Use of Decoys to Optimize an All-Atom Force Field Including Hydration

Yelena A. Arnautova and Harold A. Scheraga
Department of Chemistry and Chemical Biology, Baker Laboratory, Cornell University, Ithaca, New York

ABSTRACT A novel method of parameter optimization is proposed. It makes use of large sets of decoys generated for six nonhomologous proteins with different architecture. Parameter optimization is achieved by creating a free energy gap between sets of nativelike and nonnative conformations. The method is applied to optimize the parameters of a physics-based scoring function consisting of the all-atom ECEPP05 force field coupled with an implicit solvent model (a solvent-accessible surface area model). The optimized force field is able to discriminate near-native from nonnative conformations of the six training proteins when used either for local energy minimization or for short Monte Carlo simulated annealing runs after local energy minimization. The resulting force field is validated with an independent set of six nonhomologous proteins, and appears to be transferable to proteins not included in the optimization; i.e., for five out of the six test proteins, decoys with 1.7- to 4.0-Å all-heavy-atom root mean-square deviations emerge as those with the lowest energy. In addition, we examined the set of misfolded structures created by Park and Levitt using a four-state reduced model. The results from these additional calculations confirm the good discriminative ability of the optimized force field obtained with our decoy sets.

INTRODUCTION

Accurate prediction of protein structure when the only information provided is about amino acid sequence remains one of the greatest challenges in computational chemistry. Results of the CASP (Critical Assessment of Techniques for Protein Structure Prediction) exercises (1) demonstrated that, in many cases, the tertiary structure of a protein (especially a homologous one) can be predicted with a high degree of certainty. However, these predictions provide information about protein structure at relatively low resolution (>3 Å), which may not be sufficient for practical applications (for example, for structure-based drug design). To achieve the atomic level of detail in protein structure prediction, so called refinement methods have been introduced (2–7), and significant attention has been paid to their development (2). These methods are designed to be able to shift the low- and medium-resolution models obtained either from statistics-based methods (homology modeling and threading) or from the use of physics-based course-grained force fields closer to the native state. An important element of any refinement method is an accurate scoring function that must be able to discriminate nativelike conformations from nonnative folds. Many different scoring functions, including empirical (3), knowledge-based (4,5,8), and physics-based (2,6,7) functions, are described in the literature. One type of scoring function, i.e., the one including physics-based all-atom force fields (2,6,7), seems to be a very promising tool for refinement, because these force fields are designed to model physical interactions and may therefore help to shed light on protein folding mechanisms. As such, they are also expected to have better transferability.

According to the Anfinsen thermodynamic hypothesis (9), a necessary requirement for energy functions to produce accurate protein structure models is their ability to recognize the native state of the protein as the conformation, or a set of very similar conformations, for which the system, i.e., the protein plus its surroundings, is of lowest free energy. Several physical scoring functions, such as those based on the AMBER (10–12), OPLS (13), CHARMM (14–16), and GROMOS (17) force fields, combined with different implicit solvent models, were reported to perform well when tested on large sets of decoys generated for many proteins. Despite the significant increase in accuracy of the available all-atom force fields and treatment of solvation effects, the physics-based scoring functions still exhibit some difficulty in differentiating native structures from the sets of decoys generated using very different methods. Their performance worsens when some kind of energy relaxation method (for example, short molecular dynamics (MD) runs) is applied to the decoys (18). Limited success (19,20) has been achieved in refinement of low- and medium-resolution protein models, especially when they have significant unstructured regions and long loops (7).

The accuracy of a physics-based force field is directly related to its ability to reproduce the energetic balance between different interactions accurately. Development of all-atom force fields usually involves the use of a number of assumptions and approximations, such as a simplified form of the energy function, use of fixed charges, omission of polarization effects, and use of implicit solvent models. These approximations definitely affect the accuracy of the resulting force field; however, they are difficult to avoid without significantly increasing the computational cost of the energy calculations. The other source of inaccuracy is the lack of

Submitted March 14, 2008, and accepted for publication May 7, 2008.

Address reprint requests to Harold A. Scheraga, Dept. of Chemistry and Chemical Biology, Baker Laboratory, Cornell University, Ithaca, NY 14853-1301. Tel.: 607-255 4034; E-mail: has5@cornell.edu.

Editor: Ruth Nussinov.

0006-3495/08/09/2434/16 $2.00

doi: 10.1529/biophysj.108.133587

direct experimental data, and hence, torsional and solvation energy terms are probably the most poorly determined parts of the physics-based force fields. For example, the torsional energy terms of all-atom force fields are often optimized by using gas phase data (from high-level quantum mechanical calculations) for small molecules and peptides. It has been shown (21) that the propensity of a given residue to form a particular secondary structure in the gas phase is different from that in solution, and therefore, the torsional parameters derived in this way are not directly transferable to larger systems in a solvent environment.

There is also a lack of direct experimental data on solvation free energies of proteins, which has led to the use of experimental data for small molecules and peptides for parameterization of solvent models. The solvation free energy term is intended to capture several complex effects involved for a protein in solution, from solvent entropy to ionization effects. All these effects may manifest differently for small molecules (or peptides) than for proteins. For example, the solvent exposure of atoms of the same type may, on average, be very different in proteins and small molecules. Moreover, the solvation free energy contribution parameterized by using experimental data for small molecules is usually added to the total energy without any adjustment to take into account the differences between small molecules and proteins. All these factors may contribute to the poor performance of all-atom force fields applied to proteins in solution.

One way to improve the accuracy of all-atom force fields is to use explicit water simulations and available experimental data (conformational equilibria of peptides and small proteins) in force-field parameterization. For example, extensive folding and unfolding simulations with an explicit solvent model have been used for optimization of backbone torsional parameters alone(22) and with solvation parameters (23). Mohanty and Hansmann (24) used parallel tempering simulations with implicit water, carried out for a small β-sheet peptide, to reparameterize an empirical all-atom force field. The relatively high computational cost of this approach and the difficulties in applying it to more than one protein molecule at a time (to insure better transferability of the resulting force field) are some of the main obstacles to its wide application.

Another approach, which was also used in this work, is based on the thermodynamic hypothesis and involves the use of large sets of protein decoys to optimize the parameters of a force field. It was applied initially to parameterize coarse-grained protein models (25,26) and later to optimize all-atom force fields (27–32). Thus, Meirovitch et al. (27,31) optimized solvation parameters associated with solvent-accessible surface areas in all-atom physical energy functions intended for use in predicting surface loops in proteins. Their search of parameter space was restricted to a small number of parameters and was not systematic. A force-field optimization method, called MOPED, was used (28) to improve solvation parameters by creating an energy gap between the native conformations and a small number of decoys of two to three training proteins. Okur et al. (29) applied a genetic algorithm to optimize backbone torsional parameters using a large set of decoys generated for two peptides. Herges and Wenzel (30) parameterized their surface-area solvent model to stabilize the native structure of a single α-helical protein (the villin headpiece) against a large set of nonnative decoys. The resulting force field (30) was shown to be transferable to other α-helical proteins (but not to α/β or β-proteins because they were not considered in the parameterization). In recent work (32), the weights of the AMBER all-atom force field, supplemented by an explicit hydrogen-bond potential, were optimized to stabilize the native structures of a very large number of proteins (namely, 58) against a large number of decoy conformations. The authors also introduced additional energetic and structural criteria into their parameter optimization procedure to achieve better correlation between the energies of decoys and the similarity to the native structure. The force-field optimization led to a significant improvement in performance of the AMBER-based force field. Thus, the fraction of proteins for which the native structure had the lowest energy increased from 0.22 to 0.90. It should be mentioned that the protein decoys used by Wroblewska et al. (32) are characterized by a high degree of similarity (in terms of the secondary and tertiary structure) to the native conformation, and therefore, the ability of the force field they describe to discriminate native structures from very different compact conformations, as well as from the decoys generated by using different decoy generation procedures, remains to be established.

In this work, we introduce a new method of decoy-based force-field optimization. It is based on Anfinsen's thermodynamic hypothesis (9). Therefore, the optimization is aimed at stabilizing nativelike conformations against a large set of decoys by creating free energy gaps between the sets of nativelike and nonnative structures. The search for the best parameters of a force field is carried out with minimization in parameter space. In general, the goal of this work was to test the ability of the new optimization method to find a set of parameters of a given energy function that stabilizes the nativelike conformations of a number of proteins with different folds against large sets of nonnative decoys. To the best of our knowledge, the training set of proteins and the corresponding decoy sets used in this work are among the largest used to date for optimization of physics-based all-atom force fields.

We applied the new method to optimize the torsional and solvation parameters of the effective energy function built by using the physics-based all-atom ECEPP05 force field (33) coupled with the OONS (34) implicit surface-area (SA) solvation free energy term. The original ECEPP05/OONS force field fails to discriminate native structures from the decoys for several nonhomologous proteins (see Results and Discussion section). Although implicit solvent models, especially one as simple as a surface-area model, cannot ac-

count completely for all the effects of solvation, they are computationally very efficient and were shown to perform well when applied to the prediction of surface loops in proteins (27,31) and to the folding of small proteins (30,35) and peptides (36,37). We decided to consider this simple surface-area model and evaluate whether its accuracy can be improved by parameter optimization. We find that the optimization method succeeds in this task, and that the parameters obtained in learning from only a few proteins are transferable to other proteins. As an independent test of the optimized force field, we also considered the 4state-reduced set of decoys of Park and Levitt (38).

METHODS AND MATERIALS

Form of the scoring function

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

$$\Delta G_{\text{tot}} = \Delta G_{\text{int}} + \Delta G_{\text{solv}}, \tag{1}$$

where ΔG_{int} is the internal free energy corresponding to the intramolecular degrees of freedom of the protein. ΔG_{solv} is the solvation free energy of transfer between the gas phase and water.

The internal free energy is given by

$$\Delta G_{\text{int}} = U_{\text{int}} - T\Delta S_{\text{int}}, \tag{2}$$

where U_{int} is the internal energy of the protein and ΔS_{int} is the change in internal entropy due to translational, rotational, and vibrational motions. The entropy contribution in Eq. 2 is often omitted in protein simulation, because of the high cost of its calculation. It has also been found that the vibrational entropy contributions to the free energies of native, misfolded, or denatured conformations are small and comparable (39,40). As a result, we considered a so-called effective free energy,

$$\Delta G_{\text{eff}} = U_{\text{int}} + \Delta G_{\text{solv}}, \tag{3}$$

as a scoring function, instead of the total free energy given by Eq. 1.

The ECEPP05 force field (33) was used to compute the internal energy (U_{int}) of a protein in the absence of solvent. The ECEPP05 internal energy is a function of the torsional degrees of freedom, i.e., all the backbone and side-chain torsional angles, of a protein (all bond angles and bond lengths are fixed at standard values (41)). The U_{int} of a protein is given by

$$U_{\text{int}} = E_{\text{vdW}} + E_{\text{el}} + E_{\text{tor}}, \tag{4}$$

where E_{vdW} and E_{el} are the van der Waals and electrostatic energies, respectively.

