PROTEINS: Structure, Function, and Genetics 48:404–422 (2002)

Distinguishing Native Conformations of Proteins From Decoys With an Effective Free Energy Estimator Based on the OPLS All-Atom Force Field and the Surface Generalized Born Solvent Model

Anthony K. Felts* Emilio Gallicchio, Anders Wallqvist, and Ronald M. Levy*
Department of Chemistry and Chemical Biology, Rutgers University, Wright-Rieman Laboratories, Piscataway, New Jersey

ABSTRACT **Protein decoy data sets provide a benchmark for testing scoring functions designed for fold recognition and protein homology modeling problems. It is commonly believed that statistical potentials based on reduced atomic models are better able to discriminate native-like from misfolded decoys than scoring functions based on more detailed molecular mechanics models. Recent benchmark tests on small data sets, however, suggest otherwise. In this work, we report the results of extensive decoy detection tests using an effective free energy function based on the OPLS all-atom (OPLS-AA) force field and the Surface Generalized Born (SGB) model for the solvent electrostatic effects. The OPLS-AA/SGB effective free energy is used as a scoring function to detect native protein folds among a total of 48,832 decoys for 32 different proteins from Park and Levitt's 4-state-reduced, Levitt's local-minima, Baker's ROSETTA all-atom, and Skolnick's decoy sets. Solvent electrostatic effects are included through the Surface Generalized Born (SGB) model. All structures are locally minimized *without restraints*. From an analysis of the individual energy components of the OPLS-AA/SGB energy function for the native and the best-ranked decoy, it is determined that a balance of the terms of the potential is responsible for the minimized energies that most successfully distinguish the native from the misfolded conformations. Different combinations of individual energy terms provide less discrimination than the total energy. The results are consistent with observations that all-atom molecular potentials coupled with intermediate level solvent dielectric models are competitive with knowledge-based potentials for decoy detection and protein modeling problems such as fold recognition and homology modeling. Proteins 2002;48:404–422.**

Key words: protein decoys; protein solvation; fold recognition

INTRODUCTION

Protein structure prediction methods all must be able to discriminate native conformations successfully from any collection of misfolded ones. The various scoring schemes proposed to accomplish this task can be separated roughly into three categories: knowledge-based (or empirical), physics-based, or a combination thereof.[1] A variety of knowledge-based empirical scoring functions have been proposed for the purpose of distinguishing native folds from non-native ones.[1–9] Some of these implement statistical potentials that are "trained" to recognize native conformations. Knowledge-based potentials are well suited for fold recognition 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.[10]

Physics-based all-atom molecular mechanics force fields have not been widely considered practical for fold detection; the level of atomic detail contained in these models is considered poorly suited to the fold recognition problem with regard to both accuracy and computational cost. Physics based all-atom force-fields, however, provide some advantages over knowledge-based scoring functions. The all-atom potential is more suitable for the task of modeling proteins at higher resolution. This is an important feature for applications in studies concerned with the relationship between detailed atomic structure and function such as homology modeling, structure-based drug-design, and protein–protein recognition. Recently, studies have shown that scoring functions based on an all-atom molecular mechanics potential energy coupled with various refined solvation models can recognize native protein conforma-

Grant sponsor: National Institutes of Health; Grant number: GM-30580; Grant sponsor: Center for Biomolecular Simulations at Columbia University; Grant sponsor: High Performance Computing Project at Rutgers University.

Anders Wallqvist's current address is NCI-Frederick/SAIC Bldg.430, P.O.Box B, Frederick, MD 21702.

*Correspondence to: Ronald M. Levy, Department of Chemistry and Chemical Biology, Rutgers University, Wright-Rieman Laboratories, 610 Taylor Rd, Piscataway, NJ 08854-8087. E-mail: ronlevy@lutece.rutgers.edu or Anthony K. Felts. E-mail: felts@lutece.rutgers.edu

Received 12 December 2001; Accepted 28 March 2002

tions among a set of decoys as well as the best available knowledge-based scoring functions.[1,11–14]

Although all-atom force fields can allow for an explicit treatment of the solvent, the cost required to appropriately sample solvent configurations rapidly becomes infeasible if attempting to score hundreds or thousands of different configurations of a protein. Simplified solvation models are much more computationally efficient and they can preserve a reasonably accurate representation of the interactions between the protein and its aqueous environment. Although no continuum model can completely account for the explicit effects of solvation,[15,16] free energies of solvation of small molecules have been obtained accurately with these methods to within a fraction of a kcal/mol relative to experiments,[17–22] and recent experience suggests that their application to proteins and DNA modeling is promising.[23–25]

Solvation effects have been included using a variety of simple models.[11,26–32] 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.[33–39] 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.[18,40–43] Dielectric models based on the Born equation[44] have been developed for which free energies of hydration are comparable to the predictions of Poisson–Boltzmann and explicit solvent models.[45–51]

Due to the complexity of the protein all-atom potential surface, it is virtually impossible to consistently find the global minimum starting from an arbitrary point on the surface. Nevertheless, Liu and Beveridge[52] were able to perform ab initio protein folding by using a multicopy Monte Carlo simulated annealing technique with the AMBER all-atom force field with a generalized Born solvation model. The protein they tackled was the small, 36-residue Villin head-piece; the final folded structure was 3.5 Å rmsd from the native (1.1 Å rmsd based only on the backbone).[52] This approach is still computationally expensive and an examination with this method of several proteins of larger sizes would be infeasible. 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 are significantly dissimilar in structure from the native. Many empirical energy functions have been used to identify the correct native structure among a collection of known protein structures using fold recognition techniques.[4,53–58] A variety of scoring functions have also been used to identify native-like conformations within a large set containing native and decoy non-native conformers.[11,59–63] The use of large decoy sets to evaluate scoring functions is a more demanding test than fold recognition and is particularly challenging for the evaluation of scoring functions based on an all-atom force field due to the computational costs associated with scoring thousands of conformers.

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.[26,27,64–66] Scheraga and coworkers[67,68] have used explicit all-atom protein models in conjunction with implicit solvation models based on the molecular exposed surface area. A similar approach by Wang et al.[69,70] showed that with the inclusion of solvation effects, it is possible to successfully discriminate the native from non-native structures. Vieth and coworkers[71] 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 for generating large numbers of protein folds was also employed by Monge et al.[30] to select and evaluate structures using the AMBER all-atom force field model[72] with the generalized Born continuum solvent model of Still and coworkers[46] representing 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 used also as the basis for discriminating non-native states for a small set of twelve deliberately misfolded proteins studied by Vorobjev et al.[73] 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[11] used the CHARMM19 protein force field together with a Gaussian solvation shell model for the solvation free energy to distinguish a native conformation from a single deliberately misfolded structure in the EMBL decoy sets[27] and from a set of decoys (the 4-state-reduced decoys) generated by Park and Levitt.[60] Petrey and Honig[12] used the CHARMM19 force field along with a dielectric continuum model based on the Poisson–Boltzmann equation to distinguish native folds from misfolded ones in the EMBL decoy sets.[27] And Dominy and Brooks[13] also used the CHARMM19 force field but with the addition of the generalized Born solvation term[46] to distinguish misfolded conformations in the EMBL[27] and Park and Levitt[60] 4-state-reduced decoy sets.

In this work, we demonstrate that the all atom (OPLS-AA) force field for proteins[74] together with a surface integral formulation of the generalized Born model (SGB)[49,51] is capable of discriminating between native and non-native folds among numerous sets containing a very large number of compact decoy structures. These databases of well-packed misfolded protein conformations were generated by a variety of algorithms designed to cover exhaustively the relevant parts of conformational space.[60,75–78] All structures are minimized using the OPLS-AA force field with and without the SGB model *without any restraints* imposed on the structures. The targets of ab initio folding and homology modeling are structures that occupy the minima of a potential; therefore, assessing the locations of the local minima that a given protein structure could occupy is the relevant quantity when analyzing decoys.

We also examine the ability to discriminate a native fold from the other decoys using the various components of the potential. Unconstrained minimizations facilitate the investigation into the performance of the individual components because the total potential and its various terms do not share the same minima. It will be shown that the individual components of the energy perform worse than the total energy: it is the various terms of the OPLS-AA force field along with the SGB solvation model *acting in concert* that produces the ability to screen the native folds from the collections of decoys.

METHODS

Details of the Calculations Using IMPACT

The native protein folds from the Protein Data Bank (PDB)[79] and corresponding structures from the decoy sets described below were minimized using conjugate gradients with the OPLS-AA/SGB force field as implemented in the IMPACT modeling program (Schrödinger, Inc.).[80] No restraints were placed on these minimizations, which were allowed to proceed in most cases until a local minimum in the potential was found. This serves to relieve any bond or steric strain in the structures generated using different energy functions.

The total free energy of folding for a protein in solution can be represented approximately as the sum of two terms:

$$\Delta G_{\text{tot}} \simeq \Delta G_{\text{int}} + \Delta G_{\text{solv}}, \quad (1)$$

where ΔG_{int} is the internal free energy of folding corresponding to the intramolecular degrees of freedom of the protein, and ΔG_{solv} is the difference of solvation free energy between the folded and unfolded states. The internal free energy of folding of the protein is given by

$$\Delta G_{\text{int}} = \Delta U_{\text{int}} - T\Delta S_{\text{int}}, \quad (2)$$

where ΔU_{int} is the change in internal energy of the protein and ΔS_{int} is the change in internal entropy determined from the configurational changes in a given conformation due to translational, rotational, and vibrational motions. The internal entropy change can be estimated from MD simulations; however, calculating the internal entropy is quite expensive.[81] Nevertheless, it has been found that the internal entropy changes of native, misfolded, or denatured conformations are all roughly the same.[73,81] Since in this work different conformations of a given protein are compared, it is not necessary to include ΔS_{int} in the total free energy change;[52] therefore, an effective free energy function,

$$\Delta G_{\text{eff}} = \Delta U_{\text{int}} + \Delta G_{\text{solv}}. \quad (3)$$

can be used in lieu of ΔG_{tot}. The OPLS all-atom (OPLS-AA) force field[74] is used to model ΔU_{int}, the internal energy for all atomic interactions and intramolecular degrees of freedom. The solvation free energy, ΔG_{solv}, of each structure is estimated using the surface formulation of the generalized Born model[46,48] as implemented in the IMPACT modeling program.[49,82]

The total internal energy of folding of a protein is given by,

$$\Delta U_{\text{int}} \equiv \Delta U_{\text{OPLS-AA}} = \Delta U_{\text{bond}} + \Delta U_{\text{angle}} + \Delta U_{\text{torsion}} + \Delta U_{\text{Coulomb}} + \Delta U_{\text{vdW}}, \quad (4)$$

where the first three terms refer to intramolecular interactions arising from the connectivity of the molecule and the last two terms reflect nonlocal interactions within the protein. The term ΔU_{bond} is the bond stretching energy; ΔU_{angle}, the angle bending energy; and $\Delta U_{\text{torsion}}$, the sum of the 1,4-van der Waals, 1,4-Coulomb, and torsion angle energies. The van der Waals energy, ΔU_{vdW}, is modeled by the standard 6-12 Lennard-Jones interaction for non-bonded atoms. The term $\Delta U_{\text{Coulomb}}$ corresponds to the direct Coulomb interactions between non-bonded atoms. The free energy of folding of the protein in water calculated according to the SGB continuum solvent model is

$$\Delta G_{\text{eff}} \equiv \Delta G_{\text{OPLS-AA/SGB}} = \Delta U_{\text{OPLS-AA}} + \Delta G_{\text{SGB}} + \Delta G_{\text{cav}}, \quad (5)$$

where ΔG_{SGB} denotes the electrostatic contribution to the change in solvation energy during folding calculated using the SGB method, and the hydrophobic cavity term ΔG_{cav} is taken as $\gamma\Delta A$ where ΔA is the difference in accessible surface area between the folded and unfolded structures and $\gamma = 5$ cal/(Å^2 mol).[49] The change in solvation energy during folding is calculated using

$$\Delta G_{\text{SGB}} = G_{\text{SGB}}^{(\text{folded})} - G_{\text{SGB}}^{(\text{unfolded})}, \quad (6)$$

where G_{SGB} is the SGB solvation free energy of a given structure.

