JCTC Journal of Chemical Theory and Computation

Assessment of Semiempirical Quantum Mechanical Methods for the Evaluation of Protein Structures

Andrew M. Wollacott[†] and Kenneth M. Merz, Jr.*

University of Florida, Gainesville, Florida 32611-8435

Received November 4, 2006

Abstract: The ability to discriminate native structures from computer-generated misfolded ones is key to predicting the three-dimensional structure of a protein from its amino acid sequence. Here we describe an assessment of semiempirical methods for discriminating native protein structures from decoy models. The discrimination of decoys entails an analysis of a large number of protein structures and provides a large-scale validation of quantum mechanical methods and their ability to accurately model proteins. We combine our analysis of semiempirical methods with a comparison of an AMBER force field to discriminate decoys in conjunction with a continuum solvent model. Protein decoys provide a rigorous and reliable benchmark for the evaluation of scoring functions, not only in their ability to accurately identify native structures but also to be computationally tractable to sample a large set of non-native models.

Introduction

The three-dimensional structure of a protein is determined primarily by its amino acid sequence,[1] yet, while this principle is well established, reliable methods for the prediction of protein tertiary structure from primary structure have not yet been developed.[2] Current efforts for protein structure prediction have focused on homology modeling, threading/fold recognition, and ab initio folding, all of which share the thermodynamic hypothesis that the native three-dimensional conformation has the lowest free energy in comparison to non-native or misfolded structures.[3] While current progress has proved extremely promising,[4] ab initio folding methods cannot be consistently applied to successfully predict the fold of any given sequence.[5,6] In ab initio folding methods, not only must a very large conformational space be sampled but also it is particularly important to be able to identify native folds from those non-native structures that are generated.[7]

The development of scoring functions to be used in the studies of biological macromolecules is still ongoing, with the focus being on either fast, less accurate methods or on high accuracy methods. The choice of a potential energy function in protein modeling depends on the type of simulation performed and the size of the system being modeled. Developing reliable tests for a scoring function, therefore, remains an important aspect of computational modeling.

In order to provide an objective manner with which to evaluate scoring potentials, sets of computationally misfolded models of proteins are typically used. The ability to compare native structures to available decoy sets allows for a common and relatively unbiased benchmark for evaluating scoring functions.[8] Analyzing decoys allows a potential energy function to be tested for its ability to discriminate native protein structures from a large ensemble of non-native models,[9] which is important for structure prediction methods such as homology modeling, threading/folding recognition, and ab initio folding. In these cases, the native conformation is expected to have the lowest free energy.

Atomic-based protein scoring potentials are employed for modeling structures at higher resolutions.[10] These potentials are physics-based and fall under either a molecular mechanics (MM) type scoring function or a quantum mechanical (QM) based function. MM based functions for studying proteins include AMBER,[11] CHARMM,[12] and OPLS,[13] among others. These potentials can be used for performing either molecular

* Corresponding author e-mail: merz@qtp.ufl.edu. Corresponding author address: Department of Chemistry, Quantum Theory Project, 2328 New Physics Building, P.O. Box 118435, University of Florida, Gainesville, FL 32611-8435.

[†] Current address: University of Washington, Seattle, WA.

10.1021/ct600325q CCC: $37.00
Published on Web 05/30/2007

Table 1. Select Decoy Sets Used

decoy set	PDB	description	type	N_{decoys}	N_{res}	N_{atoms}	rmsd range (Å)	%H/%E[b]
4-state-reduced	1ctf	C-terminal domain of ribosomal protein L7/L12	X-ray	630	68	1005	2.1−9.8	53/26
	1r69	N-terminal domain of phage 434 repressor	X-ray	675	63	997	2.2−9.4	70/0
	1sn3	scorpion toxin variant 3	X-ray	660	65	948	2.5−10.3	12/22
	2cro	phage 434 Cro protein	X-ray	674	65	1081	2.0−9.5	66/0
	4pti	trypsin inhibitor	X-ray	687	58	892	2.8−10.7	16/24
	4rxn	rubredoxin	X-ray	677	54	794	2.5−9.2	0/20
Rosetta	1gb1	immunoglobulin binding domain of streptococcal protein G	NMR	999	54	823	3.1−18.0	28/33
	1hsn	high mobility group protein I box	NMR	999	62	1014	4.1−17.6	79/0
	1orc	Cro repressor (mutant)	X-ray	999	56	877	4.0−14.1	46/27
	1pgx	protein G (B2 domain)	X-ray	999	57	873	3.4−20.4	26/46
	1uxd	fructose repressor DNA-binding domain	NMR	999	43	690	2.2−12.5	81/0
	2fow	RNA binding domain of ribosomal protein LII	NMR	999	66	1009	4.0−21.6	52/9
	1hc8[a]	RNA binding domain of ribosomal protein LII	X-ray	999	66	1009	4.0−21.6	53/9
	1r69	N-terminal domain of phage 434-repressor	X-ray	999	61	976	3.1−15.5	70/0

[a] 1hc8 is the X-ray structure for the 2fow decoy set. [b] Percentage of helices and sheets in the native.

dynamics simulations or energy minimizations of structures. MM potentials have been parametrized on small molecules modeled at the quantum mechanical level or from liquid simulations, and these parameters can subsequently be extended to larger biological systems, with generally satisfactory results.

Until recently, a full QM treatment of protein structures has not been feasible. The linear-scaling semiempirical package, DivCon,[14] has been developed in our laboratory and has allowed for large biological systems to be studied at the QM level. The major problem with applying higher levels of theory to biologically relevant systems is their poor scaling. Semiempirical methods scale as N^3, where N is the number of basis functions used to represent electrons in the system.[15] With judicious use of cutoffs, and by applying a divide-and-conquer[16] approach to large systems, calculations performed with DivCon are able to scale linearly with system size.[17]

The QM treatment of proteins has several advantages over MM based methods. The point charge model used in MM packages ignores higher order effects such as charge transfer and polarization.[18] These effects have not been widely incorporated into MM functions so charge interactions can be more accurately modeled with QM methods. The utility of semiempirical methods, such as those used in DivCon, has already been shown for studying the electrostatic interactions in protein-folding and protein−ligand interactions.[19,20] However, this study marks the first large-scale investigation into the utility of semiempirical quantum mechanical methods for studying protein structures.

