PROTEINS: Structure, Function, and Genetics 41:518–534 (2000)

Discrimination of Near-Native Protein Structures From Misfolded Models by Empirical Free Energy Functions

David W. Gatchell, Sheldon Dennis, and Sandor Vajda*
Department of Biomedical Engineering, Boston University, Boston, Massachusetts

ABSTRACT **Free energy potentials, combining molecular mechanics with empirical solvation and entropic terms, are used to discriminate native and near-native protein conformations from slightly misfolded decoys. Since the functional forms of these potentials vary within the field, it is of interest to determine the contributions of individual free energy terms and their combinations to the discriminative power of the potential. This is achieved in terms of quantitative measures of discrimination that include the correlation coefficient between RMSD and free energy, and a new measure labeled the minimum discriminatory slope (MDS). In terms of these criteria, the internal energy is shown to be a good discriminator on its own, which implies that even well-constructed decoys are substantially more strained than the native protein structure. The discrimination improves if, in addition to the internal energy, the free energy expression includes the electrostatic energy, calculated by assuming nonionized side chains, and an empirical solvation term, with the classical atomic solvation parameter model providing slightly better discrimination than a structure-based atomic contact potential. Finally, the inclusion of a term representing the side chain entropy change, and calculated by an established empirical scale, is so inaccurate that it makes the discrimination worse. It is shown that both the correlation coefficient and the MDS value (or its dimensionless form) are needed for an objective assessment of a potential, and that together they provide much more information on the origins of discrimination than simple inspection of the RMSD-free energy plots. Proteins 2000;41:518–534.**

Key words: free energy function; molecular mechanics; implicit solvation; structure-based potential; protein structure determination

INTRODUCTION

Most algorithms that attempt to predict or refine protein structure are based on the thermodynamic hypothesis, and reduce the search for the native conformation to the minimization of a potential approximating the free energy of the solvated protein. Thus, an important step in this approach is the development of free energy functions that are computationally feasible and yet accurate enough to discriminate native or near-native conformations from the immense number of alternative structures generated in a conformational search.

In early studies, the free energy was frequently replaced by the conformational energy calculated from a molecular mechanics potential, which provided meaningful polypeptide geometries, correctly accounting for the effects of covalent bonding, excluded volumes, and coulombic electrostatics. However, molecular mechanics alone can not provide a valid thermodynamic description of stable, compact protein folds, and may be unable to distinguish between native and misfolded structures.[1,2] The accuracy of calculations based on molecular mechanics has been substantially improved by adding explicit models of the solvent, and by performing molecular dynamics or Monte Carlo simulations (see, e.g., Karplus and Petsko[3]). However, due to their computational burden, these methods remain restricted to the simulation of peptides or very small proteins,[4–7] and to the analysis of proteins that are near their native states.

The need for a more efficient free energy evaluation stimulated the development of structure-based, empirical potentials that implicitly include solvation effects, and are derived from a statistical analysis of known protein structures (for reviews see references[2,8,9,10]). Such potentials have been defined both for simplified protein models with one or two interaction sites per residue,[8,11–14] which are primarily used for fold recognition,[12,15,16] and for all-atom models[17–20] to be used in protein structure refinement and folding simulations.

In recent years, there has been growing interest in the use of free energy functions that combine a molecular mechanics potential with implicit solvation and entropic terms.[2,14,21–27] The most direct way to calculate electrostatic contributions to the solvation free energy is based on the well-known continuum electrostatics (CE) model in which both long-range and local electrostatic interactions are obtained by solving the linearized Poisson-Boltzmann equation.[28,29] The total solvation free energy is then obtained by adding a term representing the free energy required to form a cavity (generally assumed to be propor-

Grant sponsor: National Science Foundation; Grant number: DBI-9904834; Grant sponsor: Department of Energy; Grant number: DE-F602-96ER62263.

*Correspondence to: Sandor Vajda, Dept. of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215. E-mail: vajda@enga.bu.edu

Received 7 March 2000; Accepted 4 August 2000

tional to the change in the solvent accessible surface area), and a term that describes the solute-solvent van der Waals interactions.[30–34] The primary shortcoming of the continuum electrostatics method is its sensitivity to the position of the dielectric boundary and, hence, to small structural perturbations.[32,35,36] Vorobjev et al.[33] removed this problem by a multistep approach that generates a number of protein microstates in short molecular dynamics simulations with explicit solvent, solves the Poisson-Boltzmann equation for each, and calculates the average solvation free energy.

We study free energy potentials that can be evaluated even for a very large number of conformations, and, hence, restrict consideration to simple electrostatic and solvation models. A number of models have been developed that express the solvation free energy in terms of accessible surface areas,[37–40] hydration shell volumes,[41,42] group volumes,[25,43] atomic contacts,[44] and Born radii.[45] The two free energy expressions we study differ only in the solvation term, the first based on the atomic solvation parameter (ASP) model,[37] the second on a structure-based atomic contact energy (ACE).[46] A similar function, with a solvation model based on group volumes, has been recently developed by Lazaridis and Karplus.[25,43]

The main goal of this paper is to determine the contributions of individual terms and their combinations to the discriminatory power of combined free energy potentials. While we have already discussed the complementary roles of molecular mechanics and solvation,[24] here a much more rigorous analysis is carried out using quantitative measures of discrimination. A traditional quality measure is the correlation coefficient between RMSD and the free energy. However, as we will show, the correlation coefficient does not necessarily detect important outliers, and we therefore introduce a new measure called the minimum discriminatory slope (MDS). Because MDS depends on the scaling of the free energy function, we also propose two of its non-dimensional versions as measures of discrimination when the scaling is uncertain. As we will show, the new measures provide important information that would be difficult to obtain on the basis of the correlation coefficient alone. In particular, the internal energy, calculated by a molecular mechanics potential[47] is a relatively good discriminator on its own. However, in terms of MDS the discrimination substantially improves when a solvation free energy term is added to the internal energy. We also show that the classical ASP model performs somewhat better than the atomic contact potential, and a frequently used method for calculating the loss of side chain entropy is not accurate enough to improve the discrimination.

MATERIALS AND METHODS

Empirical Free Energy Functions

The discriminatory functions we study in this article combine molecular mechanics with empirical solvation/entropic terms to approximate the free energy, G, of the system consisting of a protein, either in a fixed conformation, or as an ensemble of equienergetic structures (e.g., with different side chain rotamers),[2,48] and the solvent, the latter averaged over its own degrees of freedom. Let G_o denote the free energy of the above system in a reference conformation. The free energy difference, $\Delta G = G - G_o$, is then decomposed according to

$$\Delta G = \Delta G_{conf} + \Delta G_{solv} \quad (1)$$

The conformational free energy change, ΔG_{conf}, is defined by $\Delta G_{conf} = \Delta E_{conf} - T\Delta S_{conf}$, where ΔE_{conf} is the conformational energy change of the protein in a reference medium (a nonpolar liquid or vacuum), and ΔS_{conf} is the change in conformational entropy. The conformational energy is calculated by a molecular mechanics potential (Version 19 of Charmm with polar hydrogens[47]). In the most general case, ΔE_{conf} includes electrostatic, van der Waals, and internal energy terms,

$$\Delta E_{conf} = \Delta E_{elec} + \Delta E_{vdW} + \Delta E_{int} \quad (2)$$

where the internal (bonded) energy, ΔE_{int}, is the sum of bond stretching, angle bending, torsional, and improper terms,

$$\Delta E_{int} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{dihedral} + \Delta E_{improper} \quad (3)$$

ΔG_{solv} is the solvation free energy defined as the free energy of transferring the proteins from the reference medium into water.[24,49,50] It includes both local solute-solvent interactions and long-range effects due to dielectric screening and polarization.

In order to assure rapid free energy evaluation, we restrict consideration to two simple continuum solvation models. The first is the classical atomic solvation parameter (ASP) model that describes the local solute-solvent interactions by the linear expression $\Delta G_{solv} = \Sigma\ \sigma_i \Delta A_i$, where ΔA_i denotes the difference between the solvent accessible surface area of the ith atomic group in the given conformation, and the surface area of the same group in the reference state. The coefficient σ_i is the corresponding ASP.[37,38,40] The second solvation model is based on the atomic level extension of the Miyazawa-Jernigan potential.[11] The local interactions are given by the sum $\Sigma_i \Sigma_j e_{ij}$, where e_{ij} denotes the atomic contact energy (ACE) of interacting atoms i and j, and the sum is taken over all atom pairs that are less than 6 Å apart.[46] By definition, e_{ij} is the effective free energy change when a solute-solute bond between two atoms of type i and j, respectively, is replaced by solute-solvent bonds. The e_{ij} values have been obtained for 18 atom types by converting frequencies of structural factors observed in protein structures into atomic contact energies (ACEs).[46] The free energy for solvating amino acid side chains obtained by this method correlates to a high degree ($r = 0.975$) with the experimentally determined free energies of transferring the side chains between water and octanol.[46] In both models, the long-range electrostatic screening is accounted for by the distance-dependent dielectric of $\epsilon = 4r$.

As mentioned, ΔG_{conf} generally describes an ensemble of equienergetic structures rather than a single conformation, and hence includes the conformational entropy change

$T\Delta S_{conf}$. As a first-order approximation, we assume that the backbone in any folded conformation has the same conformational entropy, and hence all of the entropy difference, ΔS_{conf}, comes from the side chains, i.e., $\Delta S_{conf} = \Delta S_{sc}$. The calculation of the side chain conformational entropy loss is based on an empirical entropy scale[51] in which the maximum conformational entropy, S_{sc}, of each side chain was calculated by the classical expression $S_{sc} = -R \sum_i p_i \ln(p_i)$, where p_i denotes the probability of the ith rotamer. In the free energy calculation, we assume that the entire side chain entropy is lost, i.e., $\Delta S_{sc} = S_{sc}$, if the change ΔA_t in the total solvent accessible surface area of the side chain is more than 60% of its standard side chain surface area A_t^*.[52] Otherwise the entropy loss is scaled according to $\Delta S_{sc} = \alpha S_{sc}$, where $\alpha = \Delta A_t/(0.6A_t^*)$.