The first two terms in Eq. 4 were computed as

$$E_{\text{vdW}} = \sum_{\text{ij(j>i)}} \left[-A_{\text{ij}} r_{\text{ij}}^{-6} + B_{\text{ij}} \exp(-C_{\text{ij}} r_{\text{ij}})\right] \tag{5}$$

and

$$E_{\text{el}} = \sum_{\text{ij(j>i)}} \frac{332 q_{\text{i}} q_{\text{j}}}{\varepsilon r_{\text{ij}}}, \tag{6}$$

respectively, where r_{ij} is the distance between atoms i and j separated by at least three bonds; A_{ij}, B_{ij}, and C_{ij} are nonbonded parameters; q_{i} and q_{j} are point charges (in e.u.) localized on atoms. The dielectric constant ε was taken as unity.

The torsional energy term, E_{tor}, in Eq. 4 for each dihedral angle x was computed as

$$E_{\text{tor}} = k_{\text{x}}^1[1 + \cos(x)] + k_{\text{x}}^2[1 - \cos(2x)] + k_{\text{x}}^3[1 + \cos(3x)], \tag{7}$$

where x represents the backbone and side-chain torsional angles of each decoy conformation (see Protein sets and decoy generation), and k_{x}^1, k_{x}^2, and k_{x}^3 are the torsional parameters; x varies from 0 to 180°.

It should be mentioned that there is no explicit hydrogen-bonding term in the ECEPP05 potential function. This interaction is represented by a combination of electrostatic and nonbonded interactions with the hydrogen involved in a hydrogen bond treated as a separate atom type with parameters different from those of the other types of hydrogens.

The solvation free energy, ΔG_{solv}, of each structure is estimated by using a solvent-accessible SA model,

$$\Delta G_{\text{solv}} = \sum_{\text{i}} \sigma_{\text{i}} A_{\text{i}}, \tag{8}$$

where A_{i} represents the solvent-accessible SAs of various functional groups, and σ_{i} the solvation parameters of these groups. The OONS (34) SA model, which includes the seven types of functional groups (shown in Fig. 1) and their solvation parameters (σ^1, σ^2 . . . , σ^7) derived from the free energies of transfer of small molecules from the gas phase to water, was used in this work.

Scoring methods

The performance of a given scoring function depends not only on the accuracy of its functional form and parameters but also on how it is applied. Scoring of protein decoys using physics-based functions is usually carried out through energy evaluation. Due to the roughness of the all-atom energy surface characterized by huge energy variations corresponding to small changes in structural parameters, such computations may not provide a realistic picture, especially if the decoys were generated using a very different force field. Scoring can also be carried out using local energy minimization, which relaxes a given conformation to the closest energy minimum. All local energy minimizations of the native structures of proteins from the Protein Data Bank (PDB (42)) and of the corresponding structures from the decoy sets considered in this work were carried out using the SUMSL minimizer (43) as implemented in the ECEPPAK program (44–46).

For large all-atom systems such as proteins, which have a very rugged potential energy surface, local energy minimization leads to minor changes in the structure compared with the starting conformation and does not provide information about the existence of lower energy minima corresponding to conformationally very similar structures. A conformational search, limited to

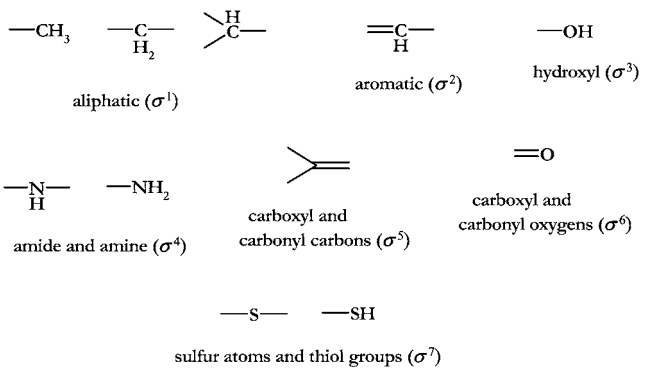

FIGURE 1 Functional groups used in the OONS (34) solvent-accessible SA model.

the vicinity of the starting conformation, should make it possible to overcome this problem; therefore, in this work, we used two types of runs, 1), local energy minimizations, followed by 2), short Monte Carlo simulated annealing (MCSA) runs, to evaluate the energies of protein decoys. The effective free energy ΔG_{eff} given by Eq. 3 was used as a scoring function for both types of runs.

Each MCSA run started at $T_0 = 1000$ K, and the system was then cooled in N cooling cycles to $T_N = 200$ K. The resulting structures at 200 K were energy-minimized. The same number of steps was carried out at each cooling cycle. The temperature of each cycle i was computed according to the formula $T_i = T_0 - i^A$ (47), with

$$A = \frac{\ln(T_0 - T_N)}{\ln(N)}. \quad (9)$$

In each cooling cycle, new conformations were generated by performing a short Monte Carlo search at a given T_i. Conformations generated during the Monte Carlo simulation were obtained by a 10% perturbation of the backbone and side-chain torsional angles.

Protein sets and decoy generation

The proteins considered in this work are listed in Tables 1–3. The training (Table 1) and test (Table 2) sets of proteins were used for force field optimization and evaluation, respectively. These sets contain a total of 12 proteins with 20–76 residues each, and without any stabilizing ligands or disulfide bonds, because these cannot be accounted for by this version of the force field. All types of secondary structure, i.e., α (5), α/β (5), or β (2), are represented in these sets of proteins. All proteins were considered with unblocked N- and C-termini. All ionizable residues and end groups were assumed to be neutral.

Decoys for all the proteins from Tables 1 and 2 were generated according to the procedure described in Ripoll et al. (48), i.e., starting from 1), native, 2), canonical α-helical ($\phi = -60.0°$, $\psi = -40.0°$, $\omega = 180.0°$), and 3), randomly generated conformations, by using the electrostatically driven Monte Carlo method (44) with the ECEPP05 force field coupled with the OONS solvent-accessible SA model. The generated conformations were clustered by using the minimal spanning tree method (49) and assuming a specific root mean-square deviation (RMSD) cutoff of 0.7 Å for all heavy atoms and no cutoff in energy. For each protein, the size of the ensemble generated from all three starting points varied from 1,203 to 7,191 conformations and is characterized for most proteins by a uniform distribution of RMSD from the native fold in the range 0.1–30.0 Å.

The decoy set generated for each protein also included the native structure. The coordinates for the native structure of each protein used in this work (listed in Tables 1 and 2) were taken from the PDB and subsequently converted to ECEPP-type geometry, i.e., with fixed (standard-value) bond lengths and bond angles. This conversion provides an all-atom representation, including hydrogen atoms, for each of the selected proteins. The RMSD for the heavy-atoms between the native structures before and after the conversion is very low, as can be seen from the 6th column in Tables 1 and 2. When more than one structure for a given protein is present in the PDB (NMR-derived structures), the one corresponding to the PDB code in Tables 1 and 2 was selected. If several conformations were submitted under the same PDB code, the model submitted as model number 1 was used.

It should be mentioned that most authors of force-field optimization methods restrict their use to the native protein structures solved only by x-ray diffraction measurements (avoiding NMR-derived models). The main reason for this choice is the lower accuracy (uncertainties up to 2 Å) of NMR structures (due to the much smaller amount of experimental data available for each atom) compared to that of the x-ray structures. Although x-ray-derived protein conformations are more accurate, uncertainties in atomic positions for high-quality structures can be up to 0.6–1.0 Å. Atomic-resolution crystal structures exhibit extensive, discrete conformational substates in which a high percentage of side chains can exist in multiple conformations (50) or are completely disordered. The main chains are more conserved, although uncertainty in the positions of the main-chain atoms can become pronounced in flexible surface loops. Kruskal (51) showed that crystal-packing effects are not a main source of structural differences between NMR and x-ray structures of the same proteins, but he suggested that the crystalline environment could have the effect of ''freezing out'' one conformation from the more diverse ensemble present in solution. This effect may be widespread, since most crystal structures reported today are determined at cryogenic temperature (~100 K). Last but not least, it is not clear what aspects of these low-temperature structures are relevant at room temperature (52). This evidence suggests that use of x-ray diffraction structures has little advantage over that of NMR-derived conformations, and we therefore considered both types of experimental structures in this work.

As an additional test of the optimized force field, we also included the 4state-reduced set of decoys of Park and Levitt (38) (Table 3). This set

TABLE 1 Results obtained for the training set of proteins using the ECEPP05/OONS and the optimized ECEPP05/SA force fields

Protein (PDB code)	Experimental method	Class	No. of residues	No. of decoys	RMSD range*	ECEPP05/OONS RMSD†	ECEPP05/SA (after optimization)‡					
							Minimization§			MCSA¶		
							RMSD‖	ΔG_{eff}**	$\Delta\Delta G_{\text{eff}}$††	RMSD‖	ΔG_{eff}**	$\Delta\Delta G_{\text{eff}}$††
1e0l	NMR	β	37	3563	0.1–18.0	11.2	1.66	−834.1	−22.5	2.01	−839.9	−4.0
1gab	NMR	α	53	7191	0.1–18.0	12.3	4.03	−701.6	−22.1	4.25	−700.8	−15.4
1igd	X-ray	α/β	61	1638	0.1–30.0	23.8	1.36	−922.6	−16.4	1.40	−946.2	−34.9
1l2y	NMR	α/β	20	4028	0.1–10.0	1.9	3.16	−450.0	−0.4	3.16	−449.6	−0.4
1csp	X-ray	β	76	1937	0.1–27.0	16.4	2.19	−1316.0	−5.1	12.8	−1315.7	3.2
1msi	X-ray	α/β	66	4229	0.1–25.0	14.5	2.35	−1397.8	−22.0	2.27	−1396.8	−29.8

*Range of RMSDs from the native structure for decoys of a given protein (Å). The first value corresponds to the RMSD of the native structure converted to the ECEPP geometry (i.e., with standard values of bond lengths and bond angles).

†RMSD (Å) from the native structure of the decoy with the lowest ECEPP05/OONS energy (after only local energy minimization).

‡Results obtained using the optimized ECEPP05/SA force field.

§Local energy minimization.

¶Monte Carlo simulated annealing run after local energy minimization.

‖RMSD of the lowest energy decoy from the native structure (Å).

**Effective free energy (Eq. 3) of the lowest energy decoy, kcal/mol.

††$\Delta\Delta G_{\text{eff}} = \Delta G_{\text{eff}}^{\text{nat}} - \Delta G_{\text{eff}}^{\text{nonnat}}$, where $\Delta G_{\text{eff}}^{\text{nat}}$ and $\Delta G_{\text{eff}}^{\text{nonnat}}$ are the effective free energies (in kcal/mol) of the lowest-energy nativelike and nonnative decoys, respectively. $\Delta\Delta G_{\text{eff}}$ was computed only for the cases in which the energy distributions in Fig. 5 had two well-defined minima corresponding to nativelike and nonnative decoys.

