Computational Methods for Protein Folding: Advances in Chemical Physics, Volume 120.
Edited by Richard A. Friesner. Series Editors: I. Prigogine and Stuart A. Rice.

ISBNs: 0-471-20955-4 (Hardback); 0-471-22442-1 (Electronic)

DETECTING NATIVE PROTEIN FOLDS AMONG LARGE DECOY SETS WITH THE OPLS ALL-ATOM POTENTIAL AND THE SURFACE GENERALIZED BORN SOLVENT MODEL

ANDERS WALLQVIST, EMILIO GALLICCHIO, ANTHONY K. FELTS, AND RONALD M. LEVY

Department of Chemistry, Rutgers University, Wright-Rieman Laboratories, Piscataway, NJ, U.S.A.

CONTENTS

I. INTRODUCTION

The ability to distinguish native protein conformations from misfolded ones is a problem of fundamental importance in the development of methods designed to predict protein structure. To this end, several empirical functions for scoring protein conformations have been proposed. [1–6]. Some of these empirical scoring functions implement knowledge-based statistical potentials that are ''trained'' to recognize native conformations. Knowledge-based potentials are best suited for ''threading'' applications where the best conformation of a protein is selected from a database of known protein conformations. Scoring functions applicable to *ab initio* folding studies, which require differentiable potentials and the inclusion of excluded volume terms, have also been developed. These are based on combinations of knowledge-based potentials and reduced atomic models sometimes augmented by simplified solvation models based on hydrophobic or hydrophilic exposure [7].

Physics-based all-atom molecular mechanics force fields have not been generally considered practical for fold detection because they are parameterized on small molecule data rather than on proteins directly; the level of atomic detail contained in these models is considered poorly matched to the fold detection problem with respect to both accuracy and computational cost. Recent studies have shown, however, that a scoring function based on the potential energy from an all-atom molecular mechanics force field can recognize native protein conformations among a set of decoys as well as the best available knowledge-based scoring functions [6].

The use of an all-atom force-field minimizes the assumptions that are inherent in an empirical scoring function; and, as will be shown, the inclusion of more refined solvation models enhances our ability to discriminate native folds. An additional value of the all-atom potential lies in its suitability for modeling proteins at higher resolution. This is an important feature for applications in structure–function relation studies such as homology modeling, drug design, and protein–protein recognition.

Although all-atom force fields allow for explicit simulations of solvent, the cost required to appropriately sample solvent configurations rapidly becomes prohibitive. Simplified solvation models are more computationally efficient while preserving a reasonably accurate representation of the interactions between the protein and the water solvent. Although no continuum model can wholly account for the explicit inclusion of solvation [8,9], free energies of solvation of small molecules have been obtained accurately to within a fraction of a kcal/mol relative to experiments using these methods [10–15].

Solvation effects have been included using a variety of simple models [16–23]. These models have been based on exposed surface area, dielectric continuum methods, and screened or modified Coulomb interactions. The validity

of a continuum representation of the solvent based on the Poisson–Boltzmann equation has been studied extensively for small and large molecules [24–30]. Continuum solvation models that treat solute and solvent as two dielectric regions with different dielectric constants have been used successfully to account for solute free energies of hydration [11,31–34]. Dielectric models based on the Born model [35] have been developed for which the free energies of hydration are comparable to the predictions of Poisson–Boltzmann and explicit solvent models [36–42].

The inclusion of solvation effects with an all-atom molecular mechanics force field has been shown to be important for the recognition of the native state [16,17,43–45]. Scheraga and co-workers [46,47] used explicit all-atom protein models in conjunction with solvation models based on the molecular exposed surface area. A similar approach by Wang et al.[48,49] showed that inclusion of solvation effects can be successful in discriminating native from non-native structures. Vieth et al. [50] generated structures of the small 33-residue GCN4 leucine zipper proteins using a simplified lattice model; promising structures were then converted to all-atom models and evaluated using a molecular mechanics force field. A hierarchical method of generating large numbers of protein folds was also employed by Monge et al. [20] to select and evaluate structures using the AMBER all-atom force field model [51]. The generalized Born continuum solvent model of Still et al. [37] has been used in this context to represent the aqueous environment. For decoy sets of three different proteins the protocol performed reasonably well in distinguishing the native structure. All-atom models with continuum solvent were also used as the basis for discrimination of non-native states for a small set of 12 deliberately misfolded proteins studied by Vorobjev et al. [52]. In their protocol, conformations for each protein are first sampled from a molecular dynamics trajectory in order to capture micro-states of the protein; this is followed by an evaluation using a dielectric continuum model. Lazaridis and Karplus [22] used the CHARMM19 protein force field together with a Gaussian solvation shell model for the solvation free energy to distinguish deliberately misfolded from native conformations considered on a pairwise basis and in large decoy sets.

Given the complexity of the protein potential surface, it is virtually impossible to consistently find the global minimum starting from an arbitrary point on the surface. Instead, tests have been designed whereby the scoring function is ''challenged'' to find the native conformation among an ensemble of conformations, most of which are compact but non-native. Many empirical energy functions have been used to identify the correct native structure among a collection of known protein structures using threading techniques [1,53–58]. Scoring functions are also used to identify native-like conformations from a large set containing native and decoy non-native conformers [22,59–63]. Due to the large ensemble of conformations available, the use of large decoy sets to

evaluate scoring functions is a more demanding test than threading and is well-suited for the evaluation of scoring functions based on an all-atom force field.

In this work we show that the all atom (OPLS-AA) force field for proteins [64] together with a surface integral formulation of the generalized Born model (SGB) [40,42] is capable of discriminating between native and non-native folds among large sets of compact decoy structures. Validation of the scoring protocol is performed on a large database of well-packed misfolded and near-native protein conformations generated by an algorithm designed to cover exhaustively the relevant parts of conformational space [65,60,66]. The inclusion of near-native decoys in these sets is important in determining whether the scoring function is well-behaved in the vicinity of an idealized native conformation, because it is unlikely that any *ab initio* method of generating conformations will generate that state exactly. In any case, the native state actually represents an ensemble of closely related conformations.

Two additional decoy data sets of misfolded proteins [17] and of predicted protein structures from the Critical Assessment of Techniques for Protein Structure Prediction (CASP) [67] are also used to illustrate the method and its utility. Individual components of the energy perform worse than the total energy; for example, for the bulk of the well-packed decoys, the van der Waals energy provides very little information about structural similarity between a well-packed non-native structure and the native state. It is also shown that some aspects of the SGB model results can be mimicked by a screened electrostatic energy, although the SGB approximation provides a better discriminatory measure between non-native and native states.

II. METHODS

A. Details of the Calculations

The energy of each protein structure investigated was calculated using the OPLS-AA/SGB force field implemented in the IMPACT modeling program (Schrödinger, Inc.) [68]. Initial structures were first minimized in order to remove any artifacts that result from the coordinates being generated with a different energy function; only minimized energies are reported here. All non-native coordinates were taken from independently generated data sets as described below; native protein coordinates were obtained from the Protein Data Bank (PDB) [69]. The force field employed in the calculation of the atomic interactions was the OPLS all-atom force field [64], including parameters for all intramolecular degrees of freedom. The surface formulation of the generalized Born model [37,39] (SGB) as coded in IMPACT was used to estimate the solvation energy [40,70].

The total energy for a protein in vacuum is given by

$$U_{\text{tot}}^{\text{vac}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{torsion}} + U_{\text{Coulomb}} + U_{\text{vdW}} \tag{1}$$

where the first three terms refer to intramolecular interactions arising from the connectivity of the molecule, and the last terms reflect nonlocal interactions within the protein. The van der Waals energy, U_{vdW}, is modeled by the standard 6-12 Lennard-Jones interaction. The energy of the protein in water calculated according to the SGB continuum solvent model is

$$U_{\text{tot}}^{\text{con}} = U_{\text{tot}}^{\text{vac}} + U_{\text{SGB}} + U_{\text{cav}} \tag{2}$$

where U_{SGB} denotes the electrostatic contribution to the solvation energy calculated using the SGB method, and the cavity term U_{cav} is taken as γA where A is the accessible surface area of the molecule and $\gamma = 5$ cal/(Å^2 mole) [40].