Since the solvent is not modeled explicitly, the calculation of solvent-solute van der Waals (vdW) interactions requires an approximation. The most straightforward strategy is to include these interactions in the solvation term ΔG_{solv}. This can be accomplished by using vacuum as the reference medium,[25,30,31,38,53] because in this case the solvation free energy (i.e., the free energy of transferring a molecule from vacuum into water) includes establishing van der Waals interactions. The solute-solute van der Waals interactions are obtained using the usual Leonard-Jones 6-12 formula. However, since the solute-solvent and solute-solute van der Waals terms are based on very different models, the free energy function will be very sensitive to small perturbations in the atomic coordinates, resulting in a rugged free energy surface.

We have shown that a relatively smooth discriminatory function can be obtained by performing van der Waals normalization.[24] This idea is based on the concept of van der Waals cancellation, which assumes that the solute-solute and solute-solvent interfaces are equally well packed, and hence the van der Waals contacts lost between solvent and solute are balanced by new solute-solute contacts formed upon protein folding.[54–56] While assuming van der Waals cancellation is a relatively good approximation for native proteins and protein-protein complexes,[57] it clearly does not apply to the misfolded conformations in the decoy sets that may have steric clashes and cavities. We remove or, at least, reduce these problems by a van der Waals normalization procedure prior to the free energy calculations. Van der Waals normalization implies that all conformations are minimized for a moderate number of steps (usually 200 steps of adopted basis Newton-Ralphson (ABNR) minimization in Charmm[47]), the structure with the lowest van der Waals energy is selected, and all other structures are further minimized to attain the same van der Waals energy value. The additional minimization removes steric clashes or cavities, and improves the validity of assuming the van der Waals cancellation.

The van der Waals cancellation implies that we can remove both the solute-solvent and the solute-solute van der Waals terms from the free energy function. Indeed, neither the ASP model with the Eisenberg-McLachlan parameters,[37] nor the ACE solvation model include van der Waals contributions, as both are based on the free energy of transfer from a hydrophobic medium (hydrocarbon and protein interior, respectively) into water. The removal of the van der Waals term reduces the conformational energy to

$$\Delta E_{conf} = \Delta E_{elec} + \Delta E_{int} \quad (4)$$

As we will show, better discrimination is obtained if, after van der Waals normalization, the electrostatic energy, ΔE_{elec}, is recalculated by assuming that all side chains, including Arg, Lys, Glu, and Asp, are in a non-ionized state, where we use the notation ΔE_{nelec} for the obtained value. In addition, we have shown that Equation 1 yields free energies of protein unfolding that are in good agreement with experimentally determined values if we assume van der Waals cancellation and neutral side chains, and neglect the change in the internal energy.[57]

The free energy functions we use in this study have been originally developed to solve conformational search problems in homology modeling.[24] More recently, Lazaridis and Karplus[25,43] described a potential that is similar in many respects. In particular, the molecular mechanics part of the free energy is based on Charmm,[47] and the charged groups are set to neutral. There are, however, a number of substantial differences. First, Lazaridis and Karplus use the distance-dependent dielectric of $\epsilon = r$ rather than $\epsilon = 4r$. Second, the solvation free energy is calculated by a Gaussian solvent exclusion model. Third, no van der Waals normalization is used, and the solvent-solute van der Waals interactions are taken into account in the solvation term. In spite of these differences, the functions provide very similar discrimination.[43]

Protein Decoys

A powerful tool for the analysis of discriminatory functions is offered by decoy sets that include the native and near-native conformations of a protein together with a large ensemble of misfolded models. Such decoys have been extensively employed in the development of potentials for protein fold discrimination and structure prediction,[18,19,33,58] and carefully generated decoy sets are available on the web, including loop libraries (http://prostar.carb.nistgov/PDec/PDecInfo.html), and protein models (http://dd.stanford.edu/), thereby facilitating the comparison of potentials developed by different groups.

In this work we study the decoy sets generated by Park and Levitt[58] for seven small proteins (http://dd.stanford.edu/). For each protein, the authors selected a set of loop residues, and enumerated all conformations obtained by the variations of four backbone conformational states for the selected residues, thereby ensuring proper sampling and a large diversity of resulting structures. The procedure yields over one million conformations, which were reduced by excluding structures that were not compact or had an excessive number of interpenetrating residues. After the application of these filters, about 500 structures were retained for each protein and made available on the Web. Lazaridis and Karplus studied the same decoys using the free energy expression we have already described.[43]

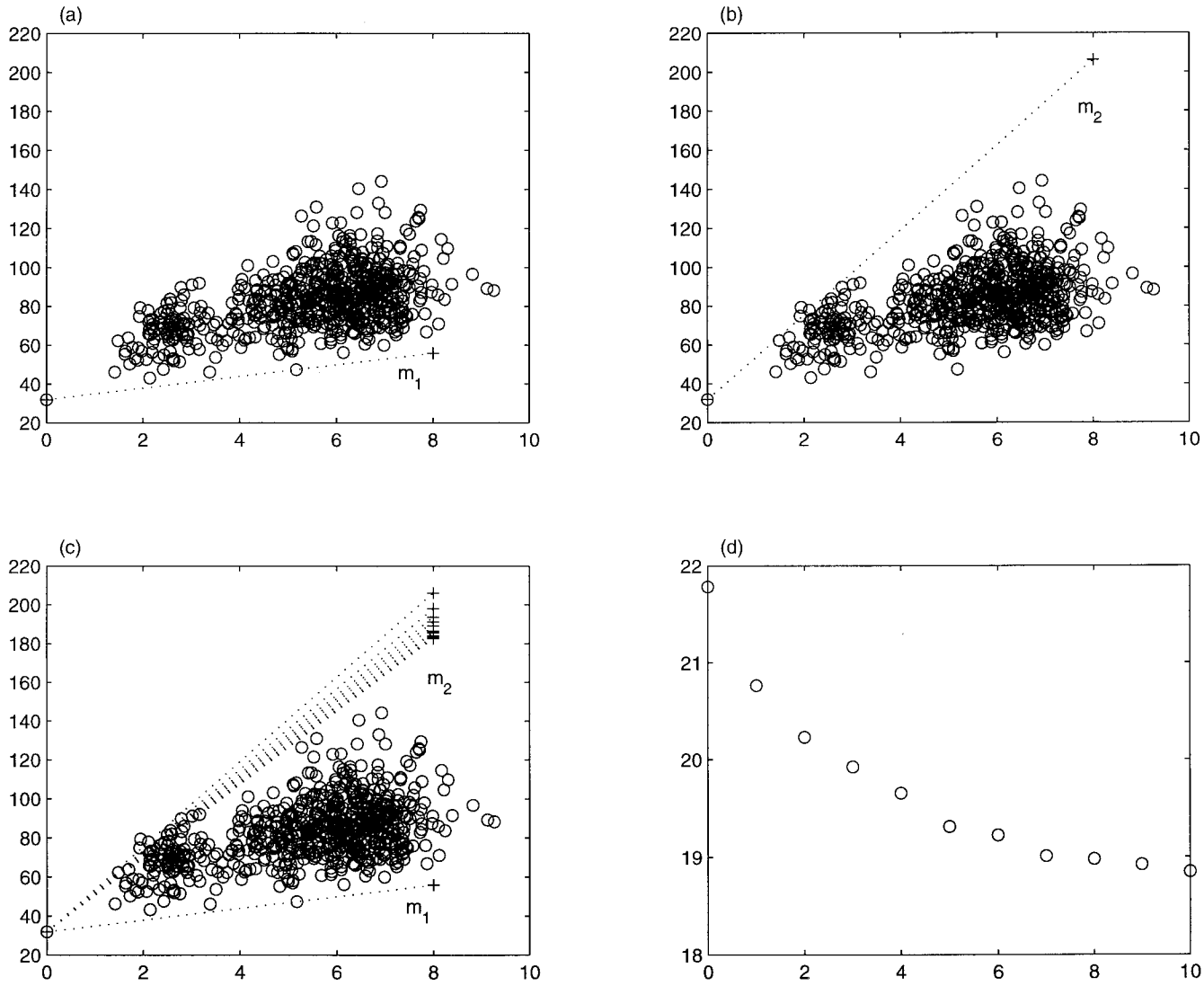

Fig. 1. Quantitative measures of discrimination. The region of interest is 2 Å < RMSD < 8 Å. **a:** The line shown represents the lower boundary of the decoy set, and its slope m_1 is the Minimum Discriminatory Slope (MDS). **b:** The line shown represents the upper boundary of the decoy set with all points included, and its slope m_2, in combination with the MDS, is used to define the Discrimination Ratio (DR), m_1/m_2. **c:** Refining the upper bound by the omission of k points, k = 1, 2, . . . , 10. **d:** Dependence of the slope m_2 on the number of points omitted. The change in m_2, $[m_2(k) - m_2(k\text{-}1)]/m_2(k)$, is less than 1% for $k \geq 7$.

Quality of Discrimination in Decoy Studies

Correlation Coefficient

The correlation coefficient, r, between the RMSD and the free energy ΔG is a traditional measure of discrimination. By definition, a correlation coefficient ranges from -1, corresponding to a linear relationship with a negative slope, to $+1$, which represents a linear relationship with a positive slope. Strong positive correlation means that for a conformation with relatively high free energy one can expect relatively high RMSD, and vice versa, but it does not exclude the existence of potential outliers and false positives, i.e., conformations with low free energy but high RMSD. In fact, adding a few false positives to a large set of decoys leaves the correlation coefficient virtually unaffected. Since the objective of a discriminatory potential is the elimination of such outliers, the correlation coefficient cannot be used as the sole measure of discrimination quality, and we introduce additional measures to better account for potentially important data points.

