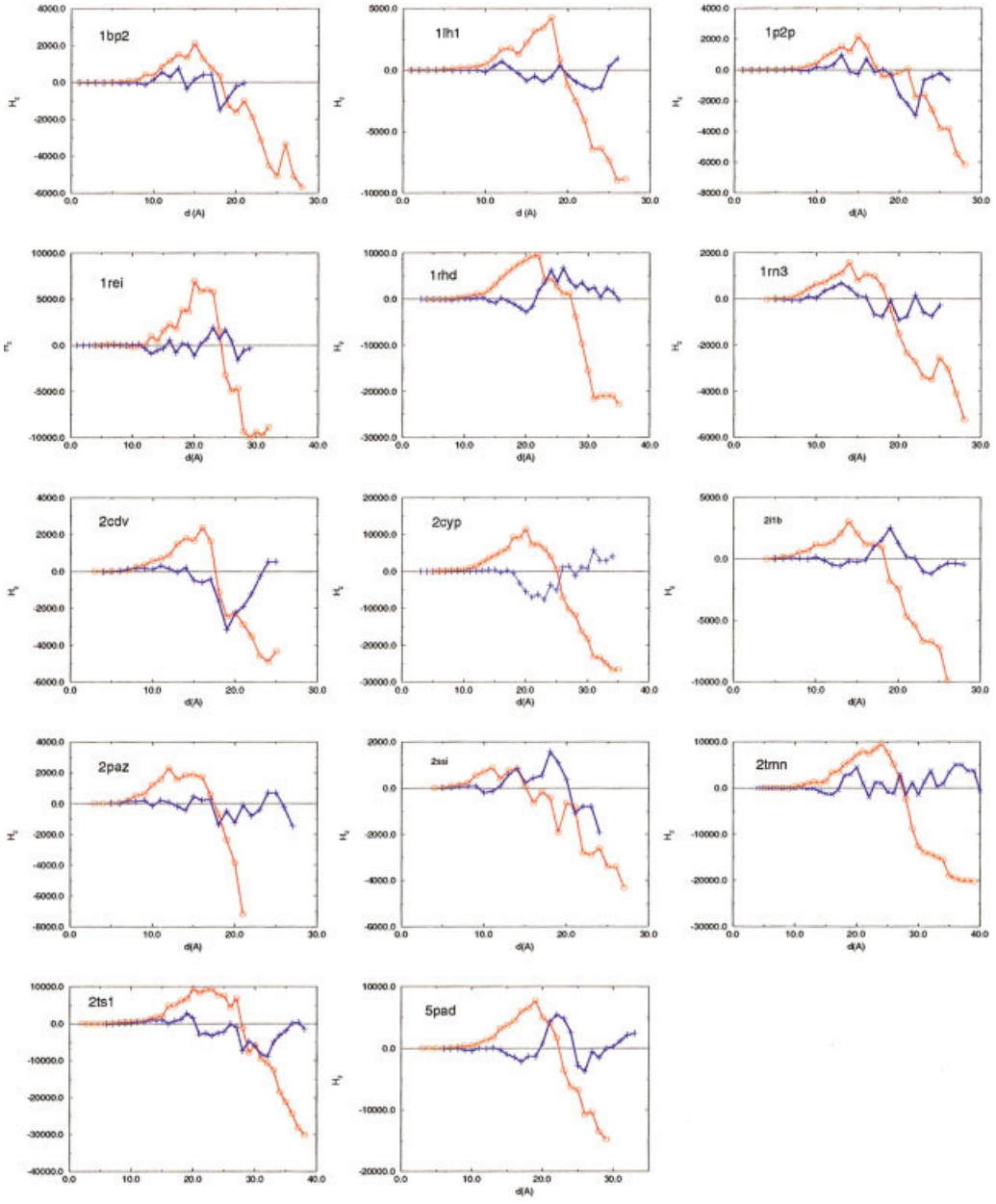

Fig. 2. Second-order moments for the native and decoy structures of the Holm & Sander single decoy sets (red circles: native; blue plus: decoy).

second-order moment profiles of the 3icb decoys from the Park and Levitt decoy set. Figure 3(a) shows several profiles of native-like decoys with RMSD < 2.0 Å, while Figure 3(b) shows several profiles of non-native like decoys with RMSD > 7.0 Å (the profile of the native structure is shown with a thick dark curve for comparison). The native-like structures show a second-order profile shape that mimics the native profile, which exhibits a strong hydrophobic core and a sharp plunge in the exterior. The non-native-like decoy structures, on the other hand, do not show the significant separation between a hydrophobic core and hydrophilic exterior. The second-order moments also fluctuate about zero on the radial axis, and the hydrophobic ratio either cannot be easily defined or cannot be defined at all for these decoy structures.

Examination of decoy and native structure profiles for an additional number of decoy sets revealed similar behavior. The native-like second-order moment profiles exhibited a pronounced hydrophobic peak and a significant plunge to negative values in the protein exterior, while the non-native-like decoys had reduced hydrophobic peaks and less prominent hydrophilic exteriors. The profiles of the decoy structures also extended out to a greater distance from the centroid of the structures. These features suggested that the total area under the second-order hydrophobic moment profile (under both the hydrophobic

