PROTEINS: Structure, Function, and Genetics 34:82–95 (1999)

Improved Recognition of Native-Like Protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins

Kim T. Simons,[1] Ingo Ruczinski,[2] Charles Kooperberg,[3] Brian A. Fox,[1] Chris Bystroff,[1] and David Baker[1]*
[1]*Department of Biochemistry, University of Washington, Seattle, Washington*
[2]*Department of Statistics, University of Washington, Seattle, Washington*
[3]*Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington*

ABSTRACT **We describe the development of a scoring function based on the decomposition *P*(structure|sequence) ∝ *P*(sequence|structure) **P*(structure), which outperforms previous scoring functions in correctly identifying native-like protein structures in large ensembles of compact decoys. The first term captures sequence-dependent features of protein structures, such as the burial of hydrophobic residues in the core, the second term, universal sequence-independent features, such as the assembly of β-strands into β-sheets. The efficacies of a wide variety of sequence-dependent and sequence-independent features of protein structures for recognizing native-like structures were systematically evaluated using ensembles of ~30,000 compact conformations with fixed secondary structure for each of 17 small protein domains. The best results were obtained using a core scoring function with *P*(sequence|structure) parameterized similarly to our previous work (Simons et al., J Mol Biol 1997;268:209–225] and *P*(structure) focused on secondary structure packing preferences; while several additional features had some discriminatory power on their own, they did not provide any additional discriminatory power when combined with the core scoring function. Our results, on both the training set and the independent decoy set of Park and Levitt (J Mol Biol 1996;258:367–392), suggest that this scoring function should contribute to the prediction of tertiary structure from knowledge of sequence and secondary structure. Proteins 1999;34:82–95.** © 1999 Wiley-Liss, Inc.

Key words: protein folding; structure prediction; knowledge-based scoring functions; fold recognition

INTRODUCTION

A scoring function capable of distinguishing native-like conformations (similar but not identical to the native structure) from non-native conformations for a given sequence is critical for protein structure prediction because it is unlikely that any method of generating structures will exactly reproduce the native structure. This paper is focused on the problem of distinguishing native-like structures from non-native structures, where native-like refers to conformations less than 4 Å rmsd (root mean squared deviation of Cα coordinates) from the native structure.

A wide variety of scoring/energy functions have been developed over the past decade.[1–10] To provide a testing ground for evaluating the ability of different scoring functions to recognize native-like structures, Park and Levitt[11] generated very large numbers of compact, self avoiding conformations with native secondary structure for eight small protein domains. Using this decoy set, a variety of scoring functions were tested.[7,11] For several of the functions, the best scoring native-like structures were in the top 100 of the ~200,000 decoy structures for each sequence (typically 100 structures were native-like for each sequence), but the scores were not consistent enough to permit unambiguous identification of the correct fold.

We recently described a new scoring function derived from the structure database using Bayes' theorem.[12] In ab initio folding simulations some success was obtained with α-helical proteins, but the scoring function was clearly insufficient for proteins with β-sheets. In this report, we further develop the scoring function by evaluating the effectiveness of descriptions of different sequence-dependent and sequence-independent features of proteins in recognizing native-like structures in large decoy sets of compact conformations generated in our laboratory. The most important new additions are terms that describe the packing of β-strands in β-sheets. We rigorously evaluate the performance of the final scoring function on the independent decoy sets of Park and Levitt,[11] herein referred to as the PL test set. It is important to note that these decoy sets were not used before the final testing.

METHODS

Development of Scoring Function

The scoring method in this report is a substantial refinement of the one used by Simons et al.[12] As in that report, the scoring function is based on a model for the probability of the structure being the native structure,

Grant sponsor: STC National Science Foundation; Grant number: DMS-9403371; Grant sponsor: Packard Foundation; Grant sponsor: U.S. Public Health Service National Research Service Award (National Institute of General Medical Sciences); Grant number: T3-GM07270.
*Correspondence to: David Baker, University of Washington, Seattle, WA 98195. E-mail: baker@ben.bchem.washington.edu
Received 28 April 1998; Accepted 17 August 1998

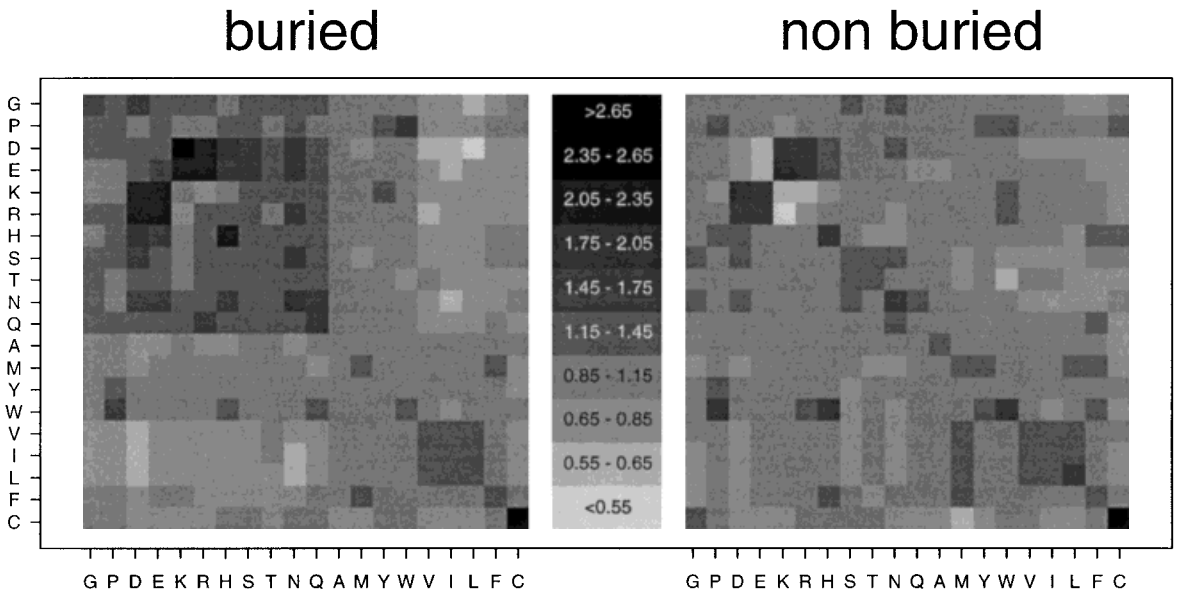

Fig. 1. P_{pair} (eq(6)) for amino acid pairs with centroids separated by <7 Å. *Left,* both residues buried (more than 16 residue centroids within 10 Å). *Right,* residues not both buried. A similar definition of environment classes was used by Kocher et al.[10] The darker the square for a particular pair of residues, the greater the frequency of contact relative to that expected given the environment term. The amino acids are ordered from hydrophilic to hydrophobic, from top to bottom and left to right. Same charged amino acids "attract" in buried environments but "repel" in nonburied environments, perhaps reflecting the presence of similarly charged amino acids in metal binding or catalytic sites. Centroid positions—the average position of all sidechain heavy atoms relative to the backbone N, Cα, and C—were computed from the pdb select 25 set of proteins (coordinates are available upon request). Use of centroid rather than Cβ distances did not significantly improve the P_{env} term but did help the P_{pair} term discriminate native-like structures from non-native structures. The amino acid on the horizontal axis is the first in sequence; sequence order is preserved to show that there is sufficient data to construct statistically significant scoring tables (upper right and lower left triangles are constructed from independent data). Because of the symmetry apparent in the panels, sequence order was removed by averaging for the P_{pair} term used in the scoring function.

given the sequence of amino acids. The various components of our model for this probability are all based on statistics from native structures except for the hard sphere repulsion term, VdW.

Using Bayes' theorem, we obtain

$$P(\text{structure}|\text{sequence}) = \frac{P(\text{sequence}|\text{structure})P(\text{structure})}{P(\text{sequence})} \quad (1)$$

The following discussion develops separate models for $P(\text{sequence}|\text{structure})$ and for $P(\text{structure})$. For the comparison of different structures with the same sequence, $P(\text{sequence})$ is a constant, and we can thus ignore this term. We first describe the terms that contribute to the core scoring function, as well as some closely related terms. The remaining terms are discussed in a later section.