Minimum Discriminatory Slope (MDS)

Figure 1a shows the free energy vs. RMSD for 1ctf, one of the proteins considered by Park and Levitt.[58] We restrict consideration to an RMSD interval, $[D_L, D_U]$, because no empirical potential can correctly rank conformations below a certain RMSD, and the RMSD-free energy relationship usually becomes flat for large RMSD values. In this work, we use $D_L = 2$ Å and $D_U = 8$ Å. The line shown in Figure 1a passes through the point corresponding to the native structure, and represents the lower boundary of all

points, i.e., within the interval $[D_L, D_U]$ no data point is below the line. Consider any two conformations with RMSD values D_1 and D_2, and free energies ΔG_1 and ΔG_2, respectively. Let $D_1 \leq D_2$. In the ideal case, i.e., when the free energies of both structures are close to their lowest possible values for the given RMSD, $\Delta G_2 - \Delta G_1 \geq m_1 * (D_2 - D_1)$ kcal/mol, where m_1 denotes the slope of the lower boundary. In view of this inequality, we define m_1 as the minimum discriminatory slope or MDS. A larger MDS means a larger free energy difference between two conformations, and hence a better discrimination between near-native and non-native structures. The most important property of MDS, at least in comparison to the correlation coefficient, is that the MDS becomes negative if there is even a single conformation with free energy below that of the native structure, whereas the correlation coefficient is minimally affected by such a point.

MDS is usually expressed in kcal/mol/Å units, and yields free energy gaps in kcal/mol. Thus, MDS is not invariant to the scaling of the potential, and cannot be used for comparing two free energy functions that may be on different thermodynamic scales or are given in arbitrary units. Therefore, we introduce two dimensionless alternatives as follows.

Dimensionless Discriminatory Slope (DDS)

An obvious method to nondimensionalize the MDS is to replace the free energy ΔG by the Z-score, $Z = (\Delta G - \overline{\Delta G})/\sigma_{\Delta G}$, where $\overline{\Delta G}$ and $\sigma_{\Delta G}$ denote the mean and standard deviation of the free energy values for the decoy set. We also introduce a nondimensional measure of RMSD, defined as $D = (RMSD - \overline{RMSD})/\sigma_{RMSD}$, where $\overline{RMSD}$ and σ_{RMSD} denote the mean and the standard deviation of RMSD values on the decoy set, and then define the dimensionless discriminatory slope (DDS) to be $DDS = Z/D$. A larger DDS value for a decoy set indicates better discrimination, independently of the units of the energy function. Notice that $DDS = MDS * \sigma_{RMSD}/\sigma_{\Delta G}$.

Discrimination Ratio (DR)

This nondimensional measure also accounts for the slope, m_2, of the upper boundary shown in Figure 1b, and is defined by $DR = m_1/m_2$. As for the MDS, DR is defined on an RMSD interval of $[D_1, D_2]$. Since $m_1 \leq m_2$, we always have $DR \leq 1.0$, and $DR = 1$ if and only if all points are on a single line with a positive slope. The latter property is similar to that of the correlation coefficient. However, unlike the correlation coefficient, DR becomes negative if there is even a single false positive.

Notice that while even a single false positive may render a potential function worthless, the presence of a few false negatives is not a major problem. In fact, the value of m_2 is generally more meaningful if the worst outliers on the false negative side are not taken into account in the calculation. To implement this filter, we define a point as an outlier if disregarding it reduces the value of m_2 for the decoy set by more than 1%. Thus, in selecting m_2 we suggest constructing a sequence of upper bounds omitting 0, 1, 2, . . . k points (Fig. 1c), and monitoring the change in the m_2 values as shown in Figure 1d until the aforementioned criterion is satisfied.

RESULTS

The seven decoy sets have been subjected to van der Waals normalization as described in the Methods. Figures 2 through 6 show plots of different combinations of free energy terms as functions of RMSD for the seven normalized decoy sets. Figure 2 shows the internal energy, which includes the usual molecular mechanics terms ΔE_{bond}, ΔE_{angle}, $\Delta E_{dihedral,}$ and $\Delta E_{improper}$, but not the van der Waals energy. Figures 3 and 4 show the non-bonded components of the free energy function using solvation terms based on the Atomic Contact Energy (ACE) and Atomic Solvation Parameters (ASP) models, respectively. This part of the free energy includes the solvation term ΔG_{solv}, and the electrostatic energy ΔE_{nelec}, calculated with non-ionized side chains. Figures 5 and 6 show the total free energies, i.e., the sums of the internal energy, the nonbonded contributions from Figures 3 and 4, respectively, and the entropic term $-T\Delta S_{sc}$, which accounts for the change in side-chain entropy. As we will show, the latter does not improve discrimination, i.e., does not improve r and MDS, and will be excluded from future versions of the target function.

Notice that all free energy values are shown without normalization or changing the reference state. For example, the internal energy values in Figure 2 are the ones given by the Charmm calculation for the individual decoys. Thus, for the internal energy the reference state is defined by $\Delta E_{int} = 0$. It is easy to adopt the native structure as the reference state, i.e., to show only the differences in all figures, and, therefore, this is how discrimination will be assessed by the various measures. However, in the plots we wanted to show typical values of the free energy components for the small globular proteins considered in this study.

We measured the quality of discrimination that can be achieved by individual free energy terms and their various combinations. The individual free energy terms are the internal energy, ΔE_{int}, the electrostatic energy ΔE_{elec} (or its neutral version, ΔE_{nelec}), and the desolvation free energy $\Delta G_{solv}(ACE)$ or $\Delta G_{solv}(ASP)$, calculated by the ACE or ASP models, respectively. An additional entropic free energy term, $-T\Delta S_{sc}$, was discussed previously. The combinations we consider are the nonbonded free energies, $\Delta E_{nelec} + \Delta G_{solv}(ACE)$ and $\Delta E_{nelec} + \Delta G_{solv}(ASP)$; two sums of internal and solvation free energies $\Delta E_{int} + \Delta G_{solv}(ACE)$ and $\Delta E_{int} + \Delta G_{solv}(ASP)$; two sums that include all free energy terms except the side chain entropy, i.e., $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE)$ and $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP)$; and finally the total free energy expressions with ACE-based and ASP-based solvation, i.e., $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE) - T\Delta S_{sc}$ and $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP) - T\Delta S_{sc}$.

Tables I, II, and III describe the quality of discrimination in terms of the correlation coefficient, the minimum discriminatory slope (MDS), its dimensionless variant (DDS), and the discrimination ratio (DR). We list these

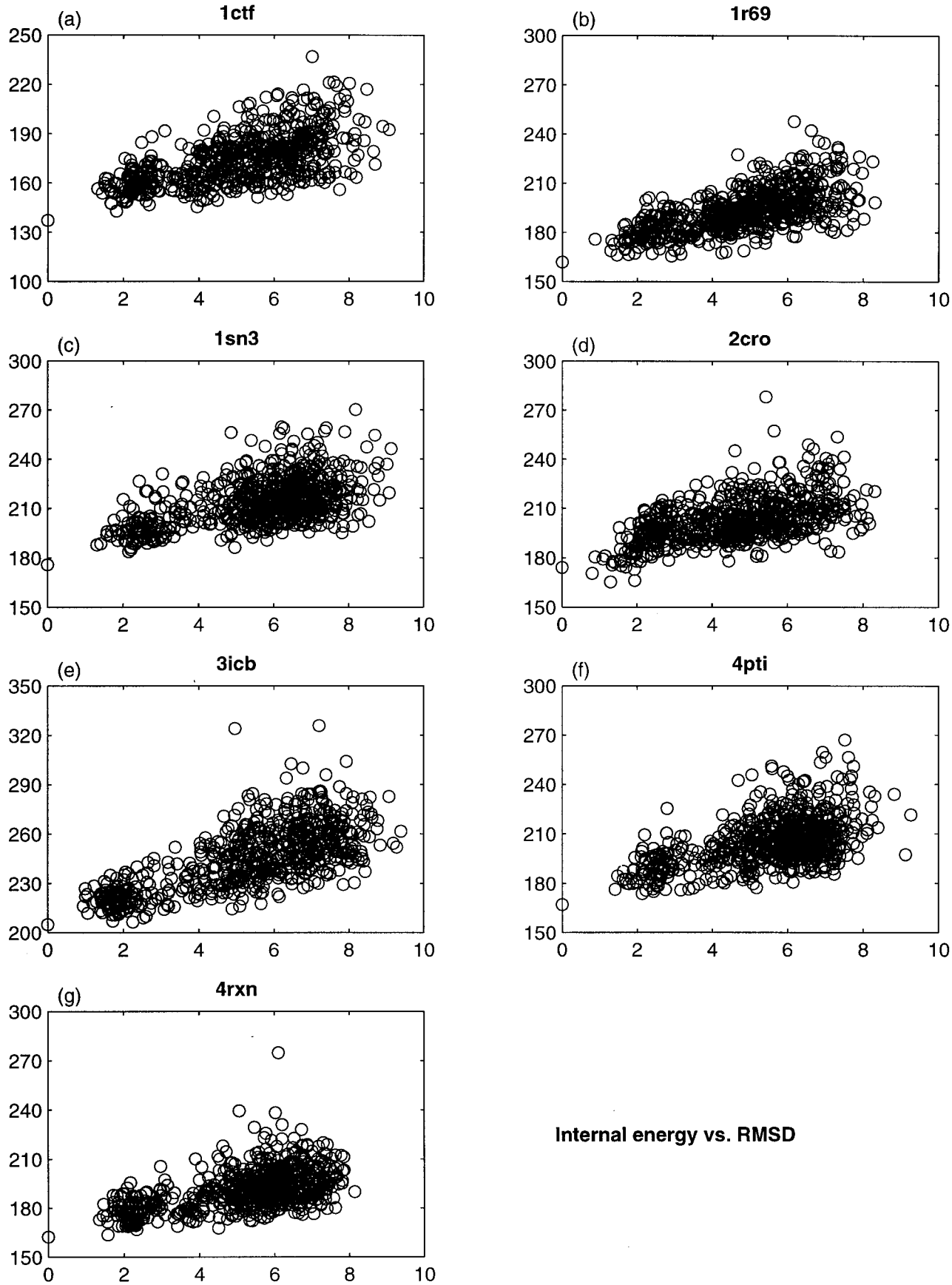

Fig. 2. Internal energy ΔE_{int} as a function of the cRMSD for the seven decoy sets of Park and Levitt.[58] The circle at cRMSD = 0 represents the native structure.