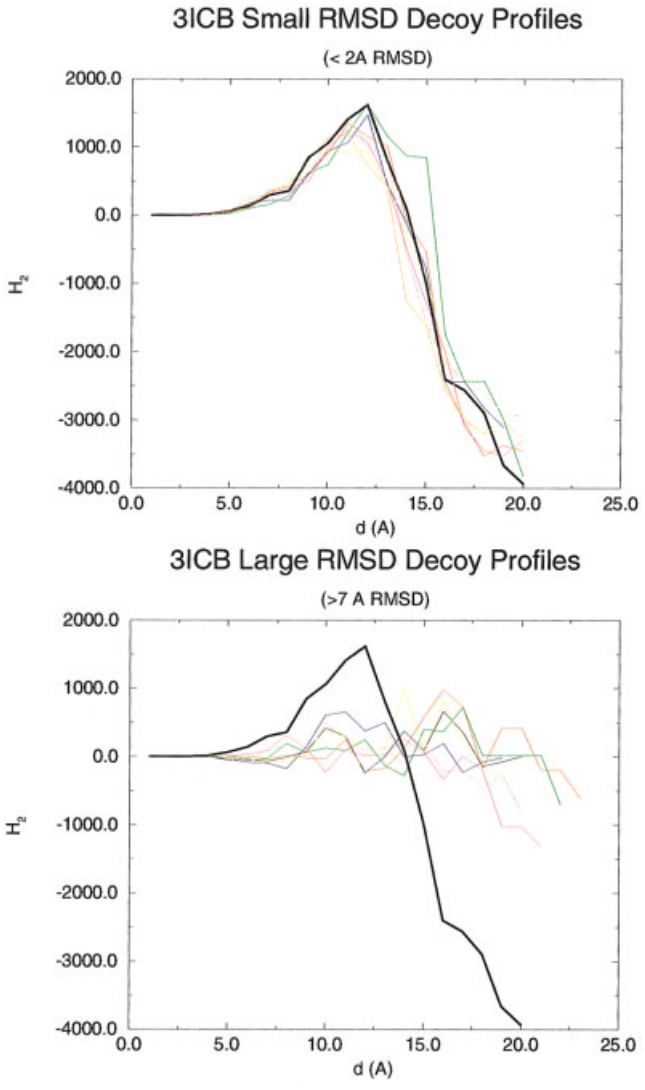

Fig. 3. Second-order hydrophobic moment profiles for some representative decoys of protein 3icb. (a) the top curve for representative decoys with small RMSDs (< 2.0 Å), (b) the bottom curve for representative decoys with large RMSDs (> 7.0 Å). The thick dark line in both figures denotes the profile of the native structure.

peak and above the hydrophilic plunge) could play a role in discriminating the native from the decoy structures. On the other hand, a significant increase in the protein extent of the decoy could yield a spurious contribution from the area under the negative moment profile. Differences due to this contribution could, however, be eliminated or reduced by scaling the native and decoy structures by the value of protein extent, namely, by d_0. The abscissa on the moment plot was, therefore, divided by d_0 and the second-order moment divided by d_0^2. Such scaling does not take differences in residue number into account. For the present case, however, the decoys and their corresponding native structures have the same number of residues.

The proposed hydrophobic score, S_H, which ranks the quality of the decoys with respect to an expected native profile, is then chosen as the integral of the area under the normalized 2nd-order hydrophobic moment profiles,

$$\tilde{H}_2 = H_2/d_0^2 \qquad s = d/d_0. \tag{6}$$

with s equal to the normalized extent of the principal major axis. The absolute value of $\tilde{H}_2$ is integrated over the normalized distance, from 0 to 1,

$$S_H = \int_0^1 |\tilde{H}_2| ds. \tag{7}$$

This score not only measures the prominence of the hydrophobic core, but also the prominence of the hydrophilic exterior. It takes into account the rapidity of decrease of the profile in the hydrophilic exterior. This hydrophobic score and the hydrophobic ratio are also extremely fast to evaluate. It takes less than a second for one structure on an IBM RS6K Power3-200MHz workstation.

Figure 4 shows the hydrophobic scores vs. the RMSDs for the four Park and Levitt decoy sets. Almost all decoys have lower hydrophobic scores (or integrated areas) than their corresponding native structures. Table III shows the number and percentage of decoys out of the total that have lower hydrophobic scores than their native proteins; 99.5, 99.4, 98.2, and 94.4% of the decoys have hydrophobic scores below their native benchmark scores of 3icb, 1ctf, 1r69, and 2cro, respectively. Proteins 3icb and 1ctf, which show native profiles accentuating the hydrophobic and hydrophilic regions (see below for more details), have fewer than 0.5–0.6% of decoys with a score that is greater than that of the native structures. One also notes a significant correlation in their decoy distributions, namely, decoys with a greater RMSD generally have smaller hydrophobic areas or scores. Proteins 2cro and 1r69, with native profiles that do not accentuate the hydrophobic and hydrophilic regions as observed for proteins 1ctf and 3icb (see below), show slightly greater numbers of decoys with greater scores than their native structures, and their distribution of decoy scores does not exhibit the correlation found for 1ctf and 3icb. The decoy scores of 1r69 and 2cro appear to be essentially uniformly distributed about the RMSD values.

Little or no correlation of hydrophobic score with RMSD might arise from native structures with profiles that do not accentuate the core and hydrophilic regions. It is then less restrictive for a decoy to score well with respect to the native structure. Figure 5 shows the native profiles of the four decoy sets of Park and Levitt, namely, 3icb, 1ctf, 1r69, and 2cro. It is clear that 1r69 and 2cro have native profiles with hydrophobic and hydrophilic regions of lesser prominence than found for 1ctf and 3icb. Thus, it is easier for decoys to score well against native proteins 1r69 and 2cro, which exhibit reduced separation of hydrophobic and hydrophilic residues, but it is still surprising that so few decoys in the Park and Levitt sets score better than the native profilings of 1r69 and 2cro. In general, if a decoy structure can manage a larger separation in hydrophobic and hydrophilic residues, it will score better than the

