PROTEINS: Structure, Function, and Bioinformatics 55:620–634 (2004)

Distinguish Protein Decoys by Using a Scoring Function Based on a New AMBER Force Field, Short Molecular Dynamics Simulations, and the Generalized Born Solvent Model

Mathew C. Lee and Yong Duan*

Department of Chemistry and Biochemistry and Center of Biomedical Research Excellence in Structural and Functional Genomics, University of Delaware, Newark, Delaware

ABSTRACT **Recent works have shown the ability of physics-based potentials (e.g., CHARMM and OPLS-AA) and energy minimization to differentiate the native protein structures from large ensemble of non-native structures. In this study, we extended previous work by other authors and developed an energy scoring function using a new set of AMBER parameters (also recently developed in our laboratory) in conjunction with molecular dynamics and the Generalized Born solvent model. We evaluated the performance of our new scoring function by examining its ability to distinguish between the native and decoy protein structures. Here we present a systematic comparison of our results with those obtained with use of other physics-based potentials by previous authors. A total of 7 decoy sets, 117 protein sequences, and more than 41,000 structures were evaluated. The results of our study showed that our new scoring function represents a significant improvement over previously published physics-based scoring functions. Proteins 2004;55:620–634.**

Key words: computational structure prediction; deliberately misfolded proteins; potential energy function; *z* scores; protein folding

INTRODUCTION

Methods of computational protein structure prediction are generally rooted in the thermodynamic hypothesis that the native-state conformation is the most stable conformation and, therefore, must occupy the lowest energetic state.[1] Although there was one recent example in which the native state of the α-lytic protein appeared to be less stable than its molten-globular intermediate and denatured states, closer examination revealed that the enthalpy of the native state is actually lower than the misfolded state by about 18 kcal/mol and that the apparent stability of the misfolded state is due to the increase in conformational entropy on unfolding.[2,3] This observation at first glance seems to nullify the underlying assumption of most structure prediction methods, but with proper interpretation, it actually supports the case of thermodynamic hypothesis. Even in this extreme example in which the native state is kinetically trapped, the effective free energy of the native state (the free energy of the protein plus solvent at a fix conformation[4,5]) remains the lowest. However, the ruggedness of the energy landscape also dictates that an ensemble of local minimum energy states around the native state exists. Therefore, effective energy functions that can accurately depict the energy landscape of protein conformation space are a common requirement for all computational approaches to the prediction problem. This discriminatory requirement was formulated as the "principle of minimum frustration" by Bryngelson and Wolynes.[6]

Three major classes of prediction methods are in use today: homology modeling, threading/fold recognition, and ab initio folding. Regardless of which class a prediction method belongs to, an effective energy function is usually required. These functions are typically used in one of two ways: they are either used as optimization criteria to drive conformational search algorithms to sift through the conformational space (folding problem) or they are used as selection criteria to select a conformation from a set of possible structures in fold recognition applications (reverse folding problem). Although the exact design of an effective energy function depends on the type of problem one wants to tackle, several factors constrain the final form that an energy function ultimately assumes.[7] For example, in fold recognition applications, one is primarily concerned with the backbone geometry of a protein; thus, one is afforded a greater freedom in simplifying the representation of the side-chains. In homology modeling or molecular dynamics-based ab initio folding in which atomic details are required, reduced representation may not be sufficient; molecular mechanics force field-based energy functions that account for full atomic details are thought to be better suited for such applications. Depending on the method from which an energy function is derived, it is classified as one of three types: knowledge or statistics-

*Correspondence to: Yong Duan, Department of Chemistry and Biochemistry and Center of Biomedical Research Excellence in Structural and Functional Genomics, University of Delaware, Newark, DE 19716. E-mail: yduan@udel.edu

Received 24 January 2003; Accepted 14 March 2003

Published online 1 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.10470

based, physics-based, and hybrids that combine elements of both knowledge and physics-based energy functions.

Until very recently, knowledge-based scoring functions have been the mainstay in structure prediction applications.[8–10] Given a database of high-quality structural information, knowledge-based scoring functions can often produce the desired results with far less computational overhead. However, it is generally recognized that molecular mechanics force fields derived from ab initio quantum calculations have several advantages over these statistically derived energy functions. For example, energy functions derived from experimental structure information are constrained by their underlying statistics, which means that their accuracy and applicability are intrinsically tied to the data source used for parameterization; if the particular data set overrepresents a certain class of structural properties (e.g., helical structures), the resulting energy function would also reflect this statistical bias in its scoring. Molecular mechanics force fields, on the other hand, do not have such inherent limitations when they are carefully parameterized. Because they are derived from ab initio quantum mechanical calculations based on the principles of physics alone, they do not have any intrinsic bias toward any particular structural properties.

Despite their perceived advantages, physics-based all-atom molecular mechanics force fields have not been widely considered practical for fold recognition or protein structure prediction types of applications. This finding was mostly due to the high-computation cost required and the cumulative inaccuracies introduced in parameterization of the force fields compounded by the fact that most of the earlier force fields were calibrated against rather sparse and often qualitative experimental data. Because of the continued improvement in computer speed and advances in force-field design, this situation has begun to change in recent years; physics-based energy functions are now showing signs of living up to their potential.[11–13]

As more new energy functions are being developed, one problem that became apparent was the lack of a standardized benchmark to allow comparisons of performances across different energy functions parameterized by using different properties and methodologies. One of the earliest studies that resembled a benchmarking test for protein potentials was the study carried out by Novotny et al.,[14] in which two proteins with the same number of residues but different folds were considered and the sequence of one was "threaded" onto the fold of the other. The resulting correct and incorrect models were then evaluated with use of the CHARMM potential. The conclusion of this study was frequently misinterpreted as supporting evidence that modern molecular mechanics-based potentials were not of sufficient accuracy to discriminate between native and non-native folds. In the spirit of the Novotny test, several groups have created decoy sets (non-native or near-native conformations) as a testing benchmark for evaluating the usefulness of a new scoring function.[15–18] Because structural information of these decoy sets are generally not used as part of the source information for parameterization of energy functions, these decoy sets provide an objective common platform on which new energy functions can be evaluated. Even though the issue of statistical bias is not a real problem for force field-based energy functions, the use of decoy tests in the evaluation process can still provide valuable insights to the characteristics of an energy function.[19,20] Furthermore, one may also view the decoys as the product of a previous step in a hierarchical structure prediction scheme; a high-quality scoring function could then be integrated as a filter to select the best candidate from among a set of low-resolution prediction of the native fold.

In this article, we present an energy-scoring function based on the new AMBER molecular mechanics force field recently developed in our laboratory (details of this force field are being prepared for a separate article to be submitted elsewhere) and its performance on the Novotny test. Although several physics-based potentials have already been published,[11,13,21–23] among them a Monte Carlo Simulated Annealing (MCSA) protocol based on the AMBER united atom potential functions,[24] our approach differs from these previous approaches in two significant ways: 1) most molecular mechanics force fields developed so far have been parameterized by using HF 6-31G* or lower level of quantum mechanical calculations with gas-phase charge fitting; to our best knowledge, our new force field is the first example parameterized with condensed-phase charge fitting at the DFT ccpVTZ level of quantum mechanical calculations; 2) although there were examples of scoring functions using molecular mechanics force field in conjunction with molecular dynamics,[23,25,26] these earlier results showed that the structures tested were generally worse after vacuum molecular dynamics compared to those obtained by vacuum minimization and that inclusion of solvation effects is required to achieve correct results. With our new force field, we want to reexamine molecular dynamics as a viable approach to developing energy scoring functions.

Another important result that has significant implications for the implementation of effective energy functions is that the internal entropy changes of native, misfolded, or denatured conformations were found to be roughly the same.[26,27] This result has allowed energy function developers to justifiably ignore the entropic contribution in their implementation of the energy function, because the entropic terms are difficult to estimate by simulation methods. In algebraic terms, the effective free energy of a protein conformation can now be expressed as follows:

$$\Delta G_{\text{eff}} \approx \Delta H_{\text{internal}} + \Delta G_{\text{solvation}} \quad (1)$$

where $\Delta H_{\text{internal}}$ is the intramolecular free energy (not including the entropic contribution) and $\Delta G_{\text{solvation}}$ accounts for the solvation free energy.

Although the conformational enthalpy due to bond angle, torsion, and so forth can be estimated by the corresponding mechanical energy terms in a molecular mechanics force field and the conformational entropy change can be ignored, the solvation free energy term presents quite a different set of challenges. In a previous study by Lazaridis et al.,[23] when vacuum molecular dynamics was applied to

the EMBL decoy set, it was found that molecular dynamics was unable to distinguish between the correctly folded and misfolded conformations due to the strong interactions between ionic side-chains in vacuum. This result made it clear that solvent effect must be included in the energy functions. Although molecular mechanics force fields can account for solvation effects by explicitly including solvent atoms, the large number of particles required to achieve adequate solvation quickly renders molecular dynamics impractical because the size of the protein and the number of conformations needed to be evaluated grew larger. The answer to this problem, many believed, was simplified solvent models that do not require explicit representation of solvent atoms.