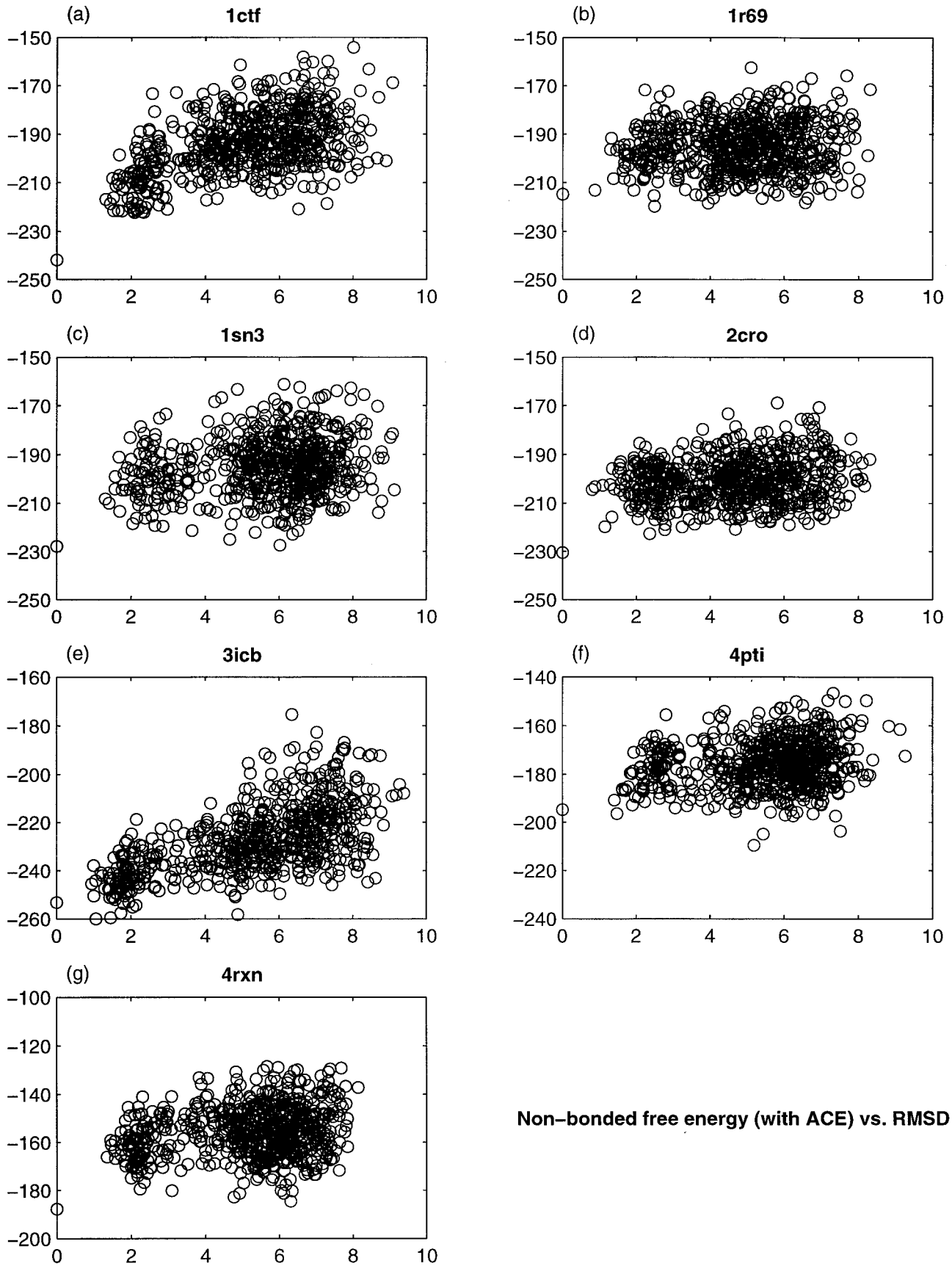

Fig. 3. The nonbonded part of the free energy, $\Delta G_{solv}(ACE) + \Delta E_{nelec} - T\Delta S_{sc}$, as a function of the cRMSD, for the seven decoy sets of Park and Levitt.[58] The circles at cRMSD = 0 represent the native structure, and $\Delta G_{solv}(ACE)$ denotes the solvation free energy term based on the atomic contact energy (ACE).

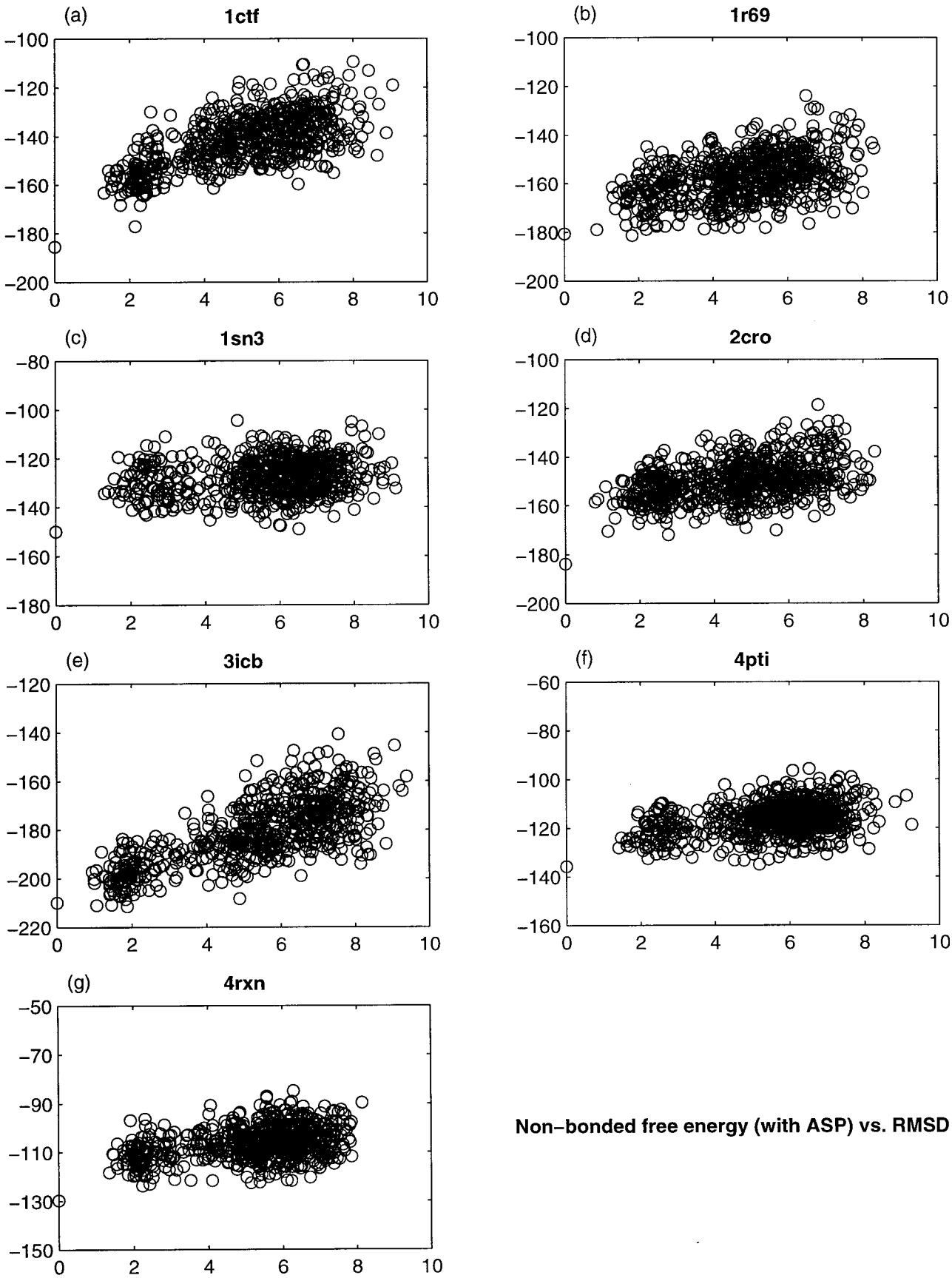

Fig. 4. The nonbonded part of the free energy, $\Delta G_{solv}(ASP) + \Delta E_{nelec} - T\Delta S_{sc}$, as a function of the cRMSD, for the seven decoy sets of Park and Levitt.[58] The circles at cRMSD = 0 represent the native structure, and $\Delta G_{solv}(ASP)$ denotes the solvation free energy term based on the atomic solvation parameter (ASP) model.