TABLE 2 Results obtained for the test set of proteins using the optimized ECEPP05/SA force fields

Protein (PDB code)	Experimental method	Class	No. of residues	No. of decoys	RMSD range	Minimization			MCSA		
						RMSD	ΔG_{eff}	$\Delta\Delta G_{\text{eff}}$	RMSD	ΔG_{eff}	$\Delta\Delta G_{\text{eff}}$
1bdd	NMR	α	46	1203	0.1–20.0	3.92	−1365.2	−72.7	3.66	−1381.0	−48.9
1vii	NMR	α	36	2271	0.9–13.0	3.07	−718.1	−18.5	5.45	−733.4	—*
1res	NMR	α	43	2824	0.1–21.0	2.69	−1494.6	—	2.66	−1503.9	—*
1fsd	NMR	α/β	28	3208	0.1–12.0	3.77	−1308.1	−46.0	6.47	−1330.4	8.3
1cc7	X-ray	α/β	72	5755	0.1–27.0	1.72	−1166.2	—	1.53	−1171.8	—*
1ail	X-ray	α	73	3622	0.1–29.0	3.09	−2440.3	−35.7	3.19	−2430.4	−19.4

See Table 1 notes for descriptions of parameters.
*$\Delta\Delta G_{\text{eff}}$ was not computed (see Table 1, last note).

contains decoy structures for seven proteins. We considered only five of these, namely, 1ctf, 1r69, 3icb, 4rxn, and 2cro, because the native structures of the remaining two proteins (4pti and 1sn3) are stabilized by disulfide bonds.

Parameter-optimization method

The parameter-optimization method described in this work makes use of Anfinsen's thermodynamic hypothesis (9). To develop a force field satisfying this hypothesis, two sets of conformations, namely, nativelike and nonnative ones, were considered, and the optimized parameters were derived by creating a free energy gap between these two sets. This approach differs from other similar optimization methods described in the literature (28,30) which make use of a single native structure instead of a set of nativelike conformations. The similarity measure that we used to define a nativelike structure is described in the next subsection.

We introduced a conformational free energy, ($F_i(\mathbf{a})$), of a native/nonnative set of structures, computed as

$$F_i(\mathbf{a}) = -\frac{1}{\beta}\ln \sum_{k\in\{i\}} \exp\left[-\beta\Delta G^k_{\text{eff}}(\mathbf{a}, x)\right], \qquad (10)$$

where **a** is a vector of force field parameters (k and σ), and ΔG^k_{eff} is the effective free energy of the kth conformation; and x represents the backbone and side-chain torsional angles of the kth conformation. $\beta = 1/RT$, where R is the universal gas constant and T is the absolute temperature. T was considered an empirical parameter. By allowing T to vary, the value of the conformational free energy of a given set of structures can be altered relative to the energy distribution of this set. The value of T used in this work (β = 0.5 mol/kcal) was chosen in such a way that the conformational free energy of each level (i.e., a set of nativelike or nonnative structures, as described in next section) was close to the energy of the lowest-energy structure from this level. The Boltzmann summation in Eq. 10 is taken over the conformations from level i (i denotes the native or nonnative level).

Using the ECEPP05/SA energy function, we modified the force-field parameters, **a**, in an attempt to satisfy the condition that, for a training set of proteins, the conformational free energy of the nativelike conformations should be lower than the conformational free energy of a set of nonnative decoys:

$$F_{\text{nat}} - F_{\text{nonnat}} < -\Delta. \qquad (11)$$

Thus, optimization of the force-field parameters was achieved by creating a negative free energy gap between the native and nonnative levels. Target gaps (Δ) were set to the same value of 5 kcal/mol for all the training proteins considered in this work. This value was chosen based on the evidence that native structures of proteins are marginally stable (53). We did not try to maximize the gap, because that would lead to a nonphysical and poorly transferable force field.

Force field parameters (**a**) were optimized by minimizing the target function

$$\Phi(\mathbf{a}) = \sum_{j}^{N} w_j g\left[F^j_{\text{nat}} - F^j_{\text{nonnat}}; -\Delta^j\right], \qquad (12)$$

where the summation runs over the number of training proteins N and

$$g(y; y_{\max}) = \left\{ \begin{array}{cc} 1/4(y - y_{\max})^4 & y > y_{\max} \\ 0 & y \le y_{\max} \end{array} \right\}. \qquad (13)$$

Δ^j is a target free energy gap for protein j, w_j is an empirical weight which was set to 1 for all training proteins, $y = F^j_{\text{nat}} - F^j_{\text{nonnat}}$, and $y_{\max} = -\Delta^j$.

TABLE 3 Results for the 4state-reduced decoy set obtained using the optimized ECEPP05/SA force field

Protein (PDB code)	Class	No. of residues	No. of decoys	Energy evaluation*		Minimization†		MCSA‡	
				RMSD§	ΔG_{eff}¶	RMSD§	ΔG_{eff}¶	RMSD§	ΔG_{eff}¶
1ctf	α/β	68	630	3.20	−430.5	1.17	−906.7	1.73	−962.0
1r69	α	63	675	0.14	−1339.7	0.76	−1914.0	1.23	−1981.3
3icb‖	α	75	653	1.80	−715.4	—	—	—	—
4rxn**	α/β	54	677	0.31	−337.8	—	—	—	—
2cro	α	65	674	0.14	−1230.1	2.65	−1909.7	1.17	−1934.3

The 4state-reduced decoy set was developed by Park and Levitt (38).
*Energy evaluation for a fixed conformation.
†Local energy minimization.
‡Monte Carlo Simulated Annealing run after local energy minimization.
§RMSD from the native structure for the lowest energy decoy, Å.
¶Effective free energy (Eq. 3) of the lowest energy decoy, kcal/mol.
‖Calcium binding protein.
**Metal-binding protein.

The energy of a given conformation, represented by a set of backbone and side-chain torsional angles, **x**, is a function of the force field parameters, **a**, i.e., all k and σ. As the force-field parameters vary, the conformations based on the initial parameters may no longer correspond to energy minima; therefore, the effective free energies (Eq. 3) computed with the new parameters will not reflect the real relative stabilities of native and nonnative levels. To solve this problem, we employed the following approach. The parameter optimization is an iterative procedure in which each iteration first involves optimization of **a** while holding the conformations fixed. Then, all conformations are energy-minimized with the resulting interim parameters, **a**. This procedure is repeated until all the free energy gaps reach the predefined value Δ^j ($\Delta^j = 5$ kcal/mol for all j). Both the minimization of the target function Φ as a function of **a** and the minimization of the effective free energy, ΔG_{eff}, for each conformation as a function of **x** are carried out by using the SUMSL minimizer (43). At each iteration, we computed the local energy minima with the current force-field parameter set by performing energy minimizations from the fixed initial native and decoy conformations. For each training protein, all decoys plus the native structure are included in the optimization procedure. The flowchart of the optimization method is shown in Fig. 2.

The method described above was applied to optimize the backbone ϕ and ψ torsional (k_x^1, k_x^2, and k_x^3 in Eq. 7) and solvation (σ_i in Eq. 8) parameters. The force field parameters for the remaining torsional angles (ω and χ) were kept fixed at the original ECEPP05 values (33). We also attempted to stabilize the nativelike conformations of the training proteins by varying the relative contributions of different energy terms of the effective energy function (Eq. 3) by optimizing the weights (w) of the equation

$$\Delta G_{\text{eff}} = w_{\text{vdW}} \times E_{\text{vdW}} + w_{\text{el}} \times E_{\text{el}} + w_{\text{tor}} \times E_{\text{tor}} + w_{\text{solv}} \times \Delta G_{\text{solv}}. \tag{14}$$

The w values were constrained to positive values. However, it was not possible to achieve the target free energy gaps by varying only the weights when more than one protein was considered. Therefore, we focused on optimization of only the torsional and solvation parameters, with all w set at 1. In the rest of this article, we report results and discuss these simulations.

Similarity measures used in parameter optimization and analysis of the results

As mentioned earlier, we considered a set of nativelike conformations (defined below) instead of a single native structure to optimize parameters of an all-atom force field. Our decision to use a set of nativelike conformations is based on the fact that a protein under physiological conditions exists as a dynamic ensemble of conformations. Ideally, NMR experiments should be able to provide information about this ensemble. Although protein structures solved by x-ray diffraction are represented by a single conformation, work on the interpretation of crystallographic data (54) showed that dynamics and heterogeneity remain even in the crystalline state, and suggested that a single conformation may not provide the best solution to the crystallographic structure-determination problem (54,55). In addition, the authors concluded that use of a single conformation may introduce a bias in the computation of protein properties such as solvent-accessible SA, total energy, etc., which are sensitive to small variations in atomic positions.

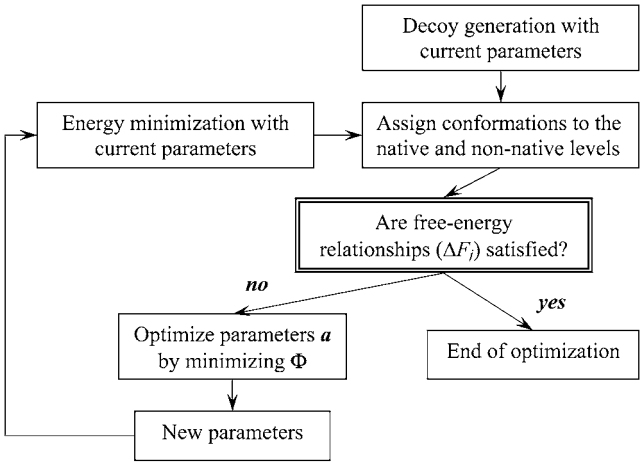

FIGURE 2 Flowchart of the parameter-optimization method.

All decoy conformations were divided into two groups (levels), namely, nativelike and nonnative, according to two criteria: 1), fraction of residues with the same conformation (according to the conformational letter code of Zimmerman et al. (56)) as in the corresponding fragment of the experimental structure, and 2), fraction of contacts in a fragment matching those in the corresponding fragment of the experimental structure. Thus, the similarity of packing of the secondary structure fragments was defined in terms of the fraction of the interfragment native contacts.

A decoy is defined as nativelike if both the secondary structure and packing of the secondary structure elements are similar to those in the native structure. The first of these requirements, i.e., similarity of the secondary structure, means that at least 60% of consecutive residues in each secondary structure element should be from the same regions (defined by Zimmerman et al. (56)) of the Ramachandran (ϕ-ψ) map as the corresponding native residues.