The SGB model is the surface implementation [40,42] of the generalized Born model [37]. The generalized Born equation

$$U_{\text{SGB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{w}}}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}(r_{ij})} \tag{3}$$

(where q_i is the charge of atom i, and r_{ij} is the distance between atoms i and j) gives the electrostatic component of the free energy of transfer of a molecule with interior dielectric ϵ_{in} from vacuum to a continuum medium of dielectric constant ϵ_{w}, by interpolating between the two extreme cases that can be solved analytically: one in which the atoms are infinitely separated and the other in which the atoms are completely overlapped. The interpolation function f_{ij} in Eq. (3) is defined as

$$f_{ij} = [r_{ij}^2 + \alpha_i \alpha_j \exp(-r_{ij}^2/4\alpha_i\alpha_j)]^{1/2} \tag{4}$$

where α_i is the Born radius of atom i defined as the effective radius that reproduces through the Born equation

$$U_{\text{single}}^{i} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{w}}}\right)\frac{q_i^2}{\alpha_i} \tag{5}$$

the electrostatic free energy, U_{single}^{i}, of the molecule when only the charge of atom i is turned on. The SGB method estimates U_{single}^{i} by integrating the interaction between atom i and the charge induced on the molecular surface by the Coulomb field of this atom:

$$U_{\text{single}}^{i} = -\frac{1}{8\pi}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{w}}}\right)\int_S \frac{(\mathbf{r} - \mathbf{r}_i)}{|\mathbf{r} - \mathbf{r}_i|^4}\cdot \mathbf{n}(\mathbf{r})\, d^2\mathbf{r} \tag{6}$$

The SGB method has been shown to compare well with the exact solution of the Poisson–Boltzmann (PB) equation. The SGB implementation used in this work includes further correction terms that bring the SGB reaction field energy even closer in agreement with exact PB results [40].

To help assess the ability of the energy function to discriminate between non-native and native protein conformations, the energy gaps between the decoy conformations and the native are evaluated:

$$\Delta U = U_{\text{tot}}^{\text{decoy}} - U_{\text{tot}}^{\text{native}} \tag{7}$$

Energy gaps of individual energy terms have also been examined [see Eqs. (1) and (2)]. Unless explicitly noted, all results presented below were performed without energy cutoffs; that is, all possible non-bonded interactions are included in the total energy. The structural similarity between two protein conformations is expressed as a root mean square deviation (RMSD) between the best overlap of the alpha-carbon (C_α) atoms of the two conformations.

B. Data Sets of Decoys

Although we are probing various energy functions for their ability to differentiate between native and non-native structures, none of the coordinate sets were originally generated by these functions. The vastness of the conformational space and the complexity of an all-atom potential energy function effectively hinders the full sampling of the appropriate degrees of freedom. Scoring conformations with the OPLS/SGB potential may be considered as a last step in the process of generating protein folds; that is, only at the end would it be appropriate to spend the time and effort to evaluate a complex all-atom potential energy function. For this study we focus on existing decoy data sets as our conformational space. These data sets have proven to be highly nontrivial to score correctly.

The first data set contains structure decoys for seven small proteins compiled by Park and Levitt [60]. The protein structures were generated by exhaustively enumerating the backbone rotamers states of 10 selected residues in each protein using an off-lattice model with four discrete dihedral angle states per rotatable bond. From this data set, containing hundreds of thousand of conformations, the authors selected for further evaluation only compact structures that scored well using a variety of scoring functions as well as those having a reasonable RMSD from the native [60]. The coordinates, available on the internet (http://dd.stanford.edu), are all-atom models built from the C_α atoms with the program SEGMOD [71]. No further refinement of these coordinates was done except for minimizing the structures using our energy function (see Eqs. 1 and 2). The decoy data sets are summarized in Table I and encompass a range of small proteins from 54 to 75 residues with varying topological folds. The number of

TABLE I
The Sequence Length, N_{res}, the Number of Decoys, N_{decoy}, and Total Charge of the Seven Proteins of the Park and Levitt Set [60]

PDB Name	N_{res}	N_{decoy}	q (e)
1ctf	68	630	−2
1r69	63	675	+4
1sn3	65	660	+1
2cro	65	674	+6
3icb	75	653	−7
4pti	58	687	+6
4rxn	54	677	−12

decoys in these sets ranged from 630 for 1ctf (the carboxy-terminal domain of L7/L12 50s ribosomal protein from *Escherichia coli*) to 687 for 4pti (bovine pancreatic trypsin inhibitor).

An extended data set for the calcium-binding protein calbindin D9K from bovine intestine (4icb) was also investigated using 2000 best-scoring conformations constructed using an *ab initio* procedure [72]. These structures were generated from an exhaustive enumeration on a tetrahedral lattice [73,74] and selected using a combination of scoring functions.

A third data set consists of 26 misfolded protein coordinates constructed by threading the original sequence on to non-native folds with the same number of residues [17]. These structures were generated by swapping main chains between folds and placing the side chains using an annealing protocol. From this data set we selected 25 misfolded structures with continuous backbone coordinates for analysis. These latter coordinate sets were also taken from the internet site listed above.

A fourth data set derived from the CASP3 [67] targets and model submissions was also investigated. CASP3 is the third experiment run by the Protein Structure Prediction Center at Lawrence Livermore National Laboratory to test how well protein structures can be predicted from amino acid sequence. Results are available on the internet at http://predictioncenter.llnl.gov/casp3/Casp3.html. For our calculations, submitted targets were chosen for which coordinates of the native structure were available from the PDB. For each target, models were chosen which had predictions over all residues given in the PDB file. We selected 11 targets and a total of 167 models, with RMS deviations ranging from 1.3 Å to 22.9 Å. The target structures investigated are given in Table IV.

The energy of each native and model structure was minimized using the full atomic model with and without the SGB dielectric continuum solvation energy term.

III. RESULTS AND DISCUSSION

The problem of differentiating non-native states from native-like states can be expressed as the ability of a scoring function, depending only on the coordinates of each structure, to score the native states better than any other structures. If such a scoring function were used also to generate structures, a further desirable property would be that in the vicinity of the native state the structural similarity to the native state would be a monotonically increasing function of improved scores.

A. Park and Levitt Decoys

Examination of minimized energies for the seven extensive data sets of protein decoys (see Fig. 1) shows that using the OPLS-AA/SGB potential, no decoy scores better than the X-ray structure. The correlation between structural similarity and score is strong only for structures with low RMSD. For RMSD $>$ 4 Å this correlation breaks down. Native-like states appear around 2 Å at low energies, with the bulk of the decoys being in non-native-like conformations with RMSD above 4 Å.

In Table II we report the statistical indicators of the quality of the scoring function. Some of the indicators depend on defining the reference structure as the native X-ray structure. It has been verified that similar results are obtained by selecting any native-like decoy as the reference structure. A global view of the results for the Park and Levitt sets is given in Fig. 2. The fraction, $P(\Delta U)$, of native-like decoys with an energy gap from the native less than ΔU is shown. A decoy conformation with an RMS less than 3 Å is considered native-like. Figure 2 indicates, for example, that structures with an energy gap from the native less than 100 kcal/mol have a $\sim$90% chance of being native-like, whereas a decoy with a +200 kcal/mol energy gap from the native has only a 20% chance of being native-like. For these data sets there are no decoy structures with a total energy, $U_{\text{tot}}^{\text{con}}$, below that of the native state (i.e., energy-minimized X-ray coordinates; see Fig. 1). This suggests that if a fold prediction program can generate protein structures within 100 kcal/mol of the native state, there should be a high ($>$90%) chance of finding native-like states in this data set.

Another measure of the fitness of the scoring functions is to evaluate the RMSD of the lowest-energy structure in each decoy set. The results are summarized in Table II. The RMSD of the lowest-energy decoy range from 0.94 Å to 2.20 Å with an average RMSD of 1.9 Å. These decoys fall within the native-like designation. The average energy deviation from the native energy is $+79.5$ kcal/mol, which represents an average deviation of $+2$% from the native total energy values. As we shall see below, not all scoring functions examined yield decoy energies consistently higher than the native energy.