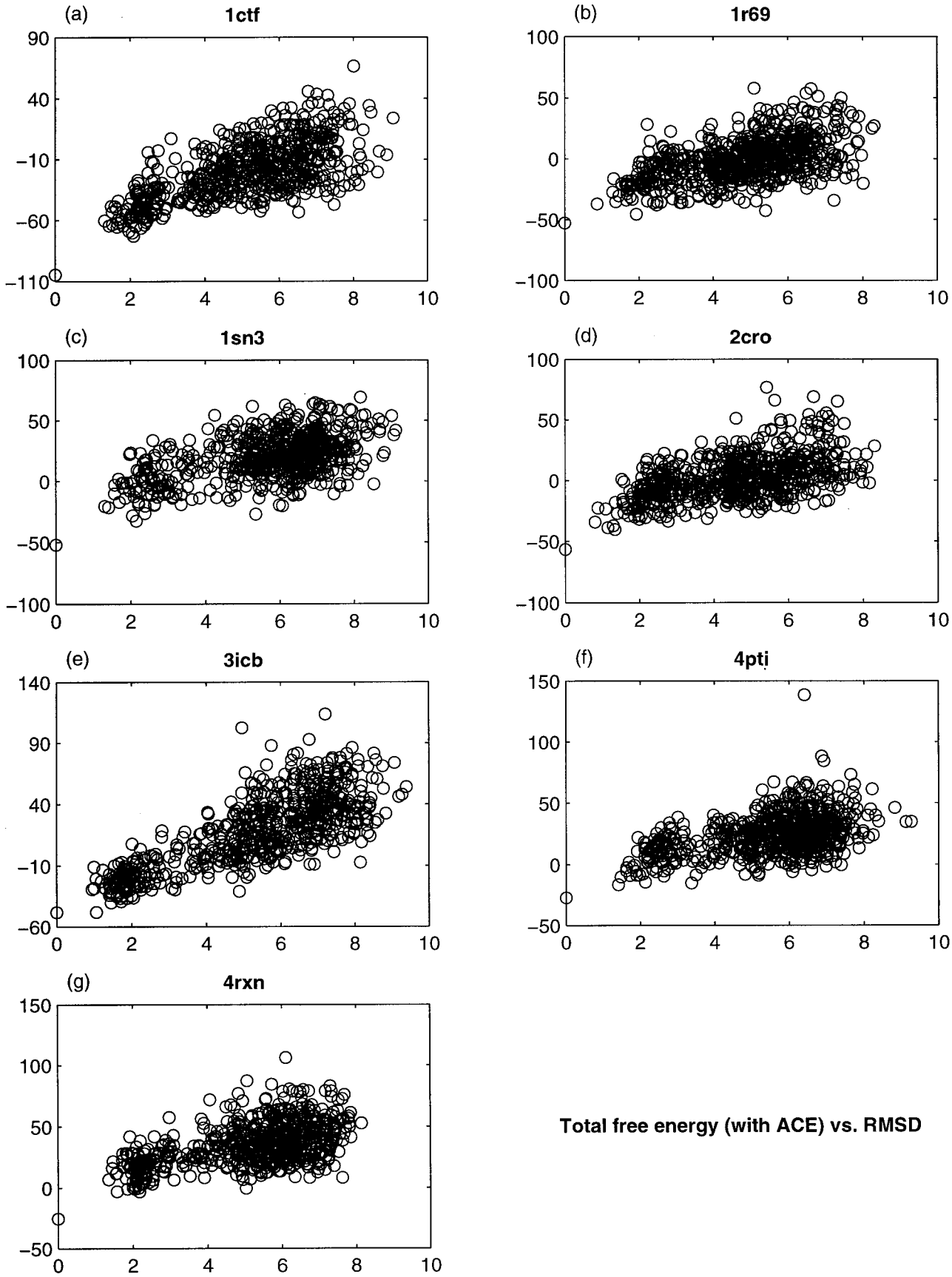

Fig. 5. The total free energy, $\Delta E_{int} + \Delta G_{solv}(ACE) + \Delta E_{nelec} - T\Delta S_{sc}$, as a function of the cRMSD, for the seven decoy sets of Park and Levitt.[58] The circles at cRMSD = 0 represent the native structure, and $\Delta G_{solv}(ACE)$ denotes the solvation free energy term based on the atomic contact energy (ACE).

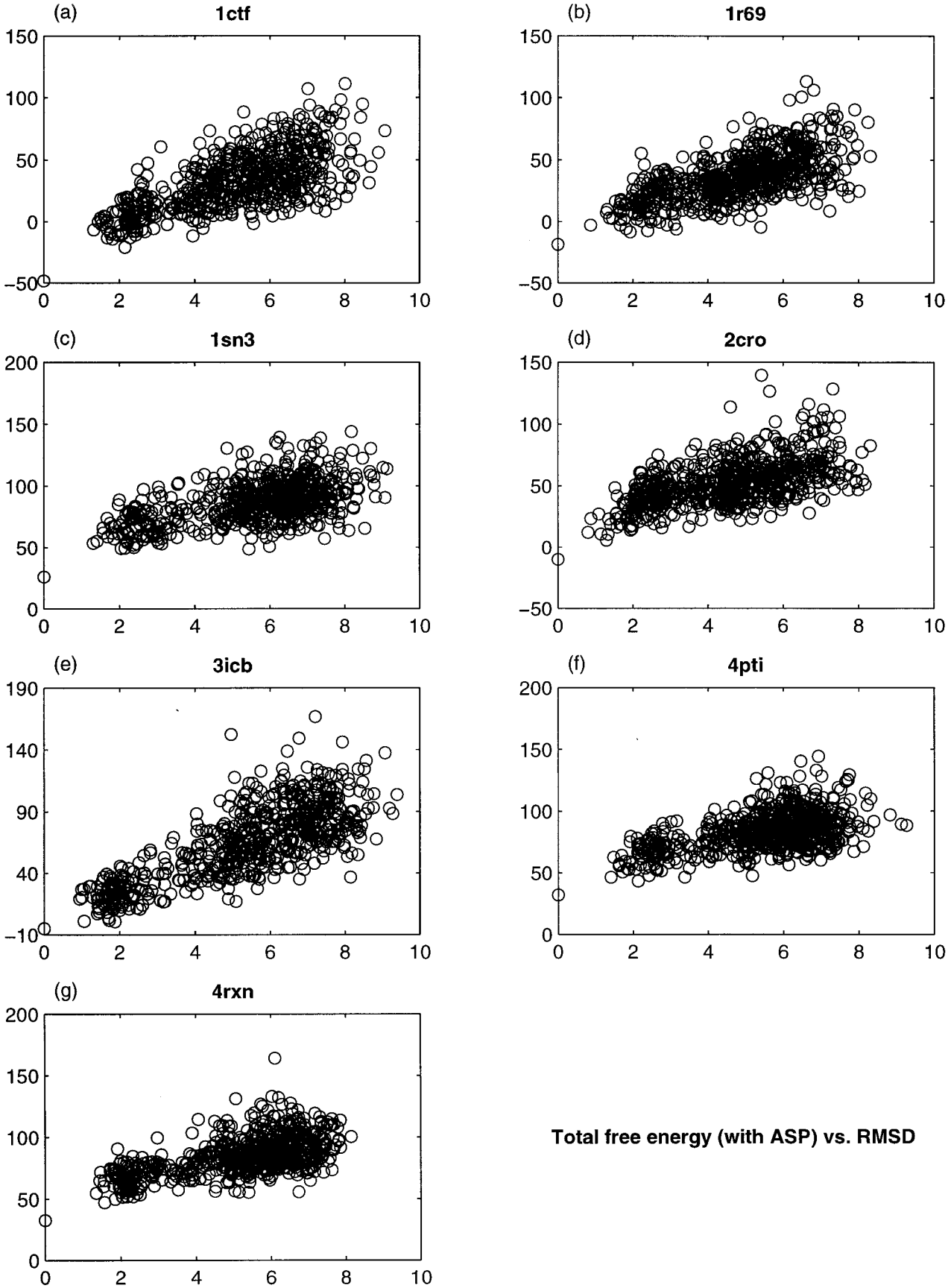

Fig. 6. The total free energy, $\Delta E_{int} + \Delta G_{solv}(ASP) + \Delta E_{nelec} - T\Delta S_{sc}$, as a function of the cRMSD, for the seven decoy sets of Park and Levitt.[58] The circles at cRMSD = 0 represent the native structure, and $\Delta G_{solv}(ASP)$ denotes the solvation free energy term based on the atomic solvation parameter (ASP) model.

TABLE I. Correlation Coefficients between Free Energy Terms and RMSD for the Decoy Sets of Park and Levitt[50]

	Correlation coefficient (r)								
Free energy terms	1ctf	1r69	1sn3	2cro	3icb	4pti	4rxn	r_{avg}	σ_r
ΔE_{int}	0.592	0.645	0.516	0.538	0.703	0.457	0.534	0.569	0.083
ΔE_{elec}	0.232	0.331	0.180	0.444	0.184	0.188	0.404	0.280	0.112
ΔE_{nelec}	0.536	0.390	0.369	0.436	0.622	0.336	0.349	0.434	0.107
$\Delta G_{solv}(ACE)$	0.222	−0.200	0.004	−0.174	0.316	0.024	0.064	0.037	0.189
$\Delta G_{solv}(ASP)$	0.406	0.332	−0.179	0.185	0.484	0.050	0.284	0.223	0.227
$\Delta E_{int} + \Delta G_{solv}(ACE)$	0.628	0.480	0.463	0.418	0.746	0.385	0.470	0.513	0.128
$\Delta E_{int} + \Delta G_{solv}(ASP)$	0.672	0.668	0.472	0.529	0.758	0.414	0.575	0.584	0.122
$\Delta E_{nelec} + \Delta G_{solv}(ACE)$	0.493	0.102	0.190	0.190	0.610	0.210	0.211	0.286	0.188
$\Delta E_{nelec} + \Delta G_{solv}(ASP)$	0.634	0.431	0.207	0.402	0.720	0.355	0.352	0.443	0.176
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE)$	0.655	0.532	0.518	0.509	0.766	0.465	0.514	0.566	0.106
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP)$	0.681	0.664	0.525	0.581	0.767	0.464	0.609	0.613	0.102
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE) - T\Delta S$	0.640	0.489	0.501	0.455	0.759	0.443	0.477	0.538	0.117
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP) - T\Delta S$	0.674	0.641	0.524	0.549	0.769	0.473	0.582	0.602	0.100

TABLE II. Minimum Discriminatory Slopes (MDS) and Dimensionless Discriminatory Slopes (DDS) for the Decoy Sets of Park and Levitt[50]