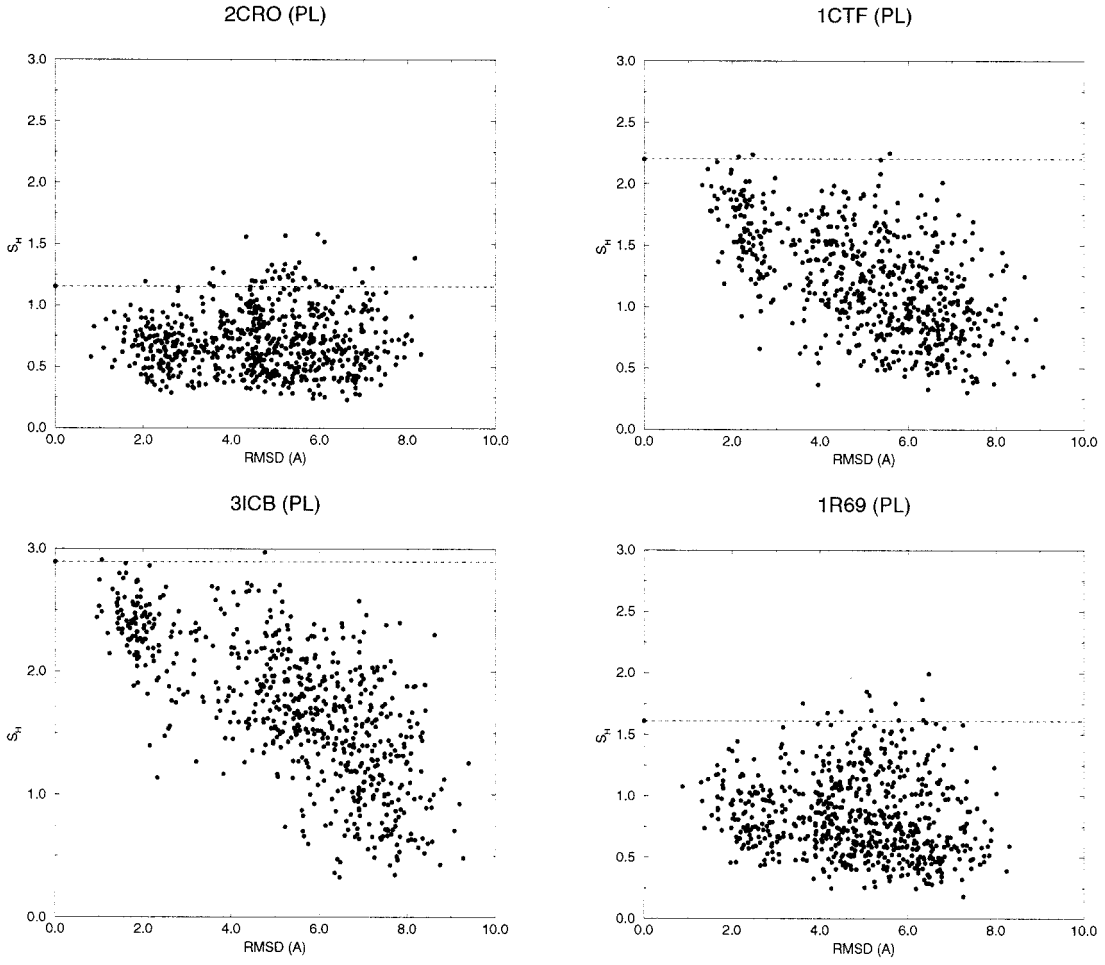

Fig. 4. Hydrophobic score versus RMSD for Park & and Levitt decoys. The dash line indicates the hydrophobic score of the native structure. The data points above this line have higher scores than the native structure, thus are false positive.

native structure as we will see below for the Baker decoys. This might explain why the hydrophobic score performs better for the 1ctf and 3icb decoy sets and also shows a higher correlation with RMSD.

It is interesting to note that there are low (good) RMSD structures that have low (bad) hydrophobic scores even among the decoys of the well-correlated sets, such as 3icb. Figure 6 shows several hydrophobic moment profiles for 3icb decoy structures with less than 3.0 Å RMSD and less than 1.5 hydrophobic score (decoy index a587, a591, and a8110, to name a few). The native score is 2.89 for this case. These decoy structures have fewer hydrophobic residues in the protein interior and consequently fewer hydrophilic residues in the protein exterior than expected for native structures. The hydrophobic residues and hydrophilic residues are more spatially mixed. Might these structures be less favorable candidates as near native structures? From the reported OPLSAA/SGB free energies,[2] they are indeed energetically unfavorable structures. The three decoys plotted, a587, a591, and a8110, are 206.98, 116.94, 110.14 kcal/mol higher than the native structure. The OPLSAA/SGB energies have been obtained from Levy and coworkers (see below for more details). This indicates that a low overall RMSD does not necessarily guarantee a good hydrophobic score, since the overall RMSD is a rather crude descriptor. It doesn't provide the detailed structural features, such as the essential hydrophobic core. The simple hydrophobic score, on the other hand, can provide useful information in discriminating decoy structures from native structures.