To quantify the similarity of the packing of secondary structure elements to that in the experimental structure, we used the Q parameter introduced in Furnham et al. (57), i.e.,

$$Q = \frac{1}{M} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \sum_{k=1}^{N(i)} \sum_{l=1}^{N(j)} \frac{|d_{kl} - d_{kl}^{\text{nat}}|}{d_{kl}^{\text{nat}}}, \tag{15}$$

where L is number of secondary structure elements; $N(i)$ is the number of residues (C^α) in the ith element; M is the total number of distances; and d_{kl} denotes the distance between the α-carbon atoms of residues k and l, respectively, of the conformation under consideration. The same quantities with the ''nat'' superscript denote the distances in the experimental structure. A conformation was assigned to the ''native'' level if the value of Q was lower than or equal to the similarity threshold θ^E. Otherwise, the conformation was added to the list of nonnative conformations. The value of the θ^E parameter was chosen empirically (by trial and error) as 0.18.

A set of nativelike conformations defined according to the twofold similarity measure introduced above includes more diverse structures than is described by an average experimentally determined native ensemble (RMSD $\leq$ 2 Å); there is no direct correspondence between the measure described above and RMSD. However, the nativelike structures correspond roughly to those with RMSD $\leq$ 4 Å from the PDB structure. Since the goal of this work was to develop a force field capable of discriminating nativelike from nonnative protein conformations (i.e., those with different tertiary and even secondary structure), we felt that the use of the definition of a nativelike structure introduced here is justified.

For analysis of the results, the structural similarity between two protein conformations was also expressed as the RMSD between the best overlap of the heavy atoms (i.e., all atoms except hydrogens) of the two conformations.

RESULTS AND DISCUSSION

In this section, we report the application of the parameter optimization method presented in this article to the development of an accurate all-atom force field including hydration. First, the accuracy of the original ECEPP05/OONS force field to score protein decoys is evaluated. Next, we report optimization of the force field parameters (torsional

and SA solvation parameters (see below)) using the procedure described in the Methods section and shown in Fig. 2. The optimization procedure minimizes the target function Φ (Eq. 12, Fig. 2) and is aimed at stabilizing nativelike conformations relative to nonnative decoys for a set of training proteins. Finally, the resulting optimized force field was used in two types of simulations: 1), local energy minimizations; and 2), MCSA runs after the local energy minimizations. Performance of the force field in these tests carried out for the training and test sets of proteins, as well as for the 4state-reduced decoy set of Park and Levitt (38), is discussed.

Performance of the original ECEPP05/OONS force field

Decoys of the training proteins (Table 1) were first energy-minimized using the all-atom ECEPP05 force field (33) combined with the OONS implicit SA solvation model (34) with the original parameters (ECEPP05/OONS). The results of the calculations are reported in Table 1, column 7, and in Fig. 3. For all the proteins but one (1l2y) from the training set, nonnative structures have lower energies than nativelike ones. The nativelike conformation with an RMSD of 1.9 Å from the experimental structure was obtained as the lowest-energy one for 1l2y. For all the other proteins from the set, the ECEPP05/OONS force field favors all-helical structures. In the case of the α-helical protein, 1gab, the lowest-energy structure (a two-helix bundle) differs from the native conformation (a three-helix bundle).

Low energies of nonnative helical conformations with ECEPP05/OONS are due to the large contribution of the sum of the nonbonded and torsional energies, which constitutes ~90% of the total energy. Since this part (i.e., $E_{vdW} + E_{el} + E_{tor}$) of the force field was parameterized to reproduce the ab initio (gas phase) ϕ-ψ map of terminally blocked alanine, which has a global minimum corresponding to the α-helical conformation, the ECEPP05 force field is expected to favor this type of conformation. On the other hand, the solvation

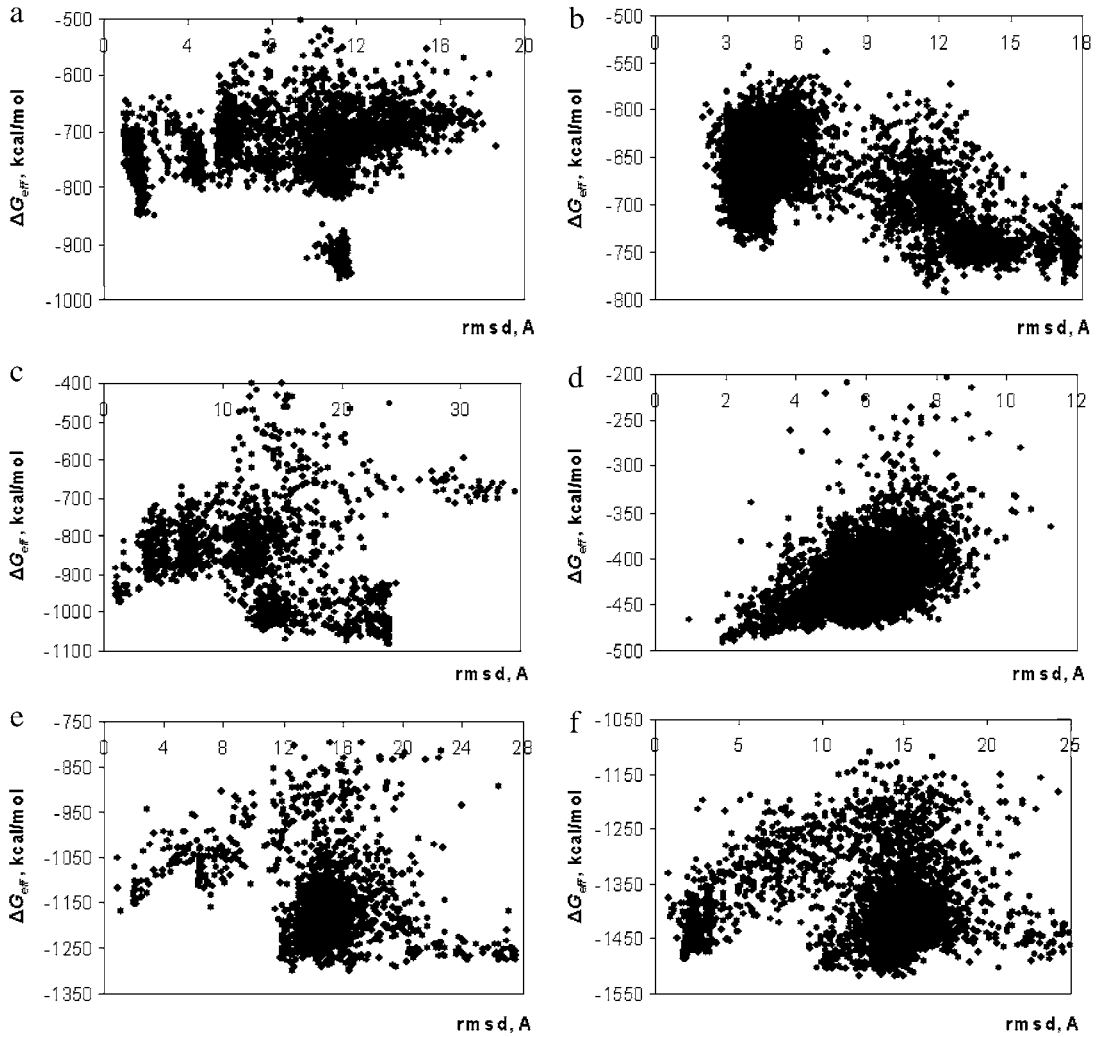

FIGURE 3 Scatter plot of the structures obtained by local minimization of the energies of the decoys from the training set with the original ECEPP05/OONS force field versus all-heavy-atom RMSDs from the experimentally determined native structure. (*a*) 1e0l. (*b*) 1gab. (*c*) 1igd. (*d*) 1l2y. (*e*) 1csp. (*f*) 1msi.

energy is more favorable for extended structures. However, the solvation energy contribution to the total energy for the OONS model is very small (~10%) and does not significantly affect the relative stabilities of different conformations. Most of the high-RMSD decoys shown in Fig. 3 are noncompact α-helical conformations, which are favored by both the gas phase and the solvation energy terms of the ECEPP05/OONS force field. Only in the case of 1l2y is the low total energy of the 1.9-Å nativelike conformation determined by the ($E_{nb} + E_{tor}$) contribution.

These results indicate that the ECEPP05 force field combined with the OONS solvent model (with their original parameters) is not able to discriminate nativelike structures and has to be improved.

Force-field optimization

In this work, we attempted to optimize the force-field parameters by minimizing a target function Φ (Eq. 12). The backbone torsional ϕ and ψ and the solvation parameters were selected as candidates for the optimization because they are the most difficult ones to derive from first principles and, therefore, their values may contain a high degree of uncertainty. First, only the torsional parameters were allowed to vary during the optimization. When more than one training protein was considered, the optimization did not lead to any significant improvement in the stability of the nativelike structures relative to the nonnative decoys. On the other hand, when the target function, Φ, was optimized as a function of either solvation alone or both solvation and torsional parameters, the target free energy gaps were achieved in a small number of iterations. There was also more than one set of parameters that minimized the target function Φ. As a result, it may be necessary to consider a very large set of proteins to obtain a unique set of parameters. Since consideration of a large number of proteins simultaneously (with a large number of decoys) is computationally very demanding, we decided to focus on optimization of the full range of σ parameters of the solvation model, allowing only a limited variation of the torsional parameters (within $\pm$10% of the gas phase values). The torsional parameters were not fixed during the optimization, because their original values were derived from ab initio (gas phase) calculations and therefore may not be adequate for a protein in solution. No restrictions were placed on possible values of the solvation parameters; however, we assumed that the OONS parameter set represents a reasonable starting point for reparameterization, and therefore, we did not attempt to explore the space of solvation parameters for alternative optimized parameters (i.e., only one starting set of parameters (OONS) was considered).

Before starting the optimization, we evaluated how the size of the training set influences the resulting values of the force-field parameters. Five sets, containing 2, 3, 4, 5, and 6 proteins, respectively, were considered. These sets were composed of the proteins from Table 1 by adding one protein at a time, e.g., 1e0l and 1gab (set 1), 1e0l, 1gab, and 1igd (set 2), and so on. Fig. 4 shows the values of the solvation parameters as a function of the size of the training set. The parameters that depend the most on the size of the training set are σ^1, σ^2, σ^5, and σ^6 (the corresponding functional groups are shown in Fig. 1). The solvation parameter for the aliphatic group (Fig. 1), σ_1, is the most sensitive to the number of proteins used for its optimization. In general, changes in the solvation parameters as a function of the training set size are small (<4%) and become even smaller when the set size reaches five proteins. Although the parameter values obtained from the optimizations carried out using five and six proteins, respectively, are very close, we decided to use six proteins to ensure better transferability of the resulting force field.