	MDS (kcal/mol-Å)										
Free energy terms	1ctf	1r69	1sn3	2cro	3icb	4pti	4rxn	MDS_{avg} (kcal/mol-Å)	σ_{MDS} (kcal/mol-Å)	DDS_{avg}	σ_{DDS}
ΔE_{int}	2.051	1.252	2.091	0.898	0.699	0.973	1.182	1.307	0.553	0.151	0.067
ΔE_{elec}	−9.211	−6.720	−7.028	−11.260	−23.601	−8.114	−6.765	−10.386	6.052	−1.174	0.647
ΔE_{nelec}	1.346	−2.808	0.168	1.239	−1.741	−2.709	0.282	−0.603	1.787	−0.157	0.456
$\Delta G_{solv}(ACE)$	−0.733	−1.630	−3.114	−2.729	−2.586	−3.505	−1.481	−2.254	0.995	−0.427	0.194
$\Delta G_{solv}(ASP)$	−1.471	0.421	−2.493	−0.448	−0.190	−0.654	−0.648	−0.783	0.944	−0.282	0.334
$\Delta E_{int} + \Delta G_{solv}(ACE)$	4.503	1.885	1.491	1.688	4.280	1.603	3.084	2.648	1.306	0.264	0.116
$\Delta E_{int} + \Delta G_{solv}(ASP)$	4.565	3.074	2.071	3.737	4.362	2.350	1.866	3.146	1.101	0.329	0.095
$\Delta E_{nelec} + \Delta G_{solv}(ACE)$	3.200	−1.948	0.061	1.823	−1.031	−2.838	0.497	−0.034	2.110	−0.009	0.327
$\Delta E_{nelec} + \Delta G_{solv}(ASP)$	2.051	1.252	0.140	2.432	0.301	0.161	1.270	1.087	0.928	0.207	0.177
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE)$	7.659	1.829	4.682	5.550	3.564	3.444	4.431	4.451	1.836	0.373	0.143
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP)$	7.451	2.968	4.127	6.361	5.148	2.778	3.992	4.689	1.735	0.402	0.128
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE) - T\Delta S$	7.763	1.456	3.860	4.828	2.851	3.133	4.387	4.040	1.981	0.322	0.144
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP) - T\Delta S$	7.555	2.594	4.153	5.623	4.286	2.998	3.388	4.371	1.721	0.365	0.121

measures of quality that have been calculated for each of the seven decoy sets, as well as their means and standard deviations. Notice that we will primarily rely on the mean values when comparing the discriminatory power of various free energy terms. As we have already mentioned, MDS provides estimates of free energy differences in kcal/mol, and thus there is an advantage in using MDS rather than DDS or DR whenever possible. Since in this paper we study free energy functions that have been carefully calibrated against well-defined thermodynamic quantities,[57,59] the scale dependence of the MDS should not be a problem. Therefore, we focus on the MDS values, but would use DDS or DR in the future for discussing empirical potentials that utilize different scales.

DISCUSSION

In order to analyze the discriminative ability of these potentials, we have plotted the relationships between the RMSD and the various free energy components as shown in Figures 2 through 6. However, based only on these plots it is difficult to draw very strong conclusions. Figure 2 shows that the internal energy ΔE_{int} is a good discriminator on its own, better than the nonbonded parts of the free energy, $\Delta E_{nelec} + \Delta G_{solv}(ACE)$ or $\Delta E_{nelec} + \Delta G_{solv}(ASP)$ (Figs. 3 and 4). However, ΔE_{int} yields only small energy gaps between native and non-native states, and for 2cro there are even false positives. The discrimination improves when the nonbonded terms are added to ΔE_{int}, i.e., the total free energy is a better discriminant than any of its

TABLE III. Discriminatory Ratios (DR) for the Decoy Sets of Park and Levitt[50]

Free energy terms	Discriminatory ratio (DR) 1ctf	1r69	1sn3	2cro	3icb	4pti	4rxn	DR_{avg}	σ_{DR}
ΔE_{int}	0.136	0.090	0.128	0.061	0.047	0.074	0.090	0.089	0.033
ΔE_{elec}	−0.681	−0.719	−0.559	−1.030	−4.159	−0.882	−0.417	−1.207	1.317
ΔE_{nelec}	0.122	−0.571	0.020	0.104	−0.290	−0.540	0.040	−0.159	0.303
$\Delta G_{solv}(ACE)$	−0.058	−0.160	−0.296	−0.353	−0.335	−0.453	−0.133	−0.256	0.141
$\Delta G_{solv}(ASP)$	−0.237	0.056	−0.549	−0.068	−0.028	−0.116	−0.114	−0.151	0.197
$\Delta E_{int} + \Delta G_{solv}(ACE)$	0.180	0.085	0.072	0.086	0.219	0.088	0.149	0.126	0.057
$\Delta E_{int} + \Delta G_{solv}(ASP)$	0.222	0.162	0.118	0.193	0.236	0.140	0.111	0.169	0.049
$\Delta E_{nelec} + \Delta G_{solv}(ACE)$	0.147	−0.149	0.003	0.108	−0.094	−0.248	0.031	−0.029	0.142
$\Delta E_{nelec} + \Delta G_{solv}(ASP)$	0.136	0.090	0.011	0.140	0.029	0.017	0.114	0.077	0.056
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE)$	0.220	0.077	0.166	0.188	0.158	0.161	0.177	0.164	0.044
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP)$	0.251	0.139	0.178	0.216	0.225	0.144	0.188	0.191	0.042
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE) - T\Delta S$	0.210	0.056	0.135	0.158	0.123	0.147	0.171	0.143	0.048
$\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP) - T\Delta S$	0.242	0.116	0.176	0.192	0.186	0.155	0.157	0.175	0.039

components, with the ASP model seeming to perform slightly better than the one based on ACE.

As we will show, the free energy components, as functions of the RMSD, contain substantial amounts of additional information concerning the nature of decoy discrimination. However, it is difficult to draw stronger conclusions on the basis of the plots, and we proceed to the analysis of Figures 2 through 6 using the qualitative measures described in Materials and Methods. Consideration will be restricted to the correlation coefficients (r values) and to the minimum discrimination slopes (MDS values), shown in Tables I and II, respectively. As we have mentioned, it is necessary to use the normalized measures DDS or DR to compare the discrimination by two potentials or energy components if the two may vary in scale, which is not the case here.

The good discrimination attained by the internal energy ΔE_{int} in Figure 2 is confirmed by the relatively high mean correlation coefficient of 0.57 (Table I). Although it is not established how large a free energy-RMSD correlation should be, values exceeding 0.5 are usually considered adequate for protein structure discrimination (Michael Levitt, personal communication). We recall that in calculating MDS, DDS, and DR values, we restricted consideration to the RMSD range of 2 to 8 Å, because most points are in this range, and we do not expect the potential to be a good discriminator beyond this interval. Since the lower bound is 2 Å, the average MDS value of 1.3 for internal energy in Table II means that there is at least 2.6 ± 1.1 kcal/mol internal energy gap between the native protein and the lowest energy structure in the region of interest.

The next observation concerns the discriminating behavior of electrostatic calculations. A comparison of correlation coefficients demonstrates that the electrostatic energy, ΔE_{nelec}, calculated with non-ionized side chains, provides much better discrimination than its counterpart ΔE_{elec} with charged side chains ($r = 0.43$ and $r = 0.28$, respectively). The MDS for ΔE_{elec} is negative in all proteins (average value −10.3 kcal/mol/Å), indicating that for each protein there are decoys that have lower electrostatic energies than the X-ray structure. Although ΔE_{nelec} also gives a negative MDS for three proteins, it clearly performs better, and thus for the remainder of the paper we restrict electrostatic consideration to ΔE_{nelec}.

The lack of a stronger correlation between ΔE_{elec} and the RMSD might be due to the fact that we restrict consideration to the C_α RMSD. However, neutral side chains also gave better discrimination in other applications where the all-atom RMSD has been considered.[24] In addition, we have shown that Equation 1 reproduces experimentally determined free energies of protein unfolding if, and only if, the side chains are assumed to be neutral.[57] The finding that removal of the charges improves the predictive power of the potential has been recently confirmed by Lazaridis and Karplus.[25,43]

There is no reason to assume that the ionizable side chains are neutral under the conditions of the crystallization. Two factors, however, may be responsible for the finding that neutral electrostatics perform better. First, most charged side chains are on the protein surface, and due to their high thermal factors or participation in crystal contacts may not be very well defined in the X-ray structure. The electrostatic energy calculated with misplaced charge positions cannot be very accurate. The second potential factor is the limited accuracy of the simple coulombic model with distance-dependent dielectrics. Elucidating the contributions of these factors and improving the calculation of the electrostatic term deserves further investigation.

The solvation energies, particularly $\Delta G_{solv}(ACE)$, provide very weak discrimination even in terms of the correlation coefficient (Table I). Solvation remains of little importance as part of the nonbonded free energies $\Delta E_{nelec} + \Delta G_{solv}(ACE)$ and $\Delta E_{nelec} + \Delta G_{solv}(ASP)$. In fact, the nonbonded free energies have the same correlation coefficients as ΔE_{nelec} alone, although the MDS values are slightly improved.

In view of the weak discriminatory power of the solvation free energies alone, it is surprising that, in terms of the MDS, the discrimination substantially improves when adding either $\Delta G_{solv}(ACE)$ or $\Delta G_{solv}(ASP)$ to ΔE_{int}, resulting in MDS values of 2.64 kcal/mol and 3.14 kcal/mol, respectively. There is no such increase in terms of the correlation coefficient. This observation can be explained in terms of a compensation effect. When we minimize the molecular mechanics energy of a conformation, all energy terms, including the internal energy, tend to decrease. Relatively low internal energies can be attained both for near-native and non-native conformations, primarily by reducing the deviations from the standard values of bond and torsional angles. In non-native structures, these changes usually expose some hydrophobic side chains to the solvent. Thus, the resulting conformations tend to have relatively high solvation free energy, and $\Delta E_{int} + \Delta G_{solv}$ discriminates better than ΔE_{int} alone. Such compensation occurs only for a fraction of the structures, thereby increasing the MDS but not the correlation coefficient. A similar compensatory effect between ΔE_{int} and ΔG_{solv} has been recently reported for a free energy expression with a solvation term based on the Poisson-Boltzmann model.[34]

In terms of the MDS, the discrimination further improves by adding ΔE_{nelec} to $\Delta E_{int} + \Delta G_{solv}$. According to Table II, the highest MDS, 4.69, is attained with the combined potential $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP)$. Thus, the free energies of two conformations that are 2 Å and 3 Å RMSD from the native, respectively, are expected to differ at least by 4.69 ± 1.7 kcal/mol, an easily detectable free energy gap. The discrimination is slightly worse if we use the ACE solvation model instead of ASP, with the sum $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ACE)$ yielding an MDS of 4.45.