Over the past few years, significant progress has been made in the development of continuum solvent models. Much of these works were based on the dielectric continuum approach and the Poisson-Boltzmann (PB) equation. Roux and Simonson[28] have written an excellent review on this topic. Several groups reported earlier that by using all-atom protein models in conjunction with molecular exposed surface area type of solvent models, it was possible to distinguish native from non-native structures using this approach.[21,29–31] Soon thereafter, a number of studies involving different types of all-atom force fields and various treatments of the solvation effects ensued. These included the study by Monge et al.[32] using the AMBER force field developed by Weiner et al.[33] in conjunction with the Generalized Born (GB) continuum solvent model,[34] and the method using molecular dynamics in conjunction with SIMS and FAMBE by Vorobjev et al.[26] Several groups also used the CHARMM19 protein force field in concert with implicit solvent models and subjected their scoring functions to decoy tests. These included the study by Lazaridis et al.[23] using a Gaussian solvation shell model, the study by Petrey and Honig[35] using a dielectric continuum model based on the PB equation, and the study by Dominy and Brooks[11] that used the GB model. In the most recent study, Felts et al.[13] have used the OPLS all-atom force field[36] in conjunction with the Surface Generalized Born (SGB)[37,38] model as the basis of their scoring function.

Among the above-mentioned methods, the use of GB appeared to be a promising approach. Therefore, in this first iteration effective energy function using our new force field parameters, we also elected to use GB as our solvent model. The implementation details and the performance results of our energy scoring function are described in the following sections.

MATERIALS AND METHODS

Protein Decoys

Several preconstructed decoy sets have been made publicly available. In this study, we tested our scoring function against two types of decoy sets. The first type is a simple true-false test in which the database contains pairs of structures, one correctly folded and one incorrectly folded. The performance of a scoring function in this case is judged by the percentage of correctly identified decoy pairs. The second type of decoy sets we tested was a best-of-the-lot type of test. In these decoy sets, each protein sequence is associated with a large number of near-native or non-native decoy structures. In this case, the performance of a scoring function is judged by its ability to identify from within the decoy set the structure having the least C-α root-mean-square deviation (RMSD) to the native structure.

For the true-false type of test, we chose the EMBL deliberately misfolded database[39] as our testing set. This database was constructed by swapping side-chains on pairs of crystallographically determined structures with the same number of residues while keeping the backbone geometry unperturbed. The resulting chimeric structures were then subjected to 500 steps of steepest descent energy minimization using the GROMOS program. The entire database contains a total of 26 decoy pairs (Tables I and II).

For the best-of-the-lot type of test, we conducted a much more extensive testing using multiple decoy sets from multiple research groups. A total of seven decoy sets were used for this test, six of which were obtained from the Decoys 'R' Us database,[16] and the other one was provided to us courtesy of Lu and Skolnick.[18]

The six decoy sets obtained from the Decoys 'R' Us database are as follows: 1) The four-state reduced set in which the CA positions for these decoys were generated by exhaustively enumerating 10 selectively chosen residues in each protein using a four-state off-lattice model. All other residues were assigned the phi/psi values based on best fit of a four-state model to the native chain.[15] 2) The fisa set, which contains decoys for four small alpha-helical proteins. The main chains for these decoys were generated by using a fragment insertion simulated annealing procedure to assemble native-like structures from fragments of unrelated protein structures with similar local sequences using Bayesian scoring functions.[40] 3) The baker_casp3 set, which contains decoys for proteins predicted by the Baker group for CASP3[41] using the same protocol as in fisa set. 4) The hg_structal set, which contains decoys for 29 globins. Each globin has been built by comparative modeling using 29 other globins as templates with the program segmod. 5) The lattice_ssfit set, which contains conformations for eight small proteins generated by ab initio methods.[42,43] 6) The local minima decoy set (LMDS) contains decoys that were derived from the experimental secondary structures of 10 small proteins that belong to diverse structural classes. Specifications of these decoy sets are summarized in Tables III–VIII.

The Lu and Skolnick decoy data set that we analyzed contains 54 unique protein sequences: 28 have NMR determined structures deposited in the PDB, and 26 have X-ray structures with resolutions ranging from 2.5 to 1.2 Å. Each of these sequences has 1333 decoy structures associated with it. The decoys in this set were generated on a lattice by using an ab initio Monte Carlo structure prediction program.[18] Full atomic details were filled later into the lattice structure. Details of the Skolnick set are summarized in Table IX.

TABLE I. AMBER Energies for the EMBL Deliberately Misfolded Database[†]

Sequence PDB	Backbone PDB	RMSD (Å)	$N_{residues}$	Misfold minimized	Native minimized	Misfold MD	Native MD	ΔE	ΔE/RMSD	ΔE/Res
1bp2	2paz	15.79	123	−2855.2	−3178.7	−1815.6	−1943.0	−127.4	−8.1	−1.0
1cbh	1ppt	10.62	36	**−837.6**	**−834.3**	−532.1	−546.0	−13.9	−1.3	−0.4
1fdx	5rxn	9.52	54	−730.9	−865.3	−443.4	−460.4	−17.0	−1.8	−0.3
1hip	2b5c	14.62	85	−2015.8	−2188.3	−1275.8	−1399.0	−123.1	−8.4	−1.4
1lh1	2i1b	17.81	153	−3447.6	−3839.9	−1944.2	−2288.1	−343.9	−19.3	−2.2
1p2p	1m3	18.97	124	−3092.6	−3115.4	−1937.3	−2013.9	−76.6	−4.0	−0.6
1ppt	1cbh	10.87	36	−1130.5	−1405.1	−995.5	−1042.1	−46.6	−4.3	−1.3
1rei*	**5pad**	**N/A**	**107**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
1rhd	2cyp	22.21	293	−7661.8	−8179.3	−5030.4	−5429.0	−398.6	−17.9	−1.4
1rn3	1p2p	18.77	124	−3395.4	−3661.4	−2308.7	−2473.3	−164.6	−8.8	−1.3
1sn3	2ci2	13.64	65	−1339.0	−1484.0	−764.8	−840.6	−75.9	−5.6	−1.2
1sn3	2cro	11.25	65	−1277.1	−1484.0	−763.1	−840.6	−77.5	−6.9	−1.2
2b5c	1hip	14.74	93	−2778.2	−2988.8	−1970.6	−2142.7	−172.1	−11.7	−1.9
2cdv	2ssi	14.60	107	−1914.7	−2194.7	−1011.7	−1107.7	−96.0	−6.6	−0.9
2ci2	1sn3	13.43	83	−2021.1	−2296.5	−1452.2	−1600.0	−147.8	−11.0	−1.8
2ci2	2cro	11.68	83	−2073.1	−2296.5	−1491.4	−1600.0	−108.6	−9.3	−1.3
2cro	1sn3	11.89	71	−2020.4	−2378.5	−1481.8	−1660.5	−178.7	−15.0	−2.5
2cro	2ci2	11.24	71	−2135.8	−2378.5	−1494.3	−1660.5	−166.2	−14.8	−2.3
2cyp	1rhd	21.91	294	−7445.7	−8858.2	−4919.1	−5614.8	−695.7	−31.8	−2.4
2i1b	1lh1	17.87	153	−4184.9	−4475.0	−2745.8	−2858.1	−112.3	−6.3	−0.7
2paz	1bp2	15.59	123	−2626.2	−2995.8	−1535.5	−1732.9	−197.4	−12.7	−1.6
2ssi	2cdv	15.10	113	−2148.0	−2170.0	−1262.5	−1370.0	−107.5	−7.1	−1.0
2tmn	2ts1	23.06	317	**−7925.4**	**−4796.0**	−5057.1	−5509.9	−452.8	−19.6	−1.4
2ts1	2tmn	23.12	317	−9535.2	−11035.0	−7085.2	−7608.1	−522.9	−22.6	−1.6
5pad	1rei	19.09	214	−3601.6	−5664.6	−3264.2	−3710.2	−446.0	−23.4	−2.1
5rxn	1fdx	9.68	54	−1314.2	−1522.8	−932.4	−1046.2	−113.8	−11.8	−2.1
Averages		15.48	129	−3180.3	−3451.5	−2140.6	−2339.9	−199.3	−11.6	−1.4

[†]Bold faced numbers indicate incorrectly predicted structures. RMSD, root-mean-square deviation between the native structure and the misfolded structure; $N_{residues}$, number of residues in the protein; Misfold minimized, minimized energy of the misfolded structure in kcal/mol; Native minimized, minimized energy of the native structures in kcal/mol; Misfold MD, AMBER/GB-MD energy score of the misfolded structure in kcal/mol; Native MD, AMBER/GB-MD energy score of the native structure in kcal/mol; ΔE, energy score difference between misfold and native.

TABLE II. EMBL Set Correlation Matrix[†]

	N_{res}	ΔE	ΔE/RMSD	RMSD	ΔE/Res
N_{res}	1	0.90	0.79	0.93	0.28
ΔE	0.89	1	0.96	0.80	0.60
ΔE/RMSD	0.75	0.95	1	0.68	0.78
RMSD	0.93	0.77	0.62	1	0.22
ΔE/Res	0.13	0.48	0.71	0.04	1

[†]Summary of the correlations between five descriptive variables for both CHARMM/GB scoring function and AMBER/GB-MD scoring function. On the upper right half of the matrix are correlation values for the AMBER/GB-MD scoring function and on the lower left half of the matrix are values for the CHARMM/GB scoring function.
N_{res}, the number of residues in each decoy; ΔE, the energy score difference between each decoy and its corresponding native structure in units of kcal/mol; RMSD, root-mean-square deviation in units of A between each decoy and its corresponding native structure; ΔE/RMSD, energy difference per unit of RMSD; ΔE/Res, energy difference per residue.