The training set containing the six proteins listed in Table 1 was considered in the parameter optimization carried out by using the procedure (Fig. 2) described in the Methods section. The force-field parameters that satisfy Eq. 11, with Δ = 5 kcal/mol, were obtained in a small number of iterations (approximately three). The resulting backbone torsional and solvation parameters are given in Tables 4 and 5, respectively. The most significant changes in the parameter values arose for three types of groups, namely, aliphatic, aromatic, and carboxyl/carbonyl oxygen. The aliphatic and aromatic groups became more ''hydrophobic'' (large positive values of σ), whereas carboxyl/carbonyl oxygen became more ''hydrophilic'' (more negative values of σ).

Fig. 5 shows energies of the decoys versus all heavy-atom RMSDs from the native structure for the six training proteins. The energies (Fig. 5, *red circles*) correspond only to the local energy minima of the optimized force field before implementation of MCSA. The optimized ECEPP05/SA force field stabilizes near-native conformations against the competing low-energy decoys for all six proteins (the energy gaps between the lowest-energy nativelike and nonnative decoys (Table 1, *column 10*) are all negative). The best result, in terms of the RMSD from the native structure, was obtained

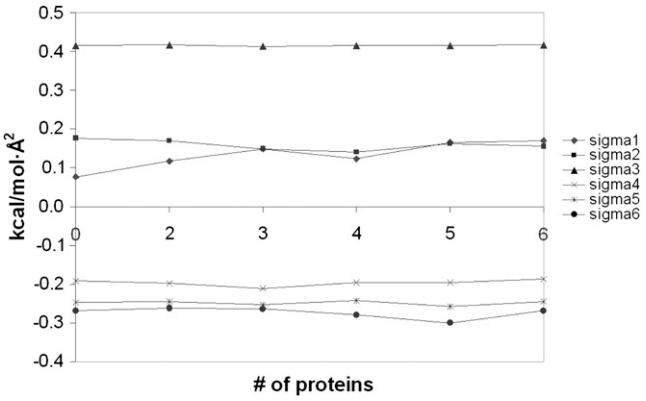

FIGURE 4 Optimized values of the solvation parameters (σ) as a function of the size of a training set. The zero in the number of proteins indicates that the values of σ are the initial OONS (34) values.

TABLE 4 Initial and optimized values of the backbone ϕ and ψ torsional parameters (kcal/mol)

	φ			ψ		
Parameters	k_1	k_2	k_3	k_1	k_2	k_3
Initial	−1.43	1.41	0.19	−1.70	1.95	−0.46
Optimized	−1.41	1.41	0.17	−1.64	1.95	−0.49

Initial values were taken from Arnautova et al. (33). Force field parameters for the torsional angles ω and χ were taken from ECEPP05, and were kept fixed during the parameter optimization.

for 1igd, for which the lowest-energy decoy was only 1.36 Å from the x-ray conformation (Fig. 5 *c*, *red circles*). For 1gab, the lowest-energy decoy (−701.6 kcal/mol) has a nativelike structure, although its RMSD from the native structure is relatively high (~4 Å, Fig. 5 *b*). The only protein for which the optimized force field performed less well than ECEPP05/OONS was 1l2y (Trp-cage, Fig. 5 *d*). The lowest-energy decoy (−450.0 kcal/mol) of 1l2y has a relatively high RMSD (3.16 Å) from the native structure and is only marginally more stable (~0.4 kcal/mol) than the low-energy nonnative conformations (Fig. 5 *d* and Table 1). At the same time, optimization of the force field led to increased stability of a set of nonnative decoys of 1l2y, with RMSDs around 6.4 Å and general flattening of the effective free energy surface for the 2–6.5 Å RMSD range (Fig. 5 *d*), indicated by the appearance of the large number of decoys in this RMSD range with similar and very low energies (Fig. 5 *d*).

Performance of the optimized ECEPP05/SA force field evaluated using short MCSA runs on the training set of proteins

The parameters obtained by using the optimization procedure described in this work (see Methods and Fig. 2) and reported in Tables 4 and 5, are not guaranteed to be optimal, even for the training proteins, because low-energy decoys, not included in the training decoy sets, can exist. The free energy relaxation carried out by performing short MCSA runs after energy minimization was used to explore the free energy surface in the vicinity of the minima corresponding to the training decoys. These simulations yielded results (Fig. 5, *blue circles*, and Table 1, *columns 11 and 13*) very similar to those obtained from local energy minimization (Table 1, *columns 8 and 10*). For five out of six training proteins, nativelike structures were scored by the optimized force field as the lowest in energy. Only in the case of 1csp did nonnative all-helical decoys have lower free energies ($\Delta\Delta G_{\text{eff}} > 0$ (Table 1, *column 13*)) than nativelike β-barrel conformations. It should be mentioned that the energies of these new nonnative conformations produced by MCSA runs are still higher (−1315.7 kcal/mol) than the energies of the lowest-energy nativelike structures obtained from local energy minimizations (−1316.0 kcal/mol) (Table 1, *columns 12* and *9*, respectively). Comparison of the free energies obtained separately by either local energy minimization or MCSA runs after local energy minimization shows that only in the case of 1e0l and 1igd did the MCSA search lead to a significant decrease in energy of the decoys compared to those obtained from local energy minimization (−839.9 vs. −834.1 and −946.2 vs. −922.6 for 1e0l and 1igd, respectively (Table 1, *columns 12* and *9*)). The MCSA runs did not yield any lower-energy nativelike decoys for 1gab and 1 csp (decoys with RMSD < 5 Å in Fig. 5 *b* and decoys with RMSD < 4 Å in Fig. 5 *e* for 1gab and 1csp, respectively) or any nativelike or nonnative conformations for 1l2y and 1msi (Fig. 5, *d* and *f*, respectively). For 1l2y, a small 20-residue protein, this result may be caused by good sampling of the conformational space during decoy generation. However, for 1gab, 1csp, and 1msi, MCSA seems to be less efficient in finding new low-energy minima. 1csp and 1msi are the largest proteins in the training set (76 and 66 residues, respectively), so it makes sense that generating near-native decoys would not be easy, because even small conformational changes in the interior residues may lead to atomic clashes and high energies.

The MCSA runs intended to explore the vicinity of free energy minima corresponding to the decoys, and therefore provide additional information about the free energy surface of a protein, did not locate any nonnative conformations with energies lower than those of the nativelike structures. In other words, considering the combined results of the local energy minimizations and the MCSA runs, the optimized ECEPP05/SA force field is able to discriminate nativelike structures for all six training proteins as the lowest in effective free energy.

TABLE 5 Initial (OONS) and optimized values of the solvation parameters (σ)

Chemical group		σ_{OONS}	σ_{opt}
Aliphatic (CH3, CH2, CH)	(σ_1)	0.008	0.171
Aromatic (=CH−)	(σ_2)	−0.008	0.155
Hydroxyl (−OH)	(σ_3)	0.427	0.416
Amide and amine (NH2, NH)	(σ_4)	−0.132	−0.187
Carboxyl and carbonyl carbon	(σ_5)	−0.172	−0.245
Carboxyl and carbonyl oxygen	(σ_6)	−0.038	−0.269
Sulfur –S− and thiol –SH	(σ_7)	−0.021	—*

Values are given in kcal/mol·Å^2.

*The training proteins did not have sulfur-containing residues in their sequences.

Evaluation of transferability of the optimized ECEPP05/SA force field using a test set of proteins

To evaluate whether the optimized ECEPP05/SA force field is transferable to other nonhomologous proteins, i.e., whether it is able to score near-native conformations as those with lowest energies, we considered sets of decoys generated for the six proteins (test set) listed in Table 2. The test set included four α-helical and two α/β proteins. Both local energy minimization and free energy relaxation (MCSA) runs after

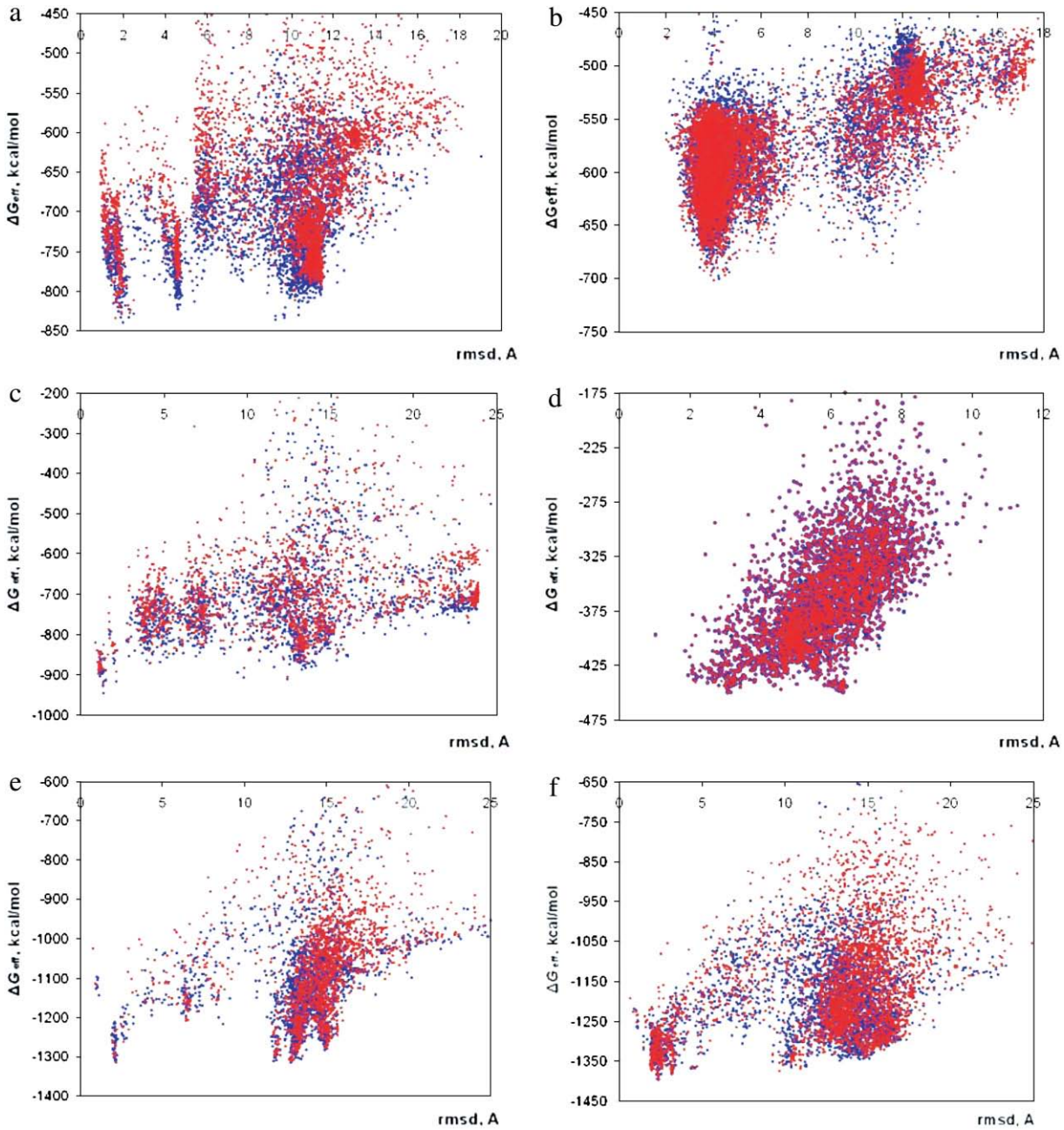

FIGURE 5 Scatter plot of the ECEPP05/SA energy (after parameter optimization) of the decoys from the training set versus the RMSD from the experimentally determined native structure. (*a*) 1e0l. (*b*) 1gab. (*c*) 1igd. (*d*) 1l2y. (*e*) 1csp. (*f*) 1msi. Red and blue circles correspond to the results obtained from either local energy minimization, or MCSA runs after local energy minimization, respectively.

energy minimization were carried out for the decoys of these six proteins. The results of the calculations are given in Table 2 and Fig. 6.