The SGB model is the surface implementation[49,51] of the generalized Born model.[46] The generalized Born equation

$$G_{\text{SGB}} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\text{in}}} - \frac{1}{\varepsilon_{\text{w}}}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}(r_{ij})}, \quad (7)$$

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: the 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. (7) is defined as

$$f_{ij} = [r_{ij}^2 + \alpha_i\alpha_j \exp(-r_{ij}^2/4\alpha_i\alpha_j)]^{1/2}, \quad (8)$$

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

$$G_{\text{single}}^i = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\text{in}}} - \frac{1}{\varepsilon_{\text{w}}}\right)\frac{q_i^2}{\alpha_i}, \quad (9)$$

the electrostatic free energy of the molecule when only the charge of atom i is present in the molecular cavity. The G_{single}^i are evaluated numerically by integrating the interaction between atom i and the charge induced on the solute-solvent boundary surface, S, by the Coulomb field of this atom

$$G^i_{\text{single}} = -\frac{1}{8\pi}\left(\frac{1}{\varepsilon_{\text{in}}}-\frac{1}{\varepsilon_{\text{w}}}\right)\int_S \frac{q_i^2}{|\mathbf{r}-\mathbf{r}_i|^4}(\mathbf{r}-\mathbf{r}_i)\cdot\mathbf{n}(\mathbf{r})d^2\,\mathbf{r}, \tag{10}$$

where **n(r)** is the normal to the surface, S, at r. The atomic radii that define the solute-solvent dielectric boundary are set to the van der Waals radii based on the Lennard-Jones σ parameters. The Born radii for Eq. 8 are calculated using Eq. (9) and G^i_{single} from Eq. (10). In this work, we set $\varepsilon_{\text{in}} = 1$ and $\varepsilon_{\text{w}} = 80$. 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.[49]

For the 4-state-reduced, local-minima, and ROSETTA all-atom decoy sets (described below), the energy of each native and decoy structure is minimized using the full atomic model with and without the SGB dielectric continuum solvation energy term. For the 4-state-reduced data sets, the minimizations are terminated using a tolerance in the energy of 1.0×10^{-2} kcal/mol and a tolerance of 0.05 rms of the gradient up to a maximum of 100 steps. While the OPLS-AA/SGB minimizations converged to the specified energy and gradient tolerances, the OPLS-AA vacuum minimizations required more steps. For the local-minima and ROSETTA all-atom data sets, the maximum number of steps was increased to 1,000 and the tolerances for the energy and gradient were set to 1.0×10^{-2} kcal/mol and 0.5 rms, respectively. Only a few of the over 16,000 minimizations for these two data sets required more than 1,000 steps to reach a local minimum using the specified convergence criteria. All calculations are performed without any energy cut-offs, i.e., all possible nonbonded interactions are included in the total energy.

Due to the large number of decoys in the Skolnick database, a different approach to scoring the decoys is taken in order to save computer time. All of the decoys are minimized only with the OPLS-AA potential to a tolerance of 0.05 rms of the gradient up to a maximum of 500 steps. After the minimizations were completed, the SGB solvation energy is calculated for the final structure. Only the native structures for the Skolnick decoy sets are minimized with the entire OPLS-AA/SGB energy function. Again, no energy cut-offs are used.

Disulfide bonds are created whenever the separation between the sulfurs of two cysteines is within a certain tolerance indicating the existence of a bond. We decided that removing disulfide bonds would only result in penalizing any native folds (and also any decoy structures) possessing them. This penalty would arise from the strain of adding the extra volume (≈ 25Å^3) created when adding two hydrogens into the space originally occupied by the disulfide bond.[83] The hydrogens are added to the bare sulfides to transform them into stable thiol groups. In order to avoid significantly perturbing the structures with the addition of these thiol hydrogens, the disulfide bonds are retained.

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,

$$\Delta\Delta G = \Delta G^{(\text{decoy})} - \Delta G^{(\text{native})}, \tag{11}$$

Energy gaps of individual energy terms have also been examined (see Eqs. 4 and 5). Because both the native and the decoys have the same unfolded state, we are not required to calculate the energy of the unfolded state since it will cancel out in Eq. 11. All energies presented in the tables and figures are presented as the gap in energy between that decoy and its corresponding native. The units reported for the energies are kcal/mol. The structural similarity between two protein conformations is expressed as a root mean square deviation (rmsd) between the best overlap of the heavy atoms (i.e., all atoms except hydrogens) of the two conformations.

Data Sets of Decoys

Our goal is to assess the effectiveness of the OPLS-AA/SGB energy function, and vacuum energy function using the OPLS-AA force field, in discriminating between native and non-native conformations. Due to the very large number of degrees of freedom possessed by a protein structure coupled with the complexity of an all-atom potential function, a full sampling of the conformational space is not practical. Another approach to testing these energy functions is to partially sample the conformational space utilizing cleverly constructed decoy sets. These decoy sets can be viewed as the product of a previous step in a hierarchical protein folding process; the application of the all-atom scoring function would serve as the next step in the overall process: to refine the set by selecting the best candidates for the native fold from the rest. Four decoy sets are used to test the effectiveness of our all-atom scoring function. They are Park and Levitt's 4-state-reduced,[60] Levitt's local-minima,[77] Baker's ROSETTA all-atom,[78] and Skolnick's[84] and Lu and Skolnick's (personal communication) decoy sets. These decoy sets are summarized in Tables I–IV.

The Park and Levitt 4-state-reduced decoy database contains structural decoys for 7 small proteins.[60] The protein structures were generated by exhaustively enumerating the backbone rotamer states of ten 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 thousands 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.[85] The decoy data sets (summarized in Table I) encompass a range of small proteins from 54–75 residues with varying topological folds. The number of decoys in these sets range 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).

TABLE I. Proteins in the 4-State-Reduced Decoy Sets[†]

PDB	Description	N_{res}	N_{decoy}	rmsd range
1ctf	C-terminal domain of ribosomal protein L7/L12	68	630	2.16–10.16
1r69	N-terminal domain of phage 434 repressor	63	675	2.28–9.50
1sn3	scorpion toxin variant 3	65	660	2.50–10.46
2cro	phage 434 Cro protein	65	674	2.05–9.72
3icb	vitamin D-dependent calcium-binding protein	75	653	1.81–10.74
4pti	trypsin inhibitor	58	687	2.83–10.79
4rxn	rubredoxin	54	677	2.58–9.28

[†]Data from Park and Levitt.[60] The PDB designation is presented along with a brief description of the protein and the number of residues (N_{res}), the number of decoys (N_{decoy}) contained in the set, and the range of rmsd from the native of the decoys. All pdb files correspond to structures derived from X-ray data unless indicated otherwise in the description.

The compactness of the 4-state-reduced structures relative to the native is assessed by considering the total number of heavy-atom contacts within 4.5 Å present in each structure relative to the total number in the native. The average total number of contacts of the decoys for each protein are shown in Table V as the ratio of the number of contacts in each structure to the number in the native. The criterion of 4.5 Å for assigning a contact is based on the greatest distance between two atoms in which a water molecule could not be inserted.[86,87] We have found that this measure of compactness has one crucial advantage over measuring the compactness based on the van der Waals energy: it is insensitive to how the different types of atoms of the structure are arranged spatially. The values of the OPLS-AA Lennard-Jones ε parameters, which measure the minimum energy two atoms achieve on close contact due to van der Waals interactions, have a range from about 0.07 to 0.2 kcal/mol for the various types of hydrogen, carbon, nitrogen, and oxygen atoms. If two structures had the same number of atoms in contact (i.e., the same level of compactness), but one had more carbons in contact while the other had more oxygens in contact, the first structure would have a lower van der Waals energy than the second. Based on the van der Waals energy, it would seem that the first structure is more compact than the other when in fact they are equally compact from a structural standpoint. While it is true that all compact structures will have low van der Waals energies, care must be taken when assessing the compactness of two structures based on the van der Waals energy alone. The measure of compactness based on the number of the 4.5 Å contacts is better in this respect because it does not depend on atom types.

The Levitt low-minima decoy sets (LMDS) also contain structural decoys for 7 small proteins.[77] These decoys are conformations that occupy a minima in a modified "classic" ENCAD force field using united and soft atoms.[88] Terms were added that favored compactness and the formation of secondary-structure hydrogen bonding and that disfavored the burial of charged residues and any other formation of hydrogen bonds. Initially ten thousand structures were generated by randomly modifying only the loop dihedral angles. These structures were minimized in torsion space to the local minima in the potential described above. Up to five hundred of the lowest energy conformations were kept to make up the decoy sets. The proteins from these decoy sets used in this study range in size from 36 to 68 residues (see Table II). The number of decoys from each set ranges from 343 to 500 decoys. The compactness of these structures is also shown in Table V.

The ROSETTA all-atom decoy sets[78] are composed of five different proteins ranging in size from 92 to 116 residues (see Table III). These decoys are generated from fragments of 3 to 9 residues from known structures matched to the targets through a multiple sequence alignments process. These fragments were assembled into the protein structures via the fragment insertion–simulated annealing strategy.[89] The terms of the scoring function included those for hydrophobic burial, electrostatics, disulfide bonds, the packing of α-helices and β-strands, and the formation of β-sheets. During annealing only, another term used to promote compactness was added based on the radius of gyration. The number of decoy structures in the final sets ranges from 994 to 999. In Table V, the compactness of the structures is summarized.

The Skolnick decoy data sets (Lu and Skolnick, personal communication) analyzed by us are composed of 18 proteins and contained 2,000 decoys each. The decoys in this set were generated on a lattice using an ab initio Monte Carlo structure prediction program. Full-atomic detail was incorporated later into the resulting lattice structures. The Skolnick decoy sets are summarized in Table IV.

RESULTS AND DISCUSSION

OPLS-AA/SGB Calculations on the 4-State-Reduced, LMDS, and ROSETTA Databases

The results of the unrestrained minimizations of the decoys from the 4-state-reduced, local minima, and ROSETTA all-atom data sets with the OPLS-AA/SGB effective free energy function are summarized in Table VI. These minimized structures differ little from their initial structures: on average by only 0.54 ± 0.14 Å rmsd. Examination of the minimized native energies and the energy gap between each native and its associated lowest energy decoy in Table VI for these data sets shows that no decoy scores better than the native structure with the exception of 1bba and 1fc2C. The native Z-score for each protein is also presented in Table VI. The Z-score of a conformation is defined as

$$Z = \frac{E - \bar{E}}{\sigma} \tag{12}$$

where E is the energy of the particular conformation, and $\bar{E}$ and σ are the average and standard deviation of the distribution of scores in the set. The magnitude of the Z-score is an indication of how far that conformation's energy is separated from the most probable energies in the distribution. A negative Z-score indicates that the confor-

TABLE II. Proteins in the Local-Minima Decoy Sets†

PDB	Description	N_{res}	N_{decoy}	rmsd range
1bba	Pancreatic hormone (ave. NMR)	36	500	3.98–9.98
1ctf	(see Table I)	68	497	4.40–12.92
1fc2C	Fragment B of protein A (complexed to immunoglobin Fc)	43	500	5.12–9.42
1igd	3rd IgG-binding domain from streptococcal protein G	61	500	4.45–13.11
2cro	(see Table I)	65	500	5.09–14.13
2ovo	3rd domain of silver pheasant ovomucoid	56	347	5.64–14.14
4pti	(see Table I)	58	343	5.93–14.06

†Data from Samudrala and Levitt.[77] For details, see Table I footnote.