Approach

Decoy Sets. There are various protein decoy sets readily available for public use, and these vary in quality depending upon the method of generation. Two of the more popular decoy sets include the 4-state-reduced set models created by Park and Levitt[7] and the Rosetta decoy set, produced by Simmons and Baker.[21]

The 4-state-reduced decoy set from Levitt has been widely used as a means of evaluating scoring potentials and is considered to be a high-quality decoy set.[22] The set consists of six small proteins whose structures have been solved by X-ray crystallography and are considered well refined. During the generation of decoys, the secondary structure of the native structure was held constant by altering only hinge regions. A set of loop residues between segments of defined secondary structure was chosen and their conformation was enumerated by varying the ϕ,ψ torsion angles. Only a subset of 10 residues was modified, with only four predefined states of the backbone torsional angles sampled. The large number of conformations generated was reduced by applying a radius of gyration cutoff as well as removing structures that contained bad clashes in the reduced representation.

The proteins in the 4-state-reduced set, listed in Table 1, possessed small compact native folds with between 54 and 68 residues and represent a diverse set of small proteins. The original set also included a calcium-binding protein (3icb) which chelates two calcium ions with carbonyl and carboxyl groups from side chains and the protein backbone. Since decoy structures were generated without calcium ions bound, this set was removed as the unbound native structure would be unfairly penalized by having a large concentration of negative charge near the binding sites.

Another high quality collection of decoys is the Rosetta decoy set, published by Baker and co-workers.[23] In the Rosetta set, nativelike structures were created by assembling fragments of nonhomologous protein structures containing similar local sequences to that of the native structure. Relative to the 4-state-reduced set, this allows for a broader range of conformational space to be explored. During decoy generation for both data sets, the structures were created in a reduced representation, with all heavy atoms for the protein backbone, and a C^β atom for the side chains. A subset of these structures, possessing more than 40 residues having structures with an rmsd less than 5.0 Å to the native, was chosen for this study (Table 1).

Scoring Functions. For this study, we are interested in the energetic differences between individual protein states, more specifically the energy gap between a protein decoy and its native structure. Since protein decoys have the same sequence as their native structure the two will have the same unfolded state, so the calculated energy of each structure

can be represented as an effective free energy as described in previous decoy studies.[10] Thus, the energy gap between a native structure and its decoy can be represented as

$$\Delta\Delta G_{\text{eff}} = \Delta G_{\text{eff}}^{\text{native}} - \Delta G_{\text{eff}}^{\text{decoy}} \quad (1)$$

Using an MM potential, in this case AMBER,[11] with the PARM94 parameter set, the energy of each structure was given as the sum of its geometric (bond lengths, bond angles, and torsions), van der Waals, electrostatic, and solvation energies (using a generalized Born approximation).

The heat of formation of a model protein as calculated using semiempirical methods was used as part of the energy term in the DivScore potential. The DivScore potential is a linear combination of the heat of formation, the solvation energy, and the attractive term of the Lennard-Jones energy function (eq 2).

$$\Delta G_{\text{tot}} = \Delta H_{\text{f}} + \Delta G_{\text{solv}} + \sum \text{LJ}_6 \quad (2)$$

The solvation energy was determined by solving the finite difference Poisson Boltzmann equation,[24] and the attractive portion of the Lennard-Jones term was taken from AMBER. This attractive potential was used to compensate for the poor treatment of dispersive effects by semiempirical methods. The heat of formation was determined at the AM1[25] and PM3[26] levels.

Here we compare the ability of AMBER and DivScore to identify native structures from non-native models. AMBER was chosen, not because it is the best MM-based method for decoy discrimination, but because it is a commonly used force field for studying proteins. While there may be other MM force fields or statistical-based potentials that outperform AMBER, this comparison is only used to validate the use of semiempirical methods for evaluating protein structures.

Results

Scoring Structures "As-Is". We have scored the available protein decoy sets using decoy heavy atom coordinates as-is (referred to as hydrogen-minimized since only hydrogen atoms were optimized). The results of scoring hydrogen-minimized decoys with both the AMBER and DivScore potentials are listed in the Supporting Information, Tables S1 and S2. Overall, both AMBER and DivScore demonstrate poor discrimination of decoys by score as a function of rmsd. Since the protein decoys have been created computationally using various force fields, scoring the decoy structures relative to the unmodified crystal structure introduces bias. The decoys may have geometries that are better suited to the force field, improving their score relative to the native structure.

In these studies, we aim to obtain a ranking of 1 for the native structure, which signifies that the native structure can be reliably identified from its decoys. The *Z*-score (eq 3), which provides a measure of the native structure's energy compared to all decoys, is also tabulated. It is desirable that the energy gap between the best scoring decoy and the native structure be sufficiently large to clearly identify the native structure as the best model. Ideally, a good scoring function should score native-like decoys more favorably than structurally dissimilar models.

$$Z_{\text{score}} = \frac{\Delta G_{\text{nat}} - \overline{\Delta G}}{\sigma} \quad (3)$$

Scoring Minimized Structures with AMBER**.** Because of the potential for bias when comparing structures generated through different means, an all-atom gradient-based minimization was performed on all decoys and native structures. Minimizing structures serves to clean up the models from any structural anomalies in a consistent fashion. van der Waals clashes are removed during optimization due to the large forces resulting from the steepness of the repulsive terms of the potential. In addition, bond lengths and angles are minimized with a consistent parametrized potential, removing any bias in the force field toward the native structure or its decoys.

The results of scoring with AMBER for the all-atom minimized structures are shown in Table 2 and illustrated in Figure 1. As indicated, the native structure was identified in all cases for the 4-state-reduced decoy sets. The native structure had a very favorable *Z*-score compared to the ensemble of decoys, in addition to a large energy gap separating the native structure from the highest scoring decoy. The rmsd of the best scoring decoy compared to the native state was low, being only 2.79 Å on average, while the lowest rmsd possible was, on average, 2.35 Å. Overall AMBER performs very well for identifying the native structure for the 4-state-reduced set, in agreement with previous studies of MM-based potentials.[10,27,28] This suggests that MM potentials with implicit solvent perform well for decoy discrimination, regardless of the parameter set used.

While AMBER performed well for Levitt's 4-state-reduced set, it was unable to identify the native structure in four cases for the Rosetta set: 1hsn, 1orc, 1uxd, and 2fow. Previous studies have shown that Rosetta decoys are a more challenging set for MM-based potentials.[29] Additionally, for the cases where AMBER was able to identify the native

Table 2. Decoy Ranking for All-Atom Minimized Structures Using AMBER

decoy set	system	rank_{nat}	$Z\text{-score}_{\text{nat}}$	$\Delta\Delta G_{\text{eff}}$ (kcal/mol)	$\text{rmsd}_{\text{best}}$[a] (Å)
4-state-reduced	1ctf	1	−3.85	−71.59	2.62
	1r69	1	−2.04	−77.79	3.38
	1sn3	1	−5.97	−102.70	3.29
	2cro	1	−3.90	−64.01	2.07
	4pti	1	−4.93	−51.62	2.78
	4rxn	1	−4.66	−61.61	2.61
Rosetta	1gb1	1	−4.53	−29.15	8.31
	1hsn	36	−1.66	38.20	10.20
	1orc	11	−2.45	14.56	8.65
	1pgx	1	−5.00	−32.45	4.55
	1uxd	4	−2.46	4.76	2.81
	2fow	80	−1.39	34.76	7.73
	1hc8	7	−2.34	13.06	7.73
	1r69	1	−6.81	−63.71	7.80

[a] rmsd of the best scoring decoy.