As a result of local energy minimization with the optimized force field, we find that for each of the six test proteins, a near-native conformation emerges as the lowest in energy when compared to other low-energy decoys (Fig. 6, *red circles*). For all the proteins, the RMSD of the lowest-energy decoys is <4.0 Å (Table 2, *column 7*). As seen in Fig.6 *e*, the most stable decoy of 1cc7 is characterized by the lowest RMSD of 1.72 Å from the native structure (the latter being shown in Fig. 7 *a*). The highest RMSD, 3.92 Å, was obtained for 1bdd (Fig. 6 *a*). The two (middle and C-terminal) helices in the lowest-energy 1bdd decoy have the same tertiary alignment as in the NMR structure, but with slightly different orientation of the N-terminal helix (shown in Fig. 7 *b*). The lowest-energy decoy of 1fsd also has a relatively high (3.77 Å) RMSD from the native structure. The main difference

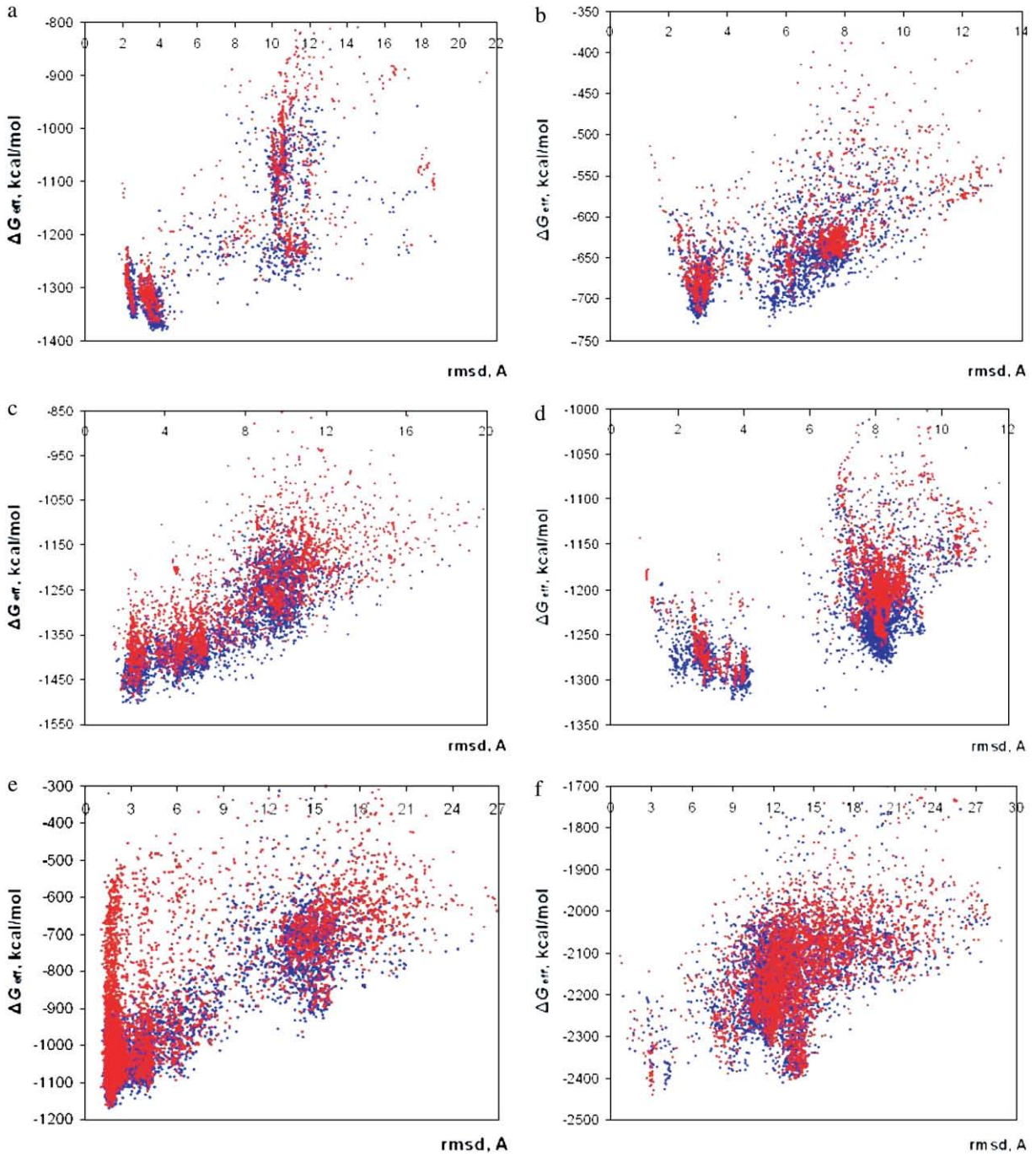

FIGURE 6 Scatter plot of the ECEPP05/SA energy (after parameter optimization) versus RMSD from the experimentally determined native structure for the proteins from the test set: (*a*) 1bdd; (*b*) 1vii; (*c*) 1res; (*d*) 1fsd; (*e*) 1cc7; (*f*) 1ail. Red and blue circles correspond to results obtained from local energy minimization, and MCSA runs after local energy minimization, respectively.

between this conformation and the NMR structure lies in the shape of the N-terminal fragment (shown in Fig. 7 *c*).

Somewhat different results were obtained for the test decoy sets using free energy relaxation (MCSA runs after energy minimization (Fig. 6, *blue circles*)). For five out of six proteins, namely 1bdd, 1vii, 1res, 1cc7, and 1ail, nativelike structures are stabilized relative to nonnative decoys (Fig. 6 and Table 2, *column 12*). For the remaining protein, 1fsd, a nonnative decoy with an RMSD of 6.47 Å (Fig. 6 *d*) was scored as the most stable (−1330.4 kcal/mol). In the case of 1vii, the lowest-energy decoy (−733.4 kcal/mol) has a high RMSD of 5.45 Å; however, it has overall nativelike topology with slightly higher helix content and, compared to the NMR conformation, a different relative orientation of helices 1 and 2

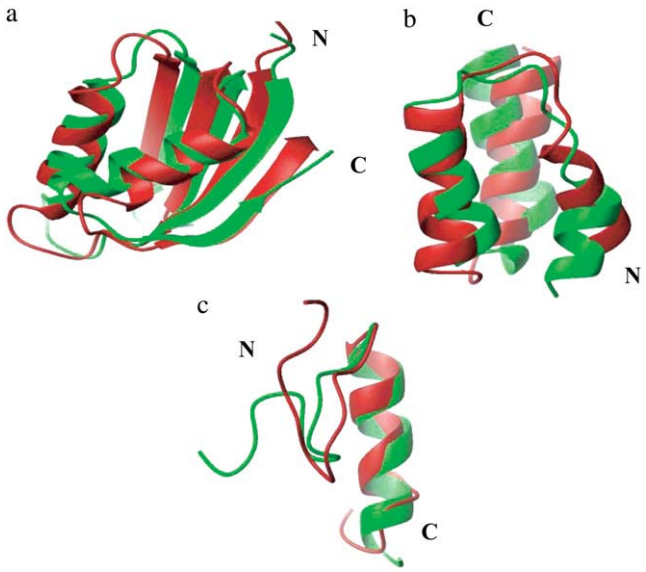

FIGURE 7 Overlay of the experimental structure (*red*) and the decoy with the lowest ECEPP05/SA energy (*green*) obtained by local energy minimization for (*a*) 1cc7, (*b*) 1bdd, and (*c*) 1fsd, with RMSDs from the native structure of 1.7, 3.9, and 3.8 Å, respectively.

(Fig. 8). It should be mentioned that, because of its small size and fast folding, the villin headpiece (1vii) has been a subject of many types of biomolecular simulations carried out using a variety of search methods, such as global optimization techniques and MD simulations, combined with different force fields and solvation models. The results presented in this work cannot be compared directly with those of MD simulations (for example, those reported by Lei et al. (58)). On the other hand, the results reported for the villin headpiece by Herges et al. (59) and Ripoll et al. (60) were obtained using the same type of force field (torsional force field with fixed valence geometry, ECEPP). A 3.3-Å backbone RMSD structure was found by Herges et al. (59) as the one with lowest energy, whereas the search carried out by Ripoll and co-workers (60) using the ECEPP03/OONS force field yielded the lowest-energy structure, with RMSD ~5 Å. It should also be mentioned that the low-energy decoys (native and nonnative) obtained in the work by Herges et al. (59) had three-helix conformations with the native secondary structure and slightly different packing of the helices. These conformations were found to be very close to one another in energy.

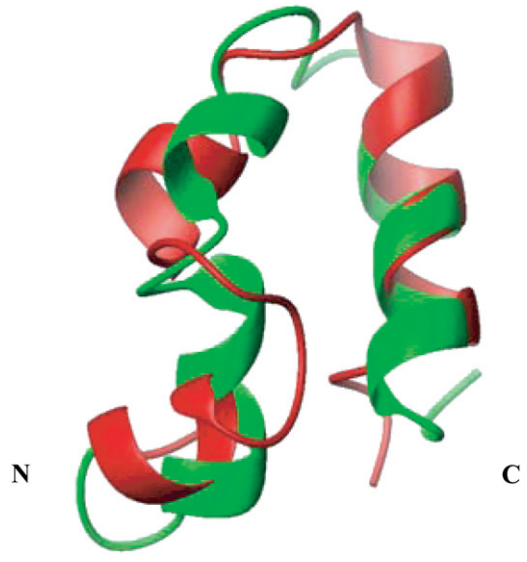

FIGURE 8 Overlay of the experimental structure (*red*) and the decoy with the lowest ECEPP05/SA energy (*green*) obtained from the MCSA run for 1vii. The RMSD from the native structure is 5.5 Å.