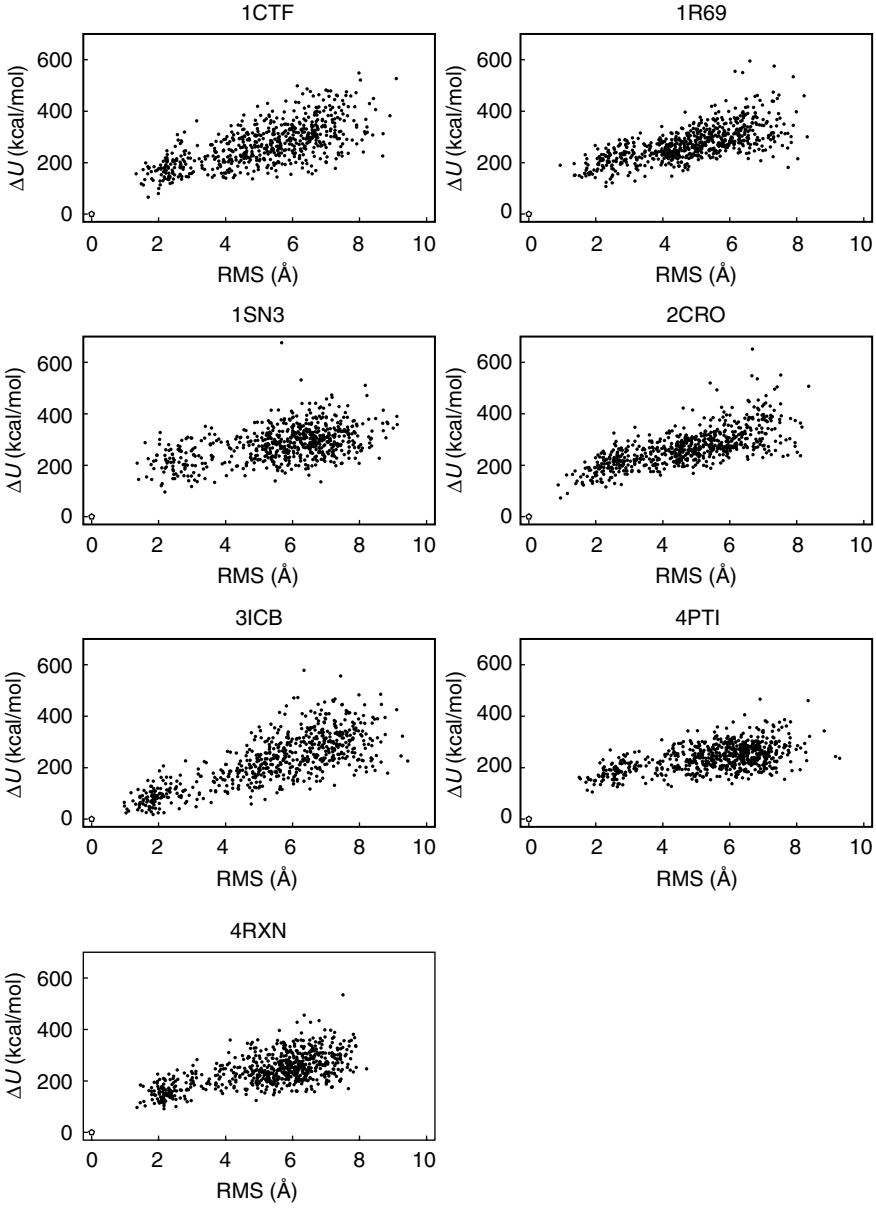

Figure 1. OPLS-AA/SGB: Energy gap/RMS correlation plots for the Park and Levitt decoy sets.

TABLE II
OPLS-AA/SGB Results: The Minimized energy, U_{native}, of the Native Conformation; the Energy Gap, min (ΔU), and the RMS Devition Between the Best-Scoring Decoy and the Native Conformation; the Native Z-Score Z_{nat} and the Average Z-Score $\bar{Z}_{nat\text{-}like}$ of the Native-like Conformations of the Park and Levitt Decoy Sets [60]

PDB Name	U_{native}	min(ΔU)	RMSD	Z_{nat}	$\bar{Z}_{nat\text{-}like}$
1ctf	−4213.92	+65.55	1.69	−3.24	−1.08
1r69	−3499.46	+107.16	2.30	−4.03	−1.01
1sn3	−3467.53	+96.08	2.19	−4.22	−1.04
2cro	−3628.30	+72.55	0.94	−3.69	−0.95
3icb	−4694.45	+18.08	1.84	−2.18	−1.34
4pti	−3055.04	+105.07	1.89	−4.53	−1.15
4rxn	−3363.51	+92.06	2.16	−3.76	−1.29

In Table II we also report the native Z score, Z_{nat}, and the average Z score of the native-like decoys, $\bar{Z}_{nat\text{-}like}$. The Z score of conformation i is defined as

$$Z_i = \frac{E_i - \bar{E}}{\sigma} \tag{8}$$

where E_i is the energy of the particular conformation, $\bar{E}$ is the average score and σ is the standard deviation of the distribution of scores in the set. The average Z score, $\bar{Z}_{nat\text{-}like}$, is obtained by averaging the Z scores of the native-like decoys. A decoy is defined as native-like if its RMSD with respect to the native is less than

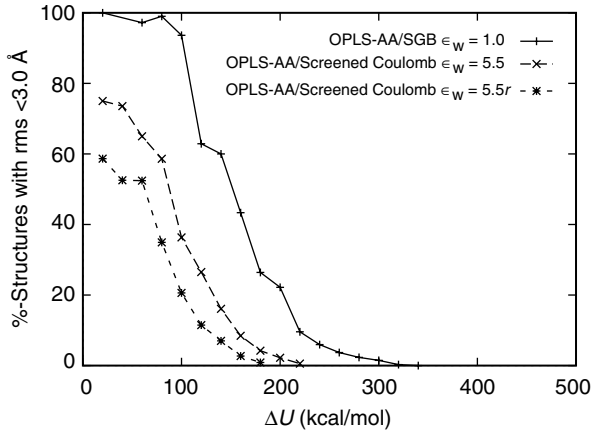

Figure 2. Fraction of the Park and Levitt decoys with energy gap from the native less than ΔU which are native-like (RMSD from native < 3 Å), using the OPLS-AA/SGB potential function and the vacuum OPLS-AA potential with screened Coulomb interactions.

3 Å. The Z score measures the ability of the scoring function to recognize native conformations. Assuming the distribution of scores is approximately Gaussian, a native Z score of, say, -2 indicates that the native structure is ranked in the best 1% in the decoy set. In general, the more negative the Z score, the better. The values of the native Z scores range from -3.2 to -4.5, indicating that the scoring function is extremely successful in finding the native structure among the decoys. The native-like average Z score represents the ability of the scoring function to discriminate the native-like conformations from the non-native conformations. The more negative the average native-like Z score, the larger the probability that a low-energy conformation is a conformation structurally similar to the native. The calculated values of the Z scores ranging from -0.95 to -1.34 indicate that, although on average the native-like conformations have lower energies than the non-native conformations, a significant number of native-like structures have a favorably low Z score. This can also be seen from Fig. 1 by looking at the vertical position of the low-RMSD structures with respect to the bulk of the decoys. This does not necessarily indicate a deficiency of the energy function but rather that for native-like conformations (i.e., those with the correct fold) the energy is also sensitive to the position and orientation of the amino acid side chains. An incorrect placement of a side chain may be enough to increase the energy of a native-like fold to the level of the misfolded conformations. A native-like energy is achieved only when all of the structural elements of the protein are placed correctly [22].

Park and Levitt [60] have evaluated six simple empirical scoring functions using the same decoy sets examined in this work. A comparison between the native and native-like Z scores calculated here with those obtained by Park and Levitt shows that the OPLS-AA/SGB energy model clearly outperforms the six empirical scoring functions examined in the Park and Levitt work. Moreover, none of the empirical scoring functions examined by Park and Levitt was able to consistently rank first the native conformation, whereas the OPLS/SGB model does.

It is instructive to evaluate the importance of each component of the OPLS-AA/SGB energy function in recognizing native conformations. Because all the decoys are well-packed, there is very little discrimination based on packing (as measured by the van der Waals energies) of the non-native states from the near-native conformations. In order to establish the role of intramolecular and solvent electrostatic interactions, we have calculated the energy scores in vacuum, $U_{\mathrm{tot}}^{\mathrm{vac}}$, using the same protocol used for the calculations in continuum solvent. The results are summarized in Table III. For several proteins the native conformation does not correspond to the minimum energy, and decoys with large RMSD from the native have very favorable scores. The native Z score and the near-native average Z scores have also significantly degraded (compare Tables II and III). This can be clearly seen in Fig. 3 showing the energy RMSD correlation plots

TABLE III
Vacuum OPLS-AA Results: The Minimized Energy, U_{native}, of the Native Conformation; the Energy Gap, min (ΔU), and the RMS Devition Between the Best-Scoring Decoy and the Native Conformation; the Native Z-Score Z_{nat} and the Average Z-Score $\bar{Z}_{nat\text{-}like}$ of the Native-like Conformations of the Park and Levitt Decoy Sets [60]

PDB Name	U_{native}	min(ΔU)	RMSD	Z_{nat}	$\bar{Z}_{nat\text{-}like}$
1ctf	−2795.74	+43.68	6.49	−2.62	−0.51
1r69	−2489.72	+76.49	1.65	−3.03	−0.42
1sn3	−2495.10	+0.04	1.42	−3.10	−0.59
2cro	−1122.06	−35.12	0.93	−2.37	−0.68
3icb	−2795.74	−282.69	1.19	−0.63	−0.84
4pti	−1324.06	+37.53	6.21	−2.97	−0.71
4rxn	−3581.88	−8.95	1.60	−2.47	−1.13

for the seven proteins studied. The gain achieved by including the solvation term is particularly noticeable for the 3icb data set. Figure 4 shows the distribution of energy gaps from the native for the 3icb decoys using either the vacuum OPLS-AA energy or the OPLS-AA/SGB energy. A shift of the distribution to positive values indicates that no decoy structures have energies lower than the native structure. Vacuum energies are scattered above and below the native state energy with little correlation between energy and structural similarity. The OPLS-AA/SGB energies produce a sharper distribution than the vacuum energies. It is clear that for this decoy set the vacuum energy is significantly poorer than the energy in solution in discriminating native folds.

An important contribution to protein stability arises from the tendency for packing nonpolar side-chains in the interior of the proteins and placing polar residues on the solvent exposed surface of the protein [75,76]. These tendencies are not represented well by the intramolecular potential in vacuum, which in general is equal to the strength of interaction between two nonpolar residues and between a nonpolar residue and polar residue and does not particularly favor the placement of a polar residue on the protein surface. The solvation energy calculated using the SGB model, however, reproduces hydrophobic interactions and favors the placement of polar residues on the protein surface where they can interact strongly with the solvent. The presence of a hydrophobic core and a polar surface is a key feature of the native protein conformation in solution. Several empirical scoring function have been designed to recognize these features [20,60,65,66,62]. A model that does not take into account solvation effects is likely to perform poorly in native fold recognition among large numbers of compact decoys.

