

ProVal: A Protein-Scoring Function for the Selection of Native and Near-Native Folds

Anders Berglund,¹ Richard D. Head,^{1,2} Eric A. Welsh,¹ and Garland R. Marshall^{1*}

¹Center for Computational Biology, Washington University Medical School, St. Louis, Missouri

²Pfizer Corporation, Computational Biology Group, St. Louis, Missouri

ABSTRACT A low-resolution scoring function for the selection of native and near-native structures from a set of predicted structures for a given protein sequence has been developed. The scoring function, ProVal (Protein Validate), used several variables that describe an aspect of protein structure for which the proximity to the native structure can be assessed quantitatively. Among the parameters included are a packing estimate, surface areas, and the contact order. A partial least squares for latent variables (PLS) model was built for each candidate set of the 28 decoy sets of structures generated for 22 different proteins using the described parameters as independent variables. The C_{α} RMS of the candidate structures versus the experimental structure was used as the dependent variable. The final generalized scoring function was an average of all models derived, ensuring that the function was not optimized for specific fold classes or method of structure generation of the candidate folds. The results show that the crystal structure was scored best in 64% of the 28 test sets and was clearly separated from the decoys in many examples. In all the other cases in which the crystal structure did not rank first, it ranked within the top 10%. Thus, although ProVal could not distinguish between predicted structures that were similar overall in fold quality due to its inherently low resolution, it can clearly be used as a primary filter to eliminate ~90% of fold candidates generated by current prediction methods from all-atom modeling and further evaluation. The correlation between the predicted and actual C_{α} RMS values varies considerably between the candidate fold sets. *Proteins* 2004; 54:289–302. © 2003 Wiley-Liss, Inc.

Key words: protein folding; empirical scoring function; structure prediction; partial least squares; PLS

INTRODUCTION

The protein structure prediction problem, inferring a three-dimensional (3D) structure from a one-dimensional (1D) amino acid sequence, can be dichotomized into two separate problems. The first problem is the search for “reasonable” candidate structures, a diverse set of protein folds that are energetically reasonable with respect to the native structure of a given sequence. Candidate structures

can be obtained by a host of means; ab initio structure generation, threading, combinatorial fragment combination, and homology modeling. These methods typically generate dozens to thousands of candidate structures falling into many different folds. Although it is not particularly difficult to bin the structures into fold classes, it is a nontrivial problem to determine which of the candidates most resembles the native fold. The problem is compounded by the fact that the native structure, or even a structure representing the native fold, is not necessarily present in the candidate set. With use of conventional scoring methods, such as energetics derived from molecular mechanics, folds that are quite dissimilar in appearance may be nearly indistinguishable in energy. Similarly, candidate structures that are very similar to that seen in the crystal structure may have unusually high energies due to van der Waals interactions because the candidate has small but significant errors in atomic coordinates.

Thus, the second problem associated with protein structure prediction is scoring (i.e., how to evaluate each candidate structure in terms of its “nearness” to the native state). There have been several papers presenting different types of scoring functions that attempt to find native or near-native structures from a set of candidates.^{1–5} There have also been several reviews on these functions for application in structure prediction.^{6–11} These functions can be divided into two major categories: those built from first principles (i.e., approximate physical energies) and those derived from statistical measures.² Energetic models can typically isolate the native structure from a set of candidates. However, given the usual nature of the protein energy surface, it can be extremely difficult to distinguish a near-native structure from more distant candidates, because very different folds may have similar energetics. For example, it is not atypical to have a 4 Å structure that has similar or higher energy than a 10 Å structure. An added complexity is that solvation must generally be accounted for to ensure accurate energy estimates. This too, is nontrivial for macromolecules and is discussed in a

A. Berglund's present address is Research Group for Chemometrics, Department of Chemistry, Umeå University, S- 901 87 Umeå, Sweden.

*Correspondence to: Garland R. Marshall, Center for Computational Biology, Washington University Medical School, 700 S. Euclid Ave., St. Louis, MO 63110. E-mail: garland@pcg.wustl.edu

Received 17 February 2003; Accepted 4 June 2003

review by Roux and Simonson.¹² Finally, energy functions are computationally expensive and may be somewhat unattractive for the evaluation of large sets of candidate structures.

Statistically derived scoring functions are built by using data from known protein structures and, in some cases, misfolded versions of those structures. The pairwise contact potential is an example of a commonly used statistical potential function. Most statistically derived scoring functions use a reduced representation of the protein. An example is the α -carbon-only representation.¹³ This type of reduction generally results in a significant timesaving for the calculations performed, but usually at the cost of accuracy. One of the strengths of statistically derived functions is the ability to combine several different and sometimes disparate terms.¹⁴ Park et al.³ evaluated a multitude of statistically derived potential functions and the factors affecting them by testing them against sets of decoys generated by multiple methods. Decoys, sets of non-native structures, are often used in both the derivation and the evaluation of energetic and statistical functions. However, the method by which these structures are generated is often of great importance. It was observed by Park that many of the functions evaluated would perform well with either certain types of proteins or on decoy sets generated by a particular methodology. The decoy sets used by Park were created by threading,¹⁵ molecular dynamics at 298K and 498K,¹⁶ and loop searching.¹⁷ One possible explanation for this method-specific performance would be the nonrandom structural characteristics of the decoy sets generated by a particular approach.

The aim of this work was to develop a scoring function that is capable of finding, when present, the native or near-native structures within a set of fold candidates. Based on the observations of Park and others, it appeared that a generalized function would need to be built from many functions. In turn, these terms could be statistically derived, physical energies, or a mixture of both. Head et al.¹⁸ used this approach in the prediction of binding affinity by combining components of molecular mechanics and binding free energy into a single scoring function termed VALIDATE.¹⁸ VALIDATE used 12 different fields that included sterics, electrostatics, hydrophobicity, and so forth and combined these terms by using a model derived by partial least squares of latent variables¹⁹ from a training set of crystallographic data and known binding affinities. Knowing that many of the existing low-resolution scoring functions work fairly well on at least some subclasses of proteins, we revisited the concept of combining several such functions into a single empirically derived function. Simons et al.⁵ presented a function of this nature for evaluating candidate protein structures. In this function, the different terms are logarithmically pooled and then weighted. Weighting was accomplished with linear regression. A typical problem with many linear regression techniques is that they are not tolerant to noisy or colinear variables. Simons et al.⁵ cited this as a reason that not all variables were used in final derivation of the scoring function. In ProVal (Protein Validate), partial least squares

of latent variables (PLS¹⁹) was the selected regression method because it allows for colinear and noisy variables and does not require user selection of the optimum set of variables.

The first objective of the ProVal function was to select the native structure (crystal/NMR structure) from a set of decoys regardless of the size of the protein, the number of decoys, type of fold, or method of decoy generation. The second objective was to obtain a favorable correlation between the root-mean-square values for the α -carbons of the amino acids (CRMS value) of decoys and the experimental structure and the calculated score. This correlation should be an indication of the function's ability to discriminate near-native structures, when present, from more distant ones and is discussed further in the Results.

MATERIALS AND METHODS

Variable selection for ProVal was not limited to published work. Novel methods were developed and known approaches were refined to capture information observed in crystallographic data from the PDB. The calculated terms range from very low-resolution, requiring nothing more than C_α coordinates, to moderate resolution, requiring either C_β coordinates, or more preferably, the coordinates of the heavy atoms on the side-chains. Currently, none of the fields require an all-atom model. Energy minimization is not required, and all of the fields can be computed rapidly, making it possible to score large numbers of structures.