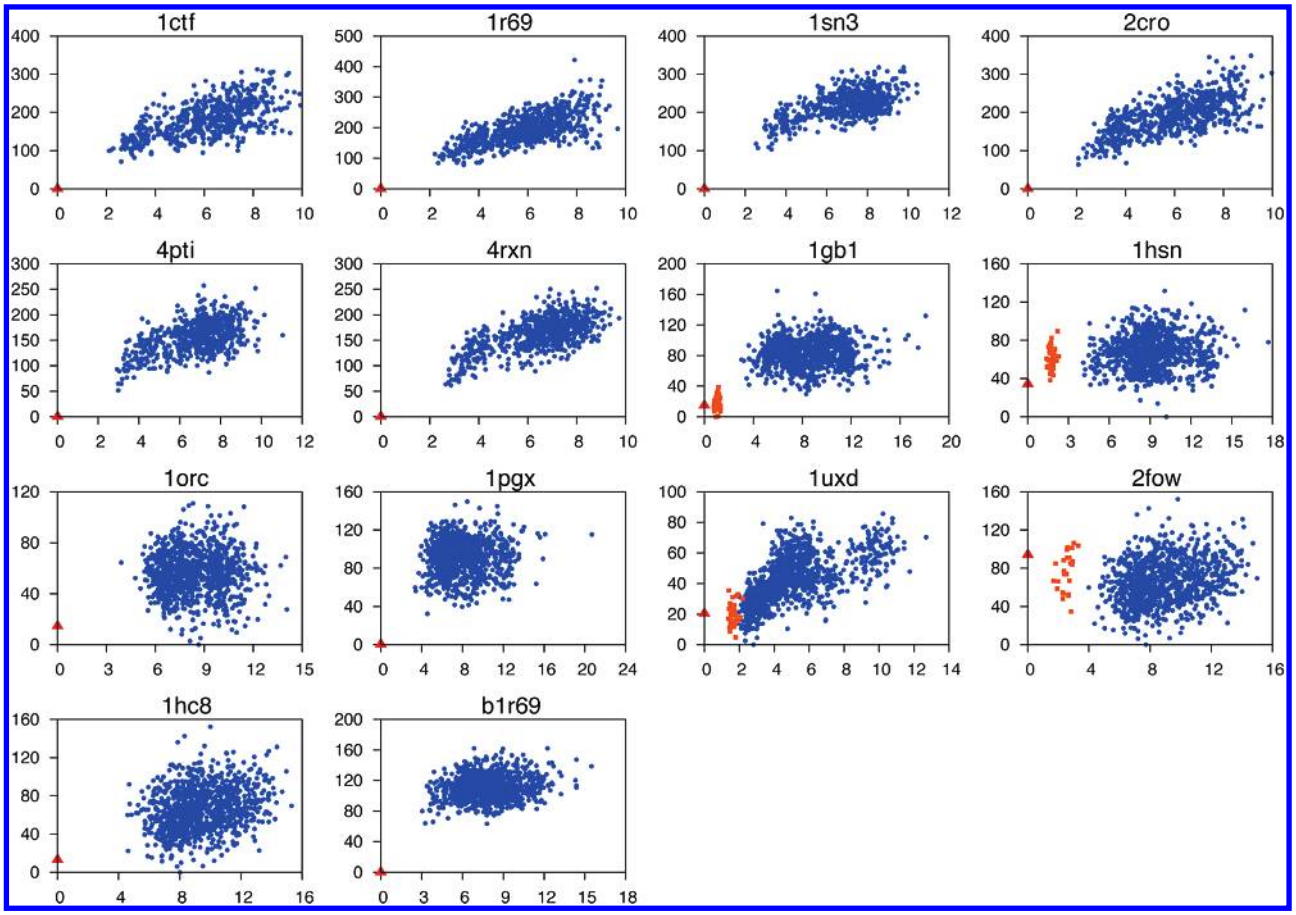

Figure 1. Energy vs rmsd plots for systems scored using the AMBER potential: (red triangle) native structure or NMR minimized mean where applicable, (orange square) individual NMR model, and (blue circle) decoy structure (*Y*-axis $\Delta\Delta G$ [kcal/mol] and *X*-axis rmsd [Å]). Energies are reported as the difference in energy for a state compared to the lowest energy structure in the decoy set.

structure, the energy gap was not as favorable as it was for the 4-state-reduced decoys. The rmsd values of the lowest scoring decoy are also generally far from the native structure.

In comparing the 4-state-reduced set to the Rosetta set, there are several possible explanations for AMBER's inability to perform well on the Rosetta decoys. All of the native structures in the 4-state-reduced set were solved using X-ray crystallography, while those in the Rosetta set are a combination of X-ray and NMR structures. Indeed, of the four cases where AMBER is unable to identify the native structure, three were NMR structures. This may suggest that AMBER structures generally score better for X-ray structures. Indeed the X-ray structure of the RNA binding domain of ribosomal protein LII (1hc8) scores significantly better than any NMR model for this system (2fow), as shown in Figure 1. It has been observed that structures are generally more stable in molecular dynamics simulations when started from an X-ray structure in comparison to an NMR structure.[30] It has also been demonstrated that NMR structures are more difficult to identify from among a set of decoys, compared to X-ray structures.[29] Rosetta decoys may also possess more nativelike structural characteristics since they were generated using fragments from the PDB.

An energy decomposition has been performed for decoys scored using the AMBER potential. As shown in Table 3, the internal geometries of the 4-state-reduced set generally favor the native state, as opposed to the Rosetta set. The energy gap for the internal geometric energy is fairly large for the Rosetta set, although most of the difference arises from the torsional energy term, with only a small fraction resulting from the bond and angle terms. Since decoys are generally less well packed than native structures, their side chains may be able to adopt more favorable torsional orientations, whereas side chains in native structures may remain strained to improve overall packing. The Lennard-Jones energy provides a reliable indicator of the native structure for the 4-state-reduced set, though less so for the Rosetta set. For the cases where the van der Waals energy was unable to identify the native structure, the energy gap to the overall-best scoring decoys was generally small. The electrostatic energy reported is a combination of the Coulombic interaction and the solvation energy. The two are generally antagonistic[28] and so have not been separated. AMBER scores well for the native structure with respect to the electrostatic energy for the 4-state-reduced set, while again performing poorly for the Rosetta set.