P(Sequence|Structure)

We first discuss the $P(\text{sequence}|\text{structure})$ term in equation (1), which may equivalently be represented as

$$P(\text{sequence}|\text{structure}) = P(aa_1, \ldots, aa_n|X) \quad (2)$$

where the sequence of length n is explicitly written as a string of amino acids and the structure is described by a vector $X = \{x_1, \ldots, x_n\}$ of three-dimensional coordinates. Now consider the expansion

$$P(aa_1, \ldots, aa_n|\mathrm{X}) = \prod_i P(aa_i|\mathrm{X}) \prod_{i<j} \frac{P(aa_i, aa_j|X)}{P(aa_i|X)P(aa_j|X)} \times \prod_{i<j<k} \frac{P(aa_i, aa_j, aa_k|X)P(aa_i|X)P(aa_j|X)P(aa_k|X)}{P(aa_i, aa_j|X)P(aa_i, aa_k|X)P(aa_j, aa_k|X)} \cdots \quad (3)$$

It seems reasonable to assume that the probability of observing a particular amino acid at position i does not depend on the complete three-dimensional structure of the protein, but only on the local structural environment E_j, defined in terms of the solvent accessibility and/or secondary structure. Some initial data analysis convinced us that the second term in equation (3) is significantly different from one (Fig. 1), and its inclusion in our scoring function considerably improves the performance of our method (see Table III), while estimation of third or higher terms in equation (3) is too unwieldy to be useful. Thus, we decided to use the following approximation for $P(\text{sequence}|\text{structure})$:

$$P(\text{sequence}|\text{structure}) \approx P_{env}P_{pair} \quad (4)$$

$$P_{env} = \prod_i P(aa_i|E_i) \quad (5)$$

TABLE I. Bins Used in Density Estimation†

Density function	Variable	Bins
P_{env}	No. of neighbors	0–3, 4, . . . , 49, 50+
P_{pair}	r_{ij}	0–7 Å, 7–10 Å, 10–12 Å, 12 Å+
$P_{xx}(\phi, \theta, r, hb, \sigma \mid Sep)$	*Sep*	1, 2–10, 11+
$P_{xx}(\phi, \theta \mid r, Sep)$	θ	0–36°, 36–72°, 72–108°, 108–144°, 144–180°
$P_{xx}(\phi, \theta \mid r, Sep)$	φ	−135–45°, −45–45°, 45–135°, −180–135°, 135–180°
$P_{xx}(hb \mid r, Sep)$	*hb*	0–0.1, 0.1–0.2, . . . , 1.9–2.0

†Density estimates for the four latter density functions were obtained by counting the number of instances of dimer pairs in each of the bins and dividing by the product of the total number of dimers and the integral of the relevant correction factor (see text) over the bin.

$$P_{pair} = \prod_{i<j} \frac{P(aa_i, aa_j \mid E_i, E_j, r_{ij})}{P(aa_i \mid E_i, r_{ij}) P(aa_j \mid E_j, r_{ij})} \quad (6)$$

where r_{ij} is the distance between the centroids of residues *i* and *j*. The environment classes E_i in equations (5) and (6) were defined solely in terms of residue burial. Because of the larger number of counts available for estimating the P_{env} term, a large number of environment classes could be used (Table I), while for the P_{pair} term, only two environment classes were used (Fig. 1 legend). Figure 1 shows P_{pair} for contacting residue pairs; the strongest interactions are between pairs of cysteine residues and between oppositely charged residues. The fact that hydrophobic residue pairs do not have ratios very different from 1 suggests that for most, the hydrophobic effect is well modeled by the P_{env} term. As expected, at long-distance separations P_{pair} approaches unity.

P(Structure)

Secondary structure packing

We used a simple vector representation to describe the packing of secondary structural elements. We initially explored representing each secondary structural element by a single vector, but this did not accurately represent the twist of β-strands. Instead, every two residue segment in helices and strands was represented by a vector. For strands, the vector is from the backbone nitrogen of the first residue to the backbone carbonyl carbon of the second residue. Helix vectors were derived using the two residues flanking each dimer: the vector is from the average of the coordinates of the first 11 backbone atoms to the average of the last 11 backbone atoms of the four residue segment centered on the dimer. This method of computing the helical vectors is a simple and accurate method for constructing a vector that runs through the helical center (data not shown).

The values of five variables must be specified to uniquely determine the relative position of two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$; we chose to describe the relative orientation of the secondary structure vectors using the coordinate system shown in Figure 2. $\mathbf{r}$ is the vector between the dimer centers, σ is the angle between $\mathbf{v_1}$ and $\mathbf{r}$, and φ and θ describe the relative orientation of $\mathbf{v_1}$ and $\mathbf{v_2}$ in a spherical coordinate system with z-axis defined by $\mathbf{v_1}$ and x-axis by $(\mathbf{v_1} \times \mathbf{r})$. *Sep* is the number of residues between the ends of the secondary structure elements containing the dimers; the distributions were somewhat sharper using this measure than using the number of residues between the dimers. For helix–helix and helix–strand pairs, we neglect the remaining degree of freedom, which corresponds to a rotation around the z-axis, because of the cylindrical symmetry of helices. For the strand–strand pairs, this symmetry breaks down because of the strict orientation of hydrogen bonding groups between paired strands. To capture this consequence of hydrogen bonding, we use the dot product between $\mathbf{r}$ and the C = 0 bond vector. This dot product is averaged over the two C = 0 bond vectors in each of the dimers, and the two averages are then summed. The more in-plane the two vectors, the closer the sum of the dot products is to 2. In the following discussion, this sum is referred to as *hb*.

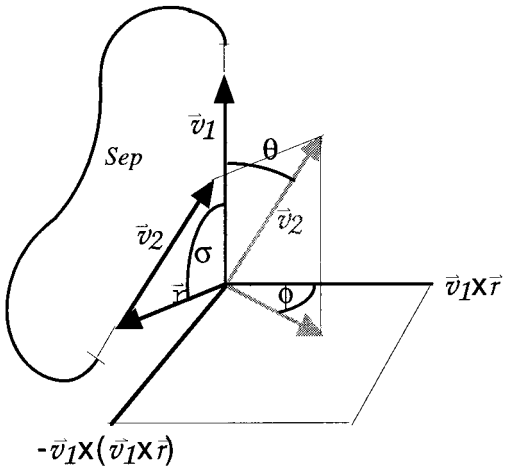

Fig. 2. Geometric description of the packing orientation between two elements of secondary structure. The two unit vectors, $\mathbf{v_1}$ and $\mathbf{v_2}$, represent the secondary structure dimers. The skew angle, σ, is the angle between the distance vector ($\mathbf{r}$) and the first vector of the pair ($\mathbf{v_1}$). θ is the angle between $\mathbf{v_1}$ and $\mathbf{v_2}$ ($arccos(\mathbf{v_1} \cdot \mathbf{v_2})$). The final angle, φ, is

$$\arctan \frac{\mathbf{v_2} \cdot (\mathbf{v_1} \times \mathbf{r})}{\mathbf{v_2} \cdot [\mathbf{v_1} \times (\mathbf{v_1} \times \mathbf{r})]}.$$

The sequence separation *Sep* is the length of the loop between the ends of the secondary structure elements containing the dimers.

Initially, we used the statistical graphics package xgobi[13] to examine the full multidimensional distributions of φ, θ, σ, *r*, and *hb* of HH (helix–helix dimer), HS (helix–strand dimer), and SS (strand–strand dimer) pairs in native protein structures. This exploratory data analysis suggested that for a given distance interval and loop length, the φθ distribution, the *hb* distribution, and the σ distribution were fairly independent of one another for each of the secondary structure pairs. Density functions for secondary structure pairs were therefore constructed using

$$P_{xx}(r, \phi, \theta, \sigma, hb \mid Sep) \approx P_{xx-\phi\theta}(\phi, \theta \mid r, Sep) P_{xx-hb}(hb \mid r, Sep) P_{xx-\sigma}(\sigma \mid r, Sep) P_{xx-dist}(r \mid Sep) \quad (7)$$

where r is the distance between the dimer centers, the hb term applies only to strand-strand pairs, and XX indicates the secondary structure of the dimers (HH, HS, or SS).