Another important function of dielectric continuum models is to dampen the strength of the electrostatic interactions between polar and charged residues. Conformations having salt bridges and intramolecular hydrogen bonds are

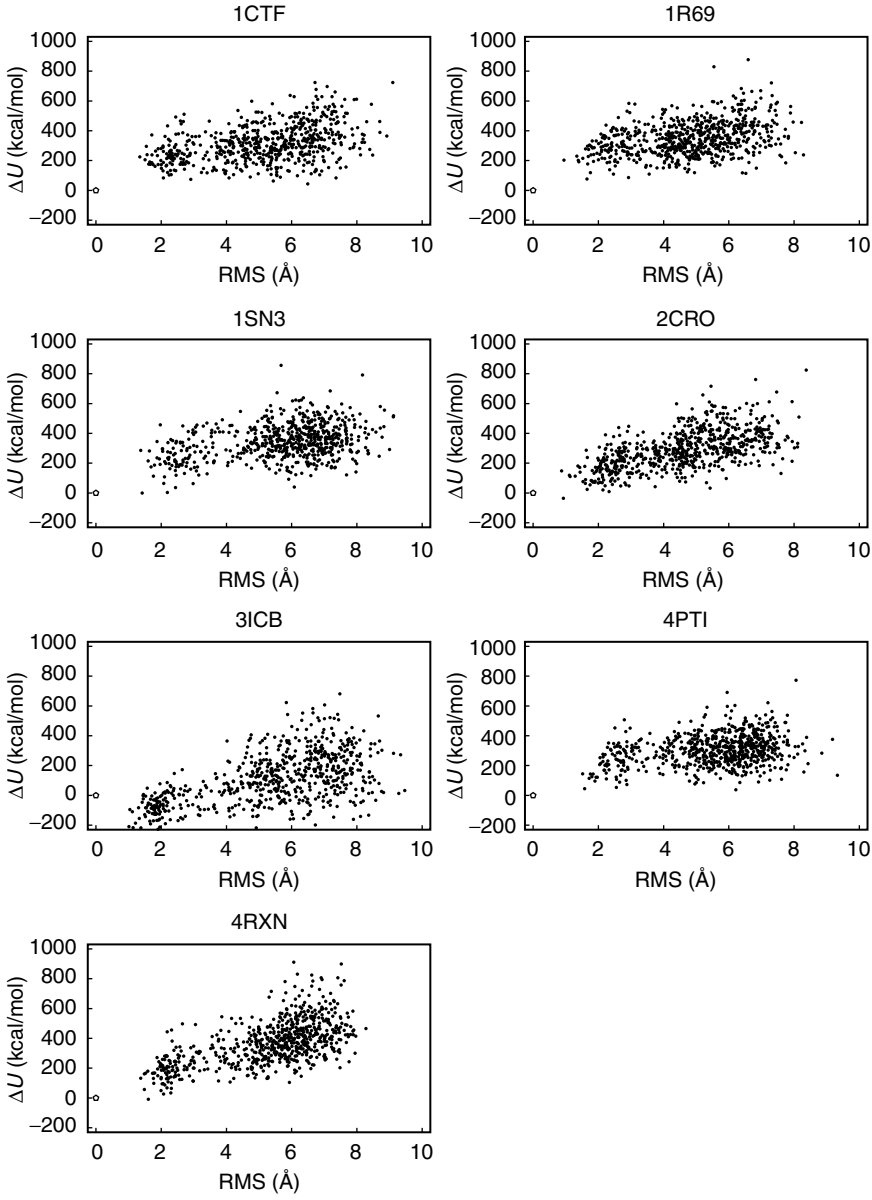

Figure 3. Vacuum OPLS-AA: Energy gap/RMS correlation plots for the Park and Levitt decoy sets.

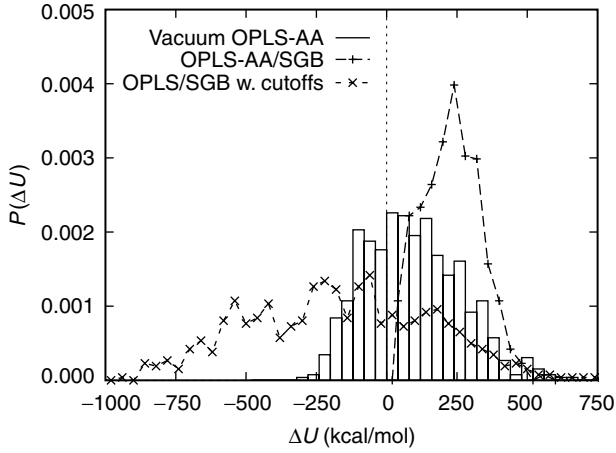

Figure 4. The distribution of energy gaps from the native for the 3icb data set of the Park and Levitt decoys using various energy functions.

strongly favored in vacuum, but much less so in solution. The SGB implicit solvent model provides a mechanism to filter out non-native conformations with artificially low intramolecular electrostatic energies that would be otherwise given a favorable score.

In these calculations, all charged interactions are included in the total energy; employing a cutoff for atom–atom interactions destroys the correlation between low energy values and native-like structures. Figure 4 shows that the proper evaluation of the long-range Coulomb interactions is crucial in selecting native conformations. If the electrostatic interactions are spatially truncated, many non-native structures assume lower total energies than do the native structure. As shown in Fig. 4, the correlation between energy and structural similarity using the OPLS-AA/SGB force field with a nonbonded cutoff of 9 Å is poor. This is a direct consequence of neglecting the long-range part of Coulomb interactions and is aggravated by the highly charged nature of some of the proteins examined (see Table I).

B. Holm and Sander Single Decoys

Recognizing single misfolded structures that have been carefully selected or devised as possible alternate folds poses a different challenge than distinguishing native-like states in large decoy data sets. Instead of picking native-like conformations among a large set of decoys, the challenge is to differentiate between two well-folded proteins, one of which corresponds to the native state. In the decoy set of Holm and Sander [17], misfolded conformations were constructed by swapping parts of the polypeptide chains with segments from

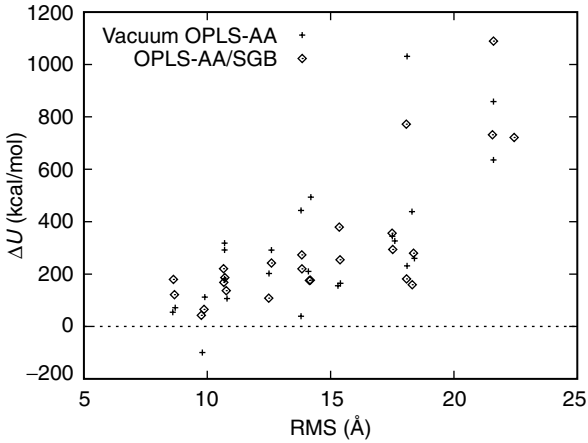

Figure 5. Energy gaps between the Holm and Sander [17] misfolded decoys and the corresponding native conformations, using the vacuum OPLS-AA and the OPLS-AA/SGB potentials.

known crystal structures. The proteins in the Holm and Sander set cover a wide range of sizes, from 36 residues for the smallest protein to over 300 residues for the largest protein. Figure 5 reports the energy gaps from the native of the misfolded proteins using the vacuum OPLS-AA energy and the OPLS-AA/SGB energy. The misfolded conformations are compact and have RMSDs from the native of 8 Å or more. Both the vacuum OPLS-AA and the OPLS-AA/SGB models are successful in ranking the native structures higher then the corresponding misfolded decoys; the only exception is for the avian pancreatic polypeptide (1ppt), a small 36 residue polypeptide, using the vacuum OPLS-AA model. Although smaller energy differences are generally correlated with higher structural similarity (see Fig. 5), the smallest ($\sim$8 Å) RMSD structure in this data set is well above the RMSD threshold of $\sim$4 Å, above which energy and structural similarity were no longer correlated for the proteins in the Park and Levitt set.

The apparent correlation between RMSD and energy gap visible in Fig. 5 is mostly due to the fact that the RMSDs and the energy gaps increase with increasing protein size. As shown in Fig. 6, the energy gaps grow roughly linearly with the sequence length of the protein (a slightly better correlation is observed when using the OPLS-AA/SGB model). The energy gaps calculated using the OPLS-AA/SGB model are generally of the same relative magnitude, when normalized by size, as the energy gaps calculated for the Park and Levitt set. This confirms that the energy function used here can discriminate between native and misfolded structures over a wide range of protein sizes.