Figure 7 shows the hydrophobic scores for the four representative Baker decoy sets: two decoys 1ctf and 1r69, which are shared with the Park and Levitt set, and the other two 2ezh and 1leb, which have the highest and lowest percentage of decoys with scores below their native structure scores. In contrast to the Park and Levitt decoy sets, the Baker decoy sets show a much broader distribution of hydrophobic scores. The percentage of decoys that have scores below their native benchmark scores ranges from 25.3% (1leb) to 95.7% (2ezh), with the majority in the range of 60–80%. Also, most of these decoy sets do not exhibit the correlation with RMSD that the 1ctf and 3icb Park and Levitt decoys show. The four plotted decoy sets 2ezh, 1ctf, 1r69, and 1leb have a percentage of decoys with

TABLE III. Performance of the HydroPhobic Score: The Percentage of Decoy Structures That Have Lower Hydrophobic Score Than Their Native Ones ("Low scores")

Decoy set	PDB entry	Low scores	Total decoys	%
ParkLevitt	3icb	651	654	99.5
	1ctf	627	631	99.4
	1r69	664	676	98.2
	2cro	637	675	94.4
Baker	2ezh	957	1000	95.7
	1mzm	864	1000	86.4
	1nkl	848	1000	84.8
	1ctf	816	1000	81.6
	1r69	656	1000	65.6
	2fow	627	1000	62.7
	2ptl	619	1000	61.9
	1sro	559	1000	55.9
	1c5a	493	991	49.8
	1hsn	245	970	25.4
	1leb	253	1000	25.3

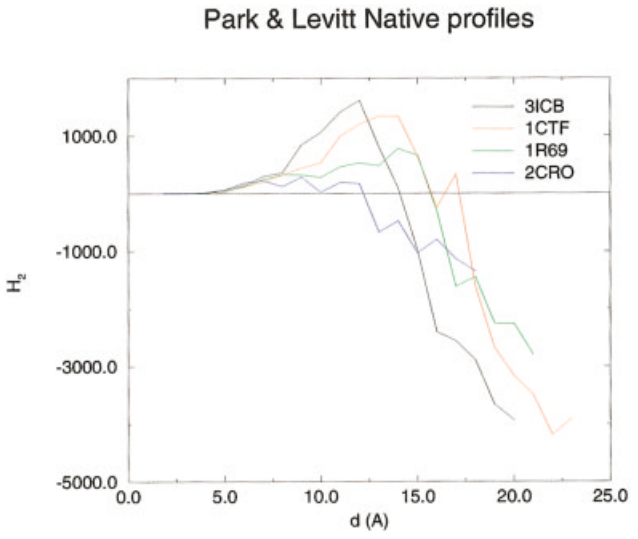

Fig. 5. The four native structure profiles in the Park & Levitt decoy set, 3icb, 1ctf, 1r69, and 2cro. Their hydrophobic scores versus RMSD are shown in Figure 4.

scores below the native at 95.7, 81.6, 65.6, and 25.3%, respectively. Interestingly, 2ezh and 1ctf (higher percentages, 95.7 and 81.6%), show a more prominent native structure profile than 1r69 and 1leb (lower percentages, 65.6 and 25.3%), as can be seen from Figure 8. Other decoys in the Baker set show similar behavior. The numbers of decoys with a higher percentage below the native score (2ezh, 1mzm, 1nkl, 1ctf, etc) show more pronounced native structure profiles than decoys with a lower percentage (1hsn, 1leb, etc). As mentioned previously for the Park and Levitt decoys, this correspondence between a higher percentage of decoys scoring well with the less prominent native profiles makes sense. It is easier for decoys to score well against native structures that exhibit reduced separation of hydrophobic and hydrophilic residues with consequent low score.

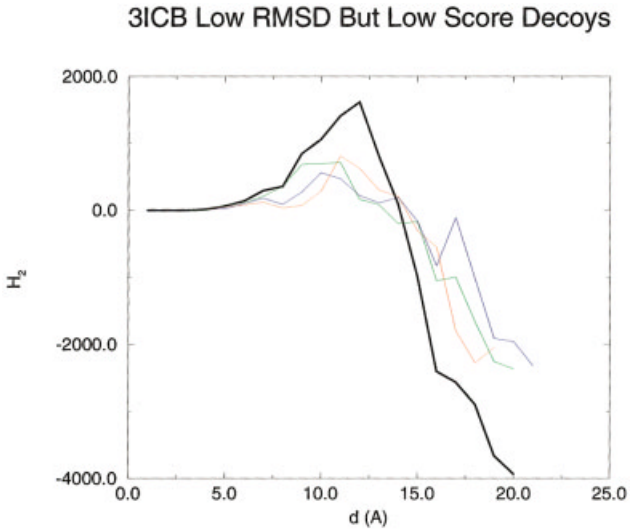

Fig. 6. Hydrophobic moment profiles for some good (low) RMSD structures but with bad (low) hydrophobic scores in Park & Levitt decoy set 3icb. The thick dark line denotes the profile of the native structure.

The relatively large number of Baker decoys with a high hydrophobic score compared with the Park Levitt decoys might be related to the manner in which the decoys were generated and selected. Examine the scores of the 1r69 and 1ctf decoys (the two common proteins in both sets) in the Park and Levitt decoy set shown in Figure 4, and in the Baker set shown in Figure 7. The Baker decoys clearly show a greater number of structures with scores that are higher than their native scores when compared with the Park and Levitt decoy scores. In particular, a significant fraction of the decoys of the 1leb Baker set clearly show greater spatial segregation of the hydrophobic and hydrophilic residues than observed for the native structure. This should be related to the way Baker and coworkers have selected these ab-initio decoys. The generation of the Baker decoys builds in a hydrophobic core. One of the fundamental assumptions underlying their program Rosetta[5,13] is that the distribution of conformations sampled for a given nine-residue segment of the chain is reasonably well approximated by the distributions in known protein structures in the PDB Databank. Fragment libraries for each 3- and 9-residue segment of the chain are extracted from the protein structure database using a sequence profile-profile comparison method. The conformational space defined by these fragments is then searched using a Monte Carlo procedure with an energy function that favors compact structures with paired β strands and buried hydrophobic residues.[21] The favoring of buried hydrophobic residues in the energy function and the hydrophobic filtering[6] should provide the Baker sets with greater segregation of hydrophobic and hydrophilic residues from the protein core to exterior[21] and consequently provide higher hydrophobic scores than achieved by the Park and Levitt decoy sets.