TABLE III. Proteins in the ROSETTA All-Atom Decoy Sets†

PDB	Description	N_{res}	N_{decoy}	rmsd range
1ksr	F-actin cross-linking gelation factor (NMR)	92	998	9.82–24.09
1lz1	Lysozyme	116	999	9.57–30.11
1ris	Ribosomal protein S6	92	999	7.38–22.43
1tul	Telokin-like protein	97	999	10.72–35.12
2acy	Acylphosphatase	92	994	8.83–28.23

†Data from Simons et al.[78] For details see Table I footnote.

mation's energy is lower than the average of the distribution. Except for 1bba and 1fc2C, the native is separated very significantly by over two standard deviations from the average of the distribution for each of the seventeen other decoy sets.

The two proteins whose natives are not successfully distinguished from their decoys are from the local-minima decoy sets. The protein 1bba is a pancreatic hormone. The native conformation is an average of 20 structures determined by NMR spectroscopy.[90] The protein contains a very flexible tail consisting of residues 33 through 36. The best-ranked decoy possesses roughly the same conformation for residues 1 to 32 as the native's, but its flexible region is folded toward the rest (see Fig. 1). This decoy, however, closely resembles one of the 20 structures used in the average[90] and could be considered as another candidate for the native conformation. Nevertheless, the native conformation is simply not very well defined for this protein.

The other protein whose native is not distinguished from its decoys by the OPLS-AA/SGB potential is 1fc2C (for both the local-minima and Skolnick decoy sets), which is the protein labelled "C" in the PDB file 1fc2. Protein "C" is fragment B of protein A and is complexed to immunoglobin Fc in the file 1fc2. Fragment B binds to immunoglobin Fc through a hydrophobic contact.[91] Burial of this hydrophobic region stabilizes the complex. With an exposed hydrophobic region, fragment B would not be as stable as a decoy that has a reduced exposed hydrophobic area. In fact, the best-ranked decoy (from LMDS), which has a lower energy than the native, has a 17.6% exposed hydrophobic surface area while the native fragment B has 21.3%. Furthermore, 1fc2C is a *fragment* of a larger protein, protein A, which is a "constituent of the cell wall of *Staphylococcus aureus*."[91] The complexed fragment of a protein cannot be expected to be stable unbound in solution; without being complexed to immunoglobin Fc, the fragment B conformation is probably not maintained in solution. For these reasons, 1fc2C does not make a good test case for the energy function.

The analysis of one of the proteins from the ROSETTA all-atom decoy sets requires some additional explanation. The "native" for the protein 1ksr consists of 20 structural models determined by NMR.[92] They differ from each other by under 1.4 Å rmsd. The OPLS-AA/SGB energies of the natives have a range of 128.22 kcal/mol. A comparison between these "native" energies and the decoy energies can be seen in Figure 2. For the following analysis, the lowest energy NMR structure (the 18th structure in the PDB file) is taken as the native. A similar protocol is used for the OPLS-AA calculations without the solvent term.

Decoy OPLS-AA/SGB energies and RMDSs

The lowest energy (i.e., best scoring) decoys' rmsd's are presented in Table VI. For the 4-state-reduced decoy set, the best decoys based on the OPLS-AA/SGB effective free energy function have rmsd's below 2.30 Å. Corresponding results are not seen for the local-minima and ROSETTA all-atom data sets because these sets simply do not possess very low rmsd decoy structures (see Tables II and III). Three proteins are shared by the 4-state-reduced and local-minima decoy sets. These proteins are 1ctf, 2cro, and 4pti. We combined the two sets for each of the three proteins creating a "super-set" that spans more of the conformational space. The results of the combined sets are shown in Table VII and Figures 3 and 4. The lowest energy decoy for 1ctf and 2cro come from the local-minima decoy sets and have an rmsd of 5.60 Å and 10.91 Å, respectively. (For 4pti, the lowest energy decoy is found in the 4-state-reduced decoy set.) We believe that lower energy decoys are found in the LMDS rather than in the 4-state-reduced decoy sets because the LMDS contains decoys that were energy-minimized in an all-atom force field while those in the 4-state-reduced decoy sets were only scored.[60,77] The minimization procedure obviously produced lower energy structures than the sampling over rotamer states with an off-lattice model. The higher rmsd's we found for the

TABLE IV. Proteins in the Skolnick Decoy Sets

PDB	Description	N_{res}	rmsd range
1cew	Proteinase inhibitor	108	8.41–18.25
1cis	Hybrid between chymotrypsin 2 and helix E from subtilisin Carlsberg (NMR)	66	6.12–13.93
1ctf	(see Table I)	68	4.43–13.97
1fas	Fasciculin 1—anti-acetylcholinesterase toxin	61	6.48–13.48
1fc2C	(see Table II)	44	4.05–10.35
1ftz	Fushi tarazu protein (NMR)	70	6.15–15.64
1gpt	γ-1-H-thionin (NMR)	47	6.23–11.51
1hmdA	Hemerythrin (1 subunit)	113	7.23–17.03
1shg	α spectrin (Sh3 domain)	57	6.53–12.60
1stfI	Inhibitor stefin B (also called crystatin B)	98	5.80–15.22
1tfi	Transcriptional elongation factor Sii (TFIIS, nucleic-acid binding domain) (NMR)	50	7.69–12.99
1thx	Thioredoxin-2	108	4.02–15.39
1tlk	Telokin	103	8.05–17.81
1ubi	Ubiquitin	76	6.28–14.19
256b	Cytochrome B_{562}	106	5.16–17.57
2aza	Azurin (oxidized)	129	5.37–16.74
2pcy	Apoplastocyanin	99	4.92–15.02
6pti	Trypsin inhibitor	57	7.74–13.94

†Data from Lu and Skolnick (personal communication). For details, see Table I footnote. The number of decoys, N_{decoy}, for each protein is 2,000.

lowest energy decoys from the LMDS are consistent with the results obtained by Samudrala and Levitt.[77]

The Spearman rank-order correlation coefficients, r_s, shown in Table VI provide a non-parametric measure of the correlation between the OPLS-AA/SGB energies and the structural similarity between the decoys and the native. Instead of finding a correlation between the values of the energies and the rmsd's, the Spearman coefficient is a measure of the correlation between the associated *ranks* of the energies and the rmsd's.[93] This avoids any potential problems due to scaling effects. When r_s is close to +1, a strong positive correlation between the ranks of the energies and rmsd's is present; when r_s is closer to 0, almost no correlation exists. Only the sets from the 4-state-reduced database show strong correlations between the energy and rmsd. This is apparent from plots of the energy vs. rmsd as seen by the black diamonds in Figures 3 and 4 for 1cft and 2cro, respectively. The rank-order coefficients ranged from 0.45 to 0.77, whereas for the other two sets, they only ranged from −0.03 to 0.28. The strong correlations between the energy and rmsd for the 4-state-reduced decoy sets have been observed by others.[11,13,14] We find, however, that with the addition of the LMDS sets for 1ctf and 2cro to the respective 4-state-reduced decoy sets (see Figs. 3 and 4), the Spearman coefficients for 1ctf and 2cro are negative showing an anti-correlation between the energy and the rmsd (see Table VII). The addition of the more competitive, high-rmsd decoys from the LMDS reveals that the high correlations for the results from the 4-state-reduced decoys appear to be due to the method by which they were generated, as mentioned above. The LMDS provides a greater challenge in validating the ability of a scoring function to distinguish the native fold from its decoys; nevertheless, the OPLS-AA/SGB effective free energy function is capable of accomplishing this for the LMDS.

The average native-like Z-score, $\bar{Z}_{nat\text{-}like}$, is obtained by averaging the Z-scores of the "native-like" decoys defined as having an rmsd less than 4.5 Å. Since the 4-state-reduced decoy sets include many native-like decoys, an average native-like Z-score is calculated to see how well the potential distinguishes the native-like distribution from the rest. (Since there are only a total of 5 native-like decoys in the local-minima decoy sets and none in the ROSETTA, no average Z-scores are calculated for these sets.) The negative average Z-scores ranging from −0.88 to −1.29 show that the native-like distributions are shifted toward lower energies than the rest, but this trend does not hold when the combined sets are considered as seen in Table VII. When the 4-state-reduced data set is combined with the local-minima set, the average Z-scores for 1ctf and 2cro drop to almost zero, showing that the native-like distributions from the 4-state-reduced decoy sets are not truly distinguishable from the rest.

Comparison of OPLS-AA/SGB results to other potentials

Park and Levitt[60] have evaluated six simple empirical scoring functions using the 4-state-reduced decoy sets examined in this work. A comparison between the native ranks and Z-scores calculated here with those obtained by Park and Levitt shows that the OPLS-AA/SGB energy function clearly outperforms the six empirical scoring functions examined in the Park and Levitt work.[60] Most

TABLE V. Average Ratios of the Total Number of Contacts Between Heavy Atoms in Each Decoy to the Total Number of Contacts in the Corresponding Native Structures and the Maximum Ratios[†]

PDB	Average ± SD	Maximum
4-state-reduced		
1ctf	0.97 ± 0.04	1.09
1r69	0.93 ± 0.04	1.08
1sn3	0.93 ± 0.05	1.06
2cro	0.95 ± 0.04	1.07
3icb	1.00 ± 0.03	1.09
4pti	0.92 ± 0.04	1.05
4rxn	0.94 ± 0.05	1.11
Total	0.95 ± 0.05	1.11
Leuitt's LMDs		
1bba	0.96 ± 0.03	1.03
1ctf	0.90 ± 0.02	0.96
1fc2	0.98 ± 0.02	1.06
1igd	0.89 ± 0.02	0.96
2cro	0.82 ± 0.02	0.88
2ovo	0.89 ± 0.03	0.98
4pti	0.81 ± 0.03	0.90
Total	0.90 ± 0.06	1.06
Rosetta all-atom decoys		
1ksr	0.86 ± 0.03	0.96
1lz1	0.79 ± 0.03	0.90
1ris	0.87 ± 0.03	0.96
1tul	0.86 ± 0.03	0.95
2acy	0.81 ± 0.03	0.90
Total	0.84 ± 0.05	0.96

[†]The structures have been minimized with the OPLS/AA-SGB potential. The contact distance is set to 4.5 Å.[86,87]

importantly, none of the empirical scoring functions examined by Park and Levitt was able to consistently rank first the native conformation, whereas the OPLS-AA/SGB energy function does. Tobi and Elber[9] performed calculations using their statistical potential on the 4-state-reduced decoy sets of 1ctf, 1r69, 1sn3, 2cro, 4pti, and 4rxn and the local-minima decoy sets of 1ctf, 1fc2C, 1igd, 2cro, and 2ovo. Excluding 1fc2C for which both scoring functions understandably fail (as mentioned above), they are able to rank the native with the best energy for six out of these ten proteins; the OPLS-AA/SGB energy function successfully ranks the native for all ten decoy sets.

A brief comparison to another all-atom force field with solvation can be made with the work by Dominy and Brooks[13] who performed calculations using the CHARMM19 force field in conjunction with the generalized Born (GB) solvation model on three (1r69, 2cro, and 3icb) of the seven 4-state-reduced decoy sets. For the proteins 1r69 and 2cro, they scored the native lowest in energy relative to the other decoys. For 3icb, a few of the decoys were found to have lower energies than the native. Their Z-scores for the native and ours (see Table VI) were similar: their Z-scores were 4.2, 3.3, and 2.2 for 1r69, 2cro, and 3icb, respectively[13] While the CHARMM19/GB model distinguishes the native in two of the three cases, the OPLS-AA/SGB does so in all three cases. A possible explanation for the somewhat improved performance of OPLS-AA/SGB is that SGB calculations of the solvation energies of protein conformations are closer to the results of accurate Poisson-Boltzmann calculations than GB. The difference in the accuracy between SGB and GB calculations is attributed to the correction terms included in the SGB model by Ghosh et al.[49]

Vacuum OPLS-AA Energy Function Calculations

It is instructive to evaluate the importance of individual components of the OPLS-AA/SGB energy function in recognizing native conformations. In order to determine the relative contributions of intramolecular and solvent electrostatic interactions, we have calculated the energy scores in vacuum with the OPLS-AA force field ($U_{OPLS\text{-}AA} \equiv U_{\text{int}}$) using the same protocol used for the calculations in continuum solvent. The minimized structures differed from the initial ones by the amounts shown in Table VIII. (The smaller deviation for the 4-state-reduced decoys is because the minimizations were limited to a maximum of 100 steps, which was not quite sufficient for completing OPLS-AA vacuum minimizations.) It should be noted that the OPLS-AA vacuum minimized structures deviate from their initial conformations by about 1 Å rmsd more on average than those minimized with the OPLS-AA/SGB potential. The results of the minimizations with the OPLS-AA energy function are summarized in Table IX. 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-scores have also significantly degraded (compare Tables VI and IX). Excluding 1bba and 1fc2C, the percentage of decoys with energies lower than the native's can be as high as 42.0% for a given decoy set. It is clear that for these decoy sets, 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 non-polar side-chains in the interior of the proteins and placing polar residues on the solvent exposed surface of the protein.[64,65,94,95] These tendencies are not represented well by the intramolecular potential in vacuum, which in general ranks equally the strength of interaction between two non-polar residues and between a non-polar 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, includes a hydrophobic interaction term 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 functions have been designed to recognize these features.[30,60,62,75,76] A model that does not take into account these solvation effects is likely to perform poorly in native fold recognition among large numbers of compact decoys.