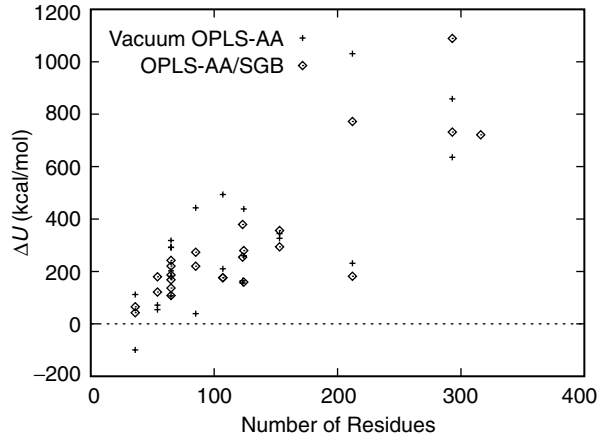

Figure 6. Protein size dependence of the energy gaps from the native of the misfolded protein structures from the Holm and Sander [17] data set.

C. CASP3 Targets

We have also analyzed some of the structures submitted to the CASP3 competition [67]. The target proteins are listed in Table IV. Our results are shown in Fig. 7, which shows the differences between the energy of each predicted structure and the energy of the corresponding native conformation. The targets can be divided into two groups: the ''easy'' targets for which the majority of the predicted models have an RMS deviation from the native of 3 Å or less, and the

TABLE IV
A Summary of the CASP3 Target Evaluated in this Study[a]

Target	N_{res}	Resolution (Å)	N_{res} Predicted	Models	Class	RMS (Å)	PDB
T0043	158	1.5	158	8	α/β	14.2–16.8	1hka
T0047	162	2.5	158	14	mostly β	1.3–1.9	1a2u
T0052	101	NMR	101	8	all β	13.7–17.1	2ezm
T0055	125	2.0	123	17	mostly β	2.8–7.4	1byf a
T0058	229	1.6	225	10	α/β	1.6–3.3	1eug
T0060	117	1.54	117	17	α/β	1.3–5.2	1dpt
T0064	111	1.9	103	22	All α	7.8–19.1	1b0n a
T0065	57	1.9	31	49	All α	2.7–10.1	1b0n b
T0068	376	1.9	376	4	Mainly β	8.9–18.5	1bhe
T0082	190	1.75	190	12	α + β	4.6–19.3	1bk7
T0085	211	2.6	211	6	Mostly α	17.8–22.9	1bvb

[a]Out of the structures predicted by the participants in CASP, we have selected those that have near- or full-length predictions only and whose PDB coordinates were available at the time of this study.

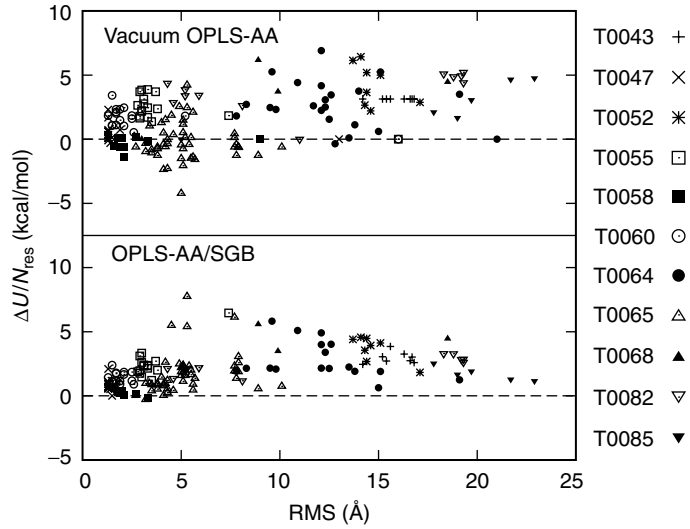

Figure 7. Energy difference per residue between native and predicted structures for a selection of targets from the CASP3 competition: T0043 (1hka), T0047 (2a2u), T0052 (2ezm), T0055 (1byf), T0058 (1eug), T0060 (1dpt), T0064 (1b0n_a), T0065 (1b0n_b), T0068 (1bhe), T0082 (1bk7), and T0085 (1bvb).

difficult targets in which none of the predicted models is native-like (RMS deviations from the native of 10 Å or more). For a few of the targets the predictions ranged from near-native (<3 Å) to non-native (>3 Å).

As shown in Fig. 7, the OPLS-AA/SGB model achieves nearly 100% discrimination of the native conformations. Only a few predictions, structurally similar to the native, score slightly better than the native. The vacuum OPLS-AA energy function does not perform as well as the OPLS-AA/SGB energy function; several high-RMS predictions for the T0055, T0058, T0064, and T0065 targets have scores significantly lower than the native. As observed for the Park and Levitt [60] decoy set, neither the vacuum OPLS-AA nor OPLS-AA/SGB energy functions are able to differentiate between models with large RMS deviations from the native; that is, a 15 Å structure can easily score better than a 10 Å structure.

D. Energy Components

The ability of a scoring function to discriminate between native and non-native conformations depends on the delicate balance between the components of the scoring function [1,20,60,66,62]. As described in this section, we find that, although some combinations of energy components show improvement over

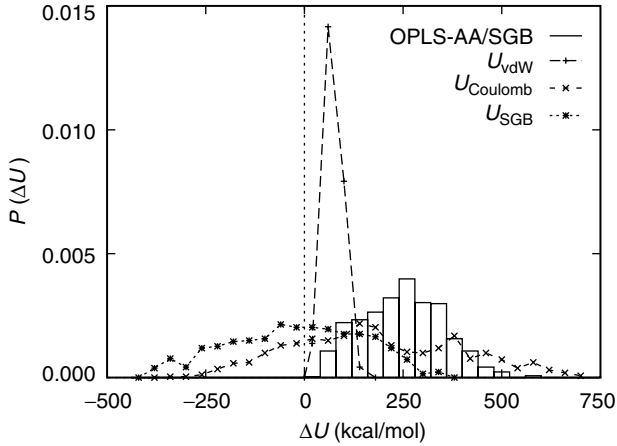

Figure 8. Distribution of energy gaps from the native of the 3icb Park and Levitt decoys for the total OPLS-AA/SGB energy and for the van der Waals, U_{vdW}, intramolecular Coulomb, U_{Coulomb}, and solvation, U_{SGB}, energy components.

each individual component, the total OPLS-AA/SGB energy is the best scoring function overall.

An analysis of the energy components of Eqs. (1) and (2) presented in Fig. 8 shows that for the Park and Levitt data set (Table I), containing only well-packed structures, the van der Waals energy difference with respect to the native is positive for most of the decoys. The van der Waals energy, however, does not strongly correlate with structural similarity to the native. This point is illustrated in Fig. 9, which shows the distribution of energy gaps from the native of both the native-like (RMSD <3 Å) and misfolded (RMSD >3 Å) 3icb decoys. In contrast, the discriminating power of the total OPLS-AA/SGB energy is indicated by the relatively small overlap between the native-like and misfolded distributions of energy gaps (see Fig. 9). A similar separation is not achieved with the van der Waals energy, indicating that the van der Waals energy alone does not provide good discrimination when used as a scoring function.

The electrostatic energy components, the intramolecular Coulomb energy, and the solvation energy, taken individually, are not effective scoring functions; the sum of the two, however, is significantly better as indicated in Figs. 10 and 11 ($\epsilon_{\mathrm{w}} = 1$ distribution). As shown in Fig. 10, the solvation energy is strongly anticorrelated with the electrostatic energy. A positive intramolecular electrostatic energy gap from the native is counteracted by a negative solvation energy gap, and vice versa. Because the solvation energy does not completely offset the intramolecular electrostatic energy, decoys having an intramolecular electrostatic energy less favorable than the native will generally continue to

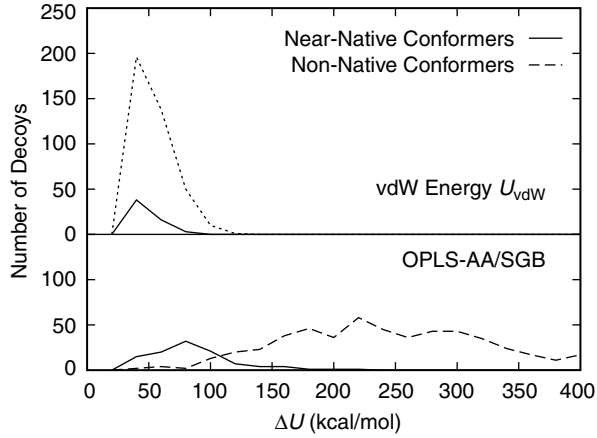

Figure 9. Near-native and non-native distributions of the OPLS-AA/SGB and van der Waals energy gaps from the native for the Park and Levitt 3icb decoys.

have a less favorable total electrostatic energy (intramolecular + solvation) with respect to the native. The contribution of the solvation energy term, however, is large enough to reverse the sign of the energy gap for those decoys having an intramolecular energy more favorable than the native, for which there are many examples in the Park and Levitt set (see Fig. 11). The native state

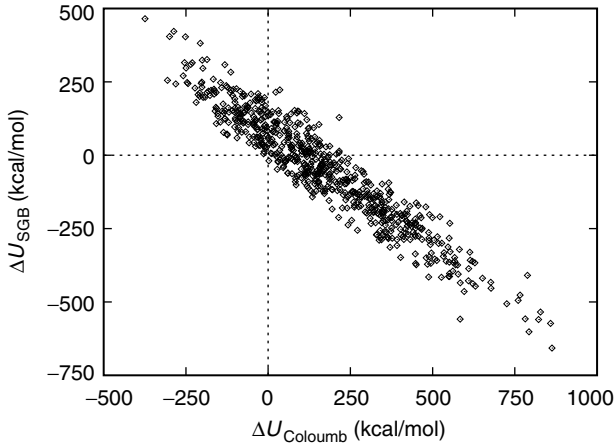

Figure 10. Correlation plot between the intramolecular Coulomb energy gap $\Delta U_{\text{Coulomb}}$ and the solvation energy gap ΔU_{SGB} for the 3icb decoys.