Scoring Minimized Structures with DivScore. As with AMBER, performing an all-atom minimization prior to scoring structures accentuates DivScore's ability to discriminate native structures from decoys. For this study, the AMBER optimized geometries are used, as it is too computationally intensive to minimize all structures at a quantum

Table 3. Energy Decomposition for the AMBER Force Field Applied to All-Atom Minimized Structures

decoy set	system	$rank_{tot}$	$rank_{int}$[a]	$\Delta\Delta G_{int}$	$rank_{vdw}$[b]	$\Delta\Delta G_{vdW}$	$rank_{ele}$[c]	$\Delta\Delta G_{ele}$
4-state-reduced	1ctf	1	4	2.47	1	−6.07	1	−24.29
	1r69	1	15	7.10	1	−35.23	1	−4.00
	1sn3	1	1	−0.53	1	−8.58	1	−26.31
	2cro	1	15	14.47	1	−24.43	2	2.79
	4pti	1	1	−7.71	2	0.29	1	−0.45
	4rxn	1	1	−11.11	4	19.21	1	−13.49
Rosetta	1gb1	1	559	23.13	1	−1.46	14	8.24
	1hsn	36	963	21.05	137	17.52	12	0.37
	1orc	11	808	27.69	162	25.75	11	14.17
	1pgx	1	865	33.74	1	−21.27	3	7.66
	1uxd	4	959	15.48	1	−19.28	37	8.56
	2fow	80	967	34.72	180	35.72	8	10.39
	1hc8	7	891	30.31	2	0.27	199	40.21
	1r69	1	977	40.30	1	−41.97	2	12.93

[a] Internal geometric energy (sum of bond, angle, and torsional energy terms). [b] van der Waals energy. [c] Electrostatic energy.

Table 4. Decoy Rankings for All-Atom Minimized Structures, Scored with DivScore Using the PM3 Hamiltonian

decoy set	system	$rank_{nat}$	Z-$score_{nat}$	$\Delta\Delta G_{eff}$ (kcal/mol)	$rmsd_{best}$[a] (Å)
4-state-reduced	1ctf	1	−4.56	−119.08	3.15
	1r69	1	−5.53	−153.98	3.37
	1sn3	1	−6.62	−145.77	4.11
	2cro	1	−5.49	−132.71	2.64
	4pti	1	−5.39	−80.18	3.02
	4rxn	1	−4.03	−55.05	3.21
Rosetta	1gb1	1	−3.03	−4.22	5.71
	1hsn	15	−2.16	22.95	7.74
	1orc	1	−4.16	−40.80	6.67
	1pgx	1	−4.64	−25.61	6.94
	1uxd	45	−1.71	24.58	2.41
	2fow	5	−2.26	26.71	6.42
	1hc8	1	−3.76	−25.77	6.37
	1r69	1	−7.04	−112.26	5.94

[a] rmsd of the best scoring decoy.

mechanical level. Previous studies have shown that AMBER-minimized protein structures should be sufficient for scoring with SE-QM methods.[31]

These results for scoring with DivScore are summarized in Table 4 for the PM3 Hamiltonian. DivScore is also able to correctly identify the native structure for all of the 4-state-reduced decoys using both PM3 (Table 4) and AM1 (data not shown). Overall, PM3 shows a slight improvement over AM1 in its ability to discriminate decoys, not only in ranking but also in Z-score and energy gap. The rmsd of the best scoring decoys as calculated with PM3 are closer to the native in general than those scored with AM1. Overall, the results indicate that PM3 performs slightly better than AMBER in regards to scoring protein structures.

Figure 2 illustrates the correlation of heats of formation as calculated by PM3 and AM1 for the 1uxd decoy set. As expected, there is a tight correlation between the two and this trend is seen for all decoy sets. Since the two Hamiltonians behave so similarly, only PM3 will be discussed below. Interestingly, there is also a very good correlation between the PM3 heats of formation and the AMBER energy, indicating that semiempirical methods may be reliably used in many of the situations where classical potentials have been successfully applied.

An energy decomposition has also been performed for the DivScore energies using the PM3 Hamiltonian (Table 5). The heat of formation of the native structure usually scores very well compared to the collection of misfolded structures. All native models, however, score very poorly with respect to the solvation energy in comparison to decoy models. While this energy gap may seem very large, the gap is reported for the native structure in relation to the decoy with the lowest free energy of solvation (not the overall best scoring decoy). Indeed, the best scoring decoys also have very unfavorable free energies of solvation, like the native state, which is usually balanced by the favorable heats of formation and dispersive interactions. All native structures possess very favorable LJ6 terms suggesting tighter packing in the native structures compared to the decoys.

In the previous discussion, the DivScore was calculated by a simple addition of the heat of formation, the solvation energy, and an attractive term to account for dispersive effects. The attractive term is taken from the AMBER force field, while the heat of formation was calculated with the PM3 Hamiltonian. Since these terms were taken from different theoretical treatments, weighted coefficients for each term in the DivScore equation have been assigned.

The coefficients were parametrized to maximize the Z-score of the native structure in relation to all decoys (or in the case of NMR structures, the lowest scoring model). The Z-score was chosen because it best represents the improvement of the native structure over the entire ensemble of decoys as opposed to $\Delta\Delta G_{eff}$, which only measures the energy gap between the native structure and the best scoring decoy. From eq 3, it can be seen that because the Z-score involves the standard deviation of the data set, it is not a linear function of the individual energy terms thus preventing a linear fitting of coefficients. A Monte Carlo method was therefore used to search the parameter space.

Of the 13 decoy sets in our study, six were chosen at random for the training set of the parametrization. A Monte Carlo method with a metropolis accepting scheme and simulated annealing was applied to optimize the parameters