Packing

Packing estimate

The first field in ProVal is a packing estimate of the folded protein. Because the native folds of most proteins are not elongated and quantum-mechanical effects prohibit atoms from existing in the same region of space at the same time, there are upper and lower limits on the average distances between residues in a folded state based solely on the total number of amino acids in the given protein. The packing estimate function was derived by calculating the sum of inter- C_α distances for the fold. For the purpose of noise reduction, the first and last two residues of the sequence were not included, and the distances were only calculated for residues separated by at least 10 positions in the linear chain.

$$PE = \sum_{i=2}^{nres-12} \sum_{j=i+10}^{nres-2} \sqrt{(\vec{r}_{\alpha_i} - \vec{r}_{\alpha_j})^2} \quad (1)$$

\vec{r}_{α_i} = position of C_α of the i 'th residue

\vec{r}_{α_j} = position of C_α of the j 'th residue

To use the packing estimate in a predictive fashion, a target function based on a diverse set of 312 protein chains was derived. The February 2000 PDBSELECT95²⁰ was used to generate a list of high-quality X-ray structures, ≤ 3.0 Å, with no chain breaks, and all of their backbone and C_β atoms resolved. Because ligand binding and multimer-

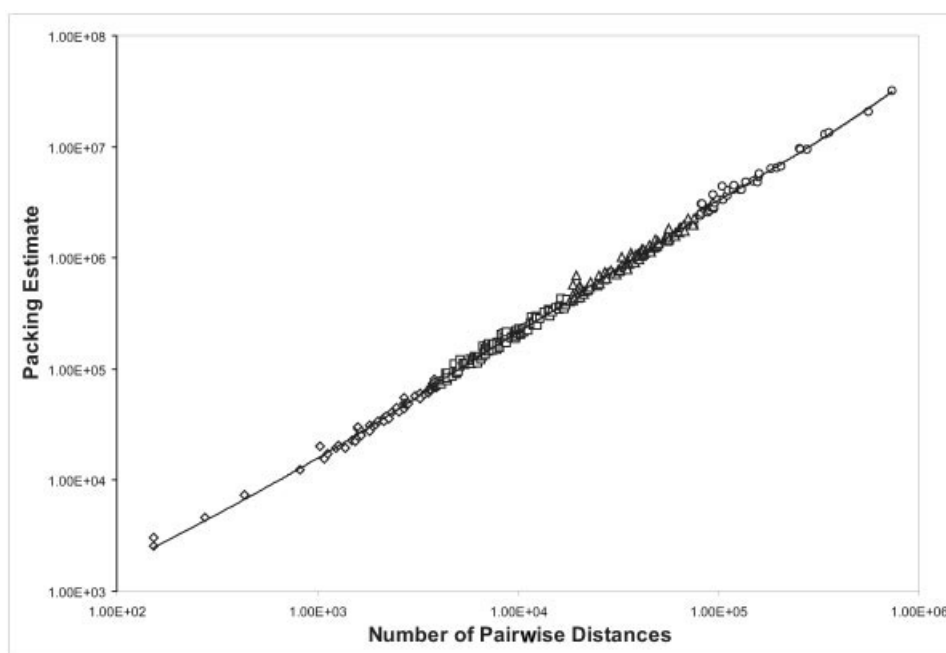


Fig. 1. This figure is a simultaneous plot of the four equations given in Table II. The y axis represents the packing estimate, which is the inter- C_{α} distance sum described in Eq. 1. The x axis is the number of pairwise distances used to generate the packing estimate for a given protein. The actual values for the 312 structures used to derive the equations have been overlaid to show the quality of the fit. \diamond , 30–99 residues, \square , 100–199 residues, \triangle , 200–399 residues, \circ , 400–1227 residues.

ization can cause large structural changes from the unbound monomeric state, all structures with bound ligands were discarded, along with all PDB^{21,22} files containing greater than two subunits. Structural diversity was achieved by keeping only the highest resolution structure from each SCOP 1.50 superfamily.²³ Several far outliers in the packing estimate plot were discarded when they could be rationalized due to high numbers of charged residues or large numbers of disulfide bridges, both of which can cause the packing to be much different from the average observed for chains of equal size. The packing estimate was calculated by using the remaining 312 proteins, and a target function was fit to it on the basis of the number of residues (see Fig. 1). Four target functions were derived on the basis of size to compensate for the nonlinear nature of the function. The functions and their corresponding r^2 value are given in Table I. A target function is then used to predict the packing estimate for a given protein sequence based on length. It is the resulting difference between the predicted packing estimate and the packing estimate calculated explicitly for the candidate structure that is used.

$$\Delta PE = PE_{\text{Predicted}} - PE_{\text{Observed}} \quad (2)$$

Hydrophobicity

In development of this parameter, there were three terms dedicated to measuring hydrophobic properties. In a sense, each is a specific extension of information gained through previous work based on statistical potentials.

TABLE I. Four Equations Used to Calculate the Packing Estimate Value (y) Based on Crystallographic Data[†]

No. of Residues (x)	Function	R^2
30–99	$1E-03x^2 + 14.439x - 249.83$	0.98
100–199	$1E-04x^2 + 21.683x - 12701$	0.97
200–399	$6E-05x^2 + 24.825x - 37417$	0.97
≥ 400	$2E-05x^2 + 30.117x + 116917$	0.99

[†]The x value is the number of pairwise distances to be used in the packing estimate (see discussion on packing). The y value is the sum of these pairwise distances.

Hydropole energy

The hydropole energy is actually a pseudoenergy function that attempts to capture the macroscopic phenomena of hydrophobicity at an atomic level. The concept of a hydropole is not unique. De Araujo described a similar hydrophobic energy function where the contribution from each monomer is a product of its hydrophobicity and the number of contacts it makes.²⁴ A hydrophobic multipole (or hydropole) was calculated at the C_{β} positions for all 20 standard amino acids. This is accomplished by first generating a conformationally averaged hydrophobic potential surface. The points on the surface are simply an average of the partition coefficients, calculated by the Hint 1.1 program,^{25,26} of the atoms that come in contact with that point for all conformations. The Estar program²⁷ was then used to fit the multipole to the potential surface. The hydropole contains monopole, dipole, and quadrupole terms.

TABLE II. Classification of Amino Acids by Hydrophobic Character

Hydropathy	Amino acids
Hydrophiles	Arg, Asn, Asp, Gln, Glu, Lys, Ser, Thr, Tyr
Hydrophobes	Ile, Leu, Phe, Trp, Val
Neutral	Ala, Cys, Gly, His, Met, Pro

All three terms were initially tested in the PLS model; however, only the monopole term appeared to contribute significantly and the dipole and quadruple terms were dropped. Thus, the calculation of hydropole energy for the entire protein became a coulomb-like potential.

$$\Phi_{hydropole} = \sum \sum \frac{q_{\beta_i} q_{\beta_j}}{r_{ij}} \quad (3)$$

where q_{β_i} is the monopole term at $C\beta$ of i 'th residue, q_{β_j} is the monopole term at $C\beta$ of j 'th residue, and $r_{i,j}$ is the distance between $C\beta$ of i 'th and j 'th residues.

Hydrophobic core potential

The hydropole energy is similar to a contact potential for residues of similar hydropathy. The hydrophobic core potential was used to capture the propensity of most soluble proteins to form a nonsolvent-accessible core consisting, almost exclusively, of hydrophobic residues. Many have developed similar functions.^{24,28-30} A pseudopotential is calculated by placing a positive "hydrophilic charge" at the centroid of a given candidate conformation and a similar charge on the atom of each residue that is nearest to the centroid. A hydrophilic residue receives a positive charge, and a hydrophobic residue receives a negative charge. Thus, candidate folds are rewarded with a negative potential for having hydrophobic residues near the centroid, which roughly represents the center of the protein.

$$\Phi_{Core} = \sum_i^{nres} \frac{q_{\beta_i} q_{Centroid}}{r_{i,Centroid}} \quad (4)$$

where q_{β_i} is the charge term at $C\beta$ of i 'th residue, $q_{Centroid}$ is the charge term at centroid of structure, and $r_{i,centroid}$ is the distance between $C\beta$ of i 'th and centroid.

Localized Hydrophobic Packing

Three localized packing scores that recognize the packing propensities of hydrophobic residues of similar nature were calculated. In this case, similar nature refers to pure hydrophobic residues (Table II) that contain rings and those that do not. The scores are simply average contact distances, and the categories are ring/ring, ring/non-ring, and non-ring/non-ring. Equation 5 below defines the localized hydrophobic packing for the hydrophobic ring-containing residues of phenylalanine and tryptophan. The equations for ring/non-ring and non-ring/non-ring are identical for the appropriate amino acid types.

$$LHP_{ring-ring} = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^N \sqrt{(\tilde{r}_{\beta_i} - \tilde{r}_{\beta_j})^2}}{N(N-1)} \quad (5)$$

\tilde{r}_{β_i} = Position of $C\beta$ of i 'th residue

\tilde{r}_{β_j} = Position of $C\beta$ of j 'th residue

N = number of residues

Electrostatics

High-resolution energy terms have been avoided in this approach because small differences in coordinates can generate large differences in energy, and the scoring function was not intended for use with all-atom structures. As was previously discussed, one of the goals of this work is to be able to identify near-native, low-resolution structures, and high-resolution functions are not well suited for this. However, a low-resolution electrostatic term was quite useful.