Conformation Clustering

In the Lu and Skolnick set, a total of 72,036 structures were to be considered. Although we only used a very short time molecular dynamics simulation in our scoring function, the shear number of decoys needed to be evaluated still posed a formidable computational challenge. Realizing that most of the structures in this decoy set were very similar in structural conformation, we applied a hierarchical clustering algorithm to reduce the number of decoys to be considered. The clustering algorithm we used was based on the semilinear clustering technique developed earlier by Duan and Kollman[44] and a pairwise method by Daura et al.[45] In this approach, based on the main-chain RMSD, each decoy is compared with the average coordinate of an existing group after rigid body alignment. A decoy may become a member of its closest cluster if the RMSD is <1.5 Å. Otherwise, a new cluster may form if the minimum RMSD exceeded the cutoff. The clusters were further filtered by removing those structures whose RMSD from the average coordinates of the clusters exceeded the cutoff. The removed snapshots were then compared with the existing clusters. In each of the resulting clusters, only the structure with minimum RMSD to the average was selected as the representative structure for scoring.

Effective Energy Score Calculations

All calculations were performed on our in-house 64 PC dual Pentium III Beowulf cluster configured with the ROCKS cluster software distribution[46] version 2.2. A set of Perl5 scripts was developed in-house to automate preparation of the decoy structures for molecular dynamics simulations. All decoy structures were prepared with the Leap

TABLE III. Contents of the Four-State Reduced Decoy Set

PDB	Description	N_{res}	RMSD range	N_{decoys}	$N_{excluded}$	$\%_{excluded}$	z	z'	Rs
1ctf	C-terminal domain of ribosomal protein L7/L12	68	2.2–10.2	631	4	0.6	−4.26	−1.15	0.35
1r69	N-terminal domain of phage 434 repressor	63	2.3–9.5	676	1	0.1	−5.35	−1.88	0.42
1sn3	Scorpion toxin variant 3	65	2.5–10.5	661	5	0.8	−6.33	−3.03	0.16
2cro	Phage 434 Cro protein	65	2.1–9.7	675	5	0.7	−5.11	−2.19	0.37
3icb	Vitamin D-dependent calcium-binding protein	75	1.8–10.7	654	4	0.6	−2.86	−0.26	0.31
4pti	Trypsin inhibitor	58	2.8–10.8	688	2	0.3	−5.35	−1.60	0.29
4rxn	Rubredoxin	54	2.6–9.3	678	1	0.1	−5.36	−2.42	0.26
	Averages	64	2.3–10.1	666	3	0.5	−4.95	−1.79	0.31

N_{res}, number of residues; RMSD range, range of RMSD between the native structure and the decoy structure; N_{decoys}, number of decoys in the set; $N_{excluded}$, number of decoys excluded from z score calculation.

TABLE IV. Contents of the fisa Decoy Set[†]

PDB	Description	N_{res}	RMSD range	N_{decoys}	$N_{excluded}$	$\%_{excluded}$	z	z'	Rs
1fc2	**Human Fc fragment**	**43**	**3.1–10.3**	**501**	**2**	**0.4**	**0.23**	**2.67**	**0.26**
1hdd-C*	Engrailed Homeodomain	57	2.8–12.9	501	5	1.0	−3.62	−0.23	0.19
2cro	Phage 434 Cro protein	65	4.3–12.6	501	7	1.4	−5.31	−2.60	0.14
4icb	**Calbindin-binding Protein**	**76**	**4.8–14.1**	**501**	**2**	**0.4**	**−2.26**	**0.73**	**0.0065**
	Averages	60	3.7–12.5	501	4	0.8	−2.74	0.14	0.15

[†]Bold faced numbers indicate incorrectly predicted structures.
N_{res}, number of residues; RMSD range, range of RMSD between the native structure and the decoy structure; N_{decoys}, number of decoys in the set; $N_{excluded}$, number of decoys excluded from z score calculation.

TABLE V. Contents of the Baker CASP3 Set[†]

PDB	Description	N_{res}	RMSD range	N_{decoys}	$N_{excluded}$	$\%_{excluded}$	z	z'	Rs
1bg8-A	**E. coli Hde A**	**76**	**6.0–15.8**	**1201**	**35**	**2.9**	**−2.17**	**0.04**	**0.11**
1bl0	**DNA binding motif in MarA**	**99**	**3.6–18.2**	**972**	**56**	**5.8**	**−1.68**	**2.20**	**−0.019**
1eh2	**Eps15 Homology domain**	**79**	**4.0–15.3**	**2414**	**152**	**6.3**	**0.58**	**4.60**	**0.095**
1jwe	E. coli Dnab Helicase	114	7.8–20.9	1408	49	3.5	−5.27	−1.63	0.0069
l30	Unknown	104	6.5–24.6	1401	N/A	N/A	N/A	N/A	N/A
smd3	D3B subcomplex of the human core Snrnp domain	71	8.5–17.0	1201	20	1.7	−7.43	−3.09	0.11
	Averages	91	6.1–18.6	1433	62	4.0%	−3.19	0.42	0.06

[†]Bold faced numbers indicate incorrectly predicted structures.
N_{res}, number of residues; RMSD range, range of RMSD between the native structure and the decoy structure; N_{decoys}, number of decoys in the set; $N_{excluded}$, number of decoys excluded from z score calculation.

module and first minimized with 500 steps of constrained minimization using the sander module. Both modules are available in the AMBER 7 package.[47] The minimized structures were then subjected to a 5-ps equilibration followed by another 5-ps molecular dynamics simulation both using 2-fs timesteps. During the equilibration run, initial velocity was assigned from a Maxwellian distribution at a temperature equal to 100 K. Both the equilibration and production runs were kept at a constant temperature of 300 K by using a weak coupling scheme.[48] Solvent effects were treated with the Generalized Born implicit solvent model implemented in the AMBER package using a cutoff value of 200 Å for the nonbonded interactions. There are four different implementations of GB in the AMBER software. We used the pairwise GB model introduced by Hawkins et al.[49,50] In this implementation, Bondi radii were used with slight modifications of the different types of hydrogen atoms, and the overlap parameters were taken from the TINKER molecular modeling package (http://dasher.wustle.edu/tinker). The effects of added monovalent salt were included at a level that approximates the solutions of the linearized Poisson-Boltzmann equation.[51]

The AMBER software reports the potential energy of a protein molecule as the sum of eight energetic components:

$$E_{\mathrm{Pot}} = E_{\mathrm{bond}} + E_{\mathrm{angle}} + E_{\mathrm{Dihedral}} + E_{\mathrm{1-4\ NB}} + E_{\mathrm{1-4\ EEL}} + E_{\mathrm{vdw}} + E_{\mathrm{elec}} + E_{\mathrm{GB}} \quad (2)$$

where E_{bond}, E_{angle}, and E_{Dihedral} are the bonded terms, $E_{\mathrm{1\text{-}4NB}}$, $E_{\mathrm{1\text{-}4EEL}}$, E_{vdw}, and E_{elec} are the nonbonded terms, and E_{GB} is the GB solvation energy. In our analysis, we

TABLE VI. Content of hg_structal Set[†]

PDB	Description	N_{res}	RMSD range	N_{decoys}	$N_{excluded}$	$\%_{excluded}$	z	z'	Rs
1ash	Ascaris hemoglobin domain I	147	2.2–7.0	30	0	0.0	−3.42	−1.70	0.69
1bab-B	Hemoglobin Thionville	146	0.7–6.9	30	0	0.0	−2.24	−0.57	0.86
1col-a	Core-forming domain of colicin A	197	12.3–30.2	30	1	3.3	−4.83	−4.14	−0.067
1cpc-A	C-phycocyanin from Fremyella diplosiphon	162	6.8–14.0	30	0	0.0	−3.92	−2.33	0.30
1ecd	Erythrocruorin	136	1.5–6.2	30	1	3.3	−4.26	−3.50	0.63
1emy	Asian elephant cyanometmyoglobin	153	0.7–9.3	30	1	3.3	−1.97	−0.12	0.92
1flp	**Sulfide-reactive hemoglobin from the clam Lucina pectinata**	**142**	**1.7–7.2**	**30**	**0**	**0.0**	**0.13**	**1.53**	**0.56**
1gdm	Leghemoglobin	153	2.6–8.4	30	1	3.3	−3.80	−2.66	0.72
1hbg	Glycera dibranchiata hemoglobin	147	2.1–6.9	30	0	0.0	−3.43	−1.97	0.73
1hbh-A	Deoxyhemoglobin of the Antarctic fish Pagothenia bernacchii	142	1.0–6.3	30	0	0.0	−1.95	−0.46	0.93
1hbh-B	Deoxyhemoglobin of the Antarctic fish Pagothenia bernacchii	146	1.0–7.3	30	2	6.7	−2.34	−0.88	0.88
1hda-A	Bovine deoxyhemoglobin	141	0.5–5.8	30	0	0.0	−1.72	−0.13	0.91
1hda-B	Bovine deoxyhemoglobin	145	0.5–5.6	30	0	0.0	−1.85	−0.26	0.92
1hlb	Hemoglobin from Caudina arenicola	157	2.9–7.0	30	1	3.3	−2.55	−1.12	0.40
1hlm	**Hemoglobin from Caudina arenicola**	**158**	**3.0–8.7**	**30**	**1**	**3.3**	**1.53**	**3.89**	**0.39**
1hsy	**Myoglobin H64T Mutant**	**153**	**0.8–9.7**	**30**	**1**	**3.3**	**−1.85**	**0.03**	**0.88**
1ith-A	Hemoglobin from Urechis caupo	141	1.6–6.1	30	1	3.3	−2.47	−0.85	0.57
1lht	**Myoglobin from Loggerhead Sea Turtle**	**153**	**0.8–9.7**	**30**	**1**	**3.3**	**−1.80**	**0.24**	**0.71**
1mba	Aplysia limacina myoglobin	146	1.8–7.3	30	1	3.3	−2.60	−1.02	0.68
1mbs	**Seal myoglobin**	**153**	**1.7–9.3**	**30**	**1**	**3.3**	**1.16**	**2.99**	**0.71**
1myg-A	**Pig metmyoglobin**	**153**	**0.5–9.6**	**30**	**1**	**3.3**	**−1.80**	**0.09**	**0.79**
1myj-A	**Aquomet Myoglobin**	**153**	**0.6–7.9**	**30**	**2**	**6.7**	**−1.77**	**0.19**	**0.77**
1myt	Myoglobin from Yellowfin Tuna	146	1.0–10.0	30	1	3.3	−2.01	−0.14	0.82
2dhb-A	**Horse deoxyhemoglobin**	**141**	**0.6–6.4**	**30**	**0**	**0.0**	**−0.57**	**1.10**	**0.91**
2dhb-B	**Horse deoxyhemoglobin**	**146**	**0.9–7.1**	**30**	**0**	**0.0**	**−0.95**	**0.78**	**0.86**
2lhb	Lamprey-hemoglobin from Petromyzon marinus	149	3.0–8.1	30	2	6.7	−3.90	−3.47	0.29
2pgh-A	**Aquomet porcine hemoglobin**	**141**	**0.7–6.5**	**30**	**1**	**3.3**	**−0.80**	**0.64**	**0.90**
2pgh-B	**Aquomet porcine hemoglobin**	**146**	**0.8–7.5**	**30**	**1**	**3.3**	**−1.10**	**0.71**	**0.93**
4sdh-A	Deoxy hemoglobin l	145	2.3–6.4	30	3	10.0	−3.55	−2.65	0.29
	Averages	150	2.0–8.6	30	1	2.8	−2.09	−0.54	0.69