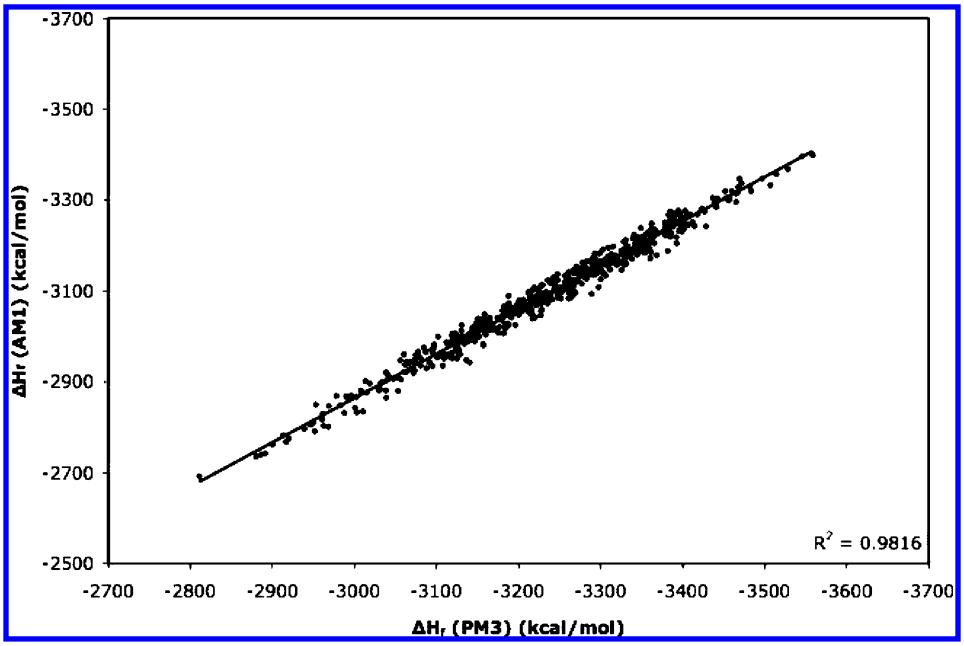

Figure 2. Correlation of heats of formation as calculated for the 1uxd decoy set using the PM3 and AM1 Hamiltonians with DivCon.

Table 5. Energy Decomposition for All-Atom Minimized Decoy Sets as Calculated with DivScore Using the PM3 Hamiltonian

decoy type	system	$rank_{tot}$	$rank_{Hf}$[a]	gap_{Hf}[d]	$rank_{solv}$[b]	gap_{solv}[d]	$rank_{LJ6}$[c]	gap_{LJ6}[d]
4-state-reduced	1ctf	1	4	44.62	471	570.51	1	−22.04
	1r69	1	2	19.53	594	559.40	1	−58.94
	1sn3	1	1	−63.47	569	613.26	5	22.06
	2cro	1	16	72.01	524	515.20	1	−59.28
	4pti	1	16	95.99	440	455.73	25	58.75
	4rxn	1	1	−32.07	646	925.86	10	34.17
Rosetta	1gb1	1	3	4.07	976	677.13	1	−13.66
	1hsn	15	154	152.73	857	447.77	1	−1.42
	1orc	1	64	183.35	812	570.97	7	6.07
	1pgx	1	1	−10.12	978	627.83	1	−60.63
	1uxd	45	21	118.11	972	343.64	1	−23.04
	2fow	5	56	173.82	894	618.72	4	16.15
	1hc8	1	24	124.08	920	644.35	1	−12.19
	1r69	1	1	−10.07	977	828.78	1	−141.21

[a] Heat of formation. [b] Solvation energy. [c] Dispersive term of the classic LJ6−12 potential. [d] Energy gaps are reported as the energetic difference between the native structure and the decoy with the lowest value for the given energetic term.

that would improve the *Z*-score of the native structure in the training sets. The training set was chosen at random, and the procedure was repeated several times to ensure that the optimal parameter set was reproduced. Equation 4 shows the weighted DivScore with appropriate coefficients for the PM3 calculations. The decoy sets that were used in the training set were 1ctf, 1gb1, 1hsn, 1pgx, and both 1r69 sets.

$$E_{tot} = 0.250 * \Delta H_f + 0.225 * \Delta G_{solv} + 0.525 * LJ_6 \quad (4)$$

Using the weighted individual energy terms, the decoy sets were rescored and reranked. Table 6 shows the results of using weighted energy terms with the PM3 Hamiltonian. As illustrated, there is a marked improvement of the rankings for 1hsn, 1uxd, and 2fow. As expected, there is a general improvement in *Z*-scores for native structures, consistent with the weighting coefficients designed to improve the discriminatory ability of the function. The energy distributions for each decoy set scored with the weighted DivScore are shown in Figure 3.

For comparison purposes, the energetic terms in the AMBER potential were also reweighted using the same procedure. The reweighted terms are shown in eq 5. The weights for the geometric energy terms (bond length, angle, and torsion) were zero and so are neglected in the equation.

$$E_{tot} = 0.378 * E_{vdw} + 0.310 * E_{ele} + 0.312 * E_{solv} \quad (5)$$

The results of scoring with the reweighted AMBER potential are shown in Table 7, and it performs well at identifying native structures from decoy models. In general, however, the *Z*-scores and free energy gaps are not as favorable as those for the weighted DivScore.

Table 6. Decoy Rankings for All-Atom Minimized Structures as Calculated with Weighted DivScore Using the PM3 Hamiltonian

decoy set	system	$rank_{nat}$	$Z\text{-score}_{nat}$	$\Delta\Delta G_{eff}$ (kcal/mol)	$rmsd_{best}$ (Å)
4-state-reduced	1ctf	1	−5.43	−55.41	3.15
	1r69	1	−6.36	−71.66	4.02
	1sn3	1	−5.70	−47.15	7.81
	2cro	1	−5.73	−58.67	2.07
	4pti	1	−4.35	−4.35	7.46
	4rxn	1	−4.39	−22.55	2.86
Rosetta	1gb1	1	−4.37	−16.77	3.66
	1hsn	1	−3.05	−2.23	9.57
	1orc	1	−3.48	−13.22	10.21
	1pgx	1	−5.26	−31.65	7.22
	1uxd	1	−3.19	−5.15	2.73
	2fow	2	−2.96	3.01	6.37
	1hc8	1	−4.30	−18.54	6.37
	1r69	1	−8.71	−76.19	6.69

The scoring results for the ribosomal protein LII decoy set are plotted in Figure 3 using the weighted DivScore with the PM3 Hamiltonian. In addition to NMR models (2fow), there is also an X-ray structure available, 1hc8. The X-ray structure obtains a ranking of 1, correctly identifying it as the native structure. None of the NMR models obtain a native ranking, although some score more favorably than others. It should be noted that, overall, there is only a small difference in energy between NMR structures. Structurally, there does not appear to be any clear defining feature distinguishing the X-ray structure from the best and worst NMR models indicating that the differences between the two may be subtle. Structural verification checks show that the X-ray structure has much better packing than the NMR models, and this trend is seen by the improved Lennard-Jones interactions of the X-ray model.

The results of scoring with DivScore show that this scoring function is particularly well suited for identifying the native structure from among all decoys. All native structures can be correctly identified, although in the case of ribosomal protein LII it is the X-ray structure (1hc8) that is identified rather than the NMR models (2fow). The *Z*-scores for all native structures are large, indicating that the potential function scores the native structure much better than the set of decoys. The energy gaps between the native structure and the best-scoring decoy are large for the 4-state-reduced set, although noticeably smaller for the Rosetta decoys.