Low-Resolution Electrostatic Energy

A very low-resolution estimate of the electrostatic potential energy of a candidate structure was calculated by placing the net charge of each residue at the $C\beta$ position and computing a standard coulombic sum. Again, the first two and the last two residues of the structure were discarded because their positions are generally less well determined.

$$\Phi = \sum_{i=2}^{Nres-3} \sum_{j=i+1}^{Nres-2} \frac{q_i q_j}{r_j} \quad (6)$$

where q_i is the formal charge at $C\beta$ of i 'th residue, q_j is the formal charge at $C\beta$ of j 'th residue, and $r_{i,j}$ is the distance between $C\beta$ of i 'th and j 'th residues.

Solvent-Accessible Surface Areas

Like the packing estimate, solvent-accessible surface area is also a pseudomeasure of the compactness of a candidate structure. However, solvent-accessible surface area can be viewed in total, or it can be broken down into hydrophobic and hydrophilic components. The distribution of total surface area into these components contains useful information as well. In this work, all surface areas were normalized by the total number of residues in the given protein sequence. This has no direct impact on a given set of candidates for a single protein of a unique size; however, its purpose lies in the later ability to merge data from several protein decoy sets into a combined model.

Solvent-accessible surface area (SASA)

The solvent-accessible surface area was calculated by computing the exposed surface area for each atom of each residue in a given candidate structure. The accessible surface area for each atom was estimated by using the SASA³¹ software with a solvent radius of 1.4 Å. The atomic

surface areas were then summed and divided by the total number of residues in the protein. Division by the total number of residues was a normalization procedure for later calculation of a generalized scoring function.

Hydrophilic solvent-accessible surface area (HSASA)

The hydrophilic solvent-accessible surface area is the hydrophilic component of the total solvent-accessible surface area.

$$\text{Normalized HSASA} = \frac{\text{Total HSASA}}{\text{Total SASA}} \quad (7)$$

The total surface area of atoms from hydrophilic amino acids defines the hydrophilic component. We admit that this was a somewhat crude determination when considering amino acids, such as tyrosine, and further refinement may be possible. Dividing by the total solvent-accessible surface area normalized this value.

Lipophilic solvent-accessible surface area (LSASA)

The lipophilic solvent-accessible surface area is the lipophilic component of the total solvent-accessible surface area. Again, the total surface area of atoms from lipophilic amino acids defines the lipophilic component. The value was normalized through division by the total solvent-accessible surface area.

$$\text{Normalized LSASA} = \frac{\text{Total LSASA}}{\text{Total SASA}} \quad (8)$$

Disulfides

The propensity of extracellular proteins to nearly always form disulfide bridges between pairs of cysteines was very useful when filtering through candidate structures. Two terms based on this feature were included in the function.

Disulfide bridge count

Disulfide bridge count is an estimate of the number of disulfide bridges formed in a given candidate structure based on the proximity of cysteine pairs. The function is basically binary in nature. Only if the C_β values of a cysteine pair are within the user-specified cutoff distance (4.5 Å for this article) were they counted as a disulfide bridge.

Cysteine minimum-distance function (CysMDF)

This function is a sum of the distances from each cysteine to the nearest neighboring cysteine in the candidate fold. The disulfide bridge count is adequate in the identification of the native structure because all bridges will be identified. However, near-native structures typically will not have all cysteine pairs in sufficiently close proximity to identify them as a cysteine bridge. The cysteine minimum distance function identified structures that have cysteine pairs that are in close proximity for forming disulfide bridges and is given by the equation below.

$$\text{CysMDF} = \sum_{i=1}^{N_{\text{Cys}}} \sqrt{(\vec{r}_{\text{Cys}i} - \vec{r}_{\text{CysNear_Neighbor}})^2} \quad (9)$$

$\vec{r}_{\text{Cys}i}$ = Position of C_β of i 'th residue

$\vec{r}_{\text{CysNear_Neighbor}}$ = Position of C_β of nearest

cysteine neighbor of i 'th residue

Contact order

The contact order is a measure of how residues in contact are distributed in the sequence in the 3D structure. The contact order correlates with the folding rate of proteins.³² If two C_α values are within 8 Å, they were defined as being in contact; this is a slight modification of the definition given by Plaxco et al.,³² where all atoms in the chain were taken into account.

C_α torsional energy

C_α torsional energy is a statistical potential function based on the probability distribution of torsional angles or pseudodihedrals defined by virtual C_α - C_α bonds. The function was developed in the Jernigan group and implemented as described by Bahar et al.¹³ The energy is calculated on a per residue basis.

It should be noted here that normalization was only applied in cases in which a nonlinear response was observed between the given field and CRMS. The localized hydrophobic terms, hydrophilic SASA and lipophilic SASA, scale with size and required normalization. However, because of the fact that terms can be both positive and negative, this treatment was not required for fields such as the hypole energy, hydrophobic core potential, and low-resolution electrostatic energy. Most of the proteins included did not contain a large number of cysteines. As a result, normalization was not necessary for the current data sets but may be required for much larger proteins.

Model Building

PLS

Partial Least Squares of Latent Variables (PLS)¹⁹ was the regression method used in the model-building procedure. The different descriptors (x variables) for each fold candidate are related to their CRMS compared with the crystal structure (y variable). PLS regression models have proven useful in similar examples where the different x variables are colinear and noisy.¹⁸ In ordinary PLS, each object (protein fold candidate in this case) is given an equal weight. On occasion, it is necessary to give certain objects a higher, or lower, weight than the rest of the objects in the training set. Weighting of the objects may also be necessary if the distribution of the objects is skewed. For example, if a certain type of object is overrepresented in the data, it will receive a much higher weight than another type with only a few representatives without intervention. The weighting of individual objects can easily be incorporated in the PLS algorithm.³³ Weighting does not affect the properties of PLS; only the resulting model is changed. Therefore, it is still possible to express the PLS model as a

TABLE III. List of the 28 Protein Decoy Sets Used to Generate the ProVal Scoring Function Together With Their CATH Classification, Number of Amino Acids, How They Were Generated, and How Many Decoys Each Set Contains

Name	Class	Architecture	Size	Generation method	No. of decoys	Rank of crystal structure	r^2
1aca	Mainly alpha	Up-down bundle	86	ab initio	500	1	0.02
1c5a	Mainly alpha	Orthogonal bundle	66	ab initio	500	1	0.03
1c5a	Mainly alpha	Orthogonal bundle	66	Rosetta	999	1	0.12
1cc5	Mainly alpha	Orthogonal bundle	83	Rosetta	999	25	0.08
1crn	Alpha beta	2-Layer sandwich	46	Threader	141	1	0.54
1csp	Mainly beta	Barrel	67	Rosetta	998	42	0.22
1ctf	Alpha beta	2-Layer sandwich	68	Rosetta	999	1	0.21
1ctf	Alpha beta	2-Layer sandwich	68	Threader	473	1	0.81
1fc2	Few secondary structures	Irregular	44	Ab initio	500	1	0.06
1gb1	Alpha beta	Roll	56	Rosetta	999	68	0.27
1hdd	Mainly alpha	Orthogonal bundle	57	Ab initio	500	1	0.18
1hoe	Mainly beta	Sandwich roll	74	Threader	554	1	0.81
1igd	Alpha beta		61	Threader	347	2	0.83
1kte	Alpha beta	3-Layer(aba) sandwich	105	Rosetta	998	1	0.31
1lsc	Mainly alpha	Orthogonal bundle	129	Threader	1178	2	0.72
1nkl	Mainly alpha	Orthogonal bundle	78	ab initio	500	1	0.03
1pou	Mainly alpha	Orthogonal bundle	71	ab initio	500	1	0.00
1pou	Mainly alpha	Orthogonal bundle	71	Rosetta	998	97	0.21
1r69	Mainly alpha	Orthogonal bundle	63	ab initio	500	1	0.09
1r69	Mainly alpha	Orthogonal bundle	63	Rosetta	999	23	0.28
1tgi	Mainly beta	Ribbon	112	Threader	1168	1	0.60
1trl	—	—	62	ab initio	500	1	0.26
1xbl	Mainly alpha	Orthogonal bundle	75	ab initio	500	11	0.00
2rat	Alpha beta	Roll	124	Threader	1147	1	0.74
2utg	Mainly alpha	Orthogonal bundle	70	ab initio	500	48	0.19
5icb	Mainly alpha	Orthogonal bundle	75	Rosetta	997	94	0.25
5pti	Few secondary structures	Irregular	58	Rosetta	999	1	0.12
5pti	Few secondary structures	Irregular	58	Threader	1147	1	0.73

linear equation by calculating the coefficients for each descriptor variable. The scoring function can then be expressed as a linear function of all the descriptor variables:

$$score = x_1 * b_1 + x_2 * b_2 + \dots + x_n * b_n \quad (10)$$

where x_i is the descriptor variable and b_i is its corresponding coefficient. This equation is then used for predicting the score for new proteins. It is important to remember that the interpretation of the regression coefficients for a PLS model has to be made with care. Because PLS maximizes the covariance, a large coefficient for a variable may come from the fact that it is highly correlated with the response and/or that it has a large variation in the descriptor matrix. This finding becomes more problematic as greater numbers of PLS components are used in the model.