[†]Bold faced numbers indicate incorrectly predicted structures.
N_{res}, number of residues; RMSD range; range of RMSD between the native structure and the decoy structure; N_{decoys}, number of decoys in the set; $N_{excluded}$, number of decoys excluded from z score calculation.

discarded all data from and before the equilibration run and took the average value of E_{Pot} from the production run as the scoring value for a protein structure.

The entire set of decoy simulations were done in a high-throughput fashion by making use of the automatic scheduling capabilities of the Batch Portable Scheduler software (http://www.openpbs.org). For a 68-residue protein, the entire calculation took about 30 min to complete. The short wall clock time required clearly showed the technical feasibility of this approach. Results were collected by the in-house Perl5 scripts and analyzed in the Microsoft® Excel 2002 electronic spreadsheet.

RESULTS AND DISCUSSION

Analysis of the EMBL Misfolded Protein Database

We began the testing of our new scoring function by applying it to the analysis of the EMBL deliberately misfolded protein database. The results of our predictions are summarized in Table I. In our analysis, we ignored all disulfide bonds and cofactors in the native structures and simply evaluated the effective energy scores based on the native folds alone. This simplification was needed to avoid introducing preferential treatment in evaluating the effective energy scores of the native structures. The misfolded

TABLE VII. Contents of the lattice ssfit Decoy Set†

PDB	Description	N_{res}	RMSD range	N_{decoys}	$N_{excluded}$	$\%_{excluded}$	z	z'	Rs
1beo	Beta-cryptogein	98	7.0–15.6	2001	4	0.2	−7.95	−4.18	−0.027
1ctf	L7/L12 50 S ribosomal protein	68	5.5–12.8	2001	4	0.2	−6.98	−2.90	−0.070
1dkt-A	Type 1 human cyclin-dependent kinase subunit	72	6.7–14.1	2001	21	1.0	−6.40	−2.66	−0.12
1fca	Ferredoxin from clostridium Acidurici	55	5.1–11.4	2001	8	0.4	−8.30	−4.06	−0.0071
1nkl	**Nk-lysin from pig**	**78**	**5.3–13.6**	**2001**	**7**	**0.3**	**−2.60**	**0.92**	**−0.030**
1pgb	Protein G (B1 lgG-binding domain)	56	5.8–12.9	2001	37	1.8	−9.55	−5.37	0.041
1trl-A	Thermolysin fragment	62	5.4–12.5	2001	11	0.5	−5.49	−1.92	0.0050
	Averages	70	5.8–13.3	2001	13	0.7	−6.75	−2.88	−0.03

†Bold faced numbers indicate incorrectly predicted structures.
N_{res}, number of residues; RMSD range, range of RMSD between the native structure and the decoy structure; N_{decoys}, number of decoys in the set; $N_{excluded}$, number of decoys excluded from z score calculation.

TABLE VIII. Contents of the Local Minima Decoy Set†

PDB	Description	N_{res}	RMSD range	N_{decoys}	$N_{excluded}$	$\%_{excluded}$	z	z'	Rs
1b0n-B	**Sinr protein/Sini protein complex**	**31**	**2.45–6.03**	**498**	**0**	**0.0**	**−0.61**	**2.07**	**0.15**
1bba	**Bovine pancreatic polypeptide**	**36**	**2.78–8.91**	**501**	**0**	**0.0**	**4.99**	**7.80**	**0.11**
1ctf	L7/L12 50 S ribosomal protein	68	3.59–12.5	498	0	0.0	−5.12	−15.44	0.31
1dtk	Dendrotoxin K	57	4.32–12.6	216	0	0.0	−6.10	−3.86	0.070
1fc2	Immunoglobulin Fc and fragment B of protein A complex	43	3.99–8.45	501	0	0.0	−3.38	−0.18	0.093
1igd	Protein G	61	3.11–12.6	501	0	0.0	−6.16	−3.14	0.12
1shf-A	Fyn proto-oncogene tyrosine kinase	59	4.39–12.3	438	0	0.0	−8.26	−5.83	0.039
2cro	434 cro protein	65	3.87–13.5	501	0	0.0	−8.03	−5.57	0.12
2ovo	Ovomucoid third domain	56	4.38–13.4	348	0	0.0	−6.00	−3.73	0.033
4pti	Trypsin inhibitor	58	4.94–13.2	344	0	0.0	−6.24	−3.89	−0.033
	Averages	53	3.78–11.3	435	0	0.0	−4.49	−3.18	0.10

†Bold faced numbers indicate incorrectly predicted structures.
N_{res}, number of residues; RMSD range, range of RMSD between the native structure and the decoy structure; N_{decoys}, number of decoys in the set; $N_{excluded}$, number of decoys excluded from z score calculation.

structures do not contain any information regarding disulfide bonds or stabilizing cofactor; therefore, accounting for these interactions in evaluating native scores would bias toward the correct answers. Thus, one may intuitively expect that the inclusion of the cofactors and disulfide bonds could improve the results. Although it is true that these interactions may play critical roles in the stability and native conformation of the proteins, our objective here was to distinguish between the grossly misfolded conformations from the native conformations. Keeping this point in mind, we hypothesized that the unfavorable interactions in the misfolded conformations would outweigh the effects of the disulfide bonds and other native fold-stabilizing interactions. This hypothesis was somewhat justified by the large RMSD (>9.6 Å) between the native conformation and misfolded conformation.

Of the 26 pairs, we were able to correctly distinguish between the native structures from the misfolded structures of 25 pairs. We excluded 1rei from our analysis because the native structure of 1rei is a homodimeric protein consisting of 2 chains of 107-residue polypeptides and the backbone from which the misfold structure of 1rei was generated is a single polypeptide consisting of 214 residues. By assuming that situations such as 1rei would not be considered in real applications, this would give the scoring function 100% accuracy. In contrast, the results obtained by using the CHARMM force field and GB solvent model was 25 of 26 pairs (96%).[11] The two results appear to be comparable, although it should be noted that the pair misidentified by CHARMM/GB (1fdx) was correctly identified by AMBER/GB-MD (CHARMM/GB correctly identified the 1rei dimer crystal structure as the native structure). Given the simplifications used in our model for scoring the EMBL set, the high success rate obtained was quite encouraging.

Because a simple energy minimization step was conducted before molecular dynamics simulations as part of our scoring scheme, we also included these minimized

TABLE IX. Content of the Lu and Skolnick Decoy Set[†]