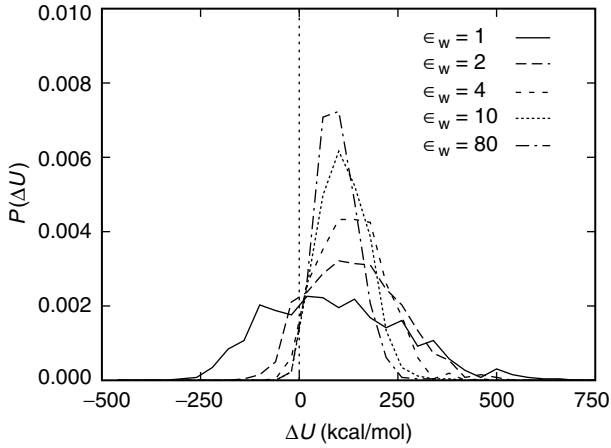

Figure 11. The distributions of the screened Coulomb OPLS-AA energy gaps from the native for the 3icb decoys as a function of dielectric constant.

corresponds to a balance between optimizing the intramolecular Coulomb interactions and the intermolecular protein–solvent interactions.

Monge et al. [20] have also studied various energy decompositions of an all-atom force field supplemented by a continuum solvation model. They analyzed a decoy data set generated by a simplified model employing a fixed, known secondary structure. The authors observe that the relative differences of both van der Waals and Coulomb energies are about 1–2% above the native values, but the total electrostatic component is the dominant factor in distinguishing non-native states from the native ones. They found that a fraction of the decoys had vdW energies lower than that of the native. Their model performed reasonably well, though some non-native conformations had better scores than the native state. This was not observed in the data sets we studied using the OPLS-AA/SGB scoring function.

E. Approximate Effective Dielectric Models

1. Screened Coulomb Approximation

As shown in Fig. 10, the solvation energy gaps with respect to the native are strongly correlated with the intramolecular Coulomb energy gaps. The equation

$$\Delta U_{\mathrm{SGB}} = \alpha + \beta \Delta U_{\mathrm{Coulomb}} \tag{9}$$

can be fitted obtaining $\beta = -0.82$ with a regression coefficient of 0.94. If we collate the total electrostatic interaction energy ΔU_{ele} as the sum of the Coulomb

and solvation energies, we find

$$\Delta U_{\text{ele}} \equiv \Delta U_{\text{Coulomb}} + \Delta U_{\text{SGB}} \cong 0.18 \Delta U_{\text{Coulomb}} \tag{10}$$

This suggests that it might be possible to employ a screened Coulomb model to account for solvation effects.

The screened Coulomb effective electrostatic interaction between two charges q a distance r apart is

$$\frac{U_{\text{Coulomb}}(r)}{\epsilon_{\text{w}}} = \frac{q^2}{\epsilon_{\text{w}} r} \tag{11}$$

The effect of the surrounding medium is accounted for by the value of ϵ_{w}, usually taken as 80 for water. Figure 11 shows the energy distributions for the 3icb decoy set relative to the native state for the vacuum case and for various values of the effective dielectric constant. A good energy function should only produce energy gap values in the positive range. It is clear that for this decoy set, a simple electrostatic energy evaluation in vacuum ($\epsilon_{\text{w}} = 1$) results in many decoy structures with energies substantially below the native values. Moreover, no correlation between the RMSD from the native and the energy is observed. Increasing the value of the effective dielectric constant removes some of the negative energy gaps and increases the propensity for the low-energy decoy structures to have low RMSD (not shown). None of the effective dielectric constants used, however, was able to differentiate all of the decoys from the native structure. This point is also illustrated in Fig. 2, which depicts the fraction of native-like structures with energy gaps from the native less than ΔU using $\epsilon_{\text{w}} = 5.5$ as suggested by the relation in Eq. (10). It is clear that the screened Coulomb scoring function provides less discrimination between decoys and native structures than does the SGB solvation model.

If a simple relationship between the reaction field energy calculated via the SGB model and the Coulomb energy as in Eq. (11) could be found, there would be no need to employ more complicated continuum models. Although the bulk of the correlation between these two terms can be explained by a screened Coulomb interaction, the discrimination between native and non-native states is degraded by such an approximation. The dispersion in the reaction field energy versus the Coulomb energy, which is not contained in the screened Coulomb model, provides a more detailed description of solvation effects which aids the discrimination of native-like conformations from misfolded ones.

Although the SGB solvation energy is correlated with the intramolecular Coulomb energy, it is not clear that the best values to use for an effective dielectric constant is given by Eq. (10). The fraction of native-like structures with energy gap less than a given energy difference calculated over all the data

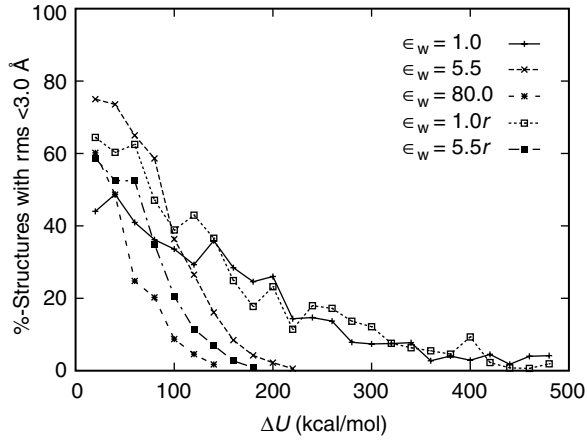

Figure 12. Fraction of the Park and Levitt decoys with energy gap from the native less than ΔU which are native-like (RMSD from native < 3 Å), using the vacuum OPLS-AA potential with screened Coulomb interactions.

sets in Table I, reported in Fig. 12, shows the efficiency achieved using different values of ϵ_w. None of the effective dielectric models achieves 100% discrimination for energy values within 20 kcal/mol of the native state energy. Using $\epsilon_w = 1$ yields a broad range of energies for both native-like and non-native states as discussed above. In comparison, using a value of ϵ_w either 5.5 or 80.0 yields distributions of energies that are like those given in Fig. 11 for the calbindin data set. The fraction of native-like structures with energies similar to the native state is around 60% for an effective dielectric constant of 80.0. This fraction increases to about 75% for an ϵ_w value of 5.5.

2. *Distance-Dependent Dielectric Approximation*

An alternative to the simple screened Coulomb interaction in protein modeling is the distance-dependent dielectric function [51]. In this approximation the effective electrostatic interaction between two partial charges q at distance r is written as

$$\frac{U_{\text{Coulomb}}(r)}{\epsilon_w r} = \frac{q^2}{\epsilon_w r^2} \tag{12}$$

Although unphysical in nature, it has been suggested that the extra screening afforded by the $1/r^2$ function can capture some of the additional polarization effects contained in higher-level implicit solvent models [51]. By calculating

the energies of the decoy conformers in Table I using the distance-dependent dielectric approximation, we obtain energy distributions similar to those obtained using the simple screened Coulomb model. Moreover, as shown in Fig. 2, both effective dielectric models produce qualitatively similar results. For both values of ϵ_w studied, 1.0 and 5.5, the fraction of native-like structures with energy similar to the native energy, is significantly less than 100%. Comparison between the distance-dependent dielectric and the non-distance-dependent dielectric function in Figs. 12 and 2 demonstrate that the distance-dependent function is less discriminatory for the decoy data sets studied here. While the distance-dependent dielectric constant has been successfully employed in some cases [77], we find that, though it is better than the vacuum Coulomb potential, a simple non-distance-dependent screened Coulomb model is more effective (Fig. 12). None of the screened Coulomb models are as effective as the SGB solvation potential for the protein decoy data sets investigated here.

F. Dependence on the Interior Dielectric Constant

The SGB solvent model requires the separation of space into an exterior region containing the solvent medium and an interior region containing the protein charge distribution. In the current implementation of the SGB model, the van der Waals surface of the protein is used to define the dividing surface. The default value for the dielectric constant of the solvent is 80, corresponding to pure water at room temperature. Up to this point, the dielectric constant of the interior region, ϵ_{in}, has been set at the value of 1, corresponding to the vacuum dielectric constant. We have also examined the cases $\epsilon_{in} = 2$ and 5.5 to see whether the OPLS-AA/SGB results can be further improved. The energy components obtained for the native conformations contained in the Park and Levitt set are given in Table V. A larger interior dielectric constant results in a lower total energy of the system due to the increase of the dielectric shielding inside the protein. The Coulomb energy and the reaction field contributions are both reduced in an amount roughly proportional to the interior dielectric constant. The van der Waals energy partly compensates for the reduction in electrostatic energy, but the variation in U_{vdW}^{native} is relatively small.