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

TABLE VI. OPLS-AA/SGB Results for the 4-State-Reduced,[60] Local-Minima (LMDS),[77] and the ROSETTA All-Atom[78] Decoy Sets†

Decoy set	PDB	Z_{nat}	min($\Delta\Delta G_{eff}$)[a]	rmsd(Å)	r_s	$\bar{Z}_{nat\text{-}like}$
4-state-reduced	1ctf	−3.24	+65.55	1.69	0.69	−1.02
	1r69	−4.03	+107.16	2.30	0.72	−0.97
	1sn3	−4.23	+96.08	2.19	0.45	−1.05
	2cro	−3.69	+72.55	0.94	0.73	−0.88
	3icb	−2.18	+18.08	1.84	0.77	−1.29
	4pti	−4.53	+105.07	1.89	0.46	−1.17
	4rxn	−3.76	+92.06	2.16	0.61	−1.19
LMDS	1bba	3.29	−168.42	5.97	0.17	NA
	1ctf	−2.63	+14.19	5.60	0.28	NA
	1fc2C	0.68	−70.83	6.16	0.07	NA
	1igd	−4.06	+42.46	9.79	0.09	NA
	2cro	−3.32	+43.94	10.91	0.15	NA
	2ovo	−2.85	+11.74	11.61	0.03	NA
	4pti	−14.42	+869.64	9.71	−0.03	NA
ROSETTA	1ksr[b]	−3.07	+6.00	14.22	0.03	NA
	1lz1	−10.64	+286.51	13.94	0.12	NA
	1ris	−6.77	+124.98	13.58	0.07	NA
	1tul	−6.28	+122.11	15.24	0.07	NA
	2acy	−7.52	+134.82	17.15	0.00	NA

†Z_{nat} is the Z-score of the native; min($\Delta\Delta G_{eff}$) is the energy gap between the best scoring decoy and minimized native; the rmsd is between the best scoring decoy and the native; r_s is the Spearman rank-order correlation coefficient. For the 4-state-reduced sets, $\bar{Z}_{nat\text{-}like}$ is the average Z-score of native-like decoys (decoys with rmsd ≤4.5 Å).
[a]kcal/mol.
[b]18th NMR model in the PDB file is defined as the native.

bonds are 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.

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.[4,30,60,62,76] To assess how the different components of the OPLS-AA/SGB and vacuum OPLS-AA energy functions serve to discriminate the native from the decoys, the differences in these terms between the natives and their best-ranked decoys are examined. The average and minimum values on a per residue basis of the following combinations of terms are presented in the top half of Table X: the sum of all of the terms (the total energy); the "bonded" terms consisting of the bond stretch energy, the angle bending energy, and the total torsional energy (1,4-van der Waals, 1,4-Coulomb, and torsion angle energy); and the total electrostatics comprised of the non-bonded Coulomb energy and the SGB solvation energy, if it is calculated. The lower half of Table X shows the average and minimum values on a per residue basis of some of the individual terms of the energy functions tested. (It should be noted that 1bba and 1fc2C were not included in this analysis for the reasons previously discussed: see OPLS-AA/SGB Calculations on the 4-State-Reduced, LMDS, and ROSETTA Databases). Also, since 1ctf, 2cro, and 4pti are found in both the 4-state-reduced and local-minima decoy sets, they are only counted once by choosing *the* lowest-energy decoy from either set. For both OPLS-AA/SGB and vacuum OPLS-AA energy functions, only the van der Waals term clearly distinguishes the native from the lowest-energy decoy. The direct non-bonded Coulomb term favors the native over the decoy in most of the OPLS-AA/SGB minimization cases (i.e., 12 out of 14 cases). But for the vacuum OPLS-AA minimizations, the decoys have a lower Coulomb energy in 13 out of the 14 cases. The other terms of the function taken by themselves are unable to distinguish the natives from the decoys, as seen in Table X. This assessment is based on examining not only the average of the deviation in energy between the decoys and the natives, but also the minimum deviations, which indicate for these other terms that at least some of the decoys possess lower energies than the natives.

van der Waals and bonded components of the potential

As mentioned previously, based on the analysis of the differences in energy between the native and the best-ranked decoy, the van der Waals component of the potential appears to be sufficient for discriminating the native from the decoys. When considering all of the decoys, not just the one with the lowest total energy, very few exceptions appear. Only five decoys out of 12,832 have van der Waals energies from the OPLS-AA/SGB energy function lower than their respective native: the magnitudes of the differences in energy between these decoys and the natives

Fig. 1. Superposition of the native and best-ranked decoy of 1bba. The native structures and best-ranked decoy (no. 6452) of 1bba are superimposed over the non-variable region from residue 1 to 32. The native conformation is in gray and the decoy is in black. The variable region is visible in the top right.

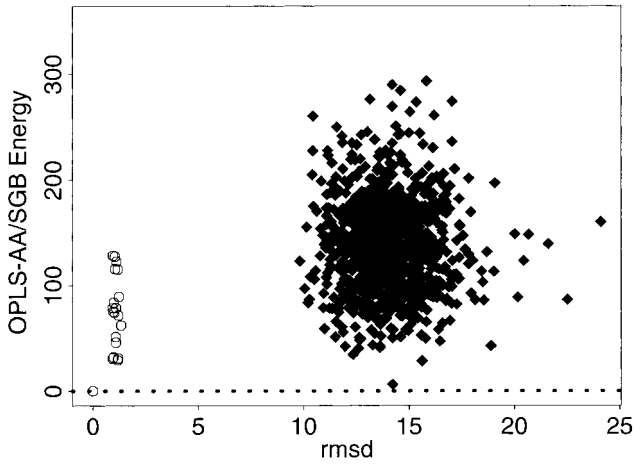

Fig. 2. OPLS-AA/SGB energies (kcal/mol) of the 1ksr decoys: The energies as calculated with the OPLS-AA/SGB energy function ($\Delta\Delta G_{eff}$) of the 20 "natives" (circles) and the decoys (solid diamonds) of the protein 1ksr are plotted with respect to their heavy-atom rmsd (in Å).

are less than 6.04 kcal/mol, as seen in Table XI. For four of these cases, the direct non-bonded Coulomb term of these decoys is large enough so that the native continues to have a lower total energy. In the fifth case, the SGB solvation energy term helps to favor the native over the decoy. In all five cases, the total electrostatics (the combination of the Coulomb and SGB solvation energies) serves to overcome

TABLE VII. OPLS-AA/SGB Results of the Combined 4-State-Reduced and Local-Minimum Decoy Sets†

PDB	Z_{nat}	min($\Delta\Delta G_{eff}$)[a]	rmsd (Å)	r_s	$\bar{Z}_{nat\text{-}like}$
1ctf	−1.70	+14.19	5.60	−0.40	−0.05
2cro	−2.06	+43.94	10.91	−0.48	0.03
4pti	−1.40	+105.07	1.89	0.73	−0.88

†Z_{nat} is the combined Z-score for the native; min($\Delta\Delta G_{eff}$) is the difference in energy between the best scoring decoy of both sets and the native; the rmsd is between the best scoring decoy and the native; $\bar{Z}_{nat\text{-}like}$ is the average Z-score of native-like decoys (i.e., their rmsd from the native is <4.5 Å); and r_s is the Spearman rank-order coefficient.
[a]kcal/mol.

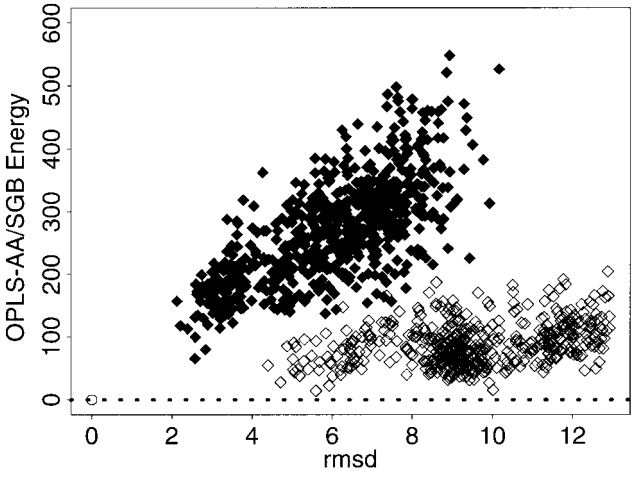

Fig. 3. OPLS-AA /SGB energies ($\Delta\Delta G_{eff}$ in kcal/mol) of the 1ctf decoys from the 4-state-reduced and local-minimum decoy sets. This is the combined decoy set of 1ctf using the 4-state-reduced decoy set (black diamonds) and the local-minimum decoy set (open diamonds). The energies are the difference between the decoy's and the native's. The native is indicated with a circle and its energy with a dotted line. The rmsd is in Å.

the favorable van der Waals energies of these decoys and to create total energies for these decoys that are less favorable than those of their natives. Nevertheless, it still would be tempting to claim that the van der Waals term would be enough to nearly always distinguish the native from the decoys.

It must be emphasized that the above van der Waals energies resulted from unrestrained minimizations *with the entire OPLS-AA/SGB energy function*. The key to adequately assessing how various individual terms of the function will perform is to carry out *unrestrained* minimizations using the corresponding potential. The minima of each individual term of the potential will differ in general from the minima of the other terms and from the minima of the total OPLS-AA/SGB energy function. For this reason, evaluating the native recognition ability of each individual term at a minimum of the total potential may be misleading. We illustrate this point with the 4-state-reduced decoys of 1ctf and the van der Waals term of the potential. Figure 5 shows the results of minimizing these decoys using only the van der Waals term and also a potential consisting of the van der Waals term and varying

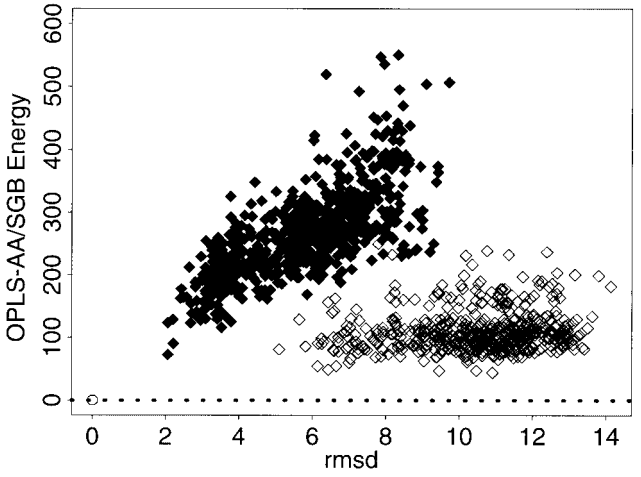

Fig. 4. OPLS-AA/SGB energies ($\Delta\Delta G_{eff}$ in kcal/mol) of the 2cro decoys from the 4-state-reduced and local-minimum decoy sets. See Figure 3 legend.