PDB	Description	Source	N_{res}	RMSD range	$N_{cluster}$	$N_{excluded}$	$\%_{excluded}$	z	z'	Rs
1a32	**Ribosomal protein S15 from Bacillus stearothermophilus**	**x-ray**	**85**	**11.0–16.2**	**96**	**2**	**2.1**	**−0.98**	**0.62**	**−0.073**
1ah9	Translational thitiation factor lf1 from E. coli	NMR	71	5.3–11.2	97	0	0.0	−4.62	−1.20	0.15
1aoy	N-terminal domain of E. coli arginine repressor	NMR	78	3.6–12.2	134	0	0.0	−2.06	−0.16	0.17
1bq9A	Rubredoxin (formyl methionine mutant) from Pyrococcus furiosus	x-ray	53	6.1–11.1	115	2	1.7	−2.95	−1.48	−0.093
1bw6A	Human centromere protein B (Cenp-B) DNA binding domain Rp1	NMR	56	4.2–12.0	120	1	0.8	−4.44	−0.85	0.17
1c5a	Porcine C5adesArg	NMR	66	4.3–12.2	110	0	0.0	−4.50	−2.09	0.19
1cewl	Chicken egg white cystatin	x-ray	108	4.8–12.3	204	1	0.5	−9.12	−6.71	0.11
1cis	Hybrid protein formed from chymotrypsin inhibitor-2 (Ci-2) from (Hordeum Vulgare) and (Subtilisin Carlsberg)	NMR	66	3.4–9.2	109	1	0.9	−6.49	−4.08	0.071
1csp	Major cold shock protein (Cspb)	x-ray	64	4.0–11.9	145	0	0.0	−9.96	−7.63	−0.16
1erv	Human thioredoxin mutant with Cys 73 replaced by Ser (reduced form)	x-ray	105	2.2–13.3	88	0	0.0	−11.63	−8.24	0.084
1fas	Fasciculin 1, an antiacetylcholinesterase toxin from green mamba snake venom	x-ray	61	3.4–8.2	114	83	72.8	−5.23	−2.18	0.071
1ftz	Fushi tarazu homeodomain from Drosophila	NMR	70	4.6–11.0	145	14	9.7	−4.46	−0.65	0.11
1gpt	Gamma1-H and gamma1-P thionins from barley and wheat endosperm	NMR	47	2.3–9.5	123	1	0.8	−4.47	−2.57	−0.071
1hlb	Hemoglobin	x-ray	138	2.4–13.8	92	1	1.1	−7.68	−4.79	0.35
1hmdA	Deoxy and oxy hemerythrin	x-ray	113	2.4–12.0	99	2	2.0	−7.99	−5.72	0.077
1hp8	Human P8-Mtcp1, a cysteine-rich protein encoded by the Mtcp1 oncogene	NMR	68	2.9–13.0	109	1	0.9	−6.14	−2.33	0.047
1ixa	**The first EGF-like module of human factor IX**	**NMR**	**39**	**3.4–10.5**	**356**	**0**	**0.0**	**−1.24**	**2.41**	**−0.014**
1kjs	C5A at pH 5.2, 303 K, 20 structures	NMR	74	4.2–13.4	119	1	0.8	−2.99	−0.59	0.10
1ksr	Actin-binding protein	NMR	100	4.0–8.9	224	1	0.4	−1.79	−0.41	−0.14
1lea	Lexa repressor DNA-binding domain	NMR	72	2.6–6.4	137	2	1.5	−6.73	−4.11	−0.083
1mba	Myoglobin	x-ray	146	2.2–13.5	93	5	5.4	−11.19	−9.42	0.21
1ner	**DNA-binding protein Ner**	**NMR**	**74**	**7.4–13.6**	**105**	**0**	**0.0**	**−0.80**	**1.84**	**−0.032**
1ngr	P75 low-affinity neurotrophin receptor	NMR	85	2.5–13.6	84	1	1.2	−2.69	−0.62	0.086
1nkl	Nk-lysin	NMR	78	2.7–16.5	78	0	0.0	−4.96	−3.01	0.29
1pdo	Mannose permease	x-ray	121	6.3–11.1	124	4	3.2	−11.07	−7.84	0.061
1pgx	Protein G type 7 (B2 domain)	x-ray	56	2.0–11.4	73	0	0.0	−8.80	−6.59	0.38
1poh	Phosphotransferase	x-ray	85	1.8–12.2	97	0	0.0	−6.36	−3.65	0.14
1pou	Oct-1 (Pou-specific domain)	NMR	71	2.9–10.6	88	0	0.0	−6.10	−3.20	−0.20
1pse	Photosystem I accessory protein E	NMR	69	7.8–13.4	329	0	0.0	−4.74	−2.44	−0.057
1rip	**Ribosomal protein S17**	**NMR**	**81**	**8.6–15.6**	**433**	**0**	**0.0**	**0.34**	**1.44**	**−0.038**
1rpo	Rop (cole1 repressor of primer)	x-ray	61	7.3–24.4	105	0	0.0	−5.03	−4.18	−0.20
1shaA	V-Src tyrosine kinase-transforming protein	x-ray	103	2.2–5.4	161	6	3.7	−8.31	−4.84	−0.039
1shg	Alpha-spectrin (Sh3 domain)	x-ray	57	3.6–9.5	86	0	0.0	−8.91	−4.84	0.14
1sro	Pnpase	NMR	66	5.2–12.4	98	0	0.0	−6.61	−3.96	−0.063
1stfl	Papain	x-ray	98	5.3–21.0	190	5	2.6	−0.93	−0.16	0.035
1thx	Thioredoxin	x-ray	108	2.2–3.7	88	0	0.0	−9.99	−6.69	−0.039
1tit	Titin	NMR	89	2.0–10.2	73	0	0.0	−8.26	−5.07	0.097
1tlk	Telokin	x-ray	103	3.4–17.2	136	2	1.5	−6.19	−2.93	−0.14

TABLE IX. (Continued)

PDB	Description	Source	N_{res}	RMSD range	$N_{cluster}$	$N_{excluded}$	$\%_{excluded}$	z	z'	Rs
1ubi	Ubiquitin	x-ray	76	1.9–4.9	122	0	0.0	−6.78	−4.20	0.0098
1vif	Dihydrofolate reductase	x-ray	60	2.6–11.5	111	2	1.8	−8.01	−4.88	0.0098
1wiu	Twitchin 18Th lgsf module	NMR	93	2.3–12.5	103	1	1.0	−4.86	−2.29	0.094
256bA	Cytochrome b562	x-ray	106	4.4–10.6	87	2	2.3	−9.56	−6.81	0.064
2af8	Actinorhodin polyketide synthase Acyl carrier protein	NMR	86	3.5–12.6	144	0	0.0	−3.03	−0.10	0.13
2azaA	Azurin (oxidized)	x-ray	129	4.2–13.0	121	1	0.8	−8.34	−5.39	−0.015
2bby	Rap30	NMR	69	2.9–12.8	92	1	1.1	−4.70	−2.25	0.25
2ezh	Transposase	NMR	65	3.5–13.0	110	2	1.8	−6.52	−2.95	0.024
2ezk	Transposase	NMR	93	4.0–15.4	82	2	2.4	−8.21	−5.78	0.11
2fmr	Fmr1 protein	NMR	65	3.3–11.1	128	1	0.8	−5.09	−2.02	0.21
2lfb	Lfb1/Hnf1 transcription factor	NMR	100	11.1–16.5	228	0	0.0	−0.58	−0.03	−0.22
2pcy	Apo-plastocyanin	x-ray	99	1.7–12.0	71	0	0.0	−7.50	−5.55	0.093
2ptl	Protein L (B1 domain)	NMR	60	2.4–12.9	122	0	0.0	−4.53	−2.60	0.060
2sarA	Ribonuclease Sa (E.C. 3.1.4.8) complex with 3′-guanylic acid	x-ray	96	4.8–13.5	125	0	0.0	−6.98	−4.11	−0.049
5fd1	Ferredoxin	x-ray	106	9.3–15.2	143	0	0.0	−7.45	−3.78	−0.11
6pti	Bovine pancreatic trypsin Inhibitor (BPTI, crystal form III)	x-ray	57	3.2–11.3	156	2	1.3	−6.31	−3.88	−0.16
	Averages		82	4.1–12.3	132	3	2.4	−5.82	−3.25	0.04

†Bold faced numbers indicate incorrectly predicted structures.
N_{res}, number of residues; RMSD range; range of RMSD between the native structure and the decoy structure; $N_{excluded}$, number of decoys excluded from z score calculation. $N_{excluded}$, number of decoys excluded from z score calculation.

energy values in Table I for comparison. From this comparison, it would appear that molecular dynamics outperforms energy minimization (23 of 25 pairs correctly distinguished). While examining these scores, the reader should bear in mind that the energy minimization performed in our study is very simple (500 steps of steepest descent); thus, this comparison is of a very qualitative nature and should not be overinterpreted as a definitive evidence to discount energy minimization methods. However, our main concern here is to evaluate the performance of this new scoring function as a whole, and a detailed comparison of energy minimization versus molecular dynamics scoring is outside the scope of our present study.

One important issue in using the average energy of molecular dynamics simulation as a scoring function is that, depending on the initial velocity, there will be a slight difference in the resulting average energy. To obtain an estimate of the variation in the average energy score due to differences in initial velocity, 100 trajectories of one structure was generated. The result is shown as a scatter plot in Figure 1(a). The average energy score of these 100 independent trajectories is −1206.8 kcal/mol, and the standard deviation is only 4.5 kcal/mol (~0.4%). In contrast, the energy fluctuation in a randomly selected trajectory has a standard deviation of 38.9 kcal/mol [Fig. 1(b)]. This is expected because the simulations are very short, and the independent trajectories are not expected to diverge much within this short simulation time. Therefore, the scores obtained from any single trajectory should be a very representative score, and multiple trajectory sampling should not be required to obtain reliable energy scores.

Another important issue for molecular dynamics simulation scoring is whether ignoring the disulfide bonds significantly altered our predictions. In the EMBL decoy set, the

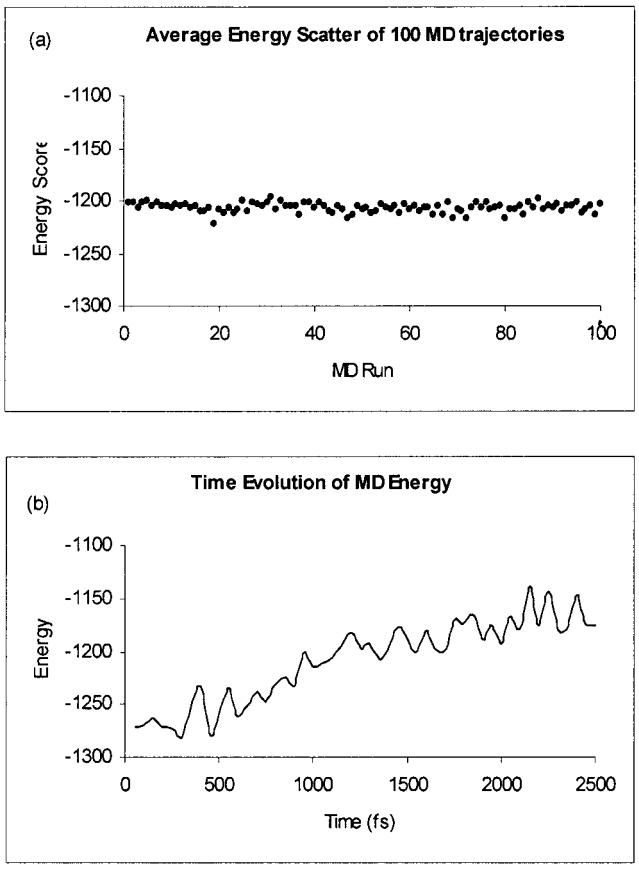

Fig. 1. **a:** Scattered plot of the energies of 100 independent molecular dynamics trajectories of the same structure (1ctf). **b:** The time course plot of one molecular dynamics trajectory randomly selected from the 100 independent trajectories, showing the fluctuation of the total energy.