The $P(\phi, \theta|r, Sep)$ distributions in native known protein structures are shown in Figures 3 and 4. For secondary structure elements adjacent in the sequence (Figs. 3a,c, 4a), the dimers are primarily antiparallel ($\theta > 90°$), while for nonlocal pairs (Figs. 3b, d, 4b), both antiparallel and parallel arrangements are observed. A number of features of helix-helix packing contribute to these distributions. The peaks at $\phi = -90°$, $\theta = 135°$, and $\phi = 90°$, $\theta = 30°$ in the helix-helix distribution (Fig. 3a,b) correspond to the preferences for interhelical dihedral angle values of 130° and −50° noted in previous studies.[14,15] More specifically, the α-α corner motif described by Efimov[16] shows up in the distribution for sequentially adjacent helices (Fig. 3a) as the mass of points centered around $\phi = -90°$ and $\theta = 135°$. The relative absence of points at $\phi = 90°$ and $\theta = 135°$ reflects the handedness of the turns between α-helices in α-α corners. The few occurrences near $\phi = 0°$ and $\theta = 90°$ represent helices that pack closer than expected due to glycines or cysteines at the interhelical interface. The packing of helix and strand pairs (Fig. 3c,d) is also very different from a random distribution (Fig. 3e) and may partially reflect the right-handed crossover preferences of known protein structures.[17] For β-strands, hydrogen bonding introduces strong preferences in particular regions of θφ space (Fig. 4a,b, $\phi = -90°$ and $\theta = 135°$ for antiparallel strand pairs and $\phi = 90°$ and $\theta = 45°$ for parallel strand pairs). The characteristic twist of sheets[17] is also captured by the φθ distribution: completely planar sheets would have $\theta = 180°$ for antiparallel strand pairs and $\theta = 0°$ for parallel strand pairs. The coplanarity of the interdimer vector and the C = 0 bond vectors is evident in the peak near 2 in the hb distribution in native protein structures (Fig. 5, solid line).

To locate the native structure in structure prediction calculations, it is important that a scoring function attribute good scores not only to native structures, but to native-like structures as well. To characterize the distributions of the different variables in native-like structures, we generated a training set consisting of large ensembles of compact structures for 21 different proteins (see below). In the construction of the probability densities, we paid close attention not only to the properties of the distributions in native protein structures, but to those of the native-like and non-native structures in these ensembles as well. Of particular interest were distributions that were markedly different for the native-like and non-native structures in the ensembles, as these are the most promising candidates for native-like structure recognition. Distributions that instead differ considerably between native and native-like structures may be useful for recognizing native structures in large sets of native-like structures, but not necessarily for the more critical native-like recognition problem. Comparisons of plots of σ vs r in native, native-like and non-native structures suggested that $P(\sigma|r, Sep)$, which is derived from these data, would have limited value for native-like structure recognition because the distribution in native-like structures resembles that of non-native structures (Fig. 6). By contrast, the strand–strand φθ distributions (Fig. 4) and $P(hb|r, \text{Sep})$ (Fig. 5) in the native-like structures differed considerably from those of the non-native structures. The use of the spherical coordinate system, rather than a generalization of the more traditional interhelical dihedral angle to describe the relative orientation of secondary structure pairs, was motivated in part by the observation of considerably broader φ distributions in the non-native structures (Fig. 4e,f).

The probabilities in equation (7) were estimated using a binning procedure. The considerations discussed in the previous paragraph are also important in choosing the coarseness of the binning. For example, the φθ distribution for β-strand pairs is considerably more diffuse in native-like structures than in native structures, but is still clearly different from that of non-native structures (Fig. 4). Because of these differences, we considered estimating the distributions directly from the native-like structures in our ensembles, rather than from native structures, but we decided against this because of the much smaller number of independent structures in these sets (17) compared with the native protein set (325) (it was impractical to generate large ensembles for a representative subset of known structures). Instead, we chose to use a coarse binning procedure, which effectively smears out the native density over the regions of high density in the native-like conformations, but not in the non-native conformations. This procedure imparts a funnel shape to the "golf course"-type landscape that would result from overly fine binning of the φθ (or any other) distribution.

Care must be taken in estimating probabilities using a binning procedure and a spherical coordinate system: constant intervals bins may have very different sizes. Correction factors were used to account for the differences in the sizes of the bins. The correction factors are readily obtained analytically for the first three terms in equation (7): $\sin\theta$, $\sin\sigma$, and hb, for $hb < 1$, and $2 - hb$ for $hb > 1$, respectively. The geometric correction term for $P(r)$ would be r^2 in the absence of the chain connectivity and compactness constraints. Because of the difficulty of accounting for the effects of these constraints analytically, we estimated the size of the different r bins using the frequency of occurrence of different values of r in native proteins averaged over all secondary structure types. Guided by the comparison between the native, non-native, and near-native distributions, and keeping in mind that bins should not be too fine to avoid noise because of low counts, we decided to bin strand–strand pairs with $r < 6.5$ Å and helix–strand and helix–helix pairs with $r < 12$ Å, as described in Table I. For $r > 6.5$ Å for strand–strand pairs and $r > 12$ Å for helix–helix and helix–strand pairs, one bin was used; thus, the probabilities in equation (7) are constants that do not depend on r, φ, θ, σ, and hb.

Assembly of strands into sheets

We found that the distribution of strands in β-sheets cannot be adequately described by the pair-density func-