Adding the entropic term $-T\Delta S_{sc}$ makes the discrimination slightly worse, both in terms of the correlation coefficient and the MDS. Notice that the ACE solvation term is defined to already include $-T\Delta S_{sc}$, and hence this finding is not a surprise.[46,59] However, the entropy change definitely should be part of the total free energy expression when the ASP solvation model is used. Thus, the unfavorable effect of $-T\Delta S_{sc}$ on the quality of discrimination indicates that our method of estimating side chain entropy loss is clearly not accurate. This is somewhat surprising given that the calculated $-T\Delta S_{sc}$ values agree very well with the side chain entropies based on calorimetric observation of temperature-induced protein unfolding,[57] and the method has been extensively used for calculating the loss of side chain entropy upon binding.[49,50] However, the approach is clearly less appropriate for calculating the relatively small difference in side chain entropy between two folded conformations. Notice that S_{sc} depends only on the backbone conformation. Nevertheless, we use the change in the solvent exposed surface area of the individual side chains to calculate $-T\Delta S_{sc}$ (see Materials and Methods). Since side chain positions can vary even with a fixed backbone, the method has an inherent error. This error may be comparable to the relatively small side chain entropy difference between two folded conformations with very similar backbones. Thus, as confirmed by both the correlation coefficient and the MDS, in protein structure discrimination it is better to neglect the entropic term than to estimate it using a method of relatively low accuracy.

By all measures, the best discrimination is attained by the sum $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP)$. However, there is a significant difference between the correlation coefficient and MDS. As shown in Table I, the mean correlation coefficient using ΔE_{int} alone is 0.57. Adding both ΔE_{nelec} and $\Delta G_{solv}(ASP)$ to ΔE_{int}, the correlation increases only to 0.61 for the ASP model, and does not change at all if the ACE solvation is used. Thus, in terms of the correlation coefficient, accounting for electrostatics and solvation does not improve the discrimination. In contrast, the MDS value of 1.3 for ΔE_{int} increases to 4.69 and 4.45 with ASP and ACE models, respectively, a more than threefold increase in the discriminatory free energy gap.

The above discussion was based on the MDS, but the same conclusions would be obtained using the dimensionless measures DDS or DR. Since the MDS values depend on the scaling of the potential, they need to be replaced by DDS or DR when comparing discriminatory functions that may be defined on different thermodynamic scales. The advantage of the MDS is that it provides estimates of free energy gaps in kcal/mol. Therefore, we believe that the MDS is the measure of choice whenever the discriminatory function and its components represent free energies, and thus are given on a well-defined thermodynamic scale.[57,59]

So far we have restricted consideration to the mean values of the correlation coefficient and the MDS. For the individual proteins, the correlation coefficients between $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP)$ and the RMSD vary from 0.46 (4pti) to 0.76 (3icb). As we mentioned, the correlation coefficients are primarily determined by ΔE_{int}. The contributions of the other free energy terms have much larger effects on the MDS values. Nevertheless, MDS also shows the worst discrimination for 4pti, although the main determinant is now ΔE_{nelec} rather than ΔE_{int}. Notice that the worst discrimination, supported by both measures, is not at all apparent from the free energy vs. RMSD plots in Figure 6.

CONCLUSIONS

We discussed the discrimination of native or near-native conformations from misfolded protein models by free energy potentials that combine molecular mechanics with empirical solvation and entropic terms. Quantitative measures have been used to study the contributions of individual free energy terms and their combinations to the discriminatory power of the potential. This analysis builds on our previous results concerning the roles of molecular mechanics and solvation terms.[24] The main conclusions can be summarized as follows.

1. The two main measures of discrimination quality are the correlation coefficient r between the free energy and RMSD, and the Minimum Discriminatory Slope (MDS). MDS is defined as the slope of the line that constitutes the lower boundary of the points in the RMSD-free energy plot, and passes through the coordinates of the

native structure. The MDS value provides an estimate of the smallest free energy gap between any two conformations with different RMSD's, and hence has a well-defined physical meaning.

The major shortcoming of MDS is its dependence on the scaling of the free energy function. To compare potentials defined on different scales or given in arbitrary units, we have introduced two dimensionless measures related to the discriminatory slope. The obvious non-dimensional version is the Dimensionless Discriminatory Slope (DDS). The second measure, the Discrimination Ratio (DR), is defined as DR $= m_1/m_2$, where m_1 and m_2 denote the slopes of the lower and upper boundary, respectively. The ideal case, in which all points in the RMSD-free energy plot are on a straight line, yields $r = 1$ and DR = 1.

2. The correlation coefficient r and the minimum discriminatory slope MDS provide highly complementary information. As a statistical measure, the correlation coefficient describes the relationship between the free energy and the expected value of the RMSD. Strong correlation means that a conformation with low free energy is *expected* to have relatively small RMSD from the native. However, it does not exclude the existence of false positives, i.e., conformations with low free energy and yet large RMSD. Since a good potential function is expected to eliminate all false positives, this is a substantial shortcoming. In contrast, MDS becomes negative if there is even a single false positive.

   The difference between the properties of the two measures, r and MDS, implies that both are needed when evaluating the discriminatory power of a potential. Conclusions supported by both measures are clearly very strong. For example, in this paper both measures show that the best discriminator is the sum $\Delta E_{int} + \Delta E_{nelec} + \Delta G_{solv}(ASP)$, that the ASP solvation model is slightly better than the ACE model, and that adding an entropic terms makes discrimination worse. Other conclusions may be supported only by one of the measures. For example, according to MDS, adding electrostatic and solvation terms, $\Delta G_{solv} + \Delta E_{nelec}$, to the internal energy ΔE_{int} substantially improves discrimination, yielding a more than threefold increase in the discriminatory free energy gap. In contrast, the correlation coefficient shows only very moderate improvement. This means that the additional terms do not strengthen the relationship between the expected values of the free energy and the RMSD, but they tend to raise the points that define the lower boundary on the free energy-RMSD plot.

   The extreme dependence of MDS on a single point defining the lowest boundary clearly provides information that is not provided by the correlation coefficient. However, this sensitivity also implies that changing the pool of decoys may completely change the MDS and thus the conclusions based on its particular value. In this work, added robustness was provided by calculations for seven decoy sets. If several data sets are not available, the robustness of the results based on MDS can be examined by performing jackknife-type tests, e.g., leaving one point out or dividing the sample into two parts. While such tests will certainly lead to substantial MDS variations, they provide the required assurances if the major conclusions remain valid at least for most subsets of decoys.

3. In spite of the significant contributions of the other free energy terms, the internal energy $\Delta E_{int} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{dihedral} + \Delta E_{improper}$ is a surprisingly good discriminator on its own. This property has also been observed for other combined free energy functions,[33,43] and suggests that adding the internal energy may improve the discriminatory power of free energy expressions that do not include any molecular mechanics energy terms.[17,18,20,58]

4. We have studied two free energy expressions that differ only in the terms representing the short-range solute-solvent interactions. These terms are based on the classical atomic solvation parameter (ASP) model,[37] and on a structure-based atomic contact energy (ACE).[46] The long-range electrostatic screening is accounted for by the distance-dependent dielectric of $\epsilon = 4r$ in both free energy expressions. By all measures of discrimination quality, the ASP model performs somewhat better than the structure-based atomic contact potential ACE. This is surprising since in ACE we consider 18 distinct atom types, in contrast to the 5 atom types used by Eisenberg and McLachlan in the ASP model.[37]

5. In terms of all measures, the discrimination significantly improves if the electrostatic energy is calculated by assuming that all side chains, including Arg, Lys, Glu, and Asp, are assumed to be neutral, since there is no reason to assume that the ionizable side chains were neutral under the conditions of the crystallization. Factors that can explain this apparent contradiction are the limited accuracy of charges in X-ray structures, or the inability of the simple coulombic model with $\epsilon = 4r$ to provide the required accuracy.

6. Since a conformation of the protein may represent an entire ensemble of equienergetic structures,[48] in the most general case the free energy expression includes a term of the form $-T\Delta S_{conf}$, where ΔS_{conf} denotes the sum of conformational and vibrational entropy. As we will further discuss, we assume that, in a folded state, the backbone contribution to this entropy are independent of the conformation. Adopting a folded conformation as the reference state, this assumption reduces $-T\Delta S_{conf}$ to $-T\Delta S_{sc}$, where ΔS_{sc} is the change in side chain entropy. This term depends on the change in the side chain degrees of freedom, and in our previous work has been calculated by an empirical entropy scale,[51] in terms of the solvent accessible surface area.[49] However, both the correlation coefficient and the MDS value showed that adding the entropic term to the potential makes the discrimination worse. This probably happens because the side chain entropy difference between two conformations with similar backbones is too small to be accurately predicted by the empirical entropy scale.

A paper on decoys can not be complete without raising three fundamental and strongly interrelated questions. What are the characteristics of good decoys? How does the information gained in a decoy study depend on the way these decoys have been generated? What are the contributions of decoy studies to the development of potentials for folding simulations and protein model refinement? Although much further work is required to fully answer these very important questions, our results provide some insight.

The principles of constructing protein structure decoys for potential function evaluation have been discussed by Park and Levitt.[58] They concluded that decoy structures must: (1) include structures that are close to the native X-ray structure; (2) be native-like in all properties of the native polypeptide chain except the overall folded conformation, otherwise they could easily be distinguished by trivial tests; (3) be diverse so as to sample all possible arrangements; and (4) be numerous for more sensitive testing. The decoys used in this paper certainly satisfy conditions (1), (2), and (4). Satisfying condition (3) is somewhat questionable, as all decoys have the correct secondary structure, and this secondary structure is always correctly aligned with that of the native state.[60] In addition, only four or five degrees of freedom have been used for repositioning helices and strands, representing a very limited search of the conformational space.[61] Therefore, while the decoys by Park and Levitt are very useful and have been utilized in other recent papers,[43] there is a need for further, publicly available decoys, in order to form a standard set that eventually will provide the tool for a comprehensive evaluation of potentials.