TABLE VIII. Average RMSD Between the OPLS-AA/SGB and OPLS-AA Vacuum Minimized and Initial Structures†

	Average δrmsd (Å)	
Decoy set	OPLS-AA/SGB	OPLS-AA vacuum
4-state-reduced	0.60 ± 0.12	0.53 ± 0.83
LMDS	0.66 ± 0.13	1.79 ± 0.34
ROSETTA	0.45 ± 0.10	1.98 ± 0.33

†For the two potential models tested, the averages of the rmsd between the minimized and initial structures (δrmsd) are shown for the 4-state-reduced, local minimum (LMDS), and ROSETTA all-atom decoy sets.

amounts of "bonded" terms (bond stretch, angle bending, torsion angle, 1,4-van der Waals, and 1,4-Coulomb). With only the van der Waals potential (Fig. 5, top left), 170 of the 631 decoys rank lower than the native in energy, whereas only one of the decoys rank lower when considering the van der Waals energy after minimizing with the entire potential (see Table XI). As more of the "bonded" terms are added to the van der Waals term, the situation improves for 1ctf such that the native once again scores better than the rest when all of the "bonded" terms are included with the van der Waals potential (Fig. 5, bottom left). A potential consisting of only "bonded" and van der Waals terms, however, does not consistently distinguish the native from a set of decoys as shown for 4rxn (Fig. 5, bottom right). Clearly the addition of the electrostatic terms is required in this case in order to discriminate the native successfully. In summary, it appears that the van der Waals term can distinguish the native successfully from the decoys only after the structures are minimized with the entire energy function.

Table X details how the various "bonded" terms (the bond stretch energy, the angle bending energy, and total torsional energy) fail to favor the native either by themselves or in combination. It is expected for the "bonded" terms to favor structures that are more open and extended instead of compact like the native. This is especially true for the torsional term, which is lower in energy for trans dihedral angles than for gauche. Open structures will have more trans dihedral angles than a compact structure like the native. Any decoys that are less compact than the native should have lower torsional energies. Since most decoys are less compact than the native (see Table V), we expect many of the decoy torsional energies to be lower than the natives'. This is supported by the torsional energy results for the best-ranked decoys: for the OPLS-AA/SGB minimizations, 9 out of 16 best-ranked decoys have lower torsional energies than the native (as seen in the negative average deviation in the torsional energy in Table X).

Electrostatic components of the potential

The direct Coulomb energy corresponding to the set minimized with the OPLS-AA/SGB energy function mostly favors the natives over the best-ranked decoys; in contrast, it mostly favors the decoys when the structures are minimized with the vacuum OPLS-AA energy function. This observation is related to the anti-correlation between the Coulomb and solvation energies; this relationship, which has been observed by others,[12–14,73,96] can be seen, for example, in the 4-state-reduced decoys of 3icb, as shown in Figure 6. When the total Coulomb energy is favorable, the solvation energy is unfavorable, and vice versa. Therefore, if a decoy has a more favorable Coulomb energy than the native, it will tend to have a more unfavorable solvation energy. During minimization, a balance is struck between the Coulomb and solvation terms of the potential.

Without including the solvation term in the energy function during minimization, as when using the vacuum OPLS-AA function, the strongly favorable Coulomb terms dominate the rankings as seen in Table X. The Coulomb energies resulting from the OPLS-AA vacuum minimizations tend to be significantly more favorable than those resulting from minimizations including solvation. This is seen, for example, in the results for the 1ctf decoys from the local-minimum database shown in Figure 7. In Figure 7, none of the OPLS-AA/SGB minimized energies of the decoys is lower than the native energy, but about 10% of the vacuum OPLS-AA minimized energies of the decoys are lower.

The OPLS-AA minimized structures deviate from the initial conformation by over 1 Å rmsd more when minimized with the vacuum potential than when minimized with OPLS-AA/SGB. By breaking down the deviation into charged and uncharged residue components (see Table XII), it can be seen clearly that most of the movement during the OPLS-AA minimization is associated with the charged residues. Without the solvation component in the energy function to screen charges that lie on the surface of the protein, the charge-charge interactions are strong enough to move charged side-chains, and thereby lower the Coulomb energy significantly.

This change in the Coulomb energy tends to be more pronounced for the decoys than for the natives as seen in Figure 7. In the case of the 1ctf local-minima decoy set, decoy no. 35665 (which is circled in Fig. 7) has a Coulomb energy 113 kcal/mol lower than the native when minimized with the vacuum OPLS-AA energy function but it

TABLE IX. Vacuum OPLS-AA Results Using the 4-state-reduced,[60] Local-Minima (LMDS),[77] and the ROSETTA All-Atom[78] Decoy Sets†

Decoy set	PDB	Z_{nat}	min($\Delta\Delta U_{int}$)[a]	rmsd (Å)	r_s	% < native
4-state-reduced	1ctf	−2.61	+43.68	6.49	0.40	0.0
	1r69	−3.18	+76.49	1.65	0.38	0.0
	1sn3	−3.05	+0.04	1.42	0.31	0.0
	2cro	−2.37	−35.12	0.93	0.60	0.1
	3icb	−0.66	−282.69	1.19	0.56	28.0
	4pti	−3.01	+37.53	6.21	0.29	0.0
	4rxn	−2.67	−8.95	1.60	0.60	0.1
LMDS	1bba	0.60	−238.82	8.00	−0.13	71.8
	1ctf	−1.51	−115.51	7.20	0.12	4.2
	1fc2C	−1.42	−97.09	6.86	0.29	7.2
	1igd	−0.77	−159.40	7.90	−0.13	23.0
	2cro	−0.24	−173.21	11.27	0.29	42.0
	2ovo	−0.11	−50.61	12.10	0.04	2.0
	4pti	−14.57	+817.84	11.34	−0.13	0.0
ROSETTA	1ksr[b]	−1.10	−177.32	24.24	0.01	0.8
	1lz1	−0.22	+81.93	15.51	0.16	0.0
	1ris	−1.38	−134.41	14.47	0.29	1.9
	1tul	−2.07	−181.91	14.48	0.28	1.3
	2acy	−0.92	−189.58	9.70	0.27	12.2

†min ($\Delta\Delta U$)[a] is the energy gap between the best scoring decoy and minimized native; "% <native" is the percentage of decoys with energies less than the native; all other columns are described in Table VI.
[a]kcal/mol.
[b]10th NMR model in the PDB file is defined as the native.

TABLE X. Average Difference in Energies Per Residue Between the Best-Scoring Decoys and the Natives†

	OPLS-AA/SGB		Vacuum OPLS-AA	
Energies	Ave. $\Delta\Delta E$[a]	(Min.)[a]	Ave. $\Delta\Delta E$[a]	(Min.)[a]
Total potential	1.096 ± 0.745	(0.065)	−1.184 ± 1.474	(−3.769)
Total intramol.[b]	−0.204 ± 0.826	(−2.170)	0.311 ± 1.175	(−2.129)
van der Waals	0.948 ± 0.416	(0.583)	1.134 ± 0.432	(0.489)
Electrostatics[c]	0.352 ± 0.700	(−1.073)	−2.628 ± 1.561	(−5.094)
Coulomb	3.133 ± 2.902	(−1.494)	−2.628 ± 1.561	(−5.094)
SGB solvation	−2.780 ± 2.454	(−8.693)		
Bond	0.013 ± 0.024	(−0.037)	0.059 ± 0.055	(−0.027)
Angle	0.015 ± 0.189	(−0.258)	0.175 ± 0.253	(−0.111)
Torsion	−0.232 ± 0.702	(−1.875)	0.076 ± 1.021	(−2.077)

†The averages for the difference in energies per residue between the best-scoring decoy and the native ($\Delta\Delta E$) for the total OPLS-AA/SGB and vacuum OPLS-AA energy functions are calculated over all proteins except for 1bba and 1fc2C. 1ctf, 2cro, and 4pti are counted only once from the combined 4-state-reduced and local minima decoy sets. Averages are also calculated for the individual components of these minimized energies and for various combinations of these components. The minimum values for all proteins are also reported.
[a]kcal/mol/residue.
[b]Sum of bond, angle, and torsion energies.
[c]Sum of the Coulomb and solvation energies.

has a Coulomb energy 437 kcal/mol *higher* than the native when minimized with the OPLS-AA/SGB function. Analysis of the deviation of the charged residues provides more insight as to why this decoy's Coulomb energy, minimized with the OPLS-AA function, is more favorable than the native's. While the native's charged residues deviated by 1.83 Å rmsd after minimization, the decoy's deviated to a greater extent, by 2.37 Å rmsd. This structural change leads to the formation of close, non-native charged pairs, that which the relaxation of the native structure does not produce, leading to a more favorable direct Coulomb energy for the decoy than for the native. Including the solvation term in the minimizations helps to dampen the effect of Coulomb potential: without it, the Coulomb energy cannot discriminate the native from the decoys when the structures are minimized without restraints.

The screening of direct Coulomb interactions can be achieved by simpler means than including the SGB reaction field in the potential. For instance, the dielectric constant can be adjusted to effectively screen the Coulomb

TABLE XI. OPLS-AA/SGB Energy Terms of Decoys With Lower van der Waals Energy Than the Native†

PDB	Decoy no.	rmsd (Å)	$\Delta\Delta U_{vdW}$ [a]	$\Delta\Delta U_{Coulomb}$ [a]	$\Delta\Delta G_{SGB}$ [a]	$\Delta\Delta G_{Electrostatics}$ [a]
1ctf	c12885	5.25	−3.89	662.70	−422.28	240.42
3icb	e13196	1.81	−5.11	0.49	60.10	60.59
	e13213	2.34	−1.33	135.16	−69.77	65.36
	g2089	6.49	−6.04	395.29	−228.72	166.57
4rxn	a5540	6.78	−1.13	933.94	−701.43	232.51

†For decoys with a non-bonded van der Waals energy lower than the native (all found in the 4-state-reduced decoy sets), the difference between theirs and the natives energies are shown for the van der Waals term ($\Delta\Delta U_{vdW}$), the non-bonded, direct Coulomb term ($\Delta\Delta U_{Coulomb}$), the SGB solvation energy ($\Delta\Delta G_{SGB}$), and the combined electrostatics ($\Delta\Delta G_{Electrostatics} = \Delta\Delta U_{Coulomb} + \Delta\Delta G_{SGB}$).
[a]kcal/mol.