The MCSA runs also led to significantly decreased energies of the 1vii and 1fsd decoys, especially the nonnative ones (RMSD $>$ 4 Å (Table 2 and Fig. 6, *b* and *d*)). The free energy relaxation for the largest test protein (1ail) did not yield any conformations with energies lower than those obtained from local minimization (Fig. 6 *f* and Table 2, *columns 8* and *11*). This result supports the earlier conclusion that the energy relaxation procedure employed here appears to be more efficient in the case of smaller proteins (1vii and 1fsd).

Analysis of the results obtained for the test proteins shows that the scoring function, which combines the optimized ECEPP05/SA force field and local energy minimization, succeeds in discriminating nativelike structures (RMSD $<$ 4 Å) from large sets of nonnative conformations for all six test proteins. At the same time, the scoring function with a short MCSA run after local energy minimization fails to identify nativelike conformations as those with the lowest free energies for 1fsd (Table 2 and Fig. 6 *d*). This indicates that free energy relaxation (MCSA), which provides additional information about the free energy surface of a protein, represents a more stringent test than local energy minimization for the accuracy of a force field and should be used to obtain a more realistic evaluation of its performance.

Evaluation of the optimized parameters using the 4state-reduced decoy set

When only one type of decoy set is used to evaluate the performance of a scoring function, good discrimination may be achieved by some special feature of this decoy set. To check whether the optimized ECEPP05/SA force field performs well for decoys generated by using a completely different method from that used in this work, we considered the 4state-reduced decoy set (38). It has been one of the most popular decoy sets used for evaluation of different scoring functions (11–15). This set not only spans the conformations with RMSD ranging from 1 to 10 Å, but also includes a large number of decoys with low RMSD ($<$4 Å) from the native conformations, and therefore is very useful for assessment of the ability of a given scoring function to discriminate both nativelike structures from nonnative decoys and native structure from nativelike conformations.

We considered only five proteins, namely, 1ctf, 1r69, 3icb, 4rxn, and 2cro from the 4state-reduced decoy set (Table 3). The native structures of 4pti and 1sn3 (the remaining proteins from the set) contain several disulfide bonds, and therefore were not considered in this work. First, energy evaluations of fixed conformations were carried out for the decoys of 1ctf, 1r69, 3icb, 4rxn, and 2cro using the optimized ECEPP05/SA force field (Table 3, *column 5*, and Fig. 9). The native structures of 1r69, 4rxn, and 2cro score lowest in effective free energy (−1339.7, −337.8, and −1230.1 kcal/mol, respectively). In the case of 3icb, the nativelike decoy with an RMSD of 1.80 Å from the native structure was scored as having the lowest free energy (−715.4 kcal/mol).

The next step in the evaluation of the optimized force field was local energy minimization and short MCSA runs after local energy minimization for the decoys from the 4state-reduced set. The experimental structures of 4rxn and 3icb were solved as complexes with metal and calcium atoms, respectively. Locations of these atoms in the loop regions suggest that they play a crucial role in defining the native conformations. It might be expected that energy minimization of the native structures carried out without considering these atoms would lead to significant conformational changes. Therefore, we did not carry out energy minimizations or Monte Carlo simulated annealing runs for the decoys of 4rxn and 3icb. The RMSDs obtained for 1ctf, 1r69, and 2cro are given in Table 3 (*columns 6 and 7*) and Fig. 9. Local energy minimization yielded the native conformations as those with the lowest energy for 1ctf and 1r69 (−906.7 and −1914.0 kcal/mol). The native structure of 2cro was obtained as the second lowest (−1896.6 kcal/mol) after the nativelike conformation with RMSD of 2.65 Å (−1909.7 kcal/mol (Fig. 9 *c*)). The MCSA runs after energy minimization discriminated nativelike conformations as those with the lowest energy for all three proteins, i.e., 1ctf, 1r69, and 2cro (Fig. 9 and Table 3, *column 7*). In the case of 2cro, the force field performed better when used for MCSA runs after energy minimization than for energy minimization alone, as seen from the lower RMSD value of the most stable decoy (1.17 vs. 2.65 Å, Table 3). The reason for the somewhat improved discriminative ability of the scoring function including free energy relaxation (MCSA runs) may be elimination of the unfavorable contacts created as a result of conversion of the original native structure to the one with the standard ECEPP geometry with fixed bond lengths and bond angles. It is likely that local energy minimization cannot relax these contacts, whereas free energy relaxation (MCSA runs) may do so.

It should be mentioned that the low RMSDs (<2 Å) from the corresponding native structure of the most stable decoys from the 4state-reduced set suggest that the resolution of the optimized force field is sufficiently high. On the other hand, relatively high RMSDs (3–4 Å) of the lowest-energy nativelike decoys of some training and test proteins considered in this work may be a result of somewhat insufficient sampling of the native region.

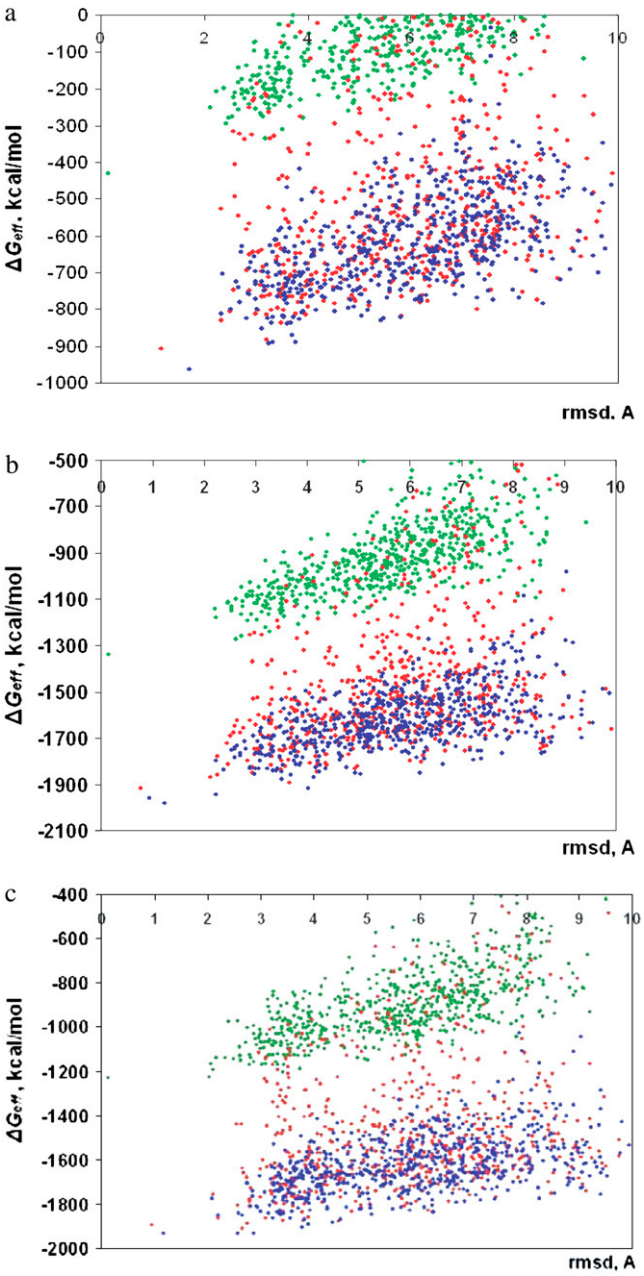

FIGURE 9 Scatter plot of the ECEPP05/SA energy (after parameter optimization) versus RMSD from the experimentally determined native structure for the proteins from the 4state-reduced set: (*a*) 1ctf, (*b*) 1r69, and (*c*) 2cro. Green, red, and blue circles correspond to results obtained from energy evaluation of fixed conformations, local energy minimization, and MCSA runs after local energy minimization, respectively.

The 4state-reduced set of decoys has been used extensively for evaluation of different all-atom physics-based scoring functions including a variety of force fields, solvent models, and scoring methods (i.e., energy evaluation, energy minimization, and short MD runs) (11–15). In practically all of these studies, the native structures scored as the most stable ones. The only cases in which some scoring functions experienced difficulty recognizing the native conformations

were 3icb and 4rxn (11,12), which is not surprising, since, as mentioned earlier, the native structures of these two proteins were solved with bound ligands, whereas ligands were not taken into consideration in either of these works (11,12).

Comparison of the results obtained for the 4state-reduced decoy set using the optimized ECEPP05/SA force field with those from other works (11–15) shows that the optimized force field performs well, even though it employs a very simple solvation model.

CONCLUSIONS

In this work, we described the development and implementation of a new parameter optimization method based on the use of protein decoys. The improved values of the parameters were obtained by creating free energy gaps between the sets of nativelike and nonnative decoys. The parameter optimization method has no restrictions on the number of optimized parameters and functional form of the force field. It can also be applied to large sets of proteins and decoys.

The new method was applied to optimize the backbone torsional and solvation parameters of the physics-based all-atom ECEPP05 force field coupled with a solvent-accessible SA model. The optimized ECEPP05/SA force field performs very well for the training proteins even after applying free energy relaxation (MCSA runs), which explores the vicinity of each decoy to locate additional low-energy conformations. Thus, the force field discriminated nativelike structures (1.5–4.0 Å RMSD) as those with the lowest energies for all six training proteins.

Tests carried out for the proteins not included in the training set showed that the optimized ECEPP05/SA force field is transferable to other proteins, e.g., it was able to identify nativelike structures as those with the lowest free energy for all six test proteins when local energy minimization was used as part of the scoring function. After the MCSA runs after energy minimization, nativelike conformations for five out of the six proteins emerged as the lowest in free energy. For the remaining protein (1fsd), competing lower-energy nonnative decoys appeared as a result of free energy relaxation. This failure to identify nativelike structures as the lowest free energy conformations for 1fsd can be a result of either deficiency of the force field or insufficient exploration of the native region during decoy generation and free energy relaxation. For example, the MCSA calculations implemented in this work may be less efficient in the case of near-native conformations that are more compact than nonnative decoys. To clarify this problem, we plan to consider decoy sets generated using different methods, as well as to modify the free energy relaxation procedure to enhance conformational sampling. Another possible explanation for why the optimized force field had difficulties recognizing the nativelike conformations of 1l2y, 1vii, and 1fsd may lie in the fact that all three proteins are small and highly flexible. Taking the flexibility of these proteins into account, as well as the small size of their hydrophobic cores, it is plausible to suggest that the relative contribution of different types of interactions and effects (for example, enthalpy versus entropy) to stabilize the native conformation of these proteins may differ from those in the other proteins considered in this work.

An independent test on the 4state-reduced decoy set of Park and Levitt (38) demonstrated that the optimized ECEPP05/SA force field is able to discriminate the native or near-native structures of the proteins from this set as those with the lowest free energy, and therefore performs in a manner comparable to the other (11–13,15) all-atom physics-based scoring functions.