TABLE X. Summary of Decoy Results

Decoy Set	N_{seq}	Average RMSD range	% correct	Average z	Average z'	Average Rs
Four-state reduced	7	2.3–10.1	100	−4.95	−1.79	0.31
fisa	4	3.7–12.5	50	−2.74	0.14	0.15
Baker CASP3	6	6.1–18.6	50	−3.19	0.42	0.06
hg_structal	29	2.0–8.6	62	−2.09	−0.54	0.69
lattice_ssfit	7	5.8–13.3	86	−6.75	−2.88	−0.03
lmds	10	3.8–11.3	80	−4.49	−3.18	0.10
Lu and Skolnick	54	4.1–12.3	93	−5.82	−3.25	0.04

N_{seq}, number of protein sequences in the decoy set; Average RMSD range, average range of RMSD between the native and the decoy structure.

smallest peptide 1cbh has 36 residues with 2 disulfide bonds. The energy difference between the native and misfolded structure without considering the disulfide bonds is about −13.9 kcal/mol, significantly lower (in absolute value) than the average −199.3 kcal/mol for the entire decoy set. This does seem to cause concern about the discriminatory power of the scoring function toward the smaller end of peptide length. However, when the two disulfide bonds in the native structure were taken into account, the energy improved only a marginal −39.4 kcal/mol, or −19.7 kcal/mol per disulfide bond. Considering that most intracelluar proteins do not have disulfide bonds and that the likelihood of small peptides having more than two disulfide bonds is mostly likely not very big, the trend of diminishing energy gap between native and decoys should not pose a real limitation for the AMBER/GB-MD scoring function.

Identification of the Native Structure From a Set of Native-Like Decoys

The EMBL misfolded protein pairs provides a simple first test to gauge the usefulness of a new scoring function. However, the large conformational differences between the decoys and native structures do not properly explore the energy landscape generated by the new force field. In this section, we present our results on scoring of seven different decoy sets with varying sizes and origins. These results cover a wider range of conformational variations and, hence, provide a more detailed statistical characterization on the performance of the new AMBER/GB-MD scoring function.

The z score

As described in Materials and Methods, the next few decoy databases we examined were all designed to challenge a scoring function by providing a large set of native-like structures; a successful scoring function in this case not only has to be able to correctly distinguish between native and native-like structures but must also do so convincingly. In this regard, the quality of a scoring function is judged by the size of energy gap it assigns to the native structure and the average energy of the rest of the native-like structures. A commonly used measure for assessing this quality is the z score.

In our study, we followed conventions in the literature and defined the z score of a structure from within a set of decoy structures as follows:

TABLE XI. Side-by-Side Comparison Against Thermodynamic Data

PDB	N_{res}	z (25°)	z (75°)	z (125°)	z_{AMBER}
1shg	57	−0.7	−2.4	−2.9	−8.91
1pgx	70	−1.3	−3.5	−4.4	−8.8

z (25°), z (75°), z (125°), experimental z scores derived from thermodynamic data at the indicated temperature. Temperature units are in degrees Celsius. z_{AMBER}, z scores as determined by AMBER/GB-MD scoring function.

$$z = \frac{E_i - \langle E \rangle}{\sigma} \tag{3}$$

where E_i is the energy score of decoy structure i derived from the last 5 ps of MD simulation, $\langle E \rangle$ is the average energy score of the decoy set, and σ is the standard deviation. By this definition, the z score is a dimensionless quantity that measures the energy separation between the native fold and the average energy of an ensemble of misfolds in the units of the standard deviation of the ensemble.

For the Lu and Skolnick decoy set, the average energy was calculated by multiplying the energy of each representative decoy by the number of decoys contained in that cluster to obtain a weighted average.

In addition to the conventional z score, we also define here a related measure z′ as follows:

$$z' = \frac{E_{\text{native}} - E_{\text{lowest}}}{\sigma} \tag{4}$$

where E_{native} is the energy score of the native fold, E_{lowest} is the lowest energy score among the respective decoys, and σ is the standard deviation of the decoy set. In contrast to the z score, the z' gives a quantitative measure of how well separated the native structure is from its lowest energy neighbor from within the decoy set.

Z score analysis of the decoy sets

Results of the z scores from the seven decoy sets are presented in Tables III–IX and summarized in Table X. Both knowledge-based and physics-based potential energy functions have typically underestimated the z score values of native structures. Using experimental thermodynamic data, Zhang and Skolnick estimated that the z score value of the native structure should be approximately −3 for a

TABLE XII. *Z* Scores of the Same Sequence Obtained From Different Decoy Sets[†]

	N_{res}	Four-state reduced	fisa	Baker CASP3	hg structal	Lattice ssfit	lMDS	Lu and Skolnick
1fc2	43	—	**0.23**	—	—	—	−3.38	—
4pti	58	−5.35	—	—	—	—	−6.24	—
2cro	65	−5.11	−5.31	—	—	—	−8.03	—
1ctf	68	−4.26	—	—	—	−6.98	−5.12	—
1nkl	78	—	—	—	—	−2.6	—	−4.96
1hlb	138	—	—	—	−2.55	—	—	−7.68
1mba	146	—	—	—	−2.6	—	—	−11.19

[†]Bold faced numbers indicate incorrectly predicted structures; N_{res}, number of residues.

small protein under 200 residues at room temperature (the sign is reversed here to conform to our *z* score definition).[52] As a general rule, the value of *z* score typically grew more negative as the temperature increased, reflecting the difference in thermal stability between the native state and the misfolded states. Direct comparisons between experimental *z* scores and scores derived from the AMBER/GB-MD scoring function are available for two protein sequences from the Lu and Skolnick decoy set and are presented in Table XI. From this comparison, it would appear that the AMBER/GB-MD scoring function has overestimated the *z* scores. This trend is not unique to the Lu and Skolnick set; it was also observed in a side-by-side comparison of *z* scores of identical sequences obtained from different decoy sets (Table XII). It should be noted that there are currently no experimental methods available for determining the true *z* scores and that the experimental *z* scores determined by Zhang and Skolnick are based on enthalpy instead of effective free energy. However, a comparison of these values to the *z* scores determined by various scoring functions is still meaningful because enthalpy has been observed as the driving component in the folding of many, if not most, proteins. In those cases, the entropic terms usually disfavor the folding and reduce the effective *z* score. Thus, the *z* scores of Zhang and Skolnick could be viewed as the upper limits of the possible experimental *z* scores.

Using a lattice model, Lu and Skolnick also estimated that a small 100-residue protein should have a *z* score in the range of −15 to −31, significantly larger (in absolute value) than the experimentally derived value of −3. The exaggerated *z* scores, although not desirable in realistic simulations, do represent an advantage for scoring functions that attempt to separate the native conformation from the rest of misfolded states. In this case, the big energetic gap between the native state and the average energy of the misfolded ensemble can boost the sensitivity of a scoring function and, therefore, is a desirable feature.

Correlation between the RMSDs and z scores

RMSD has been used widely as a measure of structural similarity; previous studies of energy scoring functions have all tried to correlate RMSD with energy differences, although the relationship between the two is less than clear. In our study, we found that despite the high degree of accuracy of the AMBER/GB-MD scoring function in identifying the native states, correlation between RMSD and *z* scores (or the raw energy scores) is generally weak to nonexistent (Fig. 2 shows a few examples of ΔE vs RMSD scattered plots). The average Spearman's rank order correlation coefficient (Rs) values for the decoy sets range from 0.04 for the Lu and Skolnick set to 0.69 for the hg_structal set (Table X). Although the 0.69 value of the hg_structal set at first glance may appear to be significant, when one takes into account that the hg_structal set only contains 29 decoy structures whereas the Lu and Skolnick set contains 1333 decoys per sequence, the significance of this correlation is substantially diminished. When one looks at the correlation between the average Rs value and the average number of decoys ($R = -0.57$), one finds that in a decoy set where there is a sufficient number of structures to cover the conformational space, there will be no real correlation between RMSD values and the AMBER/GB-MD energy scores. Given the roughness of the energy landscape, this is not a surprising result.