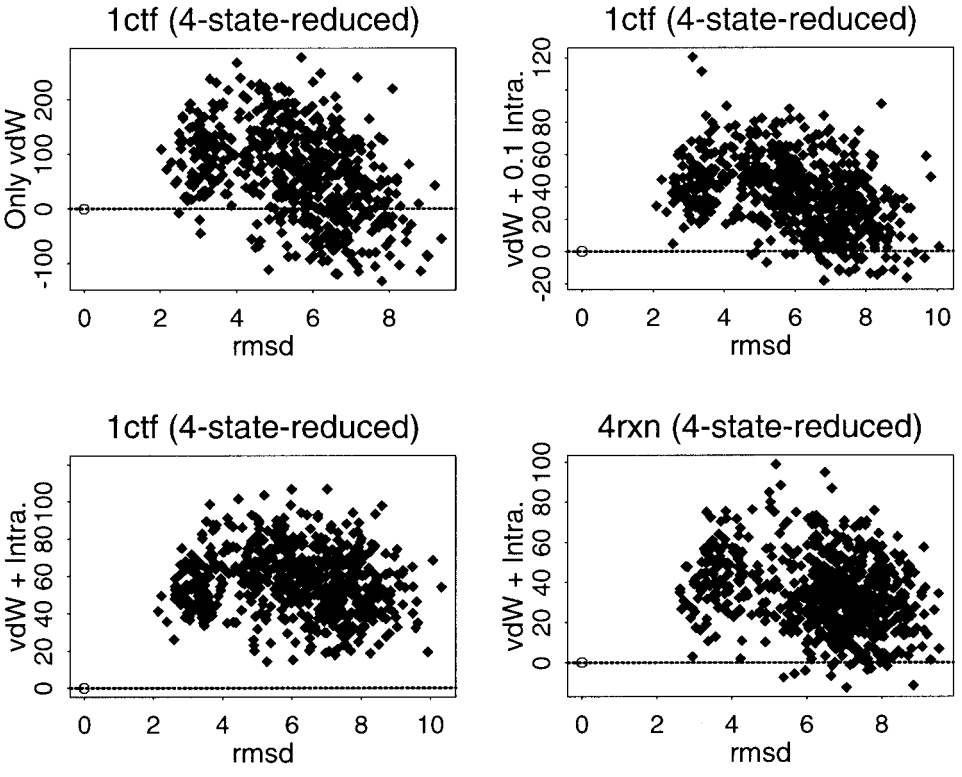

Fig. 5. The interplay between the van der Waals and intramolecular terms of the OPLS-AA potential for the 4-states-reduced decoy set of 1ctf and 4rxn. **Top left:** The differences in energy (kcal/mol) between the 1ctf decoys and the native when minimizing with only the van der Waals term ($\Delta\Delta\ U_{vdw}$) of the potential. **Top right:** The differences in energy (kcal/mol) for 1ctf when minimizing with the van der Waals and one-tenth of the combined intramolecular terms ($\Delta\Delta\ U_{vdW} + 0.1\Delta\Delta\ U_{intra.}$, where $\Delta\Delta\ U_{intra.} = \Delta\Delta\ U_{bond} + \Delta\Delta\ U_{angle} + \Delta\Delta\ U_{torsion}$). **Bottom left:** The differences in energy (kcal/mol) for 1ctf when minimizing with the van der Waals and all of the intramolecular terms ($\Delta\Delta\ U_{vdW} + \Delta\Delta\ U_{intra.}$). **Bottom right:** The differences in energy (kcal/mol) for 4rxn when minimizing with the van der Waals and all of the intramolecular terms. The native in all of these is indicated with a circle and its energy with a dotted line. The rmsd is in Å.

interactions, or a distance-dependent dielectric can be introduced to produce a similar effect. We have investigated these in conjunction with the OPLS-AA vacuum potential in a previous study. We found that none of these adaptations with the OPLS-AA force field are as effective as the OPLS-AA/SGB potential for discriminating a native fold from a set of decoys.[14]

Comparisons to other potential decompositions

Monge et al[30] have also studied various energy decompositions of the AMBER all-atom force field[72] supplemented by the GB solvation model.[46] They analyzed a data set of decoys of the proteins 1ctf, 1r69, and myoglobin (1mbo) generated by a simplified model employing fixed known secondary structure. The data sets contained less than 33 structures each. The authors observed that their energy function was capable of distinguishing the native from the non-natives for these small decoy sets. They also observed from their decomposition of the AMBER/GB energy function that the van der Waals energy alone scored the native best relative to the decoys. The natives had very low Coulomb energies with respect to most of the decoys (only

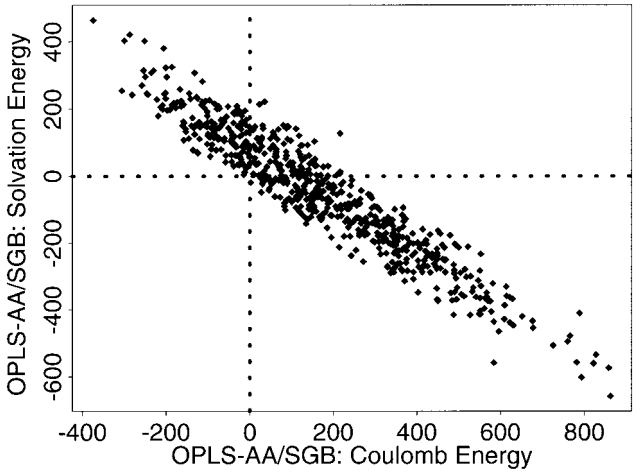

Fig. 6. Correlation plot between the OPLS-AA/SGB intramolecular Coulomb energy gap, $\Delta\Delta U_{Coulomb}$, (in kcal/mol) and the OPLS-AA/SGB solvation energy gap, $\Delta\Delta G_{SGB}$, (in kcal/mol) for the 3icb decoys.

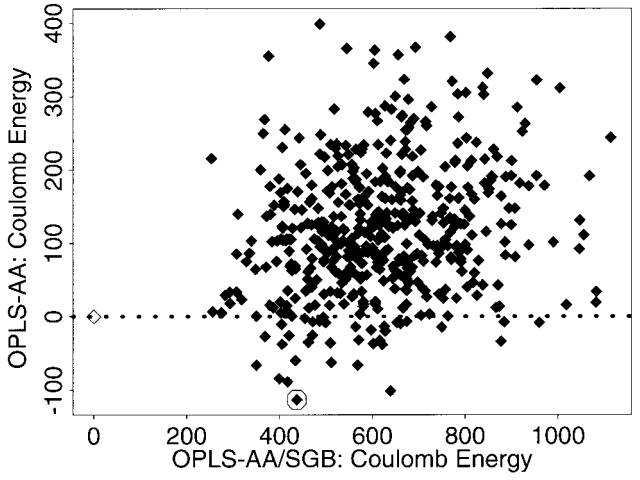

Fig. 7. Comparison of the Coulomb energies after minimizing with the OPLS-AA/SGB and OPLS-AA energy functions for the 1ctf structures from the local-minimum decoy set. Each black diamond represents one decoy structure, the abscissa and ordinate indicate the difference in Coulomb energy (in kcal/mol) from the native after minimizing with the OPLS-AA/SGB energy function and after minimizing with the vacuum OPLS-AA potential, respectively. The open diamond represents the native. The circled diamond is for the structure (no. 35665) with the lowest OPLS-AA Coulomb energy. The rmsd of this structure from the initial after OPLS-AA/SGB minimization is 0.32 Å and after vacuum OPLS-AA minimization is 1.89 Å.

4 out of a total of 75 decoys for all three proteins scored lower than the native) but had high solvation energies consistent with an anti-correlation between the direct Coulomb and reaction field solvation energies.[30]

Recently Petrey and Honig[12] 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 single decoy data sets[27,97] achieving a high level of discrimination. Based on an analysis of the individual terms of the potential function after restrained minimizations, they concluded that a simplified solvation model would be sufficient for distinguishing the native from misfolded conformations. They observed that the solvation energy often favors the misfolded conformations in the single decoy sets, concluding that the solvation energy is not useful in recognizing the native conformation. The Coulomb energy, however, tended to favor the native; therefore, they selected a simpler model consisting of only the intramolecular Coulomb energy and a hydrophobic residue burial estimator. This model was used to evaluate the 4-state-reduced decoy sets without minimizing the structures. In two cases (3icb and 4rxn), their method does not clearly rank the X-ray conformation favorably.[12] It is likely that, had Petrey and Honig carried out *unrestrained* minimizations using only the Coulomb potential for scoring the decoys without a compensating solvation term, they would have found a large number of decoys with Coulomb energies much lower than the natives' as we have observed in the vacuum OPLS-AA minimizations. We emphasize that when testing scoring functions on decoy datasets, the results can be misleading when restraints are applied to the minimization of the decoys in ways that would not be applied in the context of protein homology modeling or ab initio folding.

OPLS-AA/SGB and Vacuum OPLS-AA Calculations on the Skolnick Database

Since in excess of 36,000 structures are contained in the Skolnick decoy database analyzed by us (described in Table IV), it would have been too costly to minimize each structure with both the OPLS-AA and the OPLS-AA/SGB energy functions. In order to tackle this large database of decoys, it was decided to minimize each structure with the OPLS-AA potential followed by scoring the minimized structure with the complete OPLS-AA/SGB function. Even though the Coulomb energies are lower when minimized with only OPLS-AA than with OPLS-AA/SGB, the addition of the SGB solvation energy after minimization will tend to reduce the effect of non-native ion pairings.

Caution must be used when assessing the total energies of the native structures minimized using the OPLS-AA vacuum potential, because the deviation from the initial structure after minimization can be as high as 2 Å rmsd (for example, see Table VIII). To avoid having a reference for the OPLS-AA/SGB energy so different from the original native, the natives are minimized with the complete OPLS-AA/SGB energy function. When minimized in this manner, the final native structures differed from the initial ones on average by 0.76 Å rmsd.

The vacuum OPLS-AA scoring results for the structures in the Skolnick database are summarized in Table XIII. For five of the 17 proteins (excluding 1fc2C on grounds that it is an unreasonable protein to attempt to rank as explained in previous sections), there exist decoys which are lower in energy than the native. The average rmsd between the initial and minimized structures is 2.08 ± 0.24 Å. In Table XIV, the rmsd's between the initial native folds and their final minimized structures are shown along with the rmsd for only the charged and uncharged residues. Most of the change during the minimizations is due to the movement of charged residues just as in the OPLS-AA minimizations

TABLE XII. RMSD of Charged and Uncharged Residues Between Minimized and Initial Native Conformations†

		OPLS-AA/SGB			OPLS-AA		
Decoy set	PDB	$\delta rmsd_{chg}$	$\delta rmsd_{unchg}$	$\delta rmsd_{tot}$	$\delta rmsd_{chg}$	$\delta rmsd_{unchg}$	$\delta rmsd_{tot}$
LMDS	1ctf	0.35	0.28	0.31	1.83	1.06	1.45
	1igd	0.34	0.25	0.28	1.06	0.68	0.81
	2cro	0.41	0.40	0.40	1.14	0.84	0.94
	2ovo	0.48	0.38	0.40	1.73	1.21	1.34
	4pti	0.38	0.36	0.37	1.11	0.72	0.86
ROSETTA	1ksr	0.94	0.79	0.98	3.17	2.04	2.42
	1lz1	0.56	0.36	0.43	1.44	0.82	1.04
	1ris	0.48	0.41	0.44	1.51	0.92	1.18
	1tul	0.54	0.36	0.41	1.61	0.83	1.06
	2acy	0.37	0.27	0.30	1.42	0.80	1.01

†The rmsd (in Å) between the heavy atoms of minimized and initial native conformations of the proteins in the local-minimum and ROSETTA all-atom decoy sets are calculated using either charged (arginine, aspartate, glutamate, and lysine), uncharged, or all residues. These rmsd's are labelled $\delta rmsd_{chg}$, $\delta rmsd_{unchg}$, and $\delta rmsd_{tot}$, respectively.

of the local-minima and ROSETTA all-atom decoys (Table XII).

If the natives are scored with the OPLS-AA/SGB energy function in the same manner as the decoys (i.e., minimized with the OPLS-AA vacuum potential and subsequently scored with the OPLS-AA/SGB energy function), two of the 17 proteins (excluding 1fc2C) have one decoy that scores better than the native. With respect to the native energy, one of the 1tfi decoys has an OPLS-AA/SGB energy 13.52 kcal/mol lower, and one of the 1tlk decoys has an energy 0.57 kcal/mol lower. However, the minimized natives in these cases deviate significantly from the initial structures (by over 1.4 Å rmsd) due to the OPLS-AA vacuum minimizations. Most of the difference between minimized and initial structures was due to the movement of the charged sidechains that deviated by as much as 2.0 Å rmsd from the initial coordinates. Since the native structures for the 4-state-reduced, local-minima, and ROSETTA all-atom decoy sets changed very little from the initial structures (on average by only 0.54 Å rmsd) when minimized with the OPLS-AA/SGB function, the minima for the natives of 1tfi and 1tlk when minimized with the OPLS-AA vacuum potential clearly are not representative of the local minima corresponding to the OPSL-AA/SGB energy function. In light of this, the natives have been minimized with the entire OPLS-AA/SGB energy function. When these native OPLS-AA/SGB energies are compared to the OPLS-AA vacuum minimized decoy structures scored with the SGB solvation term included in the total energy function, all of the natives (excluding 1fc2C) once again are discriminated from the rest as seen in Table XV. Also, the native Z-scores improve considerably. Very few low-rmsd decoys, however, are present in the database.