Weighted PLS

A scheme for PLS that places an arbitrary weight on each object (structure) has been developed. The weights can then be chosen from any of the three different criteria. In each case, the weight of a given object depends on the CRMS (response) value. Objects with a large CRMS were penalized compared with objects with a small CRMS. The three weighting functions are $w = 1/\sqrt{CRMS + 1}$, $w = 1/(CRMS + 1)$, and $w = 1/(CRMS + 1)^2$. The last of these

functions penalized objects with a large CRMS heavily, whereas the first was less than linear. Thus, even if significantly less abundant, lower CRMS structures contributed as much to the final model as the more distant structures.

To be able to use variables with different metrics and variance, each variable was scaled to unite variance before any modeling was performed. The mean was also removed from each variable before the modeling was performed.

Protein Sets

For finding the best scoring function with PLS, we have used 28 different protein sets with a known crystal, or high-resolution NMR, structure and many decoys for each set. A list of the sets used is given in Table III along with the method of decoy generation. The threaded sets were generated with THREADER 2.5³⁴ with a combination of the 2.0- and 2.5-fold libraries. MODELLER4³⁵ was then used to generate all-atom structures from the threaded sequence alignments. The Rosetta-generated decoy sets were obtained from the home page of the Baker laboratory [http://depts.washington.edu/bakerpg]. The initial Rosetta library contained 92 decoy sets for which the native structure was determined by X-ray crystallography or high-resolution NMR. Roughly three fourths of these sets were initially discarded either because of small protein size, small set size, poor sampling of structures ≤ 5 Å from

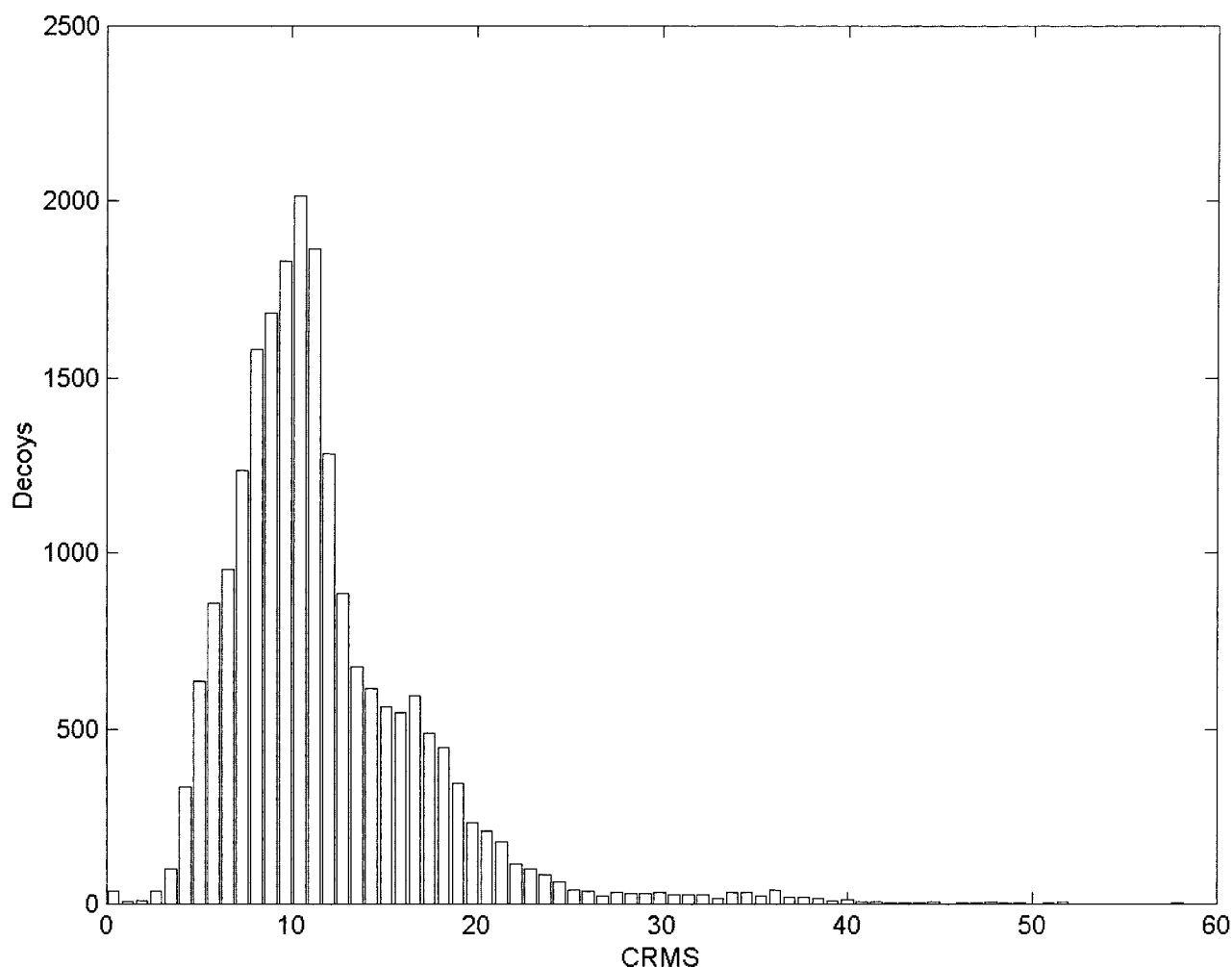


Fig. 2. Histogram for all decoys and crystal structures of the 28 different decoy sets. There are a total of 21,168 decoy folds, but only 213 decoys with a CRMS < 4 Å.

the native structure or very uneven CRMS distributions in the higher CRMS bins. Six of the remaining sets were further discarded because of missing residues from chain breaks, disordered regions of the crystal structure, or relatively large segments missing at the ends of the decoy structures. The final 10 Rosetta decoy sets were selected for their CRMS distributions. The ab initio generated decoys were the fisa set³⁶ obtained from the Stanford web site [<http://dd.stanford.edu/download.shtml>].

RESULTS

We have built and evaluated the ProVal scoring function with 28 decoy sets of structures generated for 22 different proteins. The list of proteins and methods used for creating the decoy sets associated with them are shown in Table III. Figure 2 contains a histogram of CRMS for decoys versus target for all sets used in the model-building procedure. A total of 21,168 structures are present, and it can be seen that most of the structures lie between 8 and 20 Å in CRMS. Relatively few of the structures (213) have a CRMS

lower than 4 Å. This type of distribution is not atypical for candidate structures generated by any of the methods used here, or in general. However, this is problematic for the derivation of an empirical scoring function. A distribution of this nature will lead to a model that is optimized for predictions of structures in the region of 8–20 Å CRMS. This is an unfortunate circumstance because the most interesting structures are of low CRMS. Thus, to emphasize the importance of the near-native structures, it is necessary to ensure they are considered equally in the training of the function compared with the far more numerous distant structures. This problem was overcome through the use of a weighted PLS algorithm³³ that is discussed in detail in Materials and Methods.