Levy and coworkers[2] have calculated the energies of the Park and Levitt decoys using the OPLSAA force field[22] and a Surface Generalized Born (SGB) model[23] for a con-

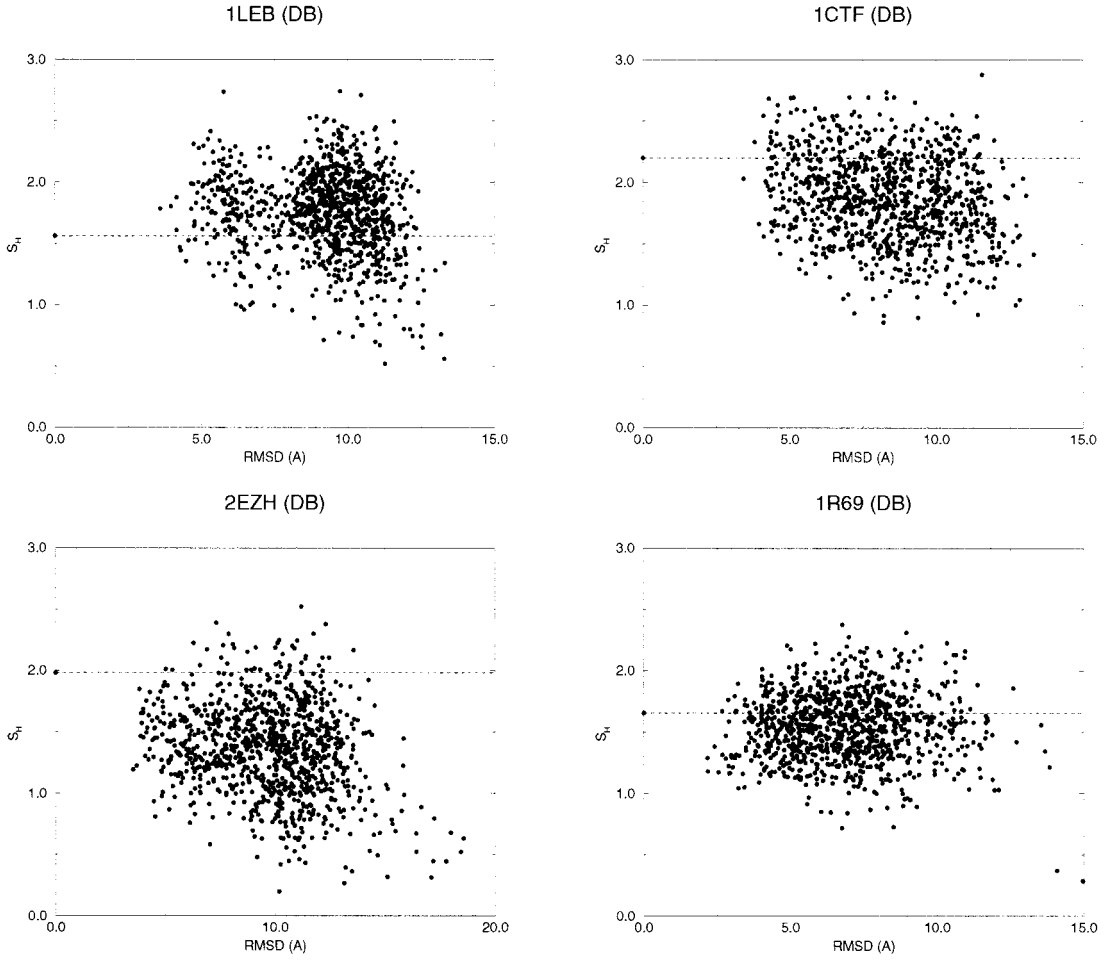

Fig. 7. Hydrophobic score versus RMSD for Baker decoys. The dash line indicates the hydrophobic score of the native structure. The data points above this line have higher scores than the native structure, thus are false positive.

tinuum solvent. They found that without the continuum solvation free energy, the OPLSAA gas phase energies are not sufficient to distinguish native-like from non-native-like structures. Kollman and coworkers[14] found similar conclusions using the AMBER force field[24] with a Poisson Boltzmann Surface Area (PBSA) continuum solvent model.[25,26] Figure 9 is a plot of the OPLSAA/SGB energy (the energy of the native structure is set at zero) vs the hydrophobic score for the protein 3icb of the Park and Levitt set. The OPLSAA/SGB energies have been kindly supplied by the Levy group. It should be noted that in the Levy energy calculations, the decoy structures are minimized first to remove bad contacts in energy space (otherwise the energies could be huge and meaningless). Thus, the structures used in the Levy energy calculations are slightly different from ours; however, this does not affect the hydrophobic scores meaningfully. This is an advantage of the method of hydrophobic scoring. Differences in structure that would affect the free energy values significantly will not affect the hydrophobic scores significantly. One need not even add hydrogen atoms to the PDB structures for most of the calculations. Free energy calculations, on the other hand, are not only sensitive to the presence or absence of hydrogen atoms, but extremely sensitive to smaller differences in structure. Figure 9 shows the correlation between the OPLSAA/SGB energy and the hydrophobic score, i.e., decoys with smaller or poorer scores have higher energies compared with the native energy, and those with higher or better scores are closer in energy to the native structures. Even though there is a good overall correlation, there are still structures having low OPLSAA/SGB energies but showing bad hydrophobic scores (more details below; Fig. 10). Similar to 3icb, protein 1ctf also shows a significant correlation between the OPLSAA/SGB energy and the hydrophobic score, whereas 1r69 and 2cro show a weaker correlation. This week correlation for the 1r69 and 2cro decoys reflects their weak correlation between the hydrophobic score and RMSD as described earlier.