There are some decoy data sets contained in the Skolnick database that deserve some extra attention. These correspond to the proteins 1cis, 1ftz, 1gpt, and 1tfi whose native conformations were determined via NMR. The OPLS-AA/SGB energies of the NMR native models and of the decoys of these proteins are shown in Figure VIII as a function of the rmsd from one of the native models. For 1cis and 1tfi, all of the native models are lower in energy than the decoys; for 1gpt and 1ftz, one of the native models has an energy higher than some of the decoys. The native models of 1ftz deviate widely from each other: the rmsd among them range up to 10 Å. A core structure, from residues 8 to 53, is roughly conserved in these native structures (with an α-carbon rmsd of less than 2 Å among the structures) while the remaining residues vary significantly (this is shown clearly in fig. 5(b) in Qian et al.[98]). Many of the corresponding decoys are more similar (lower rmsd) in structure to the lowest energy native model, yet all but one of the 20 native models have lower energies than the decoys. The likely explanation is that none of the decoys has the same structural similarity of the conserved core region as the native structures. In this conserved region

TABLE XIII. Vacuum OPLS-AA Results for the Skolnick Decoy Sets†

PDB	Z_{nat}	$\min(\Delta\Delta U_{int})$[a]	rmsd(Å)
1cew	−2.96	−7.05	15.08
1cis[b]	−6.21	306.58	11.43
1ctf	−3.47	2.26	12.66
1fas	−3.21	63.43	10.13
1fc2C	−1.48	−81.18	7.86
1ftz[b]	−3.25	−19.55	14.73
1gpt[b]	−0.97	−109.29	8.95
1hmdA	−4.25	106.50	9.04
1shg	−4.79	126.65	10.00
1stfI	−4.35	153.08	12.12
1tfi‡	−2.50	−28.84	9.61
1thx	−4.76	180.62	13.21
1tlk	−2.65	−64.36	12.26
1ubi	−3.63	55.64	13.45
256b	−4.86	165.81	14.37
2aza	−9.60	544.51	6.78
2pcy	−5.11	174.60	5.19
6pti	−4.43	113.57	11.67

†See Table IX footnote for details. The native Z-scores, Z_{nat}, are calculated excluding any energies 1000 kcal/mol above the native.
[a]kcal/mol.
[b]The PDB NMR model with the lowest energy was defined as the "native."

TABLE XIV. RMSD of Charged and Uncharged Residues Between the Vacuum OPLS-AA Minimized and Initial Native Conformations for the Skolnick Decoy Sets[†]

PDB name	$\delta rmsd_{chg}$	$\delta rmsd_{unchg}$	$\delta rmsd_{tot}$
1cew	2.05	1.01	1.38
1cis	1.33	1.08	1.19
1ctf	1.69	0.84	1.29
1fas	1.29	0.89	1.02
1ftz	1.48	1.26	1.36
1gpt	1.88	0.99	1.39
1hmdA	1.39	1.00	1.12
1shg	1.53	0.82	1.13
1stfI	1.33	0.90	1.02
1tfi	1.81	0.99	1.27
1thx	1.34	0.87	1.02
1tlk	1.98	1.03	1.44
1ubi	1.36	0.85	1.05
256b	1.67	0.83	1.23
2aza	1.29	0.82	0.97
2pcy	1.76	0.73	1.08
6pti	1.85	0.76	1.20

[†]The rmsd (in Å) between the heavy atoms of minimized and initial native conformations of the proteins in the Skolnick decoy sets are calculated using either charged (arginine, aspartate, glutamate, and lysine) ($\delta rmsd_{chg}$), uncharged ($\delta rmsd_{unchg}$), or all residues ($\delta rmsd_{tot}$).

TABLE XV. OPLS-AA/SGB Results Using the Skolnick Decoy Sets[†]

PDB	Z_{nat}	$\min(\Delta\Delta G_{eff})^a$	rmsd (Å)
1cew	−9.19	374.00	14.96
1cis[b]	−6.66	234.42	7.46
1ctf	−4.88	81.27	12.19
1fas	−4.47	81.82	9.25
1fc2C	−0.45	−75.81	7.73
1ftz[b]	−4.71	92.11	14.73
1gpt[b]	−4.31	60.95	8.95
1hmdA	−3.99	75.17	16.17
1shg	−9.71	295.92	10.00
1stfI	−12.25	559.08	12.00
1tfi[b]	−7.82	291.09	10.92
1thx	−11.99	579.27	5.09
1tlk	−9.01	339.43	13.13
1ubi	−5.28	128.03	13.42
256b	−8.30	310.20	14.49
2aza	−9.50	420.26	8.49
2pcy	−10.42	398.89	8.28
6pti	−5.01	91.11	11.49

[†]See Table VI for details. The energies are the final results after first minimizing with the OPLS-AA potential followed by scoring the minimized structure with the complete OPLS-AA/SGB energy function.
[a]kcal/mol.
[b]The PDB NMR model with the lowest energy was defined as the "native."

(residues 8 to 53), none of the decoys have an α-carbon rmsd less than 2 Å and only 119 out of 2,000 have an α-carbon rmsd within 3 Å. It is encouraging that the OPLS-AA/SGB energy function can distinguish those structures that do possess this conserved region from those that do not.

CONCLUSIONS

The OPLS-AA molecular mechanics energy function coupled with the Surface Generalized Born solvation model is found to be capable of discriminating the native structures of several proteins from a large number of decoys generated using a variety of methods from different groups. The ability of the OPLS-AA/SGB effective free energy function 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.[1] Lazaridis and Karplus[11] have shown that the CHARMM protein force field combined with their EEF1 effective solvation free energy model[99] is able to discriminate native conformations from decoy sets for 6 proteins from the 4-state-reduced data sets[60] and in pairwise comparisons between the native and a misfolded decoy for an additional 22 proteins.[27] 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 using an empirical solvation model that has the form of an effective pair potential and is simpler than the SGB solvation model.

While other studies by Monge et al.[30] and Petrey and Honig[12] have pointed out that individual terms of the potential like the van der Waals or Coulomb energy are effective in distinguishing the native from its decoys, we find that these terms alone are insufficient when the energy-minimized decoys are considered. It is a balance that is achieved among these terms acting together in the effective free energy function, which leads to the most effective discrimination between the native and the decoys.

Even though the OPLS-AA/SGB energy function can be used effectively to recognize the native conformation from a set of decoy structures, further tests are required to explore the features of the minimum occupied by the native relative to other minima. One question of interest: are there minima much lower than the native's minimum observed here? And if so, are the conformations associated with these lower minima structurally similar to the native (i.e., the rmsd is within the resolution of X-ray crystallography)? To address these questions, an exploration of the conformational phase space using the OPLS-AA/SGB energy function is needed in the vicinity of the native folds and the lowest-energy decoy structures. One very efficient and promising method to accomplish this is by simulated annealing using hybrid Monte Carlo[100,101] sampling on the native and on a set of decoys with the lowest energies. These decoys already occupy low-energy regions in the conformational space and provide a good starting point in the search for deeper minima. A second method for exploring the adjacent minima around the native and lowest-energy decoys is the Monte Carlo-minimization approach developed by Li and Scheraga for grappling with the multiple minima problem.[102] A modification of this method has been successfully employed to exhaustively search the minima of a reduced potential for a simplified protein

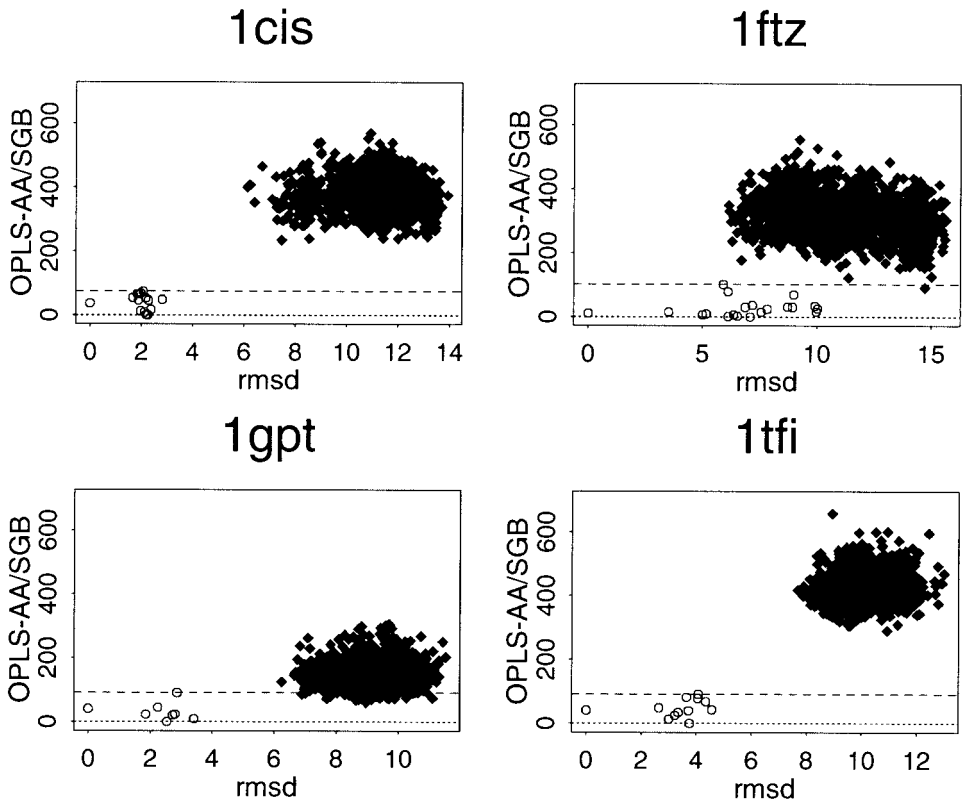

Fig. 8. OPLS-AA/SGB energies of the 1cis, 1ftz, 1gpt, and 1tfi natives and decoys. The OPLS-AA/SGB energies (in kcal/mol) of the NMR native models (circles) and of the decoys (solid diamonds) of the proteins 1cis, 1ftz, 1gpt, and 1tfi are plotted with respect to the rmsd between the structures and one of the native models. The plotted energies are the difference between the structure and lowest-energy "native." The dotted line in each plot marks the lowest energy of the "natives" and the dashed line marks the highest. The rmsd is in Å.

model.[103] While a thorough exploration of the minima possessed by an all-atom potential is intractable, a partial search around selected minima (i.e., the native and the lowest-energy decoys) is possible and would provide valuable information about the performance of the potential. Work along these lines is currently in progress.

The ability to discriminate native-like protein conformations from non-native folds is one of the fundamental problems in theoretical protein structure prediction. Knowledge-based scoring potentials, designed for reduced atomic models and derived from a combination of structural and thermodynamic data, are currently the most widely used. It is often assumed that such potentials are inherently better than all-atom force-fields. This work demonstrates the effectiveness of a physics-based all-atom potential (the OPLS-AA/SGB potential) for discriminating the native structures from a large set of decoys constructed by several groups. Thanks to their reduced atomic representation, 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

We thank Dr. Lynne Reed Murphy for help with some of the calculations. We also thank Prof. Jeffrey Skolnick and Dr. Hui Lu for providing us with their large database of decoys.

REFERENCES

1. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
2. Sippl MJJ. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. Mol Biol 1990;213: 859–883.
3. Casari G, Sippl MJJ. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structure of globular proteins is able to identify native folds. Mol Biol 1992;224:725–732.
4. Wodak SJ, Rooman MJ. Generating and testing protein folds. Curr Opin Struct Biol 1993;3:247–259.
5. Jones DT, Thornton JM. Potential energy functions for threading. Curr Opin Struct Biol 1996;6:210–216.
6. Plaxco K W, Riddle D S, Grantcharova V, Baker D. Simplified proteins: minimalist solutions to the 'protein folding problem'. Curr Opin Struct Biol 1998;8:80–85.
7. Hao M, Scheraga HA. Designing potential energy functions for protein folding. Curr Opin Struct Biol 1999;9:184–188.
8. Osguthorpe DJ. Ab Initio protein folding. Curr Opin Struct Biol 2000;10:146–152.
9. Tobi D, Elber R. Distance–dependent, pair potential for protein folding: results from linear optimization. Proteins 2000;41:40–46.