Model Building

For each protein set, a separate model was calculated by using PLS and the different weighting schemes. The descriptors, referred to as the X block, are constructed with values for the 15 variables describing different prop-

TABLE IV. Scaled and Unscaled Coefficients for the ProVal Scoring Function, Calculated by Using the Average Coefficients From all 28 Separate PLS Models for Each Decoy Set

Variable name	Scaled Coefficient	Unscaled Coefficient
Packing	0.298	0.008
Hydropole energy	0.285	1.749
Hydro vector score	0.200	1.114
HBL score	0.517	3.445
Ring-ring	0.426	2.760
Ring/non-ring	0.297	2.173
Non-ring/non-ring	0.317	2.470
Electrostatic energy	0.510	3.621
SASA	0.279	1.806
HSASA	0.056	0.391
LSASA	-0.032	-0.221
Disulfide bridges count	-0.796	-1.623
CysMDF	0.504	0.370
Contact order	-0.415	-2.465
C _α torsional energy	0.102	0.660

erties of the candidate structure. These variables are listed in Table IV and are defined in Materials and Methods. The response is the CRMS value of the structure compared to the crystal structure for each protein. Before modeling was attempted, each variable was centered columnwise and scaled to unit variance, ensuring that all the variables obtained equal weight. Otherwise, variables with a large variance influence the model more than variables of small variance. This is true even when an identical correlation to the response exists. For each set of decoy structures, four different PLS models were calculated, three weighted and one ordinary unweighted model. The results indicated that the optimal model was created by using the weighting function $\omega = 1/(\text{CRMS} + 1)^2$, which puts a large weight on the crystal structure and other low CRMS structures. The results presented here were generated from this optimum model.

Because the goal of this work was to create a generalized function capable of scoring new proteins that are unique from the 28 sets presented here, a scoring function that focused on more general aspects of protein stability was developed. This function was defined as the mean of the regression coefficients from the 28 candidate sets listed in Table IV. There are, of course, differences in the importance of specific variables with respect to the CRMS prediction for each class of protein. However, general tendencies, such as good packing, electrostatics, and hydrophobic core potential, appear to be well preserved. A generalized scoring function is typically more robust because extreme values for a specific protein set are averaged out against the entire set.

The averaged coefficients for the ProVal function are given in Table IV. The scaled coefficients give a relative measure of importance for each variable in predicting CRMS. The unscaled coefficients are directly applied to the calculated variables to obtain a predicted score for a given protein structure. The largest scaled coefficient,

disulfide bridge count, and the fourth largest scaled coefficient, cysteine minimum distance function, are both based on the premise that the presence of multiple cysteines typically indicates disulfide bridge formation. The relatively large contribution of these two terms implies a bias toward predictions of extracellular proteins. The low-resolution electrostatic energy, together with HBL score, also has a large positive correlation. It is important to note here that this function was built against the CRMS of decoy proteins as calculated versus the crystal structure. Therefore, a more negative value indicates a better structure. All results presented for a given set of candidate structures are generated with that set excluded in the calculation of the regression coefficients. When two protein decoy sets, generated by different methods, are present for a given sequence, both are left out of the training procedure and predicted. Thus, all values reported are generated from a leave-one-out cross-validation of decoy sets and are a prediction of each fold set based on the remaining 27, or 26 when a second decoy set is present, candidate sets.

Identifying the Native Structure

The first objective is to determine the accuracy of the function in selecting the native structure that we approximate through use of the crystal structure when present in a set of decoys. For 18 of the protein sets (~64%), the crystal structure ranks first. In 24 sets (~86% of the cases), including the previous 18, the crystal structure ranked in the top 5%. Finally, the crystal structure was ranked in the top 10% in all 28 cases. It is interesting to note that 6 of the 11 decoy sets for which the crystal structure did not rank first were generated by Rosetta.^{5,36} This finding may be due to the fact that Rosetta sets contained a greater percentage of structurally sound decoys (smaller CRMS); thus, it was more difficult to differentiate the crystal structure from more distant ones. The implications for this is debated further in the Discussion. For the sets generated by threading, the crystal structure scored best in all cases except for one where it ranked second. Successful selection of the crystal structure did not appear to be biased by fold classification.

Identifying Near-Native Structures

On determining that the scoring function has the ability to select the crystal structure as the native structure in a substantial percentage of the cases tested, the ability to achieve the second objective was evaluated. The second objective was to show a preference for near-native structures over non-near-native structures with some level of consistency. For reasons discussed earlier, this is by far the most difficult of the objectives stated. However, from the distribution of candidate structures, it is clear that this problem must be solved. Candidate structures within 3 Å of the crystal structure are very rare. Often, the best structure available in the prediction set is in the 4–6 Å CRMS range. Thus, for a scoring function to be truly useful, it must be able to reproducibly place structures in

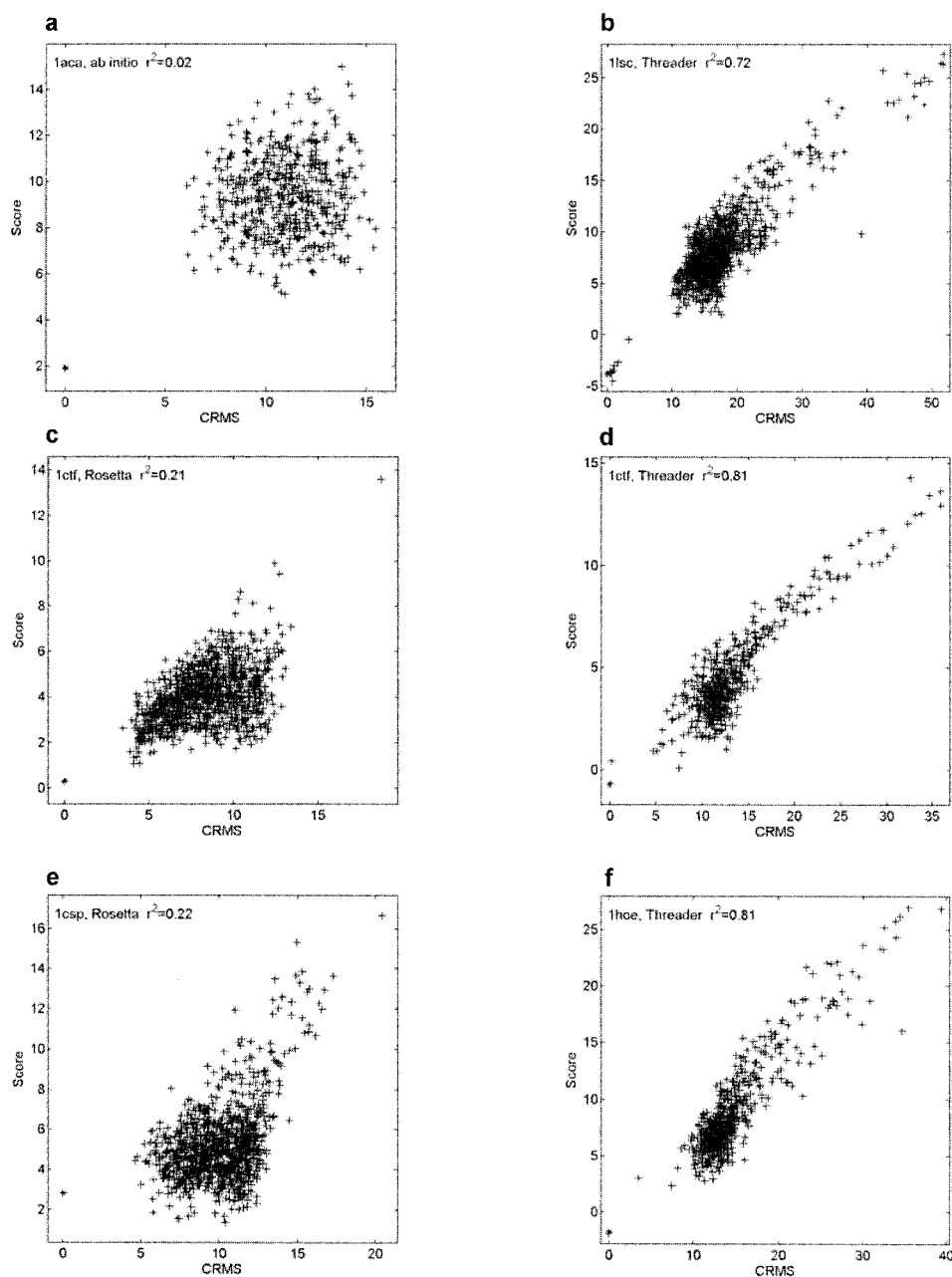


Fig. 3. Plot of score value versus CRMS for the decoy sets **a–f**: 1aca, 1lsc, 1ctf, 1ctf, 1csp, and 1hoe. Given in the figure are the PDB entry name, generation method, and the correlation coefficient. The score of the crystal structure is indicated by an asterisk that is found at 0 CRMS by definition.