Interestingly, the decoy structures with low OPLSAA/SGB free energies that do not have high hydrophobic scores are found even for the decoys of 3icb, which show a strong correlation between the hydrophobic score and RMSD. The decoy sets showing poorer correlation have a

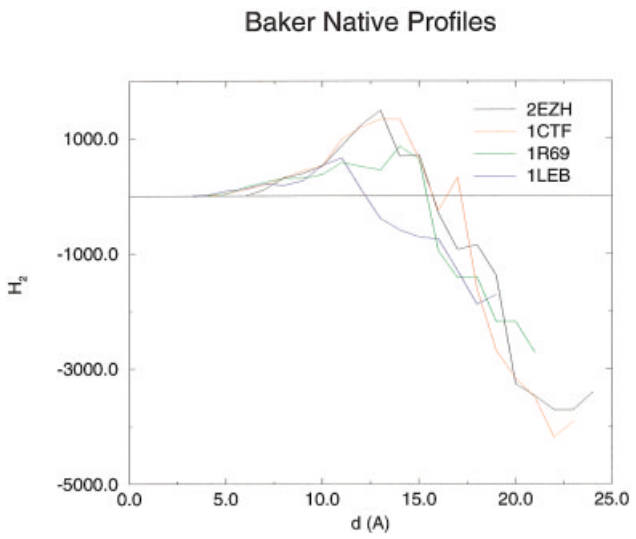

Fig. 8. The four native structure profiles in the David Baker decoy set, 2ezh, 1ctf, 1r69, and 1leb. Their hydrophobic scores versus RMSD are shown in Figure 7.

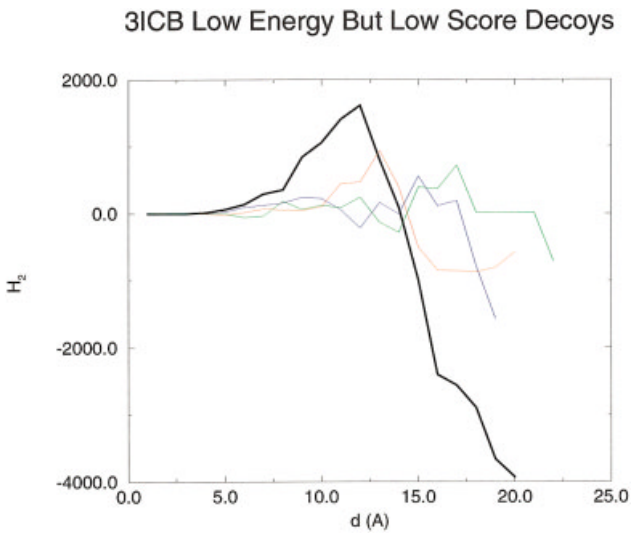

Fig. 10. Hydrophobic moment profiles for some of the low OPLSAA/SGB energy structures but with bad (low) hydrophobic scores in Park & Levitt decoy set 3icb. The thick dark line denotes the profile of the native structure.

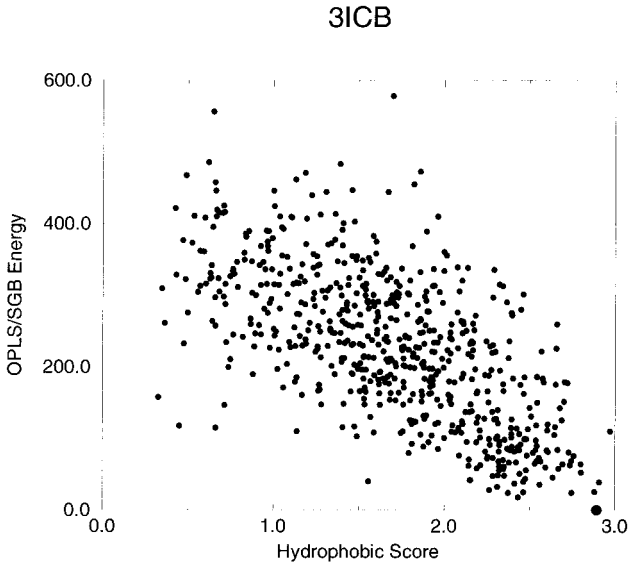

Fig. 9. Hydrophobic score versus OPLSAA/SGB energy for Park & Levitt decoy set 3icb. The OPLSAA/SGB energies are from Levy and coworkers[2] (the native one is marked with a larger circle).

greater number of decoys exhibiting this behavior. Figure 10 shows several representative profiles of 3icb decoy structures with low free energies and also low hydrophobic scores (less than 1.0). The hydrophobic score for the native structure is again 2.89. These decoys are not the same as those with low RMSD and low score as discussed previously and shown in Figure 6. The bad or low hydrophobic scores indicate that the structures have a poorly formed hydrophobic core and hydrophilic exterior even though the free energy is low. By comparison with the native profile (the dark curve in Fig. 10), it is evident that the hydrophobic core of these decoys has been "damaged." The region of positive moment that might be identified as a core region is shifted out to greater distances than found for the native structure. Furthermore, none of the decoys exhibit the sharp plunge to negative values in the protein exterior expected for a native structure. Consequently, this yields a low score or unfavorable protein structure. This example demonstrates the value of the hydrophobic score in providing complementary information to that obtained from the free energy calculations. Previously we had shown that a low RMSD does not necessarily guarantee a good hydrophobic score, and here we have shown that a low free energy does not guarantee a good hydrophobic score either. Another good point, as mentioned earlier, is that it is much faster to calculate the hydrophobic score than the force field energy minimization, which can take hours in an IBM RS6K Power3-200MHz workstation. It takes less than a second for the hydrophobic score calculation.