The fraction of native-like decoys of the Park and Levitt set as a function of energy gap is shown in Fig. 13 for the values of ϵ_{in} examined. The number of native-like conformations (RMSD <3 Å) with an energy score similar to the native increases as we decrease the dielectric constant of the interior region. It is only with an interior dielectric of 1.0 that all misfolded conformations can be eliminated based on energy alone. The discriminatory power of the OPLS-AA/SGB energy model in this fold recognition test is optimal for this choice of the internal dielectric, though it may not be optimal in other modeling contexts.

TABLE V
Selected Energy Components from Eqs. (1) and (2) for the Native State Using the Continuum Model ($\epsilon_w = 80.0$) as a Function of Interior Dielectric Constant, ϵ_{in}

PDB	ϵ_{in}	U^{native}_{total} (kcal/mol)	U^{native}_{vdW} (kcal/mol)	$U^{native}_{Coulomb}$ (kcal/mol)	U^{native}_{SGB} (kcal/mol)	U^{native}_{cav} (kcal/mol)
1ctf	1.0	−4213.9	−475.5	−5340.3	−1367.6	+37.9
	2.0	−2065.9	−519.7	−2595.2	−688.3	+38.4
	5.5	−730.6	−532.8	−925.5	−244.0	+38.7
1r69	1.0	−3499.5	−497.2	−3722.9	−1168.9	+37.2
	2.0	−1709.9	−539.0	−1781.7	−593.3	+37.7
	5.5	−599.5	−554.3	−627.9	−210.8	+38.1
1sn3	1.0	−3467.5	−465.1	−4784.2	−972.8	+36.3
	2.0	−1688.1	−499.8	−2315.2	−500.3	+36.8
	5.5	−585.3	−511.8	−821.5	−180.1	+37.1
2cro	1.0	−3628.3	−522.4	−3514.8	−1462.2	+40.4
	2.0	−1763.1	−567.2	−1662.8	−749.7	+41.0
	5.5	−604.8	−578.9	−585.2	−264.8	+41.4
3icb	1.0	−4694.5	−587.3	−5163.5	−2350.6	+45.4
	2.0	−2271.4	−641.0	−2466.5	−1195.6	+46.1
	5.5	−766.8	−656.8	−865.7	−427.2	+46.4
4pti	1.0	−3055.0	−423.9	−2542.0	−1366.9	+34.1
	2.0	−1464.2	−448.4	−1208.6	−686.6	+34.6
	5.5	−473.2	−455.1	−425.0	−240.9	+34.8
4rxn	1.0	−3363.5	−373.6	−2496.6	−2791.5	+31.3
	2.0	−1598.8	−399.3	−1190.1	−1389.9	+31.6
	5.5	−498.1	−407.6	−410.9	−489.1	+31.8

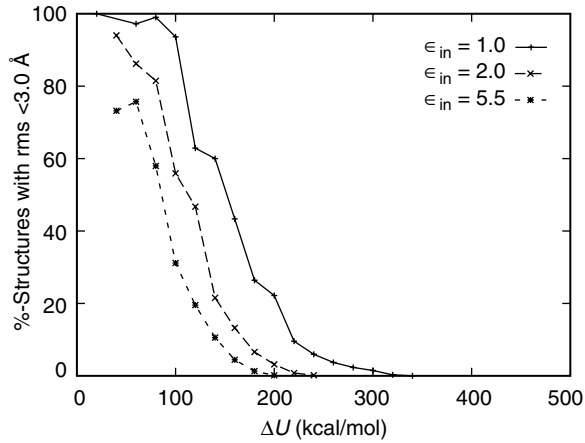

Figure 13. Fraction of the Park and Levitt decoys with energy gap from the native less than ΔU which are native-like (RMSD from native < 3 Å), using the OPLS-AA/SGB potential with various values of the interior dielectric constant.

IV. CONCLUSIONS

The OPLS-AA molecular mechanics energy function coupled with the surface generalized Born solvation model is found to be able to discriminate the native structures of several proteins from their decoys. The results show that for a number of cleverly constructed decoys the OPLS-AA/SGB scoring function correctly singles out native-like states from the bulk of the non-native conformations. Not all of the native-like structures were clearly separated in the data sets; indeed some distant non-native conformations score better than some native-like (RMSD <3 Å) conformations. This suggests that if the current scoring method is to be applied to a set of *ab initio* generated structures, it is critical that the algorithm for constructing native-like structures be such that a broad range of the relevant parts of the native-like conformational space are sampled.

The ability of the OPLS-AA/SGB model to recognize native conformations is found to be comparable, and in many cases superior, to the best knowledge-based scoring functions. Other studies have shown the usefulness of molecular mechanics force fields augmented by implicit solvation models in this area [6]. Lazaridis and Karplus [22] have shown that the CHARMM protein force field combined with their EEF1 effective solvation free energy model [78] is able to achieve 100% discrimination of the native conformations in a large decoy data set and in the single decoy data set they examined. They also observe, in agreement with our findings, that significantly poorer results are obtained by omitting the solvation free energy term. They obtain these results despite the use of a computationally fast solvation model which has the form of an effective pair potential and is simpler than the SGB solvation model. Recently, Petrey and Honig [79] have applied the CHARMM protein force field, together with a dielectric continuum model based on the Poisson–Boltzmann equation, to the problem of native fold recognition in the single decoy data set [17] (also examined in this work) achieving a discrimination level close to 100%. They also applied a simplified solvation model containing only the intramolecular electrostatic energy and a hydrophobic residue burial estimator to evaluate the Park and Levitt decoy sets. In two cases (3icb and 4rxn) their method does not clearly rank the X-ray conformation favorably. Petrey and Honig observe that the solvation energy often favors the misfolded conformation in the single-decoy sets, concluding that the solvation energy is not useful in recognizing the native conformation. However, even though the solvation energy generally favors misfolded conformations, these structures tend to be disfavored relative to the native conformation when the total electrostatic energy (sum of the direct Coulomb and solvation term) is considered. In contrast, the SGB solvation term is essential for destabilizing the relatively large number of Park and Levitt decoys for which the direct Coulomb energy is more favorable than the corresponding value for the native.

The OPLS-AA/SGB scoring function was also compared with the screened Coulomb OPLS-AA scoring function. Whereas a significant fraction of the decoys with scores within 100 kcal/mol from the native are misfolded using a screened Coulomb potential, essentially all of the decoys within 100 kcal/mol from the native are native-like using the OPLS-AA/SGB scoring function.

The ability to discriminate native-like protein conformations from non-native conformations is one of the fundamental problems in theoretical protein structure prediction. The use of knowledge-based scoring potentials, derived from a combination of structural and thermodynamic data, is currently the most widely used method. It is often assumed that such models are inherently better than all-atom force fields. This work shows the importance of correctly modeling the physical forces underlying protein folding. Thanks to their simplicity, knowledge-based scoring schemes are less costly to evaluate compared to all-atom models. In the future it should be possible to combine the best features of the two approaches to rapidly generate plausible protein conformations using knowledge-based potentials more reliably, and then discriminate between conformers using all-atom scoring functions.

Acknowledgments

This project has been supported by the National Institutes of Health Grant GM-30580, the Center for Biomolecular Simulations at Columbia University, and the High Performance Computing Project at Rutgers University. The authors thank Dr. Lynne Reed Murphy for help with some of the calculations.

References

1. S. J. Wodak and M. J. Rooman, *Curr. Opin. Struct. Biol.* **3**, 247–259 (1993).
2. D. T. Jones and J. M. Thornton, *Curr. Opin. Struct. Biol.* **6**, 210–216 (1996).
3. K. W. Plaxco, D. S. Riddle, V. Grantcharova, and D. Baker, *Curr. Opin. Struct. Biol.* **8**, 80–85 (1988).
4. M. Hao and H. A. Scheraga, *Curr. Opin. Struct. Biol.* **9**, 184–188 (1999).
5. D. J. Osguthorpe, *Curr. Opin. Struct. Biol.* **10**, 146–152 (2000).
6. T. Lazaridis and M. Karplus, *Curr. Opin. Struct. Biol.* **10**, 139–145 (2000).
7. V. Eyrich, D. Standley, A. Felts, and R. Friesner, *Proteins* **35**, 41–57 (1999).
8. S. W. Rick and B. J. Berne, *J. Am. Chem. Soc.* **116**, 3949–3954 (1994).
9. R. M. Levy and E. Gallicchio, *Annu. Rev. Phys. Chem.* **49**, 531–567 (1998).
10. A. A. Rashin and M. A. Bukatin, *J. Phys. Chem.* **98**, 386–389 (1994).
11. D. Sitkoff, K. A. Sharp, and B. Honig, *J. Phys. Chem.* **98**, 1978–1988 (1994).
12. D. J. Tannor, B. Marten, R. Murphy, R. A. Friesner, D. Sitkoff, A. Nicholls, M. Ringnalda, W. A. Goddard III, and B. Honig, *J. Am. Chem. Soc.* **116**, 11875–11882 (1994).
13. D. Sitkoff, N. Ben-Tal, and B. Honig, *J. Phys. Chem.* **100**, 2744–2752 (1996).
14. G. Hawkins, C. Cramer, and D. Truhlar, *J. Phys. Chem.* **100**, 19824–19839 (1996).
15. E. Gallicchio, L. Zhang, and R. M. Levy, submitted (2001).
16. D. Eisenberg and A. D. McLachlan, *Nature* **319**, 199–203 (1986).