this range at or near the top of the list of candidates. Correlation between CRMS and predicted score was used to evaluate the ability of the scoring function to favor near-native structures. The fit between these two values should be roughly linear. Table III lists the r^2 correlation coefficient for all of the proteins sets. The value ranges from 0 to 0.81, with 1.0 being ideal. However, we have found this measure to be somewhat misleading as an indicator of the discriminate power of the scoring function. This finding is due to the fact that, for a function of this nature, discriminate power is needed only at the low end of

the spectrum. Thus, we want to be able to isolate near-native structures from distant structures. It is not necessary to distinguish distant structures from very distant ones. Therefore, the scatter plots must be used as a complement to truly evaluate a given set of decoys. Figure 3 shows a range of fits between score and CRMS for all of the candidate sets. Discussed below are six different sets that cover a variety of performances, classes of fold, and structure generation methods. Because this is a general model, the score value is not absolute and must, therefore, be assessed in a relative way.

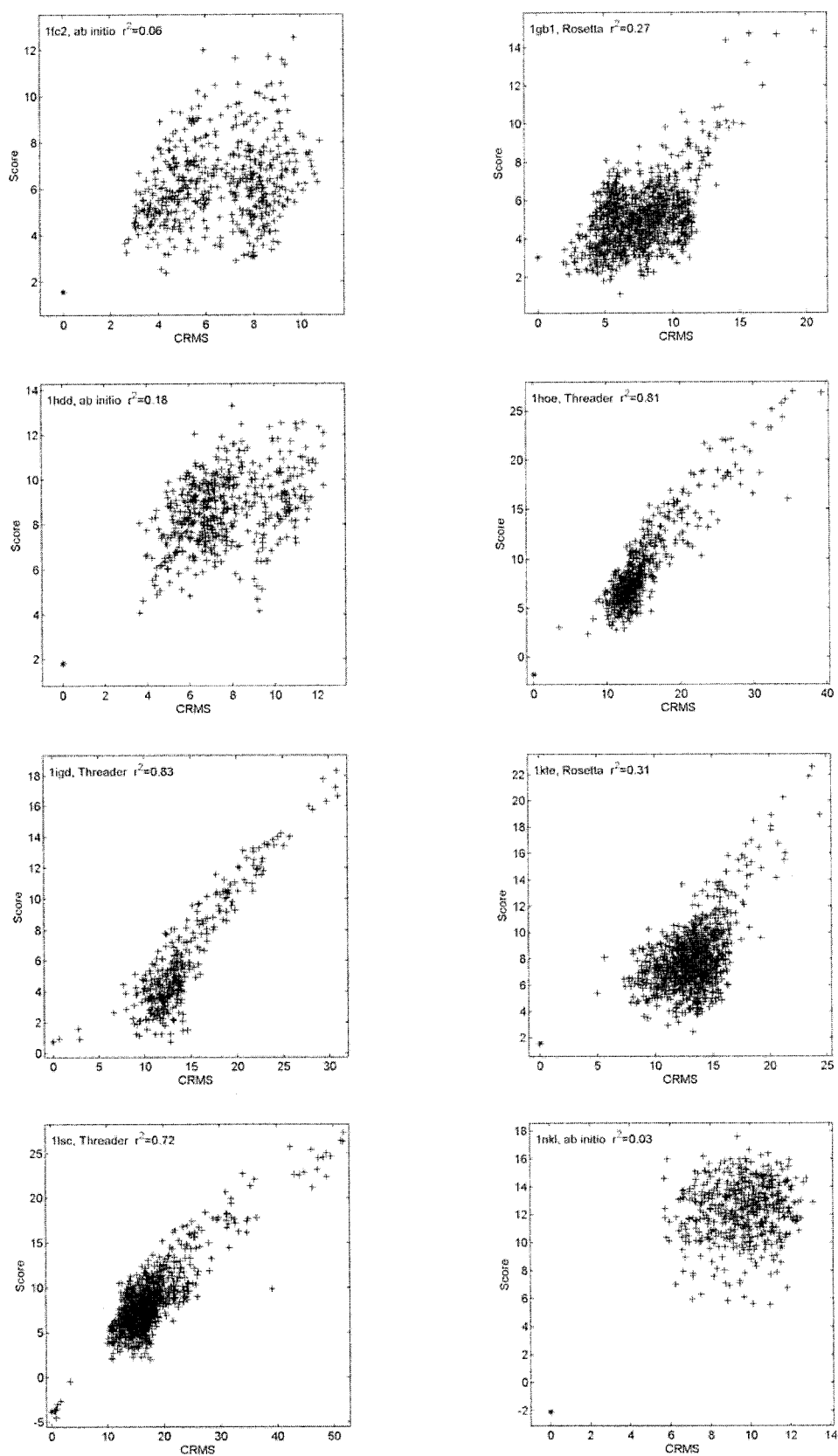


Fig. 4. Plot of score value versus CRMS for the decoy sets: 1fc2, 1gb1, 1hdd, 1hoe, 1igd, 1kle, 1lsc, and 1nkl. Given in the figure are PDB entry name, generation method, and the correlation coefficient. The score of the crystal structure is indicated by an asterisk that is found at 0 CRMS by definition.