Finally, it should be pointed out that the present hydrophobic profiling applies only to the radial distribution of hydrophobicity but not the angular distribution, thus it has limits in distinguishing the angular hydrophobicity distribution. One example that clearly shows this limitation of the profiling is the following. Figure 11(a) shows the structure of protein G in its native state and one of the decoy structures. The decoy structure was chosen from Baker's decoy set 1gb1 (qa1gb1010-low.pdb). Since 1gb1 has less than 60 residues, it wasn't included with the decoy sets previously selected for detailed examination. It does, however, provide an interesting example to exhibit the limitations of the present method. The native structure has the C-terminus and N-terminus forming an anti-parallel β-sheet, while the decoy structure has a β-sheet formed between the C-terminus with another beta strand from residue LYS-9 to THR-16, instead of the N-terminus as in the native structure [see Fig. 11(b)]. This rearrangement of the β-sheets results in a 5.62 Å RMSD from the native structure. With respect to the profiling, the radial spatial distribution of residues is hardly affected, since

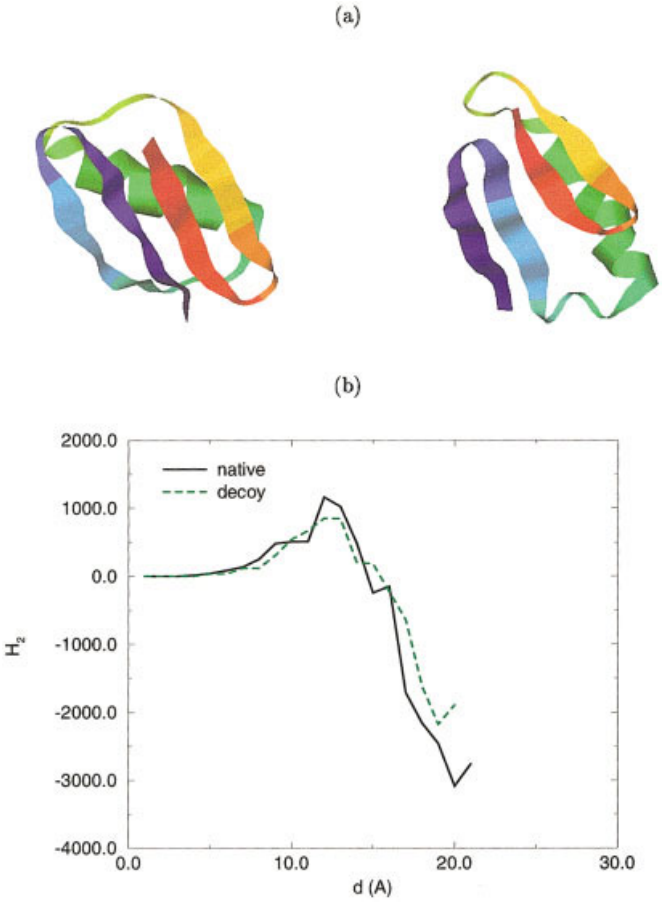

Fig. 11. (a) The native structure (left) and one decoy structure (right) for Protein G (1gb1) from Baker decoy set. The native structure has the C-terminus and N-terminus forming an anti-parallel β-sheet; while the decoy structure has a β-sheet formed between the C-terminus with another beta strand from residue LYS-9 to THR-16, instead of the N-terminus as in the native structure. See text for more detail. (b) The comparison of the hydrophobic moment profiles of the native and the decoy structures.

the interchanged residues are still distributed at roughly the same distance from the centroid. Therefore, the 2nd-order hydrophobic moment hardly changes, as shown in Figure 11(c). This shows that decoys with large RMSD deviations from their native structures may show a high hydrophobic score. On the other hand, complementary information may eliminate decoy candidate structures, e.g., the OPLSAA/SGB energy of the native structure of 1gb1 is −3,209.03 kcal/mol, while the decoy energy is −3,114.06 kcal/mol, namely, 94.97 kcal/mol higher. The OPLSAA/SGB energies (after minimization) have been obtained from the IMPACT program.[19,20] In general, it is easier to create alternate tertiary arrangements that maintain the ellipsoidal profile of hydrophobicity for small proteins. However, for larger proteins with more complex tertiary topologies, it is harder to rearrange the topology while maintaining the hydrophobicity profiles. To summarize, hydrophobic moments and scores as presently calculated will not distinguish changes in the hydrophobicity distribution that arise solely from angular changes in structure about the centroid. The hydrophobicity profiling, on the other hand, provides a picture of what the distribution should look like from protein interior to exterior when angularly averaged. This is a characteristic that can identify structures that depart from that expected for native profiles. Furthermore, since hydrophobicity profiling involves a simple calculation that needs no free energies to be calculated, no solvation models to be developed, and no force field implementation required, it should be useful as a pre-screening process in providing complementary information to approaches based on free energy calculations.[2,12,14]

CONCLUDING REMARKS

The present study has examined the hydrophobic moment profiles of all non-redundant soluble globular proteins in the entire PDB data bank, as well as the utility of hydrophobicity profiling to discriminate native and near-native protein structures from decoy structures for the widely used Holm and Sander, Park and Levitt, and Baker decoy sets. The results obtained from all the soluble globular proteins in PDB reveal a relatively invariant hydrophobic ratio of 0.71 ± 0.08.