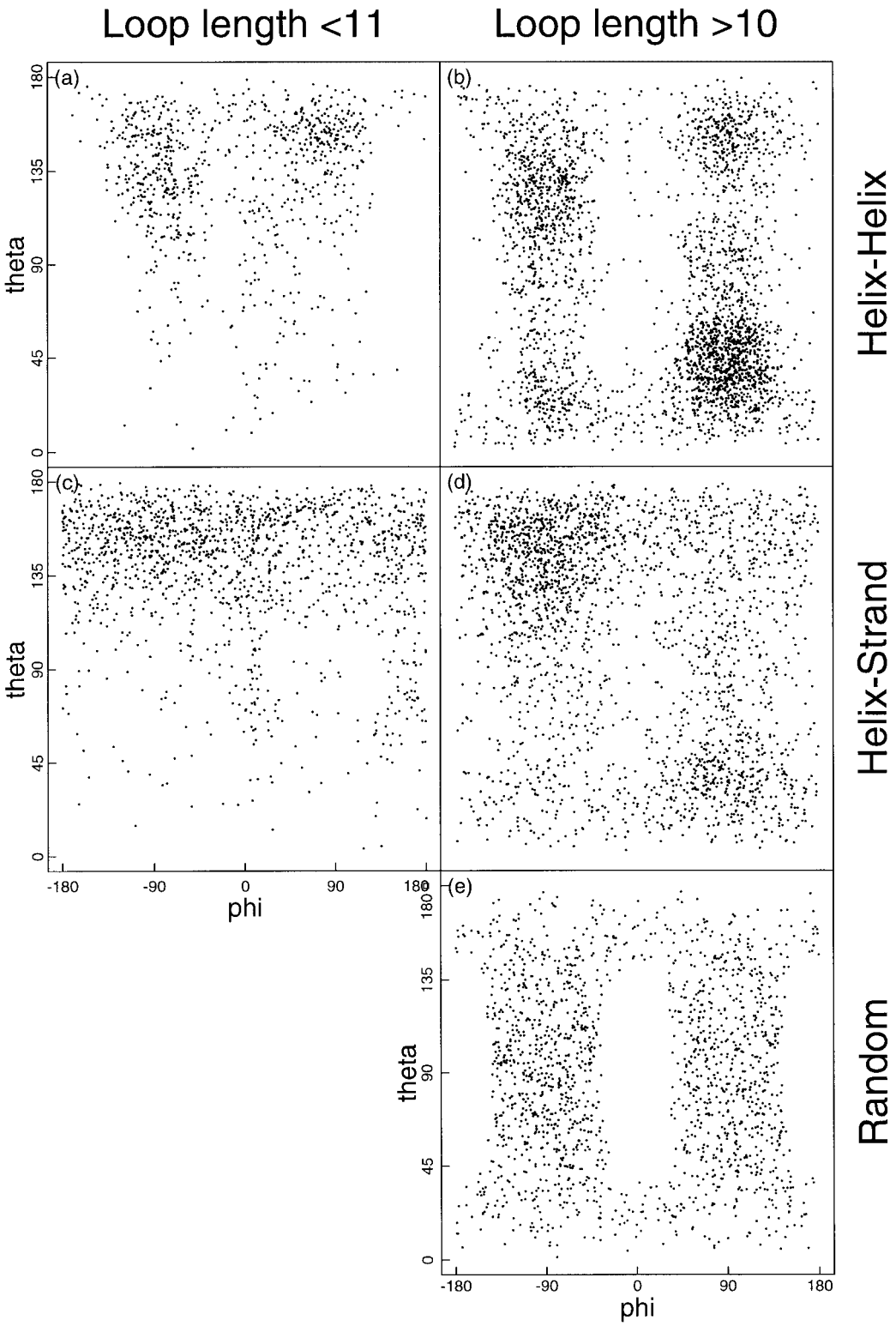

Fig. 3. $\phi\theta$ distributions in native protein structures for helix–helix and helix–strand pairs. Antiparallel helix pairs have $\theta > 90°$, and parallel pairs have $\theta < 90°$. **a:** Helical pairs separated by 2–10 residues. **b:** Helical pairs separated by >10 residues. **c:** Helix-strand pairs separated by 2–10 residues. **d:** Helix–strand pairs separated by >10 residues. **e:** Random distribution. Two vectors each the length of a helix of 12 residues were positioned by choosing their relative orientation at random and the center-to-center distance from the distribution of interhelical distances seen in known protein structures. Pseudo-atoms were placed along the vectors every 1 Å with a sphere radius of 7 Å. If the sphere of any atom intersected any atom from the other helix vector, the helical pair was rejected. A total of 3,000 accepted helical pairs were generated.

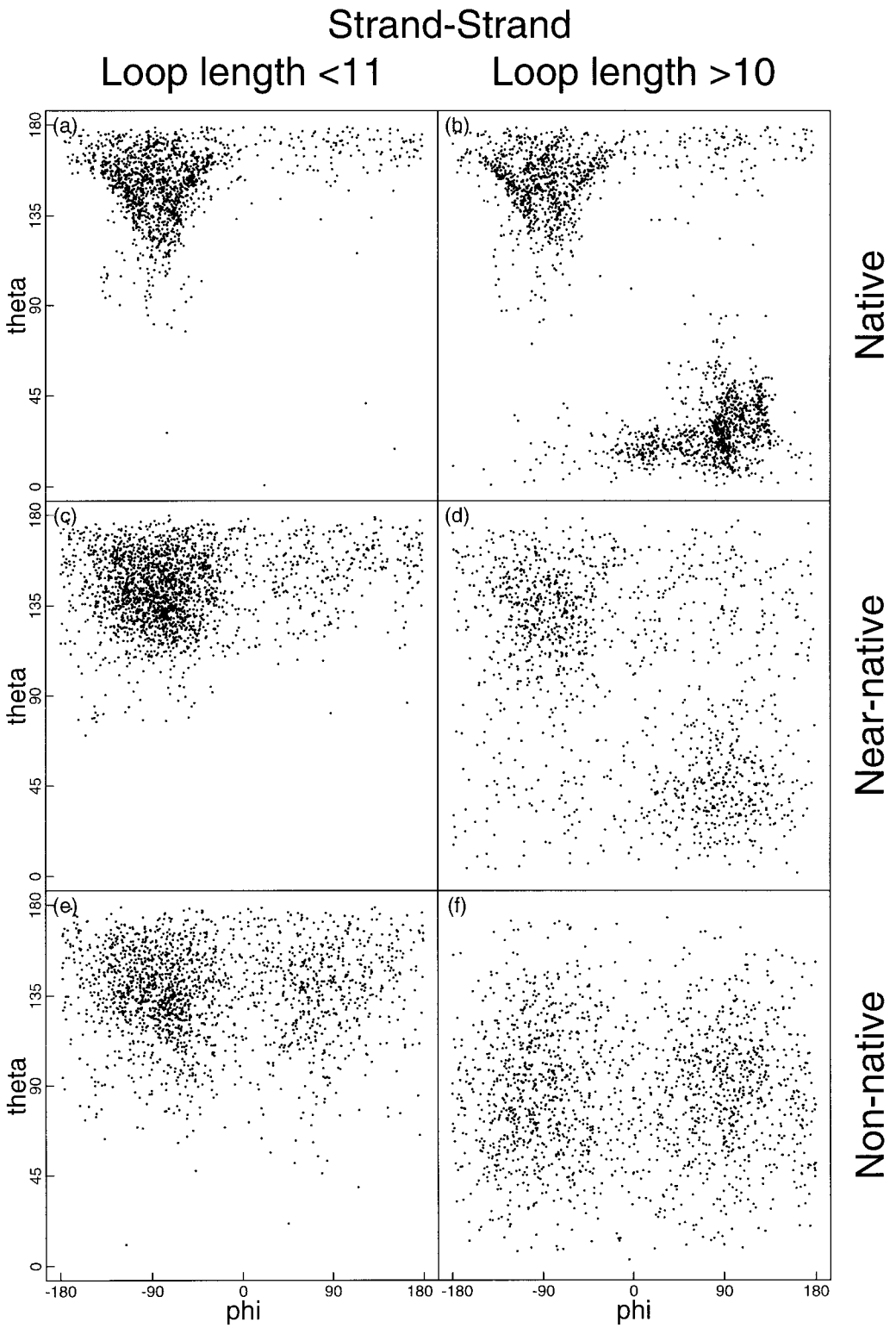

Fig. 4. $\phi\theta$ distributions for strand pairs in native, native-like, and non-native structures. Native structures were from the pdb_select 25 list, native-like and non-native structures, from the in-house decoy set. Native-like structures, $<$4 Å rmsd from the native structure; non-native structures, $>$8 Å rmsd.

tions. To describe strand assembly into sheets more accurately we collected data on how often certain configurations of strands in sheets occurred for proteins with fewer than 150 residues. These data are summarized in Table II. For example, among the 146 proteins that had any strands, 20 had five strands. Of these 20, seven had one sheet that

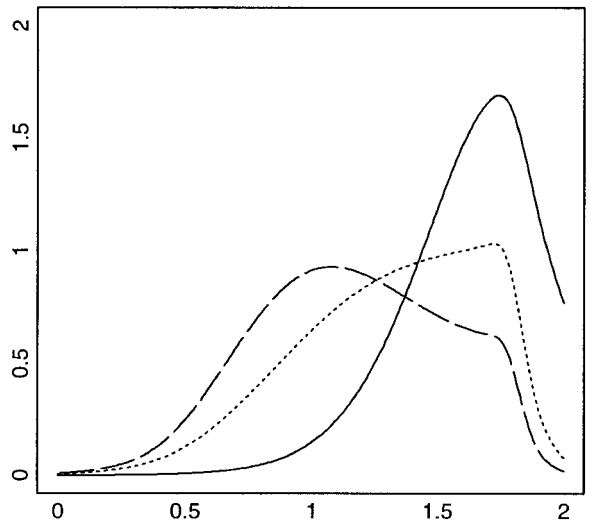

Fig. 5. Comparison of the $P(hb|r)$ distributions in native (*solid line*), native-like (*dotted line*) and non-native structures (*dashed line*).

was formed by three strands and one sheet by two strands. On the basis of these data, we decided to use the model

$$P_{\text{sheet}}(\text{sheet configuration}|\text{number of strands}) = \exp(c_n + a_1 n_1 + a_2 n_2) \quad (8)$$

where n is the number of strands, c_n is a normalizing constant that depends on the number of strands in the protein, n_1 is the number of sheets (with at least two strands) in the protein, and n_2 is the number of isolated strands. We used the method of maximum likelihood to fit this model and found $a_1 = -0.9$ and $a_2 = -2.7$. Thus, the probability that a protein with five strands has one sheet of three strands and a sheet formed by two strands is $\exp(c_4 - 1.8) = \exp(-1.19) = 0.30$. Table II summarizes the fitted values according to this model. There is reasonable agreement between the simple model and the data.