10. Eyrich V, Standley D, Felts A, Friesner R. Protein tertiary structure prediction using a branch and bound algorithm. Proteins 1999;35:41–57.
11. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1999;288:477–487.
12. Petrey D, Honig B. Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci 2000;9: 2181–2191.
13. Dominy BN, Brooks III CL. Identifying native-like protein structures using physics-based potentials. J Comput Chem 2002; 23:147–160.
14. Wallqvist A, Gallicchio E, Felts AK, Levy RM. Detecting native protein folds among large decoy sets with the OPLS all-atom potential and the surface generalized Born solvent model. Adv Chem Phys 2002;120:459–486.
15. Rick SW, Berne BJJ. The aqueous solvation of water: A comparison of continuum methods with molecular dynamics. Am Chem Soc 1994;116:3949–3954.
16. Levy RM, Gallicchio E. Computer simulation with explicit solvent: Recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. Annu Rev Phys Chem 1998;49:531–567.
17. Rashin AA, Bukatin MA. Magnitude of hydration entropies of nonpolar and polar molecules. J Phys Chem 1994;98:386–389.
18. Sitkoff D, Sharp KA, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.
19. Tannor DJ, Marten B, Murphy R, Friesner RA, Sitkoff D, Nicholls A, Ringnalda M, Goddard III WA, Honig B. Accurate first principle calculation of molecular charge distributions and solvation energies from ab initio quantum mechanics and continuum dielectric theory. J Am Chem Soc 1994;116:11875–11882.
20. Sitkoff D, Ben-Tal N, Honig B. Calculation of alkane to water solvation free energies using continuum solvent models. J Phys Chem 1996;100:2744–2752.
21. Hawkins G, Cramer C, Truhlar D. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. J Phys Chem 1996;100:19824–19839.
22. Gallicchio E, Zhang L, Levy RM. The SGB/NP hydration free energy model based on the surface generalized Born solvent reaction field and novel nonpolar hydration free energy estimators. J Comput Chem 2001;
23. Srinivasan J, Cheatham III TE, Cieplak P, Kollman PA, Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. J Am Chem Soc 1998;120:9401–9409.
24. Onufriev A, Bashford D, Case DA. Modification of the generalized Born model suitable for macromolecules. J Phys Chem B 2000;104:3712–3720.
25. Bashford D, Case DA. Generalized Born models of macromolecular solvation effects. Annu Rev Phys Chem 2000;51:129–152.
26. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
27. Holm L, Sander CJ. Evaluation of protein models by atomic solvation preference. Mol Biol 1992;225:93–105.
28. van Gunsteren WF, Luque FJ, Timms D, Torda AE. Molecular mechanics in biology: From structure to function, taking account of solvation. Annu Rev Biophys Biomol Struct 1994;23:847–863.
29. Smith PE, Pettitt BM. Modeling solvent in biomolecular systems. J Phys Chem 1994;98:9700–9711.
30. Monge A, Lathrop EJP, Gunn JR, Shenkin PS, Friesner RA. Computer modeling of protein folding: Conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
31. Schaefer M, van Vlijmen HW, Karplus M. Electrostatic contributions to molecular free energies in solution. Adv Protein Chem 1998;51:1–57.
32. Vorobjev YN, Hermans J. ES/IS: estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. J. Biophys Chem 1999;78:195–205.
33. Gilson MK, Honig B. Calculations of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. Proteins 1988;4:7–18.
34. Bashford D, Karplus M. pKa's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model: Biochemistry 1990;29:10219–10225.
35. Rashin AA. Hydration phenomena, classical electrostatics, and the boundary element method. J Phys Chem 1990;94:1725–1733.
36. Sharp KA, Honig B. Electrostatic interactions in macromolecules: Theory and applications. Annu Rev Biophys Chem 1990;19:301–332.
37. Warshel A, Åqvist J. Electrostatic energy and macromolecular function. Annu Rev Biophys Chem 1991;20:267–298.
38. Gilson MK, Davis ME, Luty BA, McCammon JA. Computation of electrostatic forces on solvated molecules using Poisson-Boltzmann equation. J Phys Chem 1993;97:3591–3600.
39. Honig B, Sharp K, Yang A-S. Macroscopic models of aqueous solutions: Biological and chemical applications. J Phys Chem 1993;97:1101–1109.
40. Mohan V, Davis ME, McCammon JA, Pettitt BM. Continuum model calculations of solvation free energies: Accurate evaluation of electrostatic contributions. J Phys Chem 1992;96:6428–6431.
41. Simonson T, Brünger AT. Solvation free energies estimated from macroscopic continuum theory: an accuracy assessment. J Phys Chem 1994;98:4683–4694.
42. Ösapay K, Young WS, Bashford D, Brooks III CL, Case DA. Dielectric continuum models for hydration effects on peptide conformational transitions. J Phys Chem 1996;100:2698–2705.
43. Edinger SR, Cortis C, Shenkin PS, Friesner RA. Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of the Poisson-Boltzmann equation. J Phys Chem B 1997;101:1190–1197.
44. Born MZ. Volumes and heats of hydration of ions. Physik 1920;1:45–48.
45. Hirata F, Rejfern P, Levy RJ. Viewing the Born model for ion hydration through a microscope. Quant Chem 1988;15:179–188.
46. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
47. Jean-Charles A, Nichols A, Sharp K, Honing B, Tempczyk A, Hendrickson TF, Still WC. Electrostatic contributions to solvation energies: Comparison of free energy perturbation and continuum calculation. J Am Chem Soc 1991;113:1454–1455.
48. Qiu D, Shenkin PS, Hollinger FP, Still WC. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J Phys Chem A 1997;101: 3005–3014.
49. Ghosh A, Rapp CS, Friesner RA. Generalized Born model based on a surface integral formulation. J Phys Chem B 1998;102: 10983–10990.
50. Roux B, Simonson T. Implicit solvent models. Biophy Chem 1999;78:1–20.
51. Zhang L, Gallicchio E, Friesner R, Levy RM. Solvent models for protein-ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. J Comp Chem 2001;22:591–607.
52. Liu Y, Beveridge DL. Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a generalized Born/solvent accessibility solvation model. Proteins: 2001; In press.
53. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
54. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
55. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
56. Miyazawa S, Jernigan RJ. Residue-residue potentials with a favorable contact pair term and and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256: 623–644.
57. Wallqvist A, Smythers GW, Covell DG. Identification of cooperative folding units in a set of native proteins. Protein Sci 1997;6: 1627–1642.
58. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. Protein 1999;36:357–369.
59. Covell D, Jernigan R. Conformations of folded proteins in restricted spaces. Biochemistry 1990;29:3287–3294.

60. 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.
61. Ozkan B, Bahar I. Recognition of native structures from complete enumeration of low-resolution models with constraints. Proteins 1998;32:211–222.
62. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
63. 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.
64. Novotny J, Bruccoleri R, Karplus M. An analysis of incorrectly folded protein models. Implications for structure prediction. J Mol Biol 1984;177:787–818.
65. Novotny J, Rashin AA, Bruccoleri R. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 1988;4:19–30.
66. Chiche L, Gregoret LM, Cohen FE, Kollman PA. Protein model structure evaluation using the solvation free energy of folding. Proc Natl Acad Sci USA 1990;87:3240–3243.
67. Vila J, Williams RL, Vasquez M, Scheraga HA. Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. Proteins 1991;10:199–218.
68. Williams RL, Vila J, Perrot G, Scheraga HA. Empirical solvation models in the context of conformational energy searches: Application to bovine pancreatic trypsin inhibitor. Proteins 1992;14:110–119.
69. Wang Y, Zhang H, Li W, Scott RA. Discriminating compact non-native structures from the native structure of globular proteins. Proc Nat Acad Sci USA 1995;92:709–713.
70. Wang Y, Zhang H, Scott RA. A new computational model for protein folding based on atomic solvation. Protein Sci 1995;4: 1402–1411.
71. Vieth M, Kolinski A, Brooks III CL, Skolnick J. Prediction of the folding pathways and structure of the GCN4 leucine zipper. J Mol Biol 1994;237:361–367.
72. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagone G, Profeta S, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984;106:765–784.
73. Vorobjev YN, Almagro JC, Hermans J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit continuum solvent. Proteins 1998;32:399–413.
74. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
75. 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.
76. 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.
77. Samudrala R, Levitt M. Decoys 'R' Us: a database of incorrect conformations to improved protein structure prediction. Protein Sci 2000;9:1399–1401.
78. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;S3:171–176.
79. Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. Protein data bank. In: Allen FH, Bergerhoff G, Sievers R, editors. Crystallographic databases: information content, software systems, scientific applications. Cambridge: Data Commission of the International Union of Crystallography; 1987.
80. Kitchen DB, Hirata F, Westbrook JD, Levy R, Kofke D, Yarmush M. Conserving energy during molecular dynamics simulations of water, proteins, and proteins in water. J Comput Chem 1990;11: 1169–1180.
81. Lee MR, Duan Y, Kollman PA. Use of MM-PB/SA in estimating the free energies of proteins: application to native, intermediates, and unfolded villin headpiece. Proteins 2000;39:309–316.
82. Zhang L, Gallicchio E, Levy RM. Implicit solvent models for protein-ligand binding: Insights based on explicit solvent simulations. In: Pratt LR, Hummer G, editors. Simulation and theory of electrostatic interactions in solution. AIP conference proceedings 492. Melville, NY: American Institute of Physics; 1999.
83. Thornton JM. Disulphide bridges in globular proteins. J Mol Biol 1981;151:261–287.
84. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 2001;98:10125–10130.
85. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 1992;226:507–533.
86. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Nat Acad Sci USA 1992;89: 2536–2540.
87. Mirny L, Domany E. Protein fold recognition and dynamics in the space of contact maps. Proteins 1996;26:391–410.
88. Levitt M, Hirshberg M, Sharon R, Daggett V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput Phys Commun 1995;91:215–231.
89. 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.
90. Li X, Sutcliffe MJ, Schwartz TW, Dobson CM. Sequence-specific ^{1}H NMR assignments and solution structure of bovine pancreatic polypeptide. Biochemistry 1992;31:1245–1253.
91. Deisenhofer J. Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from *Staphylococcus aureus* at 2.9- and 2.8-Å resolution. Biochemistry 1981;20:2361–2370.
92. Fucini P, Renner C, Herberhold C, Noegel AA, Holak TA. The repeating segments of the F-actin cross-linking gelation factor (ABP-120) have an immunoglobulin-like fold. Nat Struct Biol 1997;4:223–230.
93. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in FORTRAN: The art of scientific computing, 2nd ed. Cambridge: Press Syndicate of the University of Cambridge; 1992.
94. Dill KA. Dominant forces in protein folding. Biochemistry 1990; 29:7133–7155.
95. Dill KA. Folding proteins: Finding a needle in a haystack. Curr Opin Struct Biol 1993;3:99–103.
96. Smith KC, Honig B. Evaluation of the conformational free energies of loops in proteins. Proteins 1994;18:119–132.
97. Mosimann S, Meleshko R, James MNG. A critical assessment of comparative molecular modeling of tertiary structures of proteins. Proteins 1995;23:301–317.
98. Qian YQ, Furukubo-Tokunaga K, Resendez-Perez D, Müller M, Gerhing W J, Wüthrich K. Nuclear magnetic resonance solution structure of the *fushi tarazu* homeodomain from *Drosophila* and comparison with the *Antennapedia* homeodomain. J Mol Biol 1994;238:333–345.
99. Lazaridis T, Karplus M. Effective energy function for protein in solution. Proteins 1999;35:133–152.
100. Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid Monte Carlo. Phys Lett B 1987;195:216–222.
101. Irbäck A. Hybrid Monte Carlo simulation of polymer chains. J Chem Phys 1994;101:1661–1667.
102. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci USA 1987;84:6611–6615.
103. 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.