Scoring Near-Native Decoys. The 4-state-reduced and Rosetta decoy sets studied here lack near-native decoys with an rmsd less than 2.0 Å. As the field of structure prediction advances, it is important to clearly identify near-native decoys from those with higher RMSDs. A set of near-native, low rmsd decoys was obtained (Bradley, P., personal communication) that was used in ab initio folding studies[5] and is listed in Table 8. These decoys sets provide a more stringent test of our scoring methods, as some models are as low as 0.55 Å from the native. The results of scoring this set with DivScore are shown in Figure 4. Encouragingly, DivScore was able to discriminate native from near-native folds for all four decoy sets. In the case of 1r69 and 1di2, favorable energy funnels are observed, whereas 1af7 and 2reb demonstrate limited ability to discriminate near-native decoys from those that are more structurally divergent.

The results of these studies demonstrate the utility of semiempirical methods for studying protein structures, matching, or exceeding the AMBER force field at discriminating native structures. Given the fact that classical potentials have been parametrized for biological molecules, it is surprising that semiempirical methods performed so well—even given the fact that they are known to give poor conformational profiles. Moreover, semiempirical methods were parametrized at the element level, and some of the functional groups found in proteins were not included in the parametrization set.[26]

This study hints at the possibility of using quantum mechanical methods in large-scale folding studies, although to reduce the time taken they should be coupled with lower-resolution scoring models to remove obviously poor structures. As expected, correctly scoring protein structures is dependent upon first minimizing the structure, preferably with the same potential being used to score the model. Performing an all-atom minimization with the AMBER force field cleaned up all structures and ensured that structures could be compared without bias to their starting structures. While it would have been an interesting study to minimize all decoys at the semiempirical level before scoring with DivCon, such calculations are too costly at present.

It is worth considering why semiempirical models score protein decoys as well as we have found in the present study. Semiempirical methods are known not to give phi-psi plots that agree with high quality ab initio results,[32] while force fields are generally parametrized to reproduce these plots at some level of accuracy. This suggests that other factors play a role like long-range electrostatics or cooperativity effects observed in the folding of secondary structural elements.[33] Possibly these effects are overwhelming the conformational effects when using semiempirical methods in scoring native and decoy protein structures.

Conclusion

Here we have validated the capability of using semiempirical quantum mechanics in detecting misfolded proteins relative to the natively folded target protein. The ability of semiempirical methods to detect the native structure from a collection of decoys is quite remarkable and hints that the use of ab initio or density functional methods would also have significant potential in this regard. While the present test was for decoy detection one can envision using semiempirical approaches to facilitate the refinement of homology models as well as having an impact on the protein folding problem. Further studies along these lines are underway.

Evaluating native structures from a collection of misfolded states provides a challenging test for scoring functions and provides insights into the manner in which they fail. Energy decompositions are particularly useful because they highlight the terms in a scoring function that may require reparameterization or further study. AMBER energy decompositions indicate that for most minimized structures, there will be

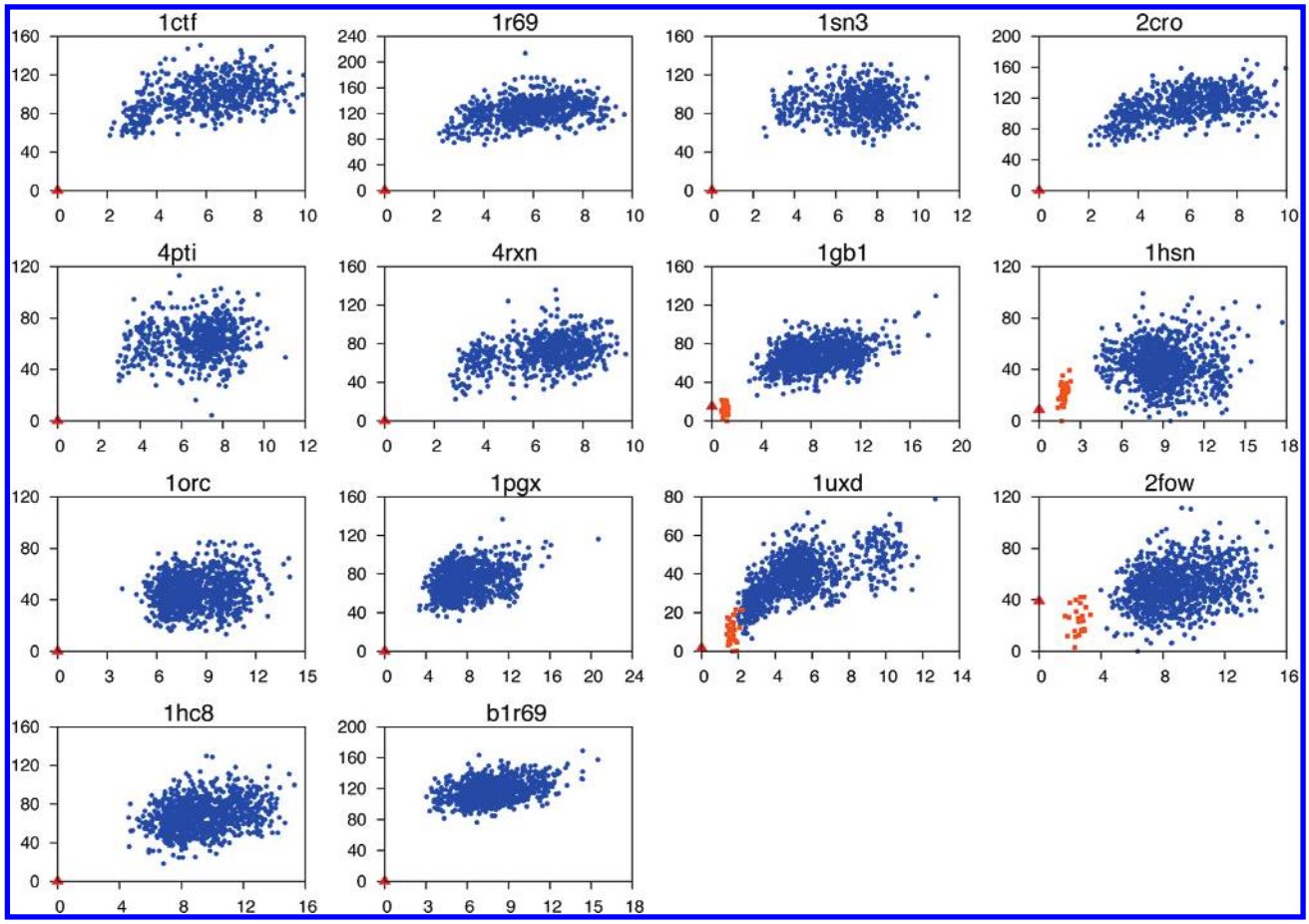

Figure 3. Energy vs rmsd plots for systems scored using the DivScore potential: (red triangle) native structure or NMR minimized mean where applicable, (orange square) individual NMR model, and (blue circle) decoy structure (*Y*-axis $\Delta\Delta G$ [kcal/mol] and *X*-axis rmsd [Å]). Energies are reported as the difference in energy for a state compared to the lowest energy structure in the decoy set.