VdW

We found that many of the decoys with good scores for the sequence dependent terms were overly compact and had many atom–atom overlaps. Atom type dependent distance cutoffs were used for assessing overlaps between atom pairs. To allow for a low frequency of errors or anomalies in pdb structures, the cutoff radius r^0_{ij} was chosen to be the 25th smallest distance between atom types i and j in the pdb_select_25 database. The VdW term is the sum of penalties for each pair of atoms separated by less than the cutoff distance; for computational simplicity, the penalty was taken to be $(r^0_{ij})^2 - (r_{ij})^2$.

Additional Features Not Part of the Core Scoring Function

$P_{packing\text{-}struct}$

Sidechain packing is known to be an important attribute of folded protein structures; molten globules and unfolded proteins lack specific sidechain interactions. A remarkable preference in the relative orientation of packing residues was found in an analysis of native protein structures.[18] Some of the anisotropy of residue packing is thought to be caused by the directional preferences of electrostatic and hydrogen bonding.[19] Sidechain coordination was described by the angular distribution of sidechain centroids separated by less than 10 Å, using a spherical coordinate system defined by the Cα-centroid and Cα-N vectors of one of the residues. The sequence-independent term $P_{\text{packing-struct}}$ is $P(\phi, \theta)/\sin(\theta)$ and the sequence-dependent term $P_{\text{packing-seq}}$, $P(aa|\phi, \theta)$.

Radius of gyration and $P_{density}$

These are alternative measures of the compactness of a conformation. The radius of gyration is the square root of the average of the squares of the distances between all pairs of Cα atoms. The contribution to P_{density} for each residue is $P(n_i)/P_{rc}(n_i)$ where n_i is the number of Cβ atoms of other residues within 10 Å, $P(n_i)$ is the frequency with which n_i neighbors are observed in protein structures, and $P_{rc}(n_i)$, the frequency in randomly generated conformations (the correction factor for the difference in the size of the bins is very difficult to estimate analytically).

ISITES

A method for predicting local structure has recently been developed that improves on conventional secondary structure prediction in turn regions. The match between local sequence and local structure of the decoys was assessed using *ISITES* local structure predictions. The *ISITES* term in Table III is the product of $P(\text{local structure}|\text{local sequence})/P(\text{structure})$ over all fragment predictions that matched the decoy, where the numerator is the confidence of the prediction.[20]

$P_{local\text{-}struct}$

The *ISITES* score formulated in the previous section assesses the match of a sequence to a particular type of local structure, but does not assess the probability of observing the structure independent of sequence. The $P_{\text{local-struct}}$ term is

$$\prod_{i=1}^{n} P(\phi_i, \Psi_i) \prod_{i=1}^{n-1} \frac{P(\phi_i, \Psi_i, \phi_{i+1}, \Psi_{i+1})}{P(\phi_i, \Psi_i)P(\phi_{i+1}, \Psi_{i+1})} \quad (9)$$

Component Weighting of the Scoring Function

It is not immediately clear how these terms should be combined to form a prior distribution on all structures, $P(\text{structure})$. If all terms were independent density functions, it would be appropriate to simply multiply them together, as in

$$P(\text{structure}) = P_A P_B P_C \quad (10)$$

However, the density functions provide partly the same information: structures in which strands are closely aligned are more likely to have sheets formed and structures in

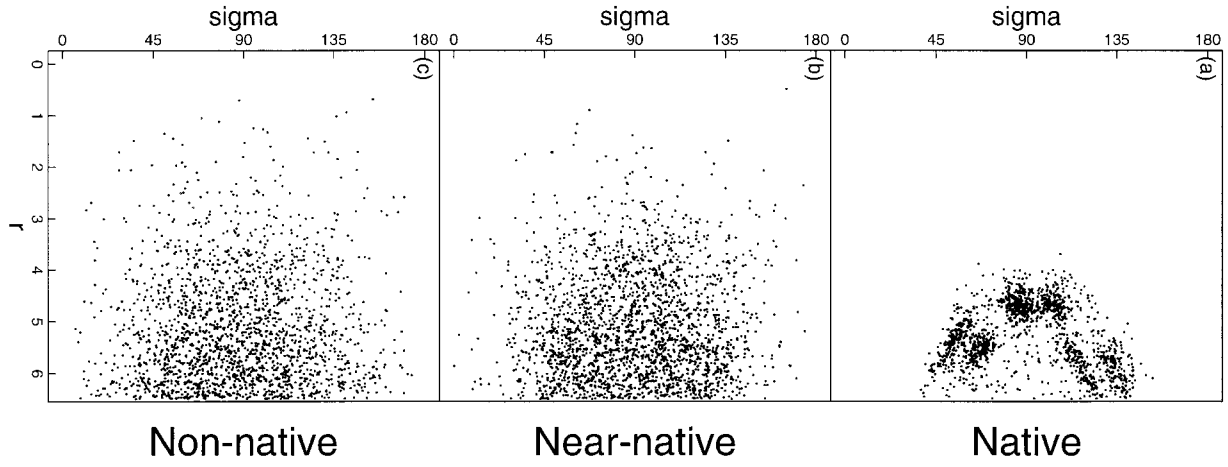

Fig. 6. Comparison of r-σ distributions in native, native-like, and non-native structures.

TABLE II. Comparison of Observed and Fitted† Sheet Configurations for Proteins With Fewer Than 150 Residues and Eight or Fewer Strands

No. of strands[a]	Sheet configuration[b]	Observed[c]	Expected[d]	c_n
1	1	0	0.0	0.07
2	2	10	10.9	0.41
	1–1	1	0.1	
3	3	6	5.6	0.43
	Others	0	0.4	
4	4	7	6.1	0.60
	2–2	2	2.5	
	Others	0	0.4	
5	5	13	13.3	0.61
	3–2	7	5.4	
	Others	0	1.3	
6	6	7	5.8	0.85
	4–2 or 3–3	3	4.7	
	Others	2	1.5	
7	7	8	7.6	0.86
	5–2 or 4–3	7	6.2	
	Others	1	2.2	
8	8	8	6.2	1.12
	6–2, 5–3 or 4–4	8	7.5	
	Others	1	3.3	

†See equation (8).
[a]The total number of strands found in the protein.
[b]Distribution of strands among separate sheets (e.g., 3–2 indicates that two separate sheets were identified, one with three strands, the other with two). Strands were grouped into sheets using single linkage clustering with a distance cutoff of 5.5 Å.
[c]The number of instances of a particular strand configuration in proteins with fewer than 150 residues and, at most, 8 strands.
[d]The number of instances predicted by equation (8) with $a_1 = -0.9$ and $a_2 = -2.7$.