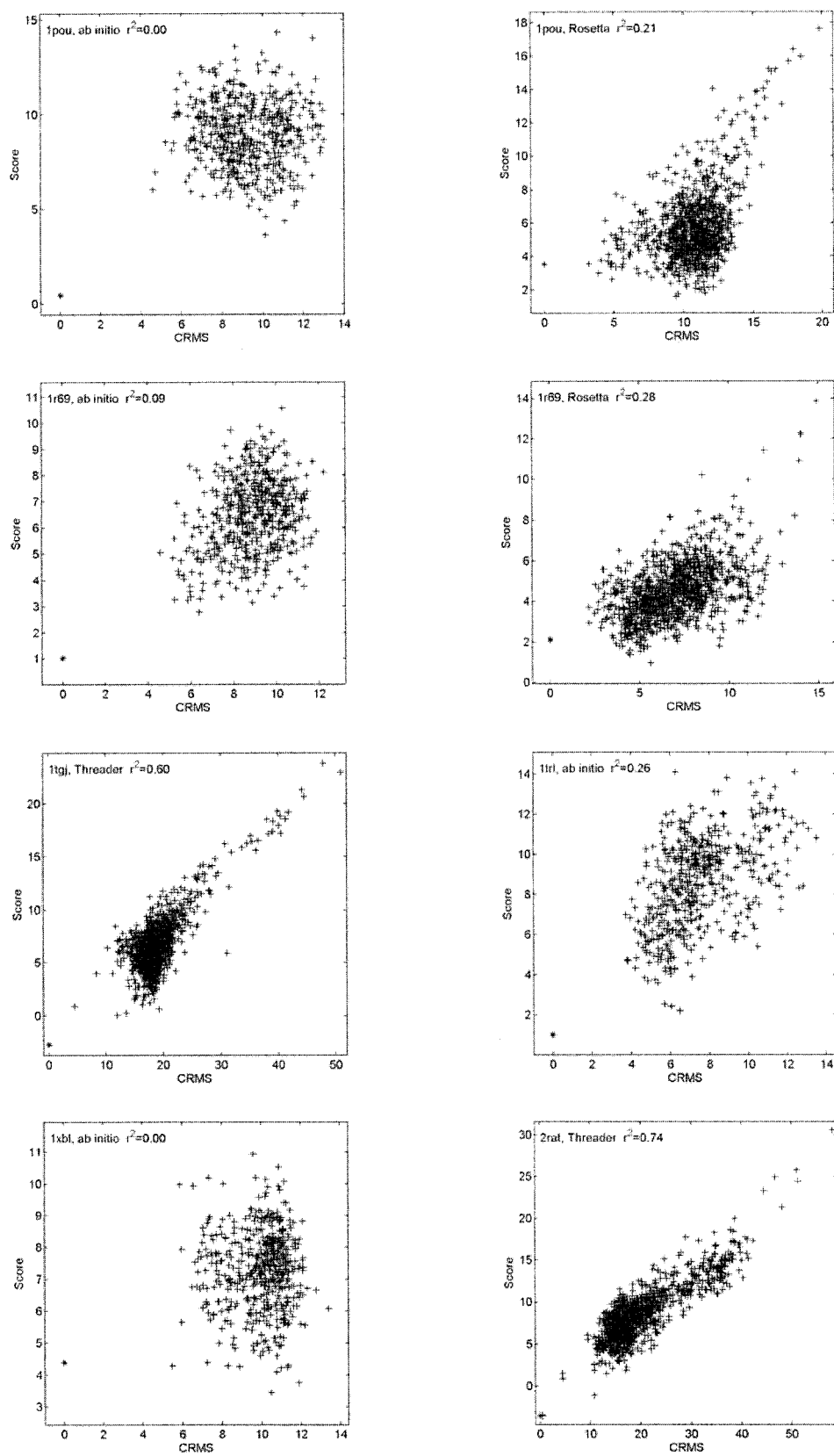


Fig. 5. Plot of score value versus CRMS for the decoy sets: 1pou (ab initio), 1pou (Rosetta), 1r69 (ab initio), 1r69 (Rosetta), 1tjg (Threader), 1trl, 1xbl, and 2rat. Given in the figure are PDB entry name, generation method, and the correlation coefficient. The score of the crystal structure is indicated by an asterisk that is found at 0 CRMS by definition.

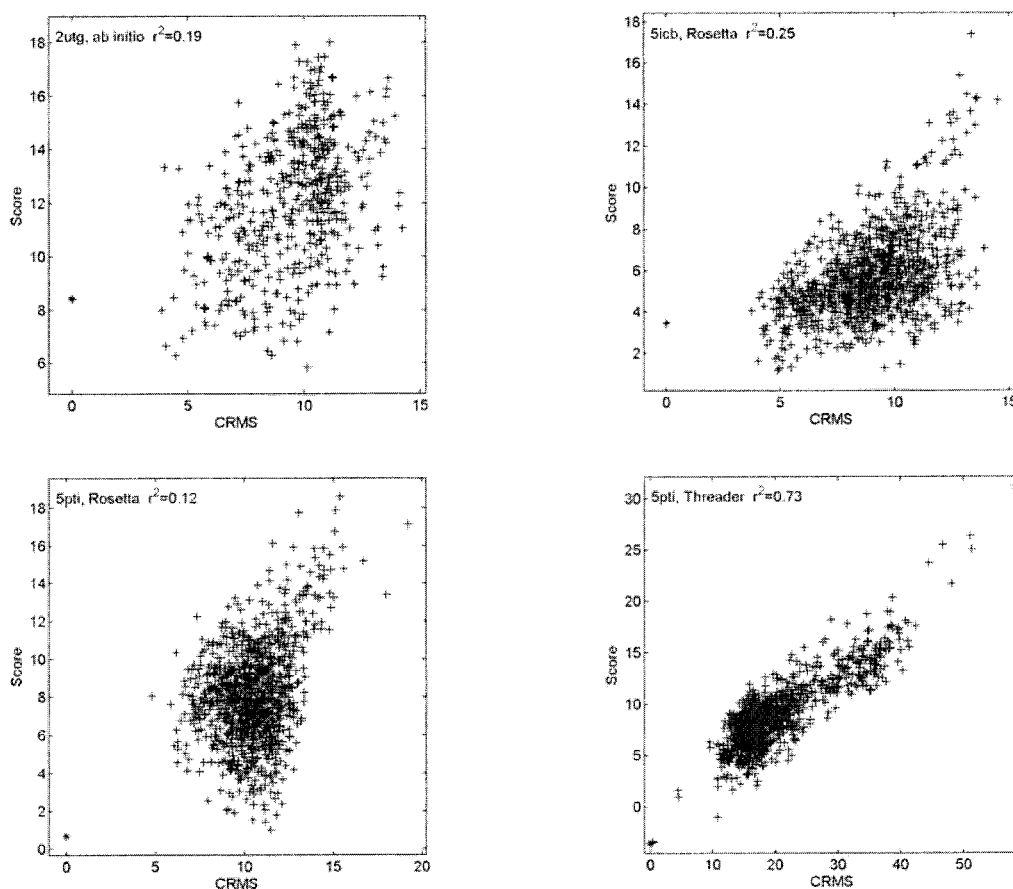


Fig. 6. Plot of score value versus CRMS for the decoy sets: 2utg, 5icb, 5pti (Rosetta), and 5pti (Threader). Given in the figure are PDB entry name, generation method, and the correlation coefficient. The score of the crystal structure is indicated by an asterisk that is found at 0 CRMS by definition.

1aca

The data for protein 1aca are an example of a poor correlation between CRMS and score ($r^2 = 0.02$) as can be seen in Figure 3(a). However, although no discriminate difference appears between 5 and 15 Å CRMS decoys, the crystal structure scores exceptionally well. It is not only selected as the top ranking structure, the difference in score between the crystal and the second-ranked structure is considerable. This finding is an example of a set where there are no candidates < 5 Å from the crystal structure. It is not likely that the difficulty in scoring this particular set lies in the distribution of the decoys, because these structures were not included when building the model used to score it. The problem may instead be attributed to the fact that 1aca is a small four-helical bundle protein. The helices may be arranged in multiple ways and still generate a well-packed structure with a reasonable hydrophobic core. However, there is clearly an optimum arrangement as shown by the relative score of the crystal structure. Similar behavior was observed in other decoy sets for small, mainly α -helical proteins such as 1c5a, 1cc5, 1nkl, 1pou, 1xbl, and 2utg. This problem of multiple comparable candidate folds was also noticed by Simons et al.⁵

1lsc

1lsc is a mainly α -helical protein with some β -strands. Although not a large protein, it is considerably larger than the four-helix bundles just discussed. A far better correlation between score and CRMS ($r^2 = 0.72$) was obtained for this decoy set, as can be seen from Figure 3(b). In this case, decoys were generated via threading, and the CRMS ranges from 0.5 to 52 Å. Eight of the decoys are within 4 Å of the crystal structure. All of these near-native structures obtain a low score. However, the high scores generated by the distant structures indicate that the procedure of down weighting high CRMS candidates did not inhibit the ability of the scoring function to predict them as such. 1hdd ($r^2 = 0.18$) and 5icb ($r^2 = 0.25$) are smaller mainly α -helical proteins that also exhibit a good correlation.

1ctf

For 1ctf, both Rosetta and threading-generated decoy sets were evaluated. In this case, both sets of decoys produced good correlations between CRMS and score as is evident in Figures 3(c) and (d). The threaded decoy set produced an exceptional correlation ($r^2 = 0.81$). 1ctf con-

sists of three helices and three β -strands and is considerably more restricted in plausible folding arrangements than a protein like 1aca because of the fact that none of the coil regions are particularly lengthy. The distribution of CRMS for the two 1ctf decoy sets is somewhat different given that the threaded set contains structures $> 15 \text{ \AA}$ from the crystal. The threaded set also has a near-native structure, which scores very well, ranking third overall, with the crystal structure being ranked first. The other α/β class proteins (1crn, 1gb1, 1igd, 1kte, and 2rat) exhibit a relatively good correlation with r^2 ranging from 0.27 to 0.83.

1csp

1csp is one of the few proteins tested that is a mainly β -protein. The decoy set for this protein was generated with Rosetta. This is an example of the correlation value ($r^2 = 0.22$) being somewhat misleading. As can be seen in Figure 3(e), the correlation is very good for high CRMS structures, but there is very little discriminate power for those of lower range. The crystal structure scores rather poorly in this set as well. 1tgj is one of only two other mainly β -proteins presented here. The major difference is that the crystal structure did rank first for 1tgj. Whether poor discriminate power is typical for a mainly β -structures or not is difficult to ascertain with such a small sample. However, the decoys were generated from different methods, so if there was a commonality responsible for this observation, it would likely be structurally related.