The second fundamental problem is the relationship between the observed performance of a potential and the method of preparing the decoys. The free energy functions were applied to various decoys, in addition to the ones used here. We always found good discrimination, but substantial differences were observed in the relative importance of the various free energy terms, particularly that of the internal energy. In most cases, the internal energy is very important, because the decoys are more strained than the native structure, and the extra strain correlates with the RMSD from the native. This applies to the protein decoys used in this paper, to the loop libraries generated by Moult and co-workers,[62] and to protein decoys obtained by torsional angle perturbations (see Janardhan and Vajda[24] and unpublished results). Similar results have been reported by other groups.[19,63,64] The internal energy may even dominate the free energy expression, as we found for buried loops[24] and for decoys obtained by high temperature molecular dynamics. However, the correlation between the internal energy and the RMSD is not universal. We found a counterexample when side chain conformations were generated for a fixed backbone using a combinatorial search algorithm,[65] and the internal energy as the target function. Some of the alternative structures found by this search had excellent packing of the side chains, and lower internal energy than that of the native state.[66] Since the solvation free energy of such "overpacked" conformations was relatively high, they were distinguishable from the native state by the complete free energy, but not by the internal energy alone. This result emphasizes that the method of decoy preparation is an important factor, and it will not be trivial to assemble a collection of decoy sets representing all possible situations.

The real goal of developing a good potential function is to use it for folding simulations and for the refinement of protein models. This takes us to the third fundamental problem. It is easy to show that the ability of a potential to discriminate in decoy studies is a necessary but not sufficient condition for its usefulness in simulations and conformational searches. Finding false positives in a decoy set implies that the particular potential is unable to discriminate the native state from misfolded models, and therefore it is not suitable for use as a target function in any application. However, the converse is not true, i.e., good discrimination in decoy studies does not exclude finding false positives when the same potential is used as the target function in a minimization. In fact, minimization can yield non-native conformations that have lower free energy than the native state, in spite of good discrimination in decoy studies. This apparent contradiction can be easily explained by recognizing that the conformational entropy of various folded states may depend on the conformation. In particular, it has been observed that the native conformation tends to have a larger number of nearby structures and hence larger conformational entropy than other low-energy conformations.[61] Thus, these alternative minima are relatively narrow and correspond to conformations with low conformational entropy. Therefore, accounting for the conformational entropy would increase the free energy of these states, thereby eliminating the narrow minima.

However, apart from an estimate of the side chain entropy, current potentials do not include any measure of conformational entropy, and without an extensive conformational sampling it would be very difficult to obtain any reliable estimate of the latter. In the absence of a conformational entropy term, the free energy surface may have deep and narrow local minima that correspond to non-native conformations. Search methods based on global or even local minimization of the potential are bound to find such false minima if they exist. In contrast, the same potential may perform very well in decoy studies, since the probability of randomly generating conformations within the narrow false minima is very small. Thus, convergence to a non-native conformation in a search algorithm does not necessarily mean that the potential is useless. For example, trapping in the narrow local minima can be avoided by adding small random perturbations to the conformation obtained by minimization, or sampling the nearby conformations can be used to derive some measure of the conformational entropy. These approaches essentially combine minimization with generating and evaluating decoys, and clearly deserve further investigation.

REFERENCES

1. Novotny J, Rashin AA, Bruccoleri RE. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 1988;4:19–30.

2. Vajda S, Sippl M, Novotny J. Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 1997;7:222–228.
3. Karplus M, Petsko GA. Molecular dynamics simulations in biology. Nature (Lond) 1990;347:631–639.
4. Tobias DJ, Mertz JE, Brooks CL III. Nanosecond time scale folding dynamics of a pentapeptide in water. Biochemistry 1991;30: 6054–6058.
5. Guo ZY, Brooks CL III, Boczko EM. Exploring the folding free energy surface of a three-helix bundle protein. Proc Natl Acad Sci USA 1997;94:10161–10166.
6. Daura X, van Gunsteren WF, Mark AE. Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. Proteins 1999;34:269–280.
7. Duan Y, Wang L, Kollman PA. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. Proc Natl Acad Sci USA 1998;95: 9897–9902.
8. Godzik A, Kolinski A, Skolnick J. Topology fingerprinting approach to the inverse folding problem. J Mol Biol 1992;227:227–238.
9. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Cur Opin Struct Biol 1996;6:195–209.
10. Moult J. Comparison of database potentials and molecular mechanics force fields. Curr Opin Struct Biol 1997;7:194–199.
11. Miyazava S, Jernigan R. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
12. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
13. Kocher JPA, Rooman MJ, Wodak S. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 1994;235:1598–1613.
14. Mohanty D, Dominy BN, Kolinski A, Brooks CL III, Skolnick J. Correlation between knowledge-based and detailed atomic potentials: Application to the unfolding of the GCN4 leucine zipper. Proteins 1999;35:444–452.
15. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
16. Lemer CMR, Rooman MJ, Wodak S. Protein structure prediction by threading methods: Evaluation of current techniques. Proteins 1995;23:337–355.
17. Melo F, Feytmans EJ. Novel knowledge-based mean force potential at atomic level. J Mol Biol 1997;267:207–222.
18. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
19. DeBolt EE, Skolnick J. Evaluation of atomic level mean force potentials via inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. Prot Eng 1996;8:175–186.
20. Rojnuckarin A, Subramaniam S. Knowledge-based interaction potentials for proteins. Proteins 1999;36:54–67.
21. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
22. Schiffer CA, Caldwell JW, Stroud RM, Kollman PA. Inclusion of solvation free energy with molecular mechanics energy: alanyl dipeptide as a test case. Prot Sci 1992;1:396–400.
23. Sun S. A genetic algorithm that seeks native states of peptides and proteins. Biophys J 1995;69:340–355.
24. Janardhan A, Vajda S. Selecting near-native conformations in homology modeling: The role of molecular mechanics and solvation terms. Prot Sci 1997;7:1772–1780.
25. Lazaridis T, Karplus M. Effective energy functions for proteins in solution. Proteins 1999;35:132–152.
26. Brasseur R. Simulating the folding of small proteins by use of the local minimum energy and the free solvation energy yields native-like structures. J Mol Graphics 1995;13:312–322.
27. Cardozo T, Totrov M, Abagyan R. Homology modeling by the ICM method. Proteins 1995;23:403–414.
28. Gilson MK, Honig B. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. Proteins 1988;4:7–18.
29. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
30. Smith KC, Honig B. Evaluation of the conformational free energies of loops on proteins. Proteins 1994;18:119–132.
31. Pellequer JL, Chen SWW. Does conformational free energy distinguish loop conformations in proteins. Biophys J 1997;73:2359–2375.
32. Novotny J, Bruccoleri RE, Davis ME, Sharp KA. Empirical free energy calculations: a blind test and further improvements to the method. J Mol Biol 1997.
33. 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. Proteins 1998;32:399–413.
34. 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.
35. Jackson RM, Sternberg MJE. A continuum model for protein-protein interactions: Application to the docking problem. J Mol Biol 1995;250:258–275.
36. Froloff N, Windemuth A, Honig B. On the calculation of binding free energies using continuum methods: Application to MHC class I protein-peptide interactions. Prot Sci 1997;6:1293–1301.
37. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature (Lond) 1986;319:199–203.
38. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Prot Sci 1992;1:227–235.
39. Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.
40. Vajda S, Weng Z, DeLisi C. Extracting hydrophobicity parameters from solute partition and protein mutation/unfolding experiments. Prot Eng 1995;8:1081–1092.
41. Kang YK, Nemethy G, Scheraga HA. Free energies of hydration of solute molecules 1. Improvement of the hydration shell modell by exact computations of overlap volumes. J Chem Phys 1987;91: 4105–4109.
42. Augspurger JD, Scheraga HA. An efficient differentiable hydration potential of peptides and proteins. J Comput Chem 1996;17: 1549–1558.
43. 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.
44. Stouten PFW, Frommel C, Nakamura H, Sander C. An effective solvation term based on atomic occupancies for use in protein simulations. Mol Simul 1993;10:97–120.
45. Still WC, Tempczyl A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
46. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystalized proteins. J Mol Biol 1996;267:707–726.
47. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
48. Gilson MK, Given JA, Head MS. A new class of models for computing receptor-ligand binding affinities. Chem Biol 1997;4:87–92.
49. Vajda S, Weng Z, Rosenfeld R, DeLisi C. Effect of conformational flexibility and solvation on receptor-ligand binding free energies. Biochemistry 1994;33:13977–13988.
50. Weng Z, Vajda S, DeLisi C. Prediction of complexes using empirical free energy functions. Prot Sci 1996;5:614–626.
51. Pickett SD, Sternberg MJE. Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 1993;231:825–839.
52. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. Biochemistry 1973;79:353–371.
53. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
54. Novotny J, Bruccoleri RE, Saul FA. On the attribution of binding energy in the antigen-antibody complexes McPC 603, D1.3 and HyHEL-5. Biochemistry 1989;28:4735–4749.
55. Adamson AW. Physical chemistry of surfaces. New York: John Wiley & Sons; 1982.
56. Nicholls A, Sharp KA, Honig B. Protein folding and association—

insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1991;11:281–296.
57. Weng Z, DeLisi C, Vajda S. Empirical free energy calculation: Comparison to calorimetric data. Prot Sci 1996 (in press)
58. 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.
59. Zhang C, Cornette JL, DeLisi C. Consistency in structural energetics of protein folding and peptide recognition. Prot Sci 1997;6:1057–1064.
60. Osguthorpe DJ. Ab initio protein folding. Curr Opin Struct Biol 2000;10:146–152.
61. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
62. Moult J, James MNG. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1986;1:146–163.
63. Pellequer J, Chen SW. Does conformational free energy distinguish loop conformations in proteins? Biophys J 1997;73:2359–2375.
64. Maiorov V, Abagyan R. Energy strain in three-dimensional protein structures. Foldinf & Design 1998;3:259–269.
65. Bruccoleri RE, Novotny J. Antibody modeling using the conformational search program CONGEN. Immunomethods 1992;1:96–106.
66. Vajda S, DeLisi C. Determining minimum energy conformations of polypeptides by dynamic programming. Biopolymers 1990;29: 1755–1772.