which secondary structure elements are closely aligned are more likely to be compact. There is also significant overcounting within each of the terms: the many dimers within each secondary structure element are clearly not independent of each other. Finally, the VdW term is not a probability density, a distinction that we ignore. Because of the overcounting, we pooled the density functions logarithmically[21]:

$$P(\text{structure}) = P_A^{w_A} P_B^{w_B} P_C^{w_C},\ w_x > 0 \tag{11}$$

We optimize weights so that our scoring function is able to distinguish between native-like and non-native structures, rather than estimate the weights based on a collection of native structures using the method of maximum likelihood. The score, $-\log P(\text{structure}|\text{sequence}) \propto -\log(P(\text{sequence}|\text{structure})\ P(\text{structure}))$, is linear in the weights, which we determine by fitting a linear regression model of the form

$$g(\text{rmsd}) = w_{\text{protein}} + w_{\text{HS}} \log P_{\text{HS}} + w_{SS} \log P_{SS} + w_{VdW} VdW + w_{\text{sheet}} \log P_{\text{sheet}} + w_{seq} (\log P_{\text{env}} + \log P_{\text{pair}}) \tag{12}$$

where rmsd is the rmsd of the decoy and $g(\text{rmsd}) = 4$ if rmsd <4 Å, $g(\text{rmsd}) = 8$ if rmsd >8 Å and $g(\text{rmsd}) = \text{rmsd}$ otherwise. The cutoffs at 4 and 8 Å are made because we consider any decoy with an rmsd of <4 Å as good, while >8 Å is poor. The w_{protein} is a different intercept for each sequence. This intercept is irrelevant for selecting the best scoring structures for a particular sequence, but the separate intercepts, to some extent, correct for the different sizes and compositions of secondary structures of the various proteins. We experimented with other weighting schemes, including logistic regression, and found that all schemes performed equivalently on our training data; linear regression was chosen, as it was the simplest approach among the procedures we investigated. In fits including all the secondary structure packing terms, the hb and HH terms had very small contributions and were dropped for simplicity. The observation that inter-helix distance is correlated with the volume of the residues at the interface between helix pairs[22] suggests that a sequence dependent helix packing term could still be useful. The small contribution of the hb term may reflect the use of rmsd as the measure of structural similarity.

TABLE III. Overview of the Scoring Function

Probability density	Functional form	Putative physical origin	z-score
I. Sequence dependent	P(sequence|structure)		
A. Residue-environment	P_{env}; eq. (5)[b]	Hydrophobic effect	−1.5
B. Residue-residue	P_{pair}; eq. (6)[b]	Electrostatics, disulfides	−1.7
C. Local sequence-structure	ISITES	Sequence-local structure	−0.4
D. Packing orientation	$P_{packing\text{-}seq}$	Packing geometry	−1.3
II. Sequence independent	P(structure)		
A. Secondary structure packing	$P_{HH\text{-}\phi\theta}$; eq. (7)	Helix-helix packing	−0.3
	$P_{HH\text{-}dist}$; eq. (7)		−0.3
	$P_{HS\text{-}\phi\theta}$; eq. (7)[b]	Helix-strand packing	−0.8
	$P_{HS\text{-}dist}$; eq. (7)[b]		−0.8
	$P_{SS\text{-}\phi\theta}$; eq. (7)[b]	Strand-strand packing	−1.1
	$P_{SS\text{-}dist}$; eq. (7)[b]		−1.4
B. Strand hydrogen bonding	P_{SShb}; eq. (7)	Hydrogen bonding	−0.3
C. Strand assembly in sheets	P_{sheet} ; eq. (8)[b]	Hydrogen bonding	−1.4
D. Hard sphere repulsion	VdW[b]	Steric repulsion	−0.5
E. Structure compactness	$P_{density}$	Hydrophobic effect,	−0.4
	Radius of gyration	Van der Waals interactions	−1.0
F. Local structure	$P_{local\text{-}struct}$; eq. (9)	Local structure preferences	−0.6
G. Packing orientation	$P_{packing\text{-}struct}$	Packing geometry	−0.3

[a]The z-score is calculated as

$$\text{z-score} = \frac{\langle \text{score} \rangle_{good} - \langle \text{score} \rangle_{all}}{\sigma_{all}} \quad (13)$$

where $\langle \text{score} \rangle_{good}$ is the average score of the structures with <4 Å rmsd, $\langle \text{score} \rangle_{all}$ is the average score of all structures, and σ_{all} is the standard deviation of the score of all structures.
[b]Core model.

Generation of Decoy Structures

A training set of ~30,000 compact self-avoiding decoy structures with fixed secondary structure for each of 21 small proteins was made by replacing native backbone torsional angles with angles from known protein structures. Starting with the native dihedral angles, a randomly chosen residue in each turn was replaced with dihedral angles randomly chosen from the protein database. If the radius of gyration of the conformation was $>2 + 3*\text{length}^{1/3}$, the structure was rejected[11] and the procedure was repeated from the start. Otherwise, if atomic overlaps between Cβ atoms (3.0 Å) could be removed and the radius of gyration could be reduced below $3*\text{length}^{1/3}$ in 10,000 moves consisting of single residue dihedral substitutions, the structure was kept. Decoy sets were created for 1ctf, 2cro, 1r69, and 4icb, but these were not used in the estimation of weights (see below) because of overlaps with the PL test set. While there is not detectable sequence similarity between the remaining proteins in the training set and the proteins in the test set, there is considerable structural similarity between 1lmb, 2cro, and 1r69 and between 1ptx and 1sn3. Removal of 1lmb and 1ptx from the training set has relatively little effect on the performance of the scoring function on the test set. The best native-like ranks for 2cro and 1r69 are unchanged, and the best native-like rank for 1sn3 increases from 1 to 4.

Generation of Backbone Coordinates From Cα Traces

Since our method of scoring structures requires the location of all heavy atom backbone atoms otherwise absent in the PL test set, we developed an algorithm similar to MaxSprout[23] for the generation of backbone coordinates: First, a database of six residue peptides (hexamers) was extracted from the high resolution (<2 Å resolution) pdb-select 25 list and then pruned to eliminate pairs with rmsd <0.1 Å to yield a set of 41,330 structurally nonredundant hexamers. Beginning with the first six residues of the Cα model and marching across the chain in four residue increments, the backbone was built from the lowest rmsd hexamer in the set. The first and last residues of each hexamer (unless it covered the first or last residue of the Cα model) were discarded because the coordinates of the first nitrogen and last carbon cannot be determined unequivocally.

Protein coordinates were taken from the Brookhaven National Archive.[24] All the statistics of this work were taken from a nonredundant subset (pdb_select25[25]) of the Brookhaven database. Protein sequences in the test set (1ctf, 1r69, 1sn3, 1ubq, 2cro, 3icb, 4pti, 4rxn) and homologues (sequences with >30% sequence identity) were removed from the database before any statistics were computed. In addition, only compact structures (radius of gyration $<3*\text{length}^{1/3}$)[11] were used yielding a final database size of 325 proteins, 83,151 positions.

RESULTS

Our approach to scoring function development may be roughly divided into three steps:

1. Identify features that may contribute to distinguishing properly folded protein conformations from non-native conformations.
2. Choose an appropriate set of variables to describe each feature and estimate probability density functions for these variables from the protein structure database. While the parameters describing the densities are estimated statistically, the physical chemistry responsible for the form of the densities is readily apparent in most cases.
3. Evaluate the effectiveness of the density functions alone and in combination in native-like structure recognition using a training set consisting of large ensembles of compact conformations for a number of small protein domains. The most effective combination of the density functions is chosen as the final scoring function.

The final test, which is carried out only once, and after steps 1–3 have been completed, is to evaluate the performance of the scoring function on a completely independent set of decoy structures.

The features explored in this paper are summarized in Table III. The first two sequence dependent terms, P_{env} and P_{pair}, are the first terms in a systematic expansion of the spatial distributions of residues in native proteins (see Methods).[12] The P_{env} term primarily describes the partitioning of hydrophobic residues to the interior and polar residues to the surface. The P_{pair} term describes specific pair interactions not accounted for by the P_{env} term, primarily electrostatic interactions and disulfide bonds. The *ISITES* term assesses the fit between local sequence and local structure using a library of sequence-structure motifs.[20] The sequence packing term describes the differences in the packing orientations of different residue pairs.[18,19] Useful sequence-independent contributions to the scoring function must capture features of protein structures that distinguish them from random compact conformations. The sequence-independent terms we considered include (1) secondary structure packing preferences, (2) hard sphere repulsion, (3) compactness, and (4) the regular packing geometry of protein interiors (Table III). The choice of an appropriate set of variables to describe each feature and the estimation of the probability density functions from the protein structure database is described in detail under Methods, with a focus on the features that contribute to the core scoring function.