Comparison with other physics-based all-atom scoring functions

To get a better sense of how well the AMBER/GB-MD scoring function performed, we have compared our results with two other physics-based scoring functions recently published in the literature. Summary of the comparisons is presented in Table XIII. In the four-state reduced decoy set, both CHARMM/GB and AMBER/GB-MD scoring functions achieved 100% accuracy in discriminating the decoys; however, AMBER/GB-MD does have a more negative average *z* score for the native structures. As mentioned earlier, the more negative *z* score values may offer the scoring function an advantage in discriminating the decoy structures. Table XIV lists the RMSD of the second lowest energy decoy structure and its corresponding RMSD ranking within each sequence of the four-state reduced decoy set. From this table, it is clear that the energy landscape around the native state is not monotonically smooth, and there are many intervening local minima between the native state and the second lowest energetic state. Both the CHARMM/GB and OPLS-AA/GB scoring functions use energy minimization algorithms as their primary means of evaluating the energetic scores. It is likely that the better performance obtained by AMBER/GB-MD is partly due to the use of molecular dynamics algorithm, which in this case can allow native structures to cross small local energy

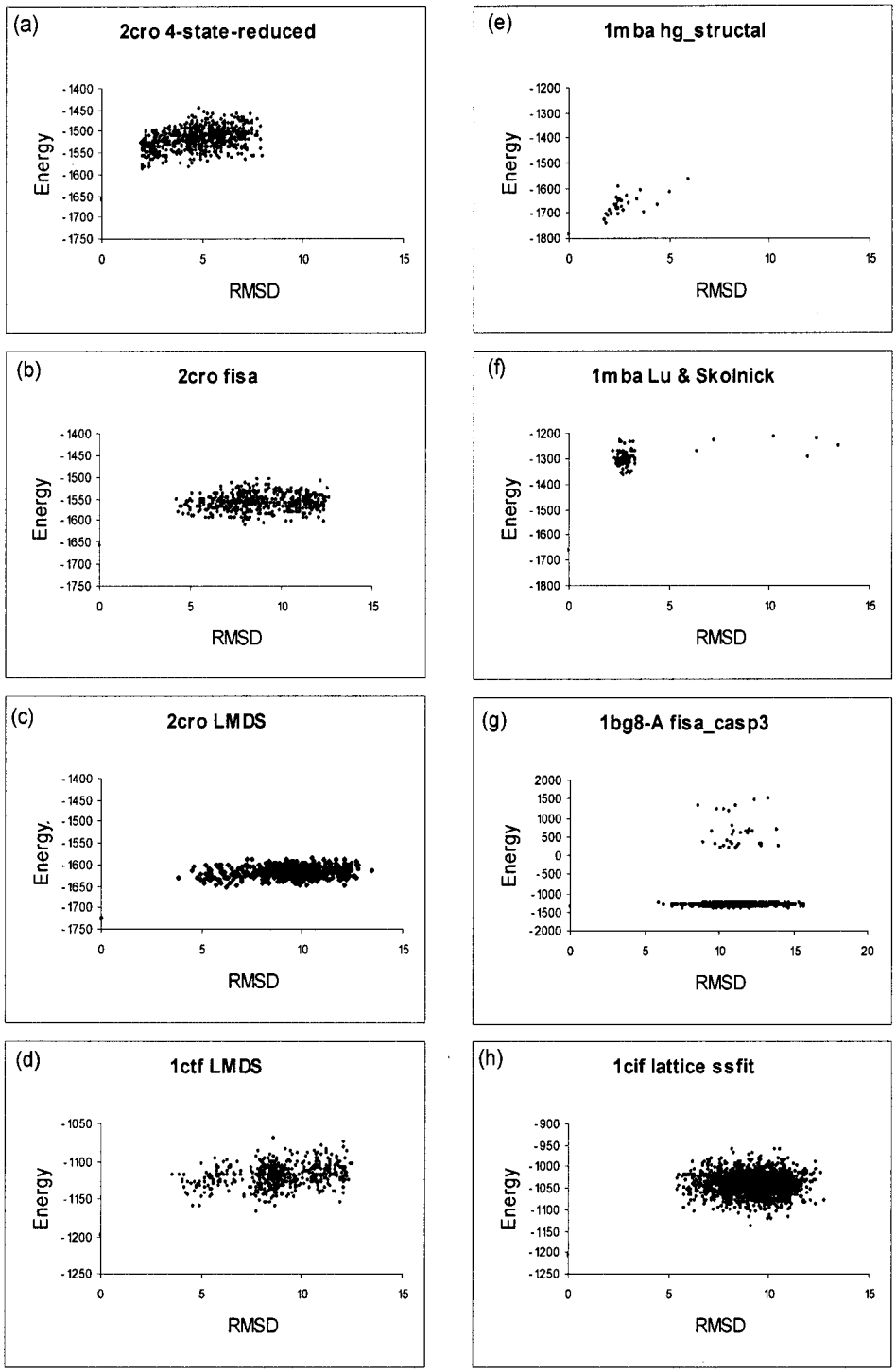

Fig. 2. Scattered plots of energy versus RMSD from the native structure. **a–c:** Plots of decoy energy/RMSD scatter plots of the same protein sequence 2cro derived from four-state reduced, fisa, and LMDS sets, respectively. **d,h:** Energy/RMSD scattered plots for the sequence 1ctf from decoy set LMDS and lattice ssfit, respectively. **e,f:** are scattered plots for the sequence 1mba from the hg_structal and Lu and Skolnick sets, respectively. **g:** Energy/RMSD scattered plot for the sequence 1bg8-A from the Baker CASP3 decoy set.

TABLE XIII. Side-by-Side Comparison of Decoy Detection Scoring Functions

Decoy sets	CHARMM/GB	OPLS-AA/GB	AMBER/GB-MD
Four-state Reduced			
% accuracy	−100%	43%	100%
Average z_{native}	−3.39	−3.67	−4.95
Average z'	n/a	n/a	−1.79
Range of z_{native}	−1.7 to −4.6	−2.18 to −4.53	−2.86 to −6.33
Average z/RMSD correlation	0.55	0.63	0.35
LMDS			
% accuracy	n/a	14%	80%
Average z_{native}	n/a	−2.57	−4.49
Average z	n/a	n/a	−3.18
Range of z_{native}	n/a	0.6 to −14.57	4.99 to −8.26
Average z/RMSD correlation	n/a	0.11	0.16
Lu and Skolnick			
% accuracy	n/a	n/a	93%
Average z_{native}	n/a	−4.02	−5.82
Average z	n/a	n/a	−3.25
Range of z_{native}	n/a	−1.48 to −9.6	0.34 to −11.63
Average z/RMSD correlation	n/a	n/a	0.22

TABLE XIV. Ranking Order of the Lowest-Energy Decoy

PDB	$RMSD_{Low}$	$RMSD_{Low}$ rank
1ctf	5.5	319
1r69	6.2	510
1sn3	3.1	87
2cro	2.0	6
3icb	8.0	536
4pti	5.8	320
4rxn	2.0	25

$RMSD_{Low}$, the RMSD between the lowest-energy decoy and the native structure; $RMSD_{Low}$ rank, the rank of the RMSD value of the lowest-energy decoy.

barriers to arrive at a more favorable energetic state. The fact that the accuracy of AMBER/GB-MD seems to degrade far less than OPLS-AA/GB in the case of LMDS where there is only very weak correlation between RMSD and energy would appear to support this view.

Energy distribution variations within decoy sets

Up to this point, we have analyzed the scoring results of AMBER/GB-MD and compared them with the results of two other physics-based scoring functions as well as results obtained from thermodynamic data. For the Novotny test, our scoring function has achieved the best set of results by far (Table XIII). However, the degree of success in discriminating decoys was not uniform across the decoy sets tested. The accuracy percentages reported in Table X at first glance seem to raise questions about the true discriminatory ability of the scoring function. In reality, most of the low percentage values were due to the inconsistent sizes of the various decoy sets we evaluated. For example, results from the four-state reduced set gave the best accuracy (100%) with an average z score of −4.95 and an average z' of −1.79. In contrast, results from the hg_structal decoy set gave only a 64% accuracy with an average z score of −2.09 and z' of −0.59, but an average Spearman's rank correlation coefficient of 0.69! In the Lu and Skolnick set, we also observed that most of the misidentified decoys are those with native structures determined by NMR (3 of 28 NMR and 1 of 27 X-ray native structures were misidentified). Although the small sample numbers are not enough to draw any definitive conclusions, it does suggest that the quality of native structure could have been the cause of misidentification.

The small sample size in several of the decoy sets we tested should also be kept in mind when viewing the predictive accuracies in Table X. In the fisa set and Baker CASP3 set, there are only four and six sequences; hence, a single misidentified sequence can significantly mar the percent accuracy values. In the Lu and Skolnick decoy set where there are 54 independent sequences, the AMBER/GB-MD scoring function has achieved 93% accuracy. We believe this value is more representative of its true discriminatory accuracy. When results of all 117 sequences are considered as a whole, the accuracy percentage is about 81%. One should note that the tests performed here with the lowest energy score as the discriminating criterion is more stringent than the criterion used in CASP,[53] where one is allowed to choose five structures for submission as the prediction. If one were to use the CASP criterion and set the discriminating boundary as the top five scoring structures, a better prediction statistics can be anticipated.