1hoe

1hoe is the third and final mainly β -protein evaluated. There is a major structural difference between 1hoe and the two other mainly β -proteins, 1csp and 1tgj. 1hoe is considerably more compact and shows the significantly improved correlation seen in Figure 3(f). This decoy set contains one structure with a CRMS of $< 4 \text{ \AA}$ from the crystal structure. This near-native structure ranks fourth overall with the crystal structure ranked first. To theorize that the improved scoring for this protein was somehow related to the compactness of the structure is enticing but probably premature.

DISCUSSION

A considerable number of scoring functions have been presented in the literature to evaluate candidate structures for a given protein sequence. Many such functions have been shown to work well in certain circumstances and poorly in others. The scoring function presented here is the linear summation of several individual functions. Each term describes some aspect of the relative distance of a candidate structure to the native one. Through the use of PLS, the terms were combined into a generalized function in which the coefficients were determined on the basis of a broad training set of decoy structures. Some of the terms are similar in nature, whereas others are quite disparate. One of the primary reasons for choosing PLS as the basic regression scheme is the ability to handle both correlated and noisy variables. In turn, the application of a weighted

PLS model helped circumvent the issue of approximately normal distributions of CRMS in the decoy sets.

With respect to the first objective, ProVal was fairly successful in identifying the native structure from a set of decoys with an overall success rate approaching 65%. Although not ranked first, the native structure was placed in the top 10% for the remaining decoy sets. Clearly, in these cases, there is considerable signal being detected, even if not optimally. Although considerably more difficult to characterize quantitatively, there appears to be an ability to rank near-native structures more favorably for a significant number of the decoy sets. Some of the decoy sets presented here show that the ability, or lack thereof, to rank the native structure first is not always a direct indicator of the ability to favor near-native structures in general. There are examples that show different combinations of accurate native structure selection and general near-native preference.^{37,38}

With respect to method of generation, ProVal appears to perform best on candidate sets created with threading methods. These sets yield high accuracy in selection of the crystal structures as well as good correlation to CRMS. This is not surprising because many of the threaded decoys are very different from the crystal structure and tend to be poorly packed. It is odd that although the scoring function did not perform particularly well in selecting the crystal structure from Rosetta sets, the correlation of score to CRMS was quite good. This finding is possibly due to the large number of low CRMS decoys present and the inability to distinguish structures that are very close to each other in CMRS due to the inherent low resolution of the model used. The ab initio generated sets show the poorest correlation overall. In some cases, only the crystal structure obtains a low score from the list of near-native structures that are present.

The fold class appears to play a significant role for identifying near-native structures, even though selection of the native structure does not exhibit the same dependence. Mainly α -proteins generated the largest number of misses in identifying the crystal structure; however, they are by far the most abundant class (14 of 28). That stated, others have observed mainly α -proteins are a difficult class for consistent selection of the native structure,⁵ which may be because the α -helices can be arranged in several ways without presenting anomalous values for any of the terms included in the scoring function and will require all-atom modeling³⁹ and more sophisticated analysis to discriminate. Proteins that are not that well packed in the native state show a similar problem, which is likely due to the need for a better description of the packing. The packing estimate used in this work tends to be biased toward compact structures.

CONCLUSION

The two objectives in the development of a global scoring function for low-resolution models of proteins relatively independent of prediction method have been achieved with ProVal. In particular, the use of weighted PLS to deal with multiple cross-correlated variables and differences in obser-

vations has been demonstrated. Validation of any scoring function obviously depends on the choice of training and test sets. By using most, if not all, of the decoy sets available during its development, one would hope that inclusion of methodological bias was minimized, but the training set and tests sets were not truly independent in this study. On the basis of the results obtained, ProVal should be used as a primary filter to significantly reduce the number of predicted structures (~90%) that are further processed by all-atom modeling regardless of the prediction method used. The ability of ProVal to help predict the structure of new classes of proteins and a thorough evaluation of its use will require its intensive testing by the community involved in protein structure prediction.

REFERENCES

- Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. *Proteins* 2000;40:71–85.
- Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol* 2000;10:139–145.
- 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.
- Eyrich VA, Standley DM, Friesner RA. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J Mol Biol* 1999;288:725–742.
- Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 1999;34:82–95.
- Moult J. Comparison of database potentials and molecular mechanics force fields. *Curr Opin Struct Biol* 1997;7:194–199.
- Vajda S, Sippl M, Novotny J. Empirical potentials and functions for protein folding and binding. *Curr Opin Struct Biol* 1997;7:222–228.
- Sippl MJ. Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 1995;5:229–235.
- Jones DT, Thornton JM. Potential energy functions for threading. *Curr Opin Struct Biol* 1996;6:210–216.
- Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol* 1996;6:195–209.
- Hao MH, Scheraga HA. Designing potential energy functions for protein folding. *Curr Opin Struct Biol* 1999;9:184–188.
- Roux B, Simonson T. Implicit solvent models. *Biophys Chem* 1999;78:1–20.
- Bahar I, Kaplan M, Jernigan RL. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. *Proteins* 1997;29:292–308.
- Eyrich VA, Standley DM, Felts AK, Friesner RA. Protein tertiary structure prediction using a branch and bound algorithm. *Proteins* 1999;35:41–57.
- Huang ES, Subbiah S, Levitt M. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J Mol Biol* 1995;252:709–720.
- 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.
- 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.
- Head RD, Smythe ML, Oprea TI, Waller CL, Green SM, Marshall GR. VALIDATE: a new method for the receptor-based prediction of binding affinities of novel ligands. *J Am Chem Soc* 1996;118:3959–3969.
- Wold S, Ruhe A, Wold H, Dunn WJ III. The collinearity problem in linear regression. The partial least squares approach to generalized inverses. *SIAM J Sci Stat Comput* 1984;5:735–743.
- Hobohm U, Sander C. Enlarged representative set of protein structures. *Protein Sci* 1994;3:522–524.
- Abola EE, Sussman JL, Prilusky J, Manning NO. Protein data bank archives of three-dimensional macromolecular structures. *Macromolecular crystallography, part B. Methods Enzymol* 1997;277:556–571.
- Sussman JL, Lin DW, Jiang JS, Manning NO, Prilusky J, Ritter O, Abola EE. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 1998;54:1078–1084.
- 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;247:536–540.
- de Araujo AFP. Folding protein models with a simple hydrophobic energy function: The fundamental importance of monomer inside/outside segregation. *Proc Natl Acad Sci USA* 1999;96:12482–12487.
- Wireko FC, Kellogg GE, Abraham DJ. Allosteric modifiers of hemoglobin. 2. Crystallographically determined binding sites and hydrophobic binding/interaction analysis of novel hemoglobin oxygen effects. *J Med Chem* 1991;34:758–767.
- Kellog GE, Semus SF, Abraham DJ. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J Comput Aided Mol Des* 1991;5:545–552.
- Breuer M. ESTAR: electrostatic properties research package. W-8033 Martinsried, Germany: Max Planck Institute for Biochemistry.
- Dudek MJ, Ramnarayan K, Ponder JW. Protein structure prediction using a combination of sequence homology and global energy minimization. II. Energy functions. *J Comput Chem* 1998;19:548–573.
- Kurochkina N, Lee B. Hydrophobic potential by pairwise surface-area sum. *Protein Eng* 1995;8:437–442.
- Vonfreyberg B, Braun W. Minimization of empirical energy functions in proteins including hydrophobic surface-area effects. *J Comput Chem* 1993;14:510–521.
- Le Grand SM, Merz KM Jr. Rapid approximation to molecular surface area via the use of boolean logic and look-up tables. *J Comput Chem* 1993;14:349–352.
- Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 1998;277:985–994.
- Wold S. Exponentially weighted moving principal components—analysis and projections to latent structures. *Chemometrics Intelligent Lab Systems* 1994;23:149–161.
- Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. *Nature* 1992;358:86–89.
- Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* 1993;234:779–815.
- 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.
- Galaktionov S, Nikiforovich GV, Marshall GR. Ab initio modeling of small, medium, and large loops in proteins. *Biopolymers* 2001;60:153–168.
- Galaktionov SG, Marshall GR. Properties of intraglobular contacts in proteins: an approach to prediction of tertiary structure. *Proc 27th Annual Hawaii International Conference on System Sciences. Vol. 5: Biotechnology Computing. Wailea, HW, USA, January 1994. IEEE Computer Society Press, 1994.*
- Pappu RV, Marshall GR, Ponder JW. A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat Struct Biol* 1999;6:50–55.