Evaluation of Individual Density Functions

To investigate the properties of the probability density functions without compromising the value of the PL test set, a training set of ~30,000 compact decoy structures with fixed native secondary structure for each of 17 different small proteins was generated as described under Methods. These decoy structures were scored using density functions corresponding to each of the features individually and the z-scores, the number of standard deviations separating the scores of the native-like structures (within 4 Å of the native structure) from the ensemble average, are listed in Table III. The three nonlocal sequence-dependent terms all had appreciable discriminatory power (average z-scores between −1.3 and −1.7). Of the sequence-independent terms, those involving strand pairing and association of strands into sheets had the most discriminatory power (Table III). The terms describing helix–helix pairing had relatively little discriminatory power.

Combination of Features: The Core Scoring Function

In combining the terms in Table III to create a scoring function that would capture both sequence dependent and sequence independent features of protein structures, we sought to use the most unrelated terms. Therefore, we began by combining the environment and pair terms, the secondary structure packing terms, and the hard sphere overlap term, which by construction are largely independent of each other (see under Methods). Linear regression was used to find a combination of these terms with optimal native-like recognition capabilities. The combined scoring function partially discriminated the native-like from the non-native conformations for all seventeen proteins (Table IV). The native-like z-scores for the different proteins ranged from −1.1 to −3.6, and a native-like structure almost always ranked in the top 10. The overall performance was considerably better for β-sheet containing proteins than for proteins containing only α-helices.

Interestingly, while a number of additional terms provided some degree of discrimination on their own, they failed to improve recognition when combined with the core scoring function. For example, the $P_{packing\text{-}seq}$ term provided modest recognition (average z-score −1.3) alone, but the z-score of the core scoring function was −2.6 with and without incorporation of this term. The most plausible explanation for such results is that these additional features are consequences of the features already included in the scoring function, together with the chain connectivity and compactness constraints.

Results With Independent Test Set

The PL test set was held in reserve until the details of the scoring function were finalized so that we could accurately assess the predictive value of the new scoring function. Proteins of detectable sequence similarity to the proteins in the PL test set (>30% sequence identity) were removed before construction of the scoring functions. As with the training set, the scoring function at least partially distinguished the native and native-like structures from the non-native conformations (Table V). The ranks of the native-like structures for the best of the functions tested by Park et al.,[7] the shell(top)m function, are also shown for comparison. The new scoring function gave the best scoring native-like structure a better rank than the shell(top)*m* function for six of the eight proteins, and the average z-score for the native-like structures was improved from −1.7 to −2.5. Native recognition was also somewhat better for the new function than for shell(top)m (Table V) (several of the functions tested by Park et al.[7]

TABLE IV. Performance of the Scoring Function on the Training Set†

Protein	2° structure	*P*(sequence|structure) z-score	*P*(structure) z-score	*P*(structure|sequence) z-score	*P*(structure|sequence) Best native-like rank	*P*(structure|sequence) Native rank
1aca	α	−2.6	−0.6	−2.6	3	4
1hdd	α	−1.1	−0.5	−1.1	1	2,691
1lmb	α	−1.6	−0.3	−1.6	25	2
1afp	β	−2.7	−2.0	−3.3	1	1
1csk	β	−1.4	−2.2	−2.4	4	1
1aba	α/β	−2.5	−2.0	−3.1	13	1
1cis	α/β	−1.5	−2.2	−2.1	2	8
1fxr	α/β	−2.3	−1.6	−2.6	1	1
1lea	α/β	−1.2	−0.6	−1.3	7	2
1orc	α/β	−2.2	−1.3	−2.3	2	2
1ptx	α/β	−3.6	−3.3	−4.7	1	1
1sap	α/β	−2.2	−2.9	−3.4	3	1
1spb	α/β	−1.7	−1.7	−2.4	3	1
1stu	α/β	−1.4	−3.2	−3.3	1	1
1tig	α/β	−2.4	−3.1	−3.5	2	1
2gb1	α/β	−1.5	−2.2	−2.7	14	1
2pt1	α/β	−1.7	−1.9	−2.4	3	1
		−2.0	−1.9	−2.6		

†Native-like structures are within 4 Å rmsd of the native structure. The best native-like rank is the rank of the native-like structure with the best score, not including the native structure. The *P*(sequence|structure) and *P*(structure) terms are those indicated in Table III. The two contributions are nearly independent: the average z-score expected, were they completely independent, is only slightly more negative than that observed (−2.75 ($\sqrt{2.0^2 + 1.9^2}$) vs −2.6).

TABLE V. Performance of the Scoring Function on the PL Test Set†

Protein	2° structure	*P*(sequence|structure) z-score	*P*(structure) z-score	*P*(structure|sequence) z-score	*P*(structure|sequence) Best native-like rank	*P*(structure|sequence) Native rank	Shell(top)*m* Best native-like rank	Shell(top)*m* Native rank
1r69	α	−1.9	−1.1	−2.0	7	1	11	30
2cro	α	−1.1	−0.9	−1.2	6	371	9	85
3icb	α	−2.7	−1.0	−2.8	2	1	1	1
4rxn	β	−2.5	−1.6	−2.8	3	21	30	31
1ctf	α/β	−2.6	−2.6	−3.5	1	1	1	1
1sn3	α/β	−1.5	−1.8	−2.2	1	1	11	370
1ubq	α/β	−2.6	−2.3	−3.1	2	1	76	1
4pti	α/β	−1.6	−1.8	−2.3	23	21	106	161
		−2.1	−1.6	−2.5				

†The function that was found to best distinguish native-like from non-native structures in Park et al.[7] was shell(top)*m*. This function had an average z-score of −1.7 compared with −2.5 for our core scoring function (z-scores for individual proteins were not reported in Park et al.[7]).

outperformed both the shell(top)*m* and our function in native recognition but were considerably worse for the native-like recognition problem, which is more relevant for ab initio structure prediction).

Plots of score vs rmsd for all the conformations in the PL test set are shown in Figure 7. In contrast to several previously tested scoring functions,[11] there is a consistent decrease in score with decreasing rmsd for all eight proteins. For the β-sheet containing proteins 1ubq and 1ctf, the scoring functions clearly distinguish the native-like structures from the vast majority of the non-native structures.

DISCUSSION

The new scoring function does substantially better at recognizing the native-like structures in the ensembles of compact decoys in the PL test set than any of the many functions previously tested.[11] To appreciate the significance of the increase in average z-score from −1.7 for the shell(top)*m* function of Park et al.[7] to −2.5 for the new scoring function, it is useful to note that were the distributions of the scores normal, this increase would indicate that the average score of the native-like structures in the top 0.6% rather than the top 4.1% of the score distributions. The performance is much better on β-sheet-containing proteins because fixing secondary structure provides stronger topological constraints for β-strands than for α-helices. While other criteria could potentially also capture the residual strand pairing in the native-like structures, standard hydrogen bonding criteria are considerably too strict (even 3-Å rmsd structures are not recognized to have paired β-strands by DSSP).