Table 7. Decoy Rankings for All-Atom Minimized Structures as Calculated with the Reweighted AMBER Potential

decoy set	system	rank_{nat}	$Z\text{-score}_{\text{nat}}$	$\Delta\Delta G_{\text{eff}}$ (kcal/mol)	$\text{rmsd}_{\text{best}}$ (Å)
4-state-reduced	1ctf	1	−4.85	−23.94	2.93
	1r69	1	−3.91	−26.81	2.65
	1sn3	1	−6.41	−26.72	2.87
	2cro	1	−4.97	−23.94	2.46
	4pti	1	−4.45	−9.04	3.67
	4rxn	1	−4.52	−9.46	2.86
Rosetta	1gb1	1	−4.06	−5.42	4.15
	1hsn	3	−3.05	3.48	9.52
	1orc	1	−2.71	−0.14	10.90
	1pgx	1	−5.37	−15.74	4.99
	1uxd	1	−3.03	−1.74	3.09
	2fow	1	−3.01	−0.10	8.04
	1r69	1	−7.95	−30.49	4.20

little energetic difference resulting from internal geometric differences. Rather, van der Waals and electrostatic contributions to the energetic state appear to be most important in specifying native folds. However, the solvation energy of the native structures usually scores poorly with respect to the decoy ensemble. While this might initially signify that the solvation energetic terms may need to be revisited, it actually highlights an important point in scoring proteins; native structures are marked by a competition of various energetic terms. No single term dominates in the energetic treatment of a protein[3] as the native state rarely has the tightest packing, the most favorable charge interactions, or the most favorable solvation energy. Instead, the native structure seems tuned to be the best combination of various energetic terms so that it can perform its function.

Methods

In this study, we investigated the ability of semiempirical methods to identify native structures from decoys in subsets of both the 4-state-reduced and Rosetta decoy sets, listed in Table 1. The Rosetta decoy set is composed of 92 different decoy sets, with systems ranging from 17 to 146 residues in length. The decoy sets also differed in their distribution of rmsd values from the native structure with some sets having all decoys over 8 Å away from the native fold. For this study, only small proteins (>45 residues long) were chosen whose decoys were distributed over a large range of conformational space and had structures with an rmsd < 5.0 Å. For sets whose native structure was solved by NMR, the minimized mean structure was taken to be the "native" structure for calculating the rmsd of a given decoy to the native conformation. However, when scoring NMR structures, each structure in the reported ensemble was treated, and the best

Table 8. Near-Native Rosetta Decoy Sets

PDB	description	type	N_{decoys}	N_{res}	N_{atoms}	rmsd range (Å)	%H/%E[a]
1af7	N-terminal domain of Chemotaxis receptor methyltransferase CheR	X-ray	999	72	1217	1.6−12.1	68/0
1di2	RNA binding protein A	X-ray	999	69	1096	0.8−9.4	42/33
1r69	N-terminal domain of phage 434 repressor	X-ray	999	61	976	1.0−7.5	70/0
2reb	RecA	X-ray	999	60	884	0.5−3.9	20/57

[a] Percentage of helices and sheets in the native.

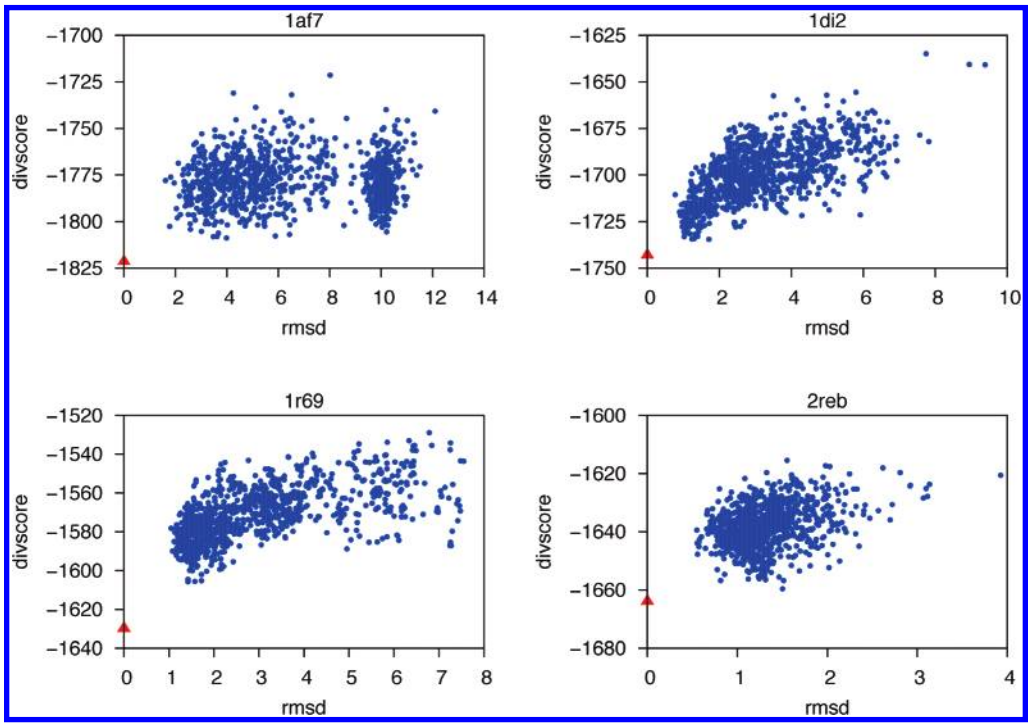

Figure 4. DivScore vs rmsd plots for systems with near-native decoys: (red triangle) native X-ray structure and (blue circle) decoy structure.

scoring conformation was reported. The RNA binding domain of ribosomal protein LII (2fow) has also been solved by X-ray crystallography, and so this structure (1hc8) has been treated as another native reference for this system.