17. L. Holm and C. Sander, *J. Mol. Biol.* **225**, 93–105 (1992).
18. W. F. van Gunsteren, F. J. Luque, D. Timms, and A. E. Torda, *Annu. Rev. Biophys. Biomol. Struct.* **23**, 847–863 (1994).
19. P. E. Smith and B. M. Pettitt, *J. Phys. Chem.* **98**, 9700–9711 (1998).
20. A. Monge, E. J. P. Lathrop, J. R. Gunn, P. S. Shenkin, and R. A. Friesner, *J. Mol. Biol.* **247**, 995–1012 (1995).
21. M. Schaefer, H. W. van Vlijmen, and M. Karplus, *Adv. Protein Chem.* **51**, 1–57 (1998).
22. T. Lazaridis and M. Karplus, *J. Mol. Biol.* **288**, 477–487 (1999).
23. Y. N. Vorobjev and J. Hermans, *Biophys. Chem.* **78**, 195–205 (1999).
24. M. K. Gilson and B. Honig, *Proteins Struct., Funct., Genet.* **4**, 7–18 (1988).
25. D. Bashford and M. Karplus, *Biochemistry* **29**, 10219–10225 (1990).
26. A. A. Rashin, *J. Phys. Chem.* **94**, 1725–1733 (1990).
27. K. A. Sharp and B.Honig, *Annu. Rev. Biophys. Chem.* **19**, 301–332 (1990).
28. A. Warshel and J. Åqvist, *Annu. Rev. Biophys. Chem.* **20**, 267–298 (1991).
29. M. K. Gilson, M. E. Davis, B. A. Luty, and J. A. McCammon, *J. Phys. Chem.* **97**, 3591–3600 (1993).
30. B. Honig, K. Sharp, and A.-S. Yang, *J. Phys. Chem.* **97**, 1101–1109 (1993).
31. V. Mohan, M. E. Davis, J. A. McCammon, and B. M. Pettitt, *J. Phys. Chem.* **96**, 6428–6431 (1992).
32. T. Simonson and A. T. Brünger, *J. Phys. Chem.* **98**, 4683–4694 (1994).
33. K. Ösapay, W. S. Young, D. Bashford, C. L. Brooks III, and D. A. Case, *J. Phys. Chem.* **100**, 2698–2705 (1996).
34. S. R. Edinger, C. Cortis, P. S. Shenkin, and R. A. Friesner, *J. Phys. Chem. B* **101**, 1190–1197 (1997).
35. M. Born, *Z. Physik* **1**, 45–48 (1920).
36. F. Hirata, P. Rejfern, and R. Levy, *J. Quantum Chem.* **15**, 179–188 (1988).
37. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *J. Am. Chem. Soc.* **112**, 6127–6129 (1990).
38. A. Jean-Charles, A. Nichols, K. Sharp, B. Honing, A. Tempczyk, T. F. Hendrickson, and W. C. Still, *J. Am. Chem. Soc.* **113**, 1454–1455 (1991).
39. D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, *J. Phys. Chem. A* **101**, 3005–3014 (1997).
40. A. Ghosh, C. S. Rapp, and R. A. Friesner, *J. Phys. Chem. B* **102**, 10983–10990 (1998).
41. B. Roux and T. Simonson, *Biophys. Chem.* **78**, 1–20 (1999).
42. L. Zhang, E. Gallicchio, R. Friesner, and R. M. Levy, *J. Comp. Chem.* **22**, 591–607 (2001).
43. J. Novotny, R. Bruccoleri, and M. Karplus, *J. Mol. Biol.* **177**, 787–818 (1984).
44. J. Novotny, A. A. Rashin, and R. Bruccoleri, *Proteins Struct. Funct. Genet.* **4**, 19–30 (1988).
45. L. Chiche, L. M. Gregoret, F. E. Cohen, and P. A. Kollman, *Proc. Natl. Acad. Sci. USA* **87**, 3240–3243 (1990).
46. J. Vila, R. L. Williams, M. Vasquez, and H. A. Scheraga, *Proteins Struct. Funct. Genet.* **10**, 199–218 (1991).
47. R. L. Williams, J. Vila, G. Perrot, and H. A. Scheraga, *Proteins Struct. Funct. Genet.* **14**, 110–119 (1992).
48. Y. Wang, H. Zhang, W. Li, and R. A. Scott, *Proc. Natl. Acad. Sci. USA* **92**, 709–713 (1995).
49. Y. Wang, H. Zhang, and R. A. Scott, *Protein Sci.* **4**, 1402–1411 (1995).

50. M. Vieth, A. Kolinski, C. L. Brooks III, and J. Skolnick, *J. Mol. Biol.* **237**, 361–367 (1994).
51. S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagone, S. Profeta, and P. Weiner, *J. Am. Chem. Soc.* **106**, 765–784 (1984).
52. Y. N. Vorobjev, J. C. Almagro, and J. Hermans, *Proteins Struct. Funct. Genet.* **32**, 399–413 (1998).
53. M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl, *J. Mol. Biol.* **216**, 167–180 (1990).
54. M. J. Sippl, *Curr. Opin. Struct. Biol.* **5**, 229–235 (1995).
55. R. L. Jernigan and I. Bahar, *Curr. Opin. Struct. Biol.* **6**, 195–209 (1996).
56. S. Miyazawa and R. J. Jernigan, *J. Mol. Biol.* **256**, 623–644 (1996).
57. A. Wallqvist, G. W. Smythers, and D. G. Covell, *Protein Sci.* **6**, 1627–1642 (1997).
58. S. Miyazawa and R. L. Jernigan, *Proteins Struct. Funct. Genet.* **36**, 357–369 (1999).
59. D. Covell and R. Jernigan, *Biochemistry* **29**, 3287–3294 (1990).
60. B. Park and M. Levitt, *J. Mol. Biol.* **258**, 367–392 (1996).
61. B. Ozkan and I. Bahar, *Proteins Struct. Funct., Genet.* **32**, 211–222 (1998).
62. R. Samudrala and J. Moult, *J. Mol. Biol.* **275**, 895–916 (1998).
63. K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker, *Proteins Struct. Funct. and Genet.* **34**, 82–95 (1999).
64. W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).
65. E. S. Huang, S. Subbiah, J. Tsai, and M. Levitt, *J. Mol. Biol.* **257**, 716–725 (1996).
66. B. H. Park, E. S. Huang, and M. Levitt, *J. Mol. Biol.* **266**, 831–846 (1997).
67. J. Moult, T. Hubbard, K. Fidelis, and J. T. Pedersen, *Proteins Struct. Funct. Genet. Suppl.* **3**, 2–6 (1999).
68. D. B. Kitchen, F. Hirata, J. D. Westbrook, R. Levy, D. Kofke, and M. Yarmush, *J. Comp. Chem.* **11**, 1169–1180 (1990).
69. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, Protein data bank, in *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, F. H. Allen, G. Bergerhoff, and R. Sievers, eds., Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987.
70. L. Zhang, E. Gallicchio, and R. M. Levy, Implicit solvent models for protein–ligand binding: Insights based on explicit solvent simulations, in *Simulation and Theory of Electrostatic Interactions in Solution, AIP Conference Proceedings 492*, L. R. Pratt and G. Hummer, eds., American Institute of Physics, New York, 1999.
71. M. Levitt, *J. Mol. Biol.* **226**, 507–533 (1992).
72. R. Samudrala, Y. Xia, M. Levitt, and E. S. Huang, *Proc. Pacific Symp. Biocomput.* **4**, 505–516 (1999).
73. D. A. Hinds and M. Levitt, *Proc. Natl. Acad. Sci. USA* **89**, 2536–2540 (1992).
74. D. A. Hinds and M. Levitt, *J. Mol. Biol.* **243**, 668–682 (1994).
75. K. A. Dill, *Biochemistry* **29**, 7133–7155 (1990).
76. K. A. Dill, *Curr. Opin. Struct. Biol.* **3**, 99–103 (1993).
77. M. Schaefer, C. Bartels, and M. Karplus, *Theor. Chem. Acc.* **101**, 194–204 (1998).
78. T. Lazaridis and M. Karplus, *Proteins* **35**, 133–152 (1999).
79. D. Petrey and B. Honig, *Protein Sci.* **9**, 2181–2191 (2000).