In this work, we demonstrated that the decoy-based parameter optimization method represents a useful tool for development of accurate scoring functions applicable to proteins with different folds (α, β, or α/β). Thus, the optimized all-atom ECEPP05 force field coupled with a SA solvation model is capable of discriminating near-native from nonnative folds for numerous protein sets containing a very large number of decoy structures. It is worth noting that the good performance of the force field was achieved, first of all, by using a very simple, but computationally efficient, solvation model containing only a few parameters, and, second, without introducing any additional empirical or ad hoc terms or parameters.

The ability to discriminate nativelike structures from a large set of nonnative conformations is a necessary but not sufficient requirement for an accurate scoring function. Although the free energy relaxation used in this work represents a stricter test for a force field, it explores only the closest vicinity of a given decoy and is less efficient for larger proteins. An additional test, which will help to better assess the accuracy of the force field, and which we plan to carry out in the future, is to use the force field for folding of peptides and small proteins with different architecture. Since our ultimate goal is to obtain high-resolution protein models, we also plan to use the force field, optimized in this work, for refinement of low- and medium-resolution models produced by using the UNRES/MD (61) and other methods.

We thank Dr. C. Czaplewski and Dr. J.A. Vila for helpful discussions.

This work was supported by grants from the National Science Foundation (MCB05-41633) and the National Institutes of Health (GM-14312). This research was conducted using the resources of our 818-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology (Cornell University) and the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, Pittsburgh, Pennsylvania.

REFERENCES

1. Kryshtafovych, A., K. Fidelis, and J. Moult. 2007. Progress from CASP6 to CASP7. *Proteins*. 69:194–207.
2. Fan, H., and A. E. Mark. 2004. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci*. 13:211–220.

3. Misura, K., and D. Baker. 2005. Progress and challenges in high-resolution refinement of protein structure models. *Proteins.* 59:15–29.
4. Zhu, J., L. Xie, and B. Honig. 2006. Structural refinement of protein segments containing secondary structure elements: local sampling, knowledge-based potentials, and clustering. *Proteins.* 65:463–479.
5. Summa, C., and M. Levitt. 2007. Near-native structure refinement using in vacuo energy minimization. *Proc. Natl. Acad. Sci. USA.* 104: 3177–3182.
6. Verma, A., and W. Wenzel. 2007. Protein structure prediction by all-atom free-energy refinement. *BMC Struct. Biol.* 7:12.
7. Chen, J., and C. L. Brooks III. 2007. Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins.* 67:922–930.
8. Lu, H., and J. Skolnick. 2003. Application of statistical potentials to protein structure refinement from low resolution ab initio models. *Biopolymers.* 70:575–584.
9. Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science.* 181:223–230.
10. Lee, M. C., and Y. Duan. 2004. Distinguishing protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized Born solvent model. *Proteins.* 55:620–634.
11. Hsieh, M.-J., and R. Luo. 2004. Physical scoring function based on AMBER force field and Poisson-Boltzmann implicit solvent for protein structure prediction. *Proteins.* 56:475–486.
12. Lee, M. C., R. Yang, and Y. Duan. 2005. Comparison between generalizad-Born and Poisson-Boltzmann methods in physics-based scoring functions for protein structure prediction. *J. Mol. Model.* 12: 101–110.
13. Felts, A. K., E. Gallicchio, A. Wallqvist, and R. M. Levy. 2002. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized Born solvent model. *Proteins.* 48: 404–422.
14. Lazaridis, T., and M. Karplus. 1998. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* 288:477–487.
15. Dominy, B. N., and C. L. Brooks III. 2002. Identifying native-like protein structures using physics-based potentials. *J. Comput. Chem.* 23:147–160.
16. Hassan, S. A., and E. L. Mehler. 2002. A critical analysis of continuum electrostatics: the screened Coulomb potential-implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. *Proteins.* 47:45–61.
17. Zhu, J., Q. Zhu, Y. Shi, and H. Liu. 2003. How well can we predict native contacts in proteins based on decoy structures and their energies? *Proteins.* 52:598–608.
18. Wroblewska, L., and J. Skolnick. 2007. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. *J. Comput. Chem.* 28:2059–2066.
19. Moult, J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* 15:285–289.
20. Valencia, A. 2005. Protein refinement: a new challenge for CASP in its 10^{th} anniversary. *Bioinformatics.* 21:277.
21. Jarrold, M. F. 2007. Helices and sheets in vacuo. *Phys. Chem. Chem. Phys.* 9:1659–1671.
22. Jang, S., E. Kim, and Y. Pak. 2006. Free energy surfaces of miniproteins with a $\beta\beta\alpha$ motif: Replica exchange molecular dynamics simulation with an implicit solvation model. *Proteins.* 62:663–671.
23. Chen, J., W. Im, and C. L. Brooks III. 2006. Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field. *J. Am. Chem. Soc.* 128:3728–3736.
24. Mohanty, S., and U. H. E. Hansmann. 2007. Improving an all-atom force field. *Phys. Rev.* E76:012901.
25. Liwo, A., P. Arlukowicz, C. Czaplewski, S. Oldziej, J. Pillardy, and H. A. Scheraga. 2002. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application to the UNRES force field. *Proc. Natl. Acad. Sci. USA.* 99:1937–1942.
26. Fujitsuka, Y., S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes. 2004. Optimizing physical energy functions for protein folding. *Proteins.* 54:88–103.
27. Das, B., and H. Meirovitch. 2001. Optimization of solvation models for predicting the structure of surface loops in proteins. *Proteins.* 43:303–314.
28. Seok, C., J. B. Rosen, J. D. Chodera, and K. A. Dill. 2003. MOPED: Method for optimizing physical energy parameters using decoys. *J. Comput. Chem.* 24:89–97.
29. Okur, A., B. Strockbine, V. Hornak, and C. Simmerling. 2003. Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins. *J. Comput. Chem.* 24:21–31.
30. Herges, T., and W. Wenzel. 2004. An all-atom force field for tertiary structure prediction of helical proteins. *Biophys. J.* 87:3100–3109.
31. Szarecka, A., and H. Meirovitch. 2006. Optimization of the GB/SA solvation model for predicting the structure of surface loops in proteins. *J. Phys. Chem.* B110:2869–2880.
32. Wroblewska, L., A. Jagielska, and J. Skolnick. 2008. Development of a physics-based force field for the scoring and refinement of protein models. *Biophys J.* 94:3227–3240.
33. Arnautova, Y. A., A. Jagielska, and H. A. Scheraga. 2006. A new force field (ECEPP05) for peptides, proteins and organic molecules. *J. Phys. Chem.* B110:5025–5044.
34. Ooi, T., M. Oobatake, G. Nemethy, and H. A. Scheraga. 1987. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA.* 84:3086–3090.
35. Vila, J. A., D. R. Ripoll, and H. A. Scheraga. 2003. Atomically-detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc. Natl. Acad. Sci. USA.* 100:14812–14816.
36. Trebst, S., M. Troyer, and U. H. E. Hansmann. 2006. Optimized parallel tempering simulations of proteins. *J. Chem. Phys.* 124:174903.
37. Mohanty, S., and U. H. E. Hansmann. 2007. Folding of a miniprotein with mixed fold. *J. Chem. Phys.* 127:035102.
38. Park, B., and M. Levitt. 1996. Energy functions that discrimination x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* 258:367–392.
39. Vorobjev, Y. N., J. C. Almagro, and J. Hermans. 1998. 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.* 32:399–413.
40. Lee, M. R., Y. Duan, and P. A. Kollman. 2000. Use of MM-PB/SA in estimating the free energies of proteins: application to native, intermediates, and unfolded villin headpiece. *Proteins Struct. Funct. Genet.* 39:309–316.
41. Némethy, G., K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga. 1992. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to praline-containing peptides. *J. Phys. Chem.* 96:6472–6484.
42. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
43. Gay, D. M. 1983. Subroutines for unconstrained minimization using a model trust-region approach. *ACM Trans. Math. Softw.* 9:503–524.
44. Ripoll, D. R., and H. A. Scheraga. 1988. On the multiple-minima problem in the conformational-analysis of polypeptides. II. An electrostatically driven Monte Carlo method—tests on poly(L-alanine). *Biopolymers.* 27:1283–1303.
45. Ripoll, D. R., M. S. Pottle, K. D. Gibson, A. Liwo, and H. A. Scheraga. 1995. Implementation of the ECEPP algorithm, the Monte Carlo minimization method, and the electrostatically driven Monte Carlo

method on the Kendall Square research KSR1 computer. *J. Comput. Chem.* 16:1153–1163.

46. Ripoll, D. R., A. Liwo, and C. Czaplewski. 1999. ECEPP package for conformational analysis of polypeptides. *T.A.S.K Quart.* 3:313–331.
47. http://fconyx.ncifcrf.gov/~lukeb/simanf1.html.
48. Vila, J. A., D. R. Ripoll, Y. A. Arnautova, Y. Vorobjev, and H. A. Scheraga. 2005. Relevance of the charge distribution on the discrimination of native folds in proteins. *Proteins.* 61:56–68.
49. Kruskal, J. B., Jr. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* 7:48–50.
50. Rejto, P. A., and S. T. Freer. 1996. Protein conformational substates from x-ray crystallography. *Prog. Biophys. Mol. Biol.* 66:167–196.
51. Andrec, M., D. A. Snyder, Z. Zhou, J. Young, G. T. Montelione, and R. M. Levy. 2007. A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins.* 69: 449–465.
52. Juers, D. H., and B. W. Matthews. 2004. Cryo-cooling in macromolecular crystallography: advantages, disadvantages and optimization. *Q. Rev. Biophys.* 37:1105–1119.
53. Privalov, P. L. 1982. Stability of proteins—proteins which do not present a single cooperative system. *Adv. Protein Chem.* 35:1–104.
54. DePristo, M. A., P. I. W. de Bakker, and T. L. Blundell. 2004. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure.* 12:831–838.
55. Furnham, N., T. L. Blundell, M. A. DePristo, and T. C. Terwilliger. 2006. Is one solution good enough? *Nat. Struct. Mol. Biol.* 13:184–185.
56. Zimmerman, S. S., M. S. Pottle, G. Nemethy, and H. A. Scheraga. 1977. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. *Macromolecules.* 10:1–9.
57. Holm, L., and C. Sander. 1993. Protein-structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233:123–138.
58. Lei, H., C. Wu, and Y. Duan. 2007. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA.* 104:4925–4930.
59. Herges, T., and W. Wenzel. 2005. Free-energy landscape of the villin headpiece in an all-atom force field. *Structure.* 13:661–668.
60. Ripoll, D. R., J. A. Vila, and H. A. Scheraga. 2004. Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH. *J. Mol. Biol.* 339:915–925.
61. Liwo, A., M. Khalili, and H. A. Scheraga. 2005. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. USA.* 102:2362–2367.