Using the available decoy sets, several analyses have been performed: (1) using protein decoy heavy atom coordinates as-is (referred to as hydrogen-minimized since only hydrogen atoms were optimized) and (2) minimizing all structures with the AMBER force field prior to analysis (referred to as all-atom minimized structures). In all cases, hydrogen atoms were removed from all structures and then added with the LEaP module of AMBER ensuring that all hydrogen atoms were modeled in a consistent manner. All histidine residues were treated as a singly protonated species with the N^{δ} atom being protonated. Several decoy systems contained multiple cysteine residues. These were treated as disulfide bonds when the geometry was perceived to be favorable to disulfide bond formation. Upon hydrogen addition, hydrogen atoms were minimized using the Sander package in AMBER using the PARM94 force field for 300 steepest descent steps followed by 700 conjugate gradient steps. The resulting structures were then scored using the AMBER MM potential as well as with semiempirical QM methods found in DivCon, at both the AM1[25] and PM3[34] levels. On current hardware (2.4 GHz AMD opteron) a DivCon single point energy evaluation takes on average 15 min.

Beyond scoring decoys "as-is", an all-atom minimization was performed on all decoys. First, a restrained all-atom minimization was performed for 1500 steps with a force constant of 2.0 kcal/mol $Å^2$ applied, followed by an unrestrained minimization for 5000 steps with implicit solvation. The initial restrained minimization was performed to limit instabilities during the minimization caused by nonoptimal starting structures. Minimization with the AMBER/GBSA module resulted in structures that did not deviate significantly from the starting structure, with an average heavy-atom rmsd of 1.1 Å.

Abbreviations: molecular mechanics (MM); quantum mechanics (QM); root mean squared deviation (rmsd).

Acknowledgment. We thank Kenneth Ayers for his assistance with managing the computational resources and for valuable discussions. We would also like to thank Philip Bradley for supplying us with near-native decoy sets. We thank the NSF (MCB-0211639) and the NIH (GM44974 and GM066859) for financial support of this research and the National Center for Supercomputer Applications (NCSA) and the Pittsburgh Supercomputer Center for generous allocations of supercomputer time.

Supporting Information Available: Rankings of preminimized decoy sets scored with AMBER and DivScore are available. Scoring decoys as-is accentuated differences between the scoring force field and the method used to generated the structures. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **1973**, *181* (96), 223−30.

(2) Schonbrun, J.; Wedemeyer, W. J.; Baker, D. Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* **2002**, *12* (3), 348−54.

(3) Lazaridis, T.; Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000**, *10* (2), 139−45.

(4) Moult, J. A. decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **2005**, *15* (3), 285−9.

(5) Bradley, P.; Misura, K. M.; Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **2005**, *309* (5742), 1868−71.

(6) Zhang, Y.; Arakaki, A. K.; Skolnick, J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* **2005**, *61 Suppl 7*, 91−8.

(7) Park, B.; Levitt, M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* **1996**, *258* (2), 367−92.

(8) Park, B. H.; Huang, E. S.; Levitt, M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **1997**, *266* (4), 831−46.

(9) Hendlich, M.; Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J.; 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* (1), 167−80.

(10) Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. 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. *Proteins* **2002**, *48* (2), 404−22.

(11) Cornell, W. D.; Cieplak, P.; Baylay, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field For the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc* **1995**, *117*, 5179−5197.

(12) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy Minimization and Dynamical Calculations. *J. Comput. Chem.* **1983**, *4*, 182−217.

(13) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-atom Force Field on Conformational Eneregetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(14) Wang, B.; Liao, K. R. N.; Peters, M. B.; Kim, H.; Westerhoff, L. M.; Wollacott, A. M.; van der Vaart, A.; Gongonea, V.; Suarez, D.; Dixon, S. L.; Vincent, J. J.; Brothers, E. N.; Merz, K. M., Jr. *DivCon* 2005.

(15) Dixon, S. L.; Merz, K. M. Semiempirical molecular orbital calculations with linear system size scaling. *J. Chem. Phys.* **1996**, *104* (17), 6643−6649.

(16) Yang, W. T.; Lee, T. S. A Density-Matrix Divide-and-Conquer Approach for Electronic-Structure Calculations of Large Molecules. *J. Chem. Phys.* **1995**, *103* (13), 5674−5678.

(17) Dixon, S. L.; Merz, K. M. Fast, accurate semiempirical molecular orbital calculations for macromolecules. *J. Chem. Phys.* **1997**, *107* (3), 879−893.

(18) van der Vaart, A.; Merz, K. M., Jr. The Role of Polarization and Charge Transfer in the Solvation of Biomolecules. *J. Am. Chem. Soc.* **1999**, *121*, 9182−9190.

(19) van der Vaart, A.; Suarez, D.; Merz, K. M. Critical assessment of the performance of the semiempirical divide and conquer method for single point calculations and geometry optimizations of large chemical systems. *J. Chem. Phys.* **2000**, *113* (23), 10512−10523.

(20) Raha, K.; Merz, K. M., Jr. A Quantum Mechanics Based Scoring Function: Study of Zinc-ion Mediated Ligand Binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020−1021.

(21) Simons, K. T.; 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* (1), 209−25.

(22) McConkey, B. J.; Sobolev, V.; Edelman, M. Discrimination of native protein structures using atom-atom contact scoring. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (6), 3215−20.

(23) Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins: Struct., Funct., Genet.* **2003**, *53* (1), 76−87.

(24) Gogonea, V.; Merz, K. M., Jr. Fully Quantum Mechanical Description of Proteins in Solution. Combining Linear Scaling Quantum Mechanical Methodologies with the Poisson-Boltzmann Equation. *J. Phys. Chem. A* **1999**, *103*, 5171−5188.

(25) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(26) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. 2. Applications. *J. Comput. Chem.* **1989**, *10* (2), 221−264.

(27) Lazaridis, T.; Karplus, M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **1999**, *288* (3), 477−87.

(28) Dominy, B. N.; Brooks, C. L. Identifying native-like protein structures using physics-based potentials. *J. Comput. Chem.* **2002**, *23* (1), 147−60.

(29) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. Molecular dynamics in the endgame of protein structure prediction. *J. Mol. Biol.* **2001**, *313* (2), 417−30.

(30) Lee, M. R.; Kollman, P. A. Free-energy calculations highlight differences in accuracy between X-ray and NMR structures and add value to protein structure prediction. *Structure (London)* **2001**, *9* (10), 905−16.

(31) Wollacott, A. M.; Merz, K. M. Development of a Parametrized Force Field To Reproduce Semiempirical Geometries. *J. Chem. Theory Comput.* **2006**, *2* (4), 1070−1077.

(32) Mohle, K.; Hofmann, H. J.; Thiel, W. Description of peptide and protein secondary structures employing semiempirical methods. *J. Comput. Chem.* **2001**, *22* (5), 509−520.

(33) Morozov, a. V.; Tsemekhman, K.; Baker, D. Electron density redistribution accounts for half the cooperativity of alpha helix formation. *J. Phys. Chem. B* **2006**, *110* (10), 4503−4505.

(34) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods. 1. Method. *J. Comput. Chem.* **1989**, *10* (2), 209−220.

CT600325Q