Furthermore, subject to the conditions that limit the type of small structures examined, the moment profiling enables one to distinguish differences in the radial hydrophobicity distribution of the decoys and near native structures. Overall, the hydrophobic score is found to be very discriminating for the Holm and Sander and Park and Levitt decoys, but less significantly discriminating for the Baker decoys, since the Baker decoys already have the hydrophobic core bias built in their procedure. It is also found that the hydrophobic score, based on moment profiling, can suggest that certain structures with relatively small overall RMSD from the native structure can be eliminated as candidates due to profiles displaced significantly from their native hydrophobicity profiles. Interestingly, some decoys with low free energies, such as OPLSAA/SGB energy, can also be eliminated by the hydrophobic moment profiling and consequent hydrophobic score, since they show little or no hydrophobic core and hydrophilic exterior compared with their native profiles. This shows that the simple hydrophobic score can provide information that complements that obtained by the more rigorous free energy approach.

The hydrophobic ratio and score could also be useful for guiding protein folding simulations. This could be implemented by eliminating the simulations that evolved to deviant values of the ratio and to low values of the score. Such a strategy could also be applied in the case of thousands of the parallel kinetic simulations as generated by Pande et al. at folding@home.[27] Examinations of the ratio and score could also supply a guiding potential to penalize the structures with bad hydrophobic profiles in the umbrella sampling.

It is generally agreed that more than a single attribute may be required to significantly discriminate between near native and incorrect decoys. This is particularly true of the dense decoy sets used for ab-initio validation. Such sets involve numerous minor structural modifications. The decoys with large RMSD and high hydrophobic score found

in the present study emphasize that the present procedure will not always identify good decoy candidates. The more pronounced the native decoy profile, however, the fewer such decoys. It should be emphasized once again that the choice of small decoys was dictated predominantly by the interest generated in the evaluation of such decoy sets. Small decoy proteins do not trade on the strength of the profiling procedure. One expects its discrimination to increase significantly with an increase in protein and decoy size.

ACKNOWLEDGMENTS

We thank Prof. Ron Levy and his group for supplying us with the OPLSAA-SGB energies for the Park and Levitt decoy sets. We also thank Prof. Bruce Berne for useful discussions.

REFERENCES

1. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139.
2. Felts AK, Wallqvist A, Gallicchio E, Levy R. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the opls all-atom force field and the surface generalized born solvent model. Proteins 2002, 48, 404.
3. Kihara D, Lu H, Kolinski A, Skolnick J. Touchstone: An ab initio protein structure prediciton method that uses threading-based tertiary restraints. Proc. Natl. Acad. Sci. 2001;98:10125.
4. Hao M, Scheraga HA. Designing potential energy functions for protein folding. Curr. Opin. Struct. Biol. 1999;9:184.
5. Bonneau R, Strauss CEM, Baker D. Improving the performance of rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins: Structure, Function and Genetics 2001;43:1.
6. Shortle D, Simons KT, Baker D. Clustering of low energy conformations near the native structures of small proteins. Proc. Natl. Acad. Sci. 1998;95:11158.
7. Huang ES, Subbiah S, Levitt M. Recognizing native folds by the arrangement of hydrophobic residues. J. Mol. Biol. 1995;252:709.
8. Holm L, Sander C. Evaluation of protein models by atomic solvation preference. J. Mol. Biol. 1992;225:93.
9. Jones DT, Thornton JM. Potential energy functions for threading. Curr. Opin. Struct. Biol. 1996;6:210.
10. Silverman DB. Hydrophobic moments of protein structures: Spatially profiling the distribution. Proc. Natl. Acad. Sci 2001;98: 4996.
11. Zhou R, Silverman DB. Detecting native protein folds among large decoy sets with hydrophobic moment profiling. Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE, editors. Proceedings of Pacific Symposium on Biocomputing. Singapore: World Scientific; 2002. p 673–684.
12. Park B, Levitt M. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. J. Mol. Biol. 1996;258:367.
13. Simons KT, Bonneau R, Ruczinski I, Baker D. Structure prediction of casp iii targets using rosetta. Proteins: Structure, Function and Genetics 1999;37 S3:171.
14. Kollman PA. personal communication 2001.
15. Eisenberg D, Weiss RM, Terwilliger TC, Wilcox W. Hydrophobic moments and protein structure. Faraday Symp. Chem. Soc. 1982;17:109.
16. Eisenberg D, Weiss RM, Terwilliger TC. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature (London) 1982;299:371.
17. Silverman BD. A two-component model of protein hydrophobicity: Spatially profiling the distribution. J. Theor. Biol. 2002;216:139.
18. Murzin AG, Brenner SE, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 1995;247:536.
19. Kitchen DB, Hirata F, Westbrook JD, Levy RM, Kofke D, Yarmush M. Conserving energy during molecular dynamics simulations of water, proteins and proteins in water. J. Comp. Chem. 1990;11: 1169.
20. Figueirido F, Zhou R, Levy R, Berne BJ. Larger scale simulation of macro-molecules in solution: Combining the periodic fast multipole method with multiple time step integrators. J. Chem. Phys. 1997;106:9835.
21. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Science 2002;11:1937.
22. Jorgensen WL, Maxwell D, Tirado-Rives J. Development and testing of the opls all-atom froce field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. 1996;118: 11225.
23. Ghosh A, Rapp CS, Friesner RA. Generalized born model based on a surface integral formulation. J. Phys. Chem. 1998;102:10983.
24. Cornell W, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 1995;117:5179.
25. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144.
26. Friedrichs M, Zhou R, Edinger S, Friesner RA. Poisson-Boltzman analytical gradients for molecular modeling calculations. J. Phys. Chem. 1999;103:3057.
27. Pande M, Shrits F.S. Screen savers of the world, unite! Science 2000, 290, 1903.