Furthermore, in evaluating the z scores, normal distribution analysis was performed to obtain the value of the standard deviation. The original energy score distribution patterns exhibited substantial irregularity, some with relatively high energies that significantly alter the true ensemble average. Because we were primarily interested in discriminating the decoys that are close energetic neighbors of the native state, we graphically plotted the energy distribution of the decoy set first and then eliminated the high-energy clusters from the data set before evaluating the final z score. Figure 3 shows an example of

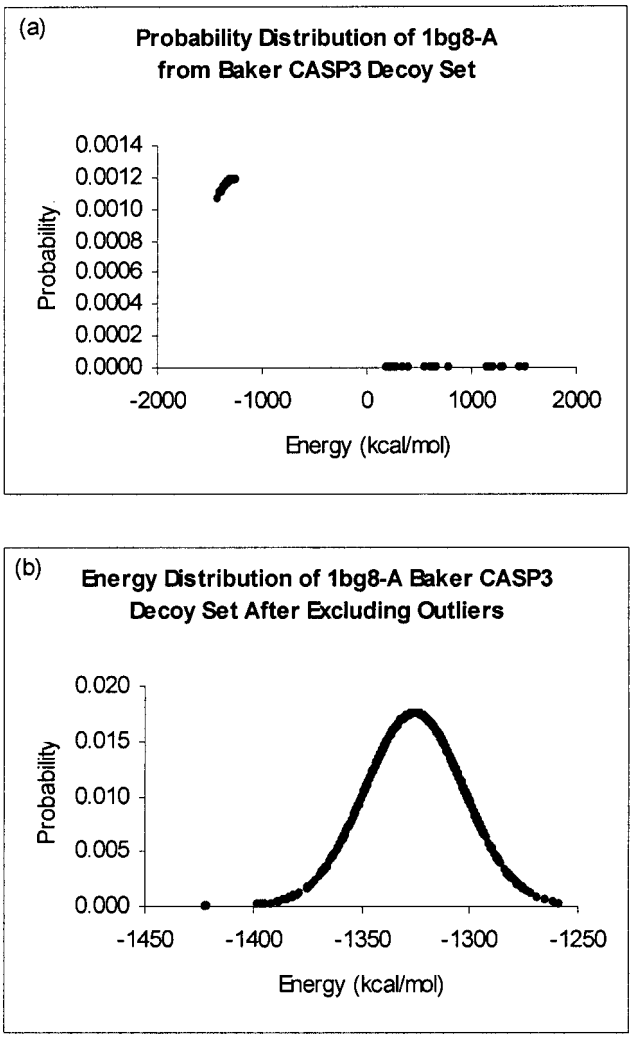

Fig. 3. The energy distribution plots of the 1201 decoys for the sequence 1bg8-A from the Baker CASP3 decoy set. **a:**The distribution when all of the 1201 data points were included, giving a *z* score of −0.32. **b:** The recalculated distribution after excluded 35 high-energy data points, which shows a normal distribution pattern. The *z* score for the native structure in this restricted set is −2.17.

the energy distribution plot. This procedure does not in any way alter the outcome of the Novotny test; it only limits the scope of the *z* score measurements to a more meaningful subset of the decoy data set.

CONCLUSION

In the past, molecular dynamics and physics-based energy potentials had been dismissed as viable approaches to the fold recognition problem. The data presented in this article reinforce recent results using CHARMM and OPLS-AA that physics-based energy potentials are capable of discriminating decoys with high accuracy, challenging a long-standing preconception that molecular mechanics force fields are incapable of such demanding applications. More importantly, by putting the AMBER/GB-MD combination through the Novotny test, we demonstrated that molecular dynamics is not only a viable but potentially a superior method for protein structure prediction applications. The strength of MD-based scoring scheme lies in its ability to cross energy barriers, which is difficult in the simple energy minimization methods. One would anticipate with the advancement in computer technology that increasingly longer timescales will be assessable in the near future even for large decoy sets. Given the short wall clock time (30 min) needed to complete a set in this initial attempt, we can realistically expect the increase in simulation time by one to two orders of magnitude soon. By then, we expect a wide range of applications of MD in protein structure prediction, including protein structure refinement. Thus, the MD-based scoring scheme would be consistent with the refinement scheme. The ability to differentiate the native structures from the decoys would be directly translated into its ability to move the non-native structures closer to the native structures. The latter would be one of the important applications of this rather mature, yet powerful method.

However, despite the large data set examined in this study and the encouraging results stated above, we found that detailed comparison and systematic characterization of scoring functions is a difficult task at best, even with the help of the current generation of publicly available decoy datasets. We believe that the development of physics-based energy potentials has caught up with the statistics-based potentials, but quantitatively characterizing these energy potentials remains a challenge. A case in point is the fact that the decoy sets examined in this study all exhibited quite irregular distributions of energetic states even for decoy sets of the same sequences (e.g., 4pti in four-state reduced and local minima decoy set). It is difficult to tell whether the distribution pattern was due to inaccuracies in the scoring function or due to the particular ensemble of decoys; hence, it is difficult to draw quantitative conclusions that might aide in the refinement of the energy-scoring function. We believe that better calibration tools to aide in the development of the next generations of high-performance energy functions are needed at this point.

ACKNOWLEDGMENT

This work was supported by research grants from NIH (RR15588, PI Lenhoff and GM64458 to YD).

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Sohl JL, Jaswal SS, Agard DA. Unfolded conformations of alpha-lytic protease are more stable than its native state. Nature 1998;395:817–819.
3. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
4. Lazaridis T, Archontis G, Karplus M. Enthalpic contribution to protein stability: insights from atom-based calculations and statistical mechanics. Adv Protein Chem 1995;47:231–306.
5. Karplus M, Shakhnovich EI. Protein folding: theoretical studies of thermodynamics and dynamics. In: Creighton T, editor. Protein folding. New York: WH Freeman; 1992. p 127–195.
6. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy model (with applications to protein folding). J Phys Chem B 1989;93:6902–6915.
7. Park BH, Huang ES, Levitt M. Factors affecting the ability of

energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
8. Vajda S, Sippl M, Novotny J. Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 1997;7:222–228.
9. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213: 859–883.
10. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
11. Dominy BN, Brooks CL. Identifying native-like protein structures using physics-based potentials. J Comput Chem 2002;23:147–160.
12. Price DJ, Brooks CL III. Modern protein force fields behave comparably in molecular dynamics simulations. J Comput Chem 2002;23:1045–1057.
13. Felts AK, Gallicchio E, Wallqvist A, Levy RM. 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;48:404–422.
14. Novotny J, Bruccoleri R, Karplus M. An analysis of incorrectly folded protein models. Implications for structure predictions. J Mol Biol 1984;177:787–818.
15. 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.
16. Samudrala R, Levitt M. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000;9:1399–1401.
17. Simons KT, Bonneau R, Ruczinski II, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37:171–176.
18. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 2001;98:10125–10130.
19. Vorobjev YN, Hermans J. Free energies of protein decoys provide insight into determinants of protein stability. Protein Sci 2001;10: 2498–2506.
20. Vorobjev YN, Hermans J. Free energies of protein decoys provide insight into determinants of protein stability. Protein Sci 2002;11: 994.
21. Vila J, Williams RL, Vasquez M, Scheraga HA. Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. Proteins 1991;10:199–218.
22. Vieth M, Kolinski A, Brooks CL, 3rd, Skolnick J. Prediction of the folding pathways and structure of the GCN4 leucine zipper. J Mol Biol 1994;237:361–367.
23. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 999;288:477–487.
24. Liu Y, Beveridge DL. Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. Proteins 2002;46:128–146.
25. Vorobjev YN, Hermans J. ES/IS: estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. Biophys Chem 1999;78: 195–205.
26. Vorobjev YN, Almagro JC, Hermans J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 1998;32:399–413.
27. 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.
28. Roux R, Simonson T. Implicit solvent models. Biophys Chem 1999;78:1–20.
29. Williams RL, Vila J, Perrot G, Scheraga HA. Empirical solvation models in the context of conformational energy searches: application to bovine pancreatic trypsin inhibitor. Proteins 1992;14:110–119.
30. Wang Y, Zhang H, Li W, Scott RA. Discriminating compact nonnative structures from the native structure of globular proteins. Proc Natl Acad Sci USA 1995;92:709–713.
31. Wang Y, Zhang H, Scott RA. A new computational model for protein folding based on atomic solvation. Protein Sci 1995;4:1402–1411.
32. Monge A, Lathrop EJ, Gunn JR, Shenkin PS, Friesner RA. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
33. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagone G, Profeta S, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984;106: 765–784.
34. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
35. Petrey D, Honig B. Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci 2000;9:2181–2191.
36. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118: 11225–11236.
37. Zhang L, Gallicchio E, Friesner RA, Levy RM. Solvent models for protein-ligand binding: comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. J Comput Chem 2001;22:591–607.
38. Ghosh A, Rapp CS, Friesner RA. Generalized Born model based on a surface integral formulation. J Phys Chem B 1998;102:10983–10990.
39. Holm L, Sander C. Evaluation of protein models by atomic solvation preference. J Mol Biol 1992;225:93–105.
40. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
41. Sternberg MJ, Bates PA, Kelley LA, MacCallum RM. Progress in protein structure prediction: assessment of CASP3. Curr Opin Struct Biol 1999;9:368–373.
42. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
43. Samudrala R, Xia Y, Levitt M, Huang ES. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Pac Symp Biocomput 1999:505–516.
44. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
45. Daura X, van Gunsteren WF, Mark AE. Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. Proteins 1999;34:269–280.
46. Katz M, Papadopoulos P, Bruno G. Leveraging standard core technologies to programmatically build linux cluster appliances. IEEE International Conference on Cluster Computing 2002;47–54.
47. Case DA, Pearlman DA, Caldwell JW, Cheatham TE III, Wang J, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, et al. AMBER 7. San Francisco: University of California San Francisco; 2002.
48. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684–3690.
49. Hawkins GD, Cramer CJ, Truhlar DG. Pairwise solute descreening of solute charges from a dielectric medium. Chem Phys Lett 1995;246:122–129.
50. Hawkins GD, Cramer CJ, Truhlar DG. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute charges from a dielectric medium. J Phys Chem 1996;100: 19824–19839.
51. Srinivasan J, Trevathan MW, Beroza P, Case DA. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. Theor. Chem Acc 1999;101:416–434.
52. Zhang L, Skolnick J. What should the Z-score of native protein structures be? Protein Sci 1998;7:1201–1207.
53. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins 1995;23:ii–v.