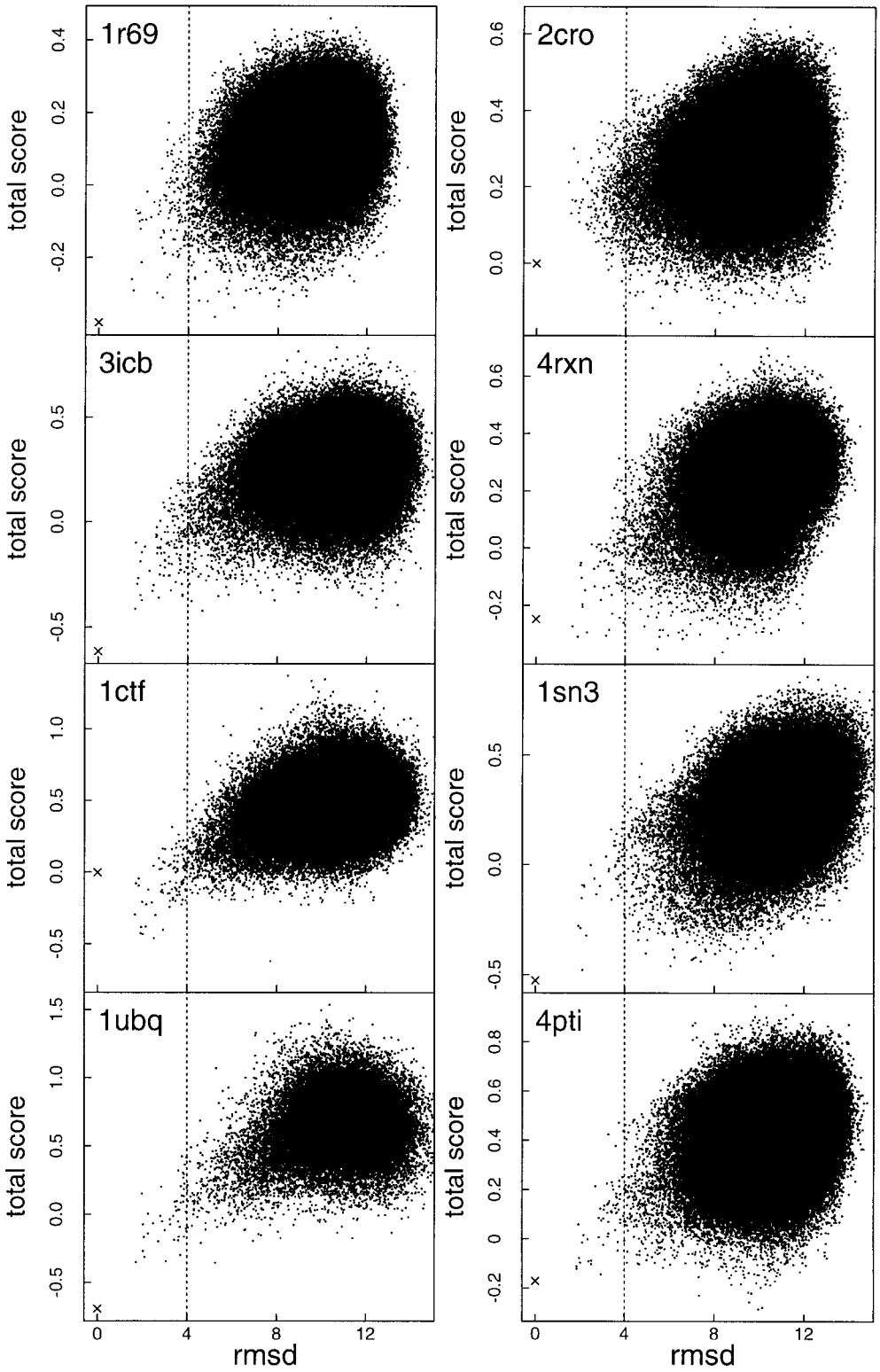

Fig. 7. Plots of $-\log P(\text{structure}|\text{sequence})$ vs rmsd for the eight decoy sets of Park and Levitt.[11] Native-like structures (<4 Å rmsd) are distinguished from the ensemble of non-native structures for most of the proteins.

The separation of the scoring function into components provides insights into the relationships between different features of protein structures. For example, the ineffectiveness of the helix–helix packing terms compared with the strand-strand and strand-helix terms suggests that the patterns evident in the helix–helix distributions (Fig. 3a,b) are fully accounted for by the chain connectivity and compactness constraints, together with terms already included in the scoring function, notably hydrophobic burial (a similar conclusion was reached by Bowie[26]). The pronounced orientational packing preferences of residue pairs in protein structures also appear to be a secondary consequence of more basic features since, despite having a significant z-score alone (−1.3), the $P_{\text{packing-seq}}$ term did not improve the performance of the core scoring function. The local interactions based *ISITES* and $P_{\text{local-struct}}$ terms provided relatively little discrimination; this may reflect the method by which the decoy set was generated or simply the insensitivity of the overall rmsd to details of local structure; *ISITES* predictions may be more useful as a move set in ab initio folding simulations than in structure evaluation. Finally, measures of chain compactness such as the overall Cβ density and the radius of gyration had little discriminatory power in the compact decoy sets studied here, suggesting that the cutoff on the radius of gyration proposed by Park and Levitt[11] together with *VdW* and hydrophobic burial terms capture the high packing density of protein structures to the extent that it can be captured in a simplified model. This result is useful for ab initio folding simulations because the appropriate functional form for a compacting force is not at all apparent; the radius of gyration-based functions may bias simulations toward spherical structures.

There is still considerable room for improvement of the scoring function, particularly for the all helical proteins such as the cro repressor (2cro) for which many non-native conformations scored as well as, or better than, native-like conformations. Recent results using atom pair distributions[27] suggest that improved discrimination can be obtained by scoring full atom models following the addition of explicit side chains. The systematic decomposition strategy used in this report should be applicable to scoring functions based on full atom models and may further improve performance on the helical proteins.

The improved performance on β-sheet containing proteins contrasts sharply with our earlier results in ab initio folding simulations using a scoring function lacking the secondary structure packing terms: α-helical proteins folded much more readily than β-sheet containing proteins. The current success with β-sheet containing proteins (Fig. 7) suggests that ab initio folding simulations using the method of Simons et al.[12] and the new scoring function should yield better results with a broad range of small proteins provided that the secondary structure is known. We look forward to testing the approach in the CASP3 structure prediction challenge.

ACKNOWLEDGMENTS

We thank Britt Park and Michael Levitt for making their decoy set available to us, and Jerry Tsai and Enoch Huang for considerable help with using it. David Shortle and Kevin Plaxco provided stimulating discussion and critical reading of the manuscript. This work was supported by an STC NSF grant and by Young Investigator awards (to D.B.) from the NSF and the Packard Foundation. K.T.S. was supported by PHS NRSA T32 GM07270 from NIGMS. C.K. was supported in part by NSF grant DMS-9403371.

REFERENCES

1. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
2. Dandekar T, Argos P. Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. J Mol Biol 1996;256:645–660.
3. Huang ES, Subbiah S, Tsai J, Levitt M. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. J Mol Biol 1996;257:716–725.
4. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 1994;18:338–352.
5. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256: 623–644.
6. Monge A, Lathrop EJ, Gunn JR, Shenkin PS, Friesner RA. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
7. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
8. Sippl MJ. 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 1990;213: 859–883.
9. Srinivasan R, Rose GD. LINUS: a hierarchic procedure to predict the fold of a protein. Proteins 1995;22:81–99.
10. Kocher JA, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 1994;235:1598–1613.
11. Park B, Levitt M. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
12. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 1997;268:209–225.
13. Swayne DF, Cook D, Buja A. XGobi: Interactive dynamic graphics in the X Window system with a link to S. Proceedings of the 1991 American Statistical Association Meetings. Alexandria, VA: ASA; 1992.
14. Harris NL, Presnell SR, Cohen FE. Four helix bundle diversity in globular proteins. J Mol Biol 1994;236:1356–1368.
15. Walther D, Eisenhaber F, Argos P. Principles of helix-helix packing in proteins: the helical lattice superposition model. J Mol Biol 1996;255:536–553.
16. Efimov AV. A novel super-secondary structure of proteins and the relation between structure and the amino acid sequence. FEBS 1984;166:33–38.
17. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167–339.
18. Bahar I, Jernigan RL. Coordination geometry of nonbonded residues in globular proteins. Folding Design 1996;28:357–370.
19. Mitchell JBO, Laskowski RA, Thornton JM. Non-randomness in

side-chain packing: the distribution of interplanar angles. Proteins 1996;29:370–380.
20. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
21. Genest C, Zidek J. Combining probability distributions: a critique and an annotated bibliography. Stat Sci 1986;1:114–148.
22. Reddy BVB, Blundell TL. Packing of secondary structural elements in proteins. J Mol Biol 1993;233:464–479.
23. Holm L, Sander C. Database algorithm for generating protein backbone and side-chain co-ordinates from a Cα trace. J Mol Biol 1991;218:183–194.
24. Bernstein FC, Koetzle TF, Williams GJ, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
25. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
26. Bowie JU. Helix packing angle preferences. Nature Struct Biol 1997;4:915–917.
27. Samudrala R, Moult J. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
28. Reid LS, Thornton JM. Rebuilding flavodoxin from C alpha coordinates: a test study. Proteins 1989;5:170–182.