Full Optimization of Linear Parameters of a United Residue Protein Potential

Julian Lee, Kibeom Park,† and Jooyoung Lee*

School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-012, Korea

Received: May 13, 2002; In Final Form: July 24, 2002

We apply the general protocol of parameter optimization (Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7291) to the UNRES potential. In contrast to earlier works where only the relative weights of various interaction terms were optimized, we optimize all linear parameters of the potential. The method exploits the high efficiency of the conformal space annealing method in finding distinct low-energy conformations. For a given training set of proteins, the parameters are modified to make the nativelike conformations energetically more favorable than the non-native ones. Linear approximation is used to estimate the energy change due to the parameter modification. The parameter change is followed by local energy reminimization and new conformational searches to find the energies of nativelike and non-native local minima of the energy function with new parameters. These steps are repeated until the potential predicts a nativelike conformation as one of the low-energy conformations for each protein in the training set. We consider a training set of crambin (PDB ID 1ejg), 1fsd, and the 10−55 residue fragment of staphylococcal protein A (PDB ID 1bdd). As the first check for the feasibility of our protocol, we optimize the parameters separately for these proteins and find an optimal set of parameters for each of them. Next, we apply the method simultaneously to these three proteins. By refining all linear parameters, we obtain an optimal set of parameters from which the nativelike conformations of the all three proteins are retrieved as the global minima, *without introducing additional multibody energy terms*.

I. Introduction

The prediction of the 3D structure of a protein solely from its amino acid sequence is one of the most challenging problems in computational science today. Popular approaches to this problem have included comparative modeling and fold recognition, which can be classified as knowledge-based methods.[1−10] These methods use statistical relationships between the sequences and the 3D structures of the proteins in the Protein Data Bank (PDB) in order to predict the unknown structure of a protein sequence without a deep understanding of the protein folding.

However, the ab initio method,[11−17] which is also called the energy-based or physics-based method, is based on the thermodynamic hypothesis that postulates that proteins adopt native structures that minimize their free energies[18]. Because it attempts to understand the fundamental principles of the protein folding itself, the success of this method will lead not only to successful structure prediction but also to the clarification of the protein-folding mechanism.

However, there have been two major obstacles to the successful application of energy-based methods to the protein-folding problem. First, the energy landscape of a protein is riddled with an astronomical number of local minima, making it difficult to search. Second, there are inherent inaccuracies in potential energy functions that attempt to describe the energetics of proteins. The first problem has been largely alleviated to some extent by recent developments of efficient search algorithms such as the conformational space annealing (CSA) method.[19−22] The second problem is the one that is addressed in this paper. The accuracy of a given potential energy function can be improved by modifying its functional form as well as its parameters. In this work, we will refine the parameters of the potential energy without changing the functional form.

Physics-based potentials are generally parametrized from quantum mechanical calculations and experimental data on model systems.[23] However, such calculations and data do not determine the parameters with perfect accuracy. The residual errors in potential energy functions may have significant effects on simulations of macromolecules such as proteins, where the total energy is the sum of a large number of interaction terms. Moreover, these terms are known to cancel each other to a high degree, making their systematic errors even more significant. Thus, it is crucial to refine the parameters of a potential energy function before it is applied to the protein-folding problem.

In fact, an iterative procedure that systematically refines the parameters of a given potential energy function was presented by Lee et al.[23] Because the CSA method can efficiently sample a wide range of the conformational space of a protein, the strategy is to apply this method to the proteins with known structures in order to refine the potential. We refine the parameters so that nativelike conformations of these proteins have lower energies than non-native conformations. The set of proteins used for the parameter refinement is called the training set. It would be desirable to include many proteins in the training set that belong to representative structural classes of proteins. However, it is quite a nontrivial issue to check whether this procedure is feasible, even for a small number of proteins in the training set.

* Corresponding author. E-mail: jlee@kias.re.kr.
† Present address: Electronics and Telecommunications Research Institute, Daejon 305-350, Korea.

10.1021/jp026108d CCC: $22.00
Published on Web 10/16/2002

The 10−55 fragment of staphylococcal protein A (1bdd) was used by Lee et al.[23] to refine a coarse-grained potential called the UNRES potential,[24−28] where each residue is approximated by two interaction sites. This potential was successfully applied to the prediction of the unknown structures of proteins in CASP3,[17,29,30] and its basic version consists of seven interaction terms. Lee et al.[23] optimized six relative weights of these interaction terms. Because only an α protein was used for the refinement, the resulting potential was suitable for α proteins.

However, Pillardy et al.[31] used three training sets consisting of one protein 1pou, one protein 1tpm, and two proteins betanova and 1bdd to optimize the potentials for predicting α, β, and α/β proteins, respectively.[42] This potential was extended to include 6 additional multibody terms, which increased the number of relative weights to be optimized from 6 to 12.[43] Therefore, in these works, the functional form of the UNRES potential as well as a part of the parameters, which are the twelve relative weights, was modified. The introduction of the six additional multibody terms was necessary in order to incorporate proteins with β strands.

However, one should note that each of the seven interaction terms in the original version of the UNRES potential contains its own parameters. Therefore, it is natural to ask whether one can optimize the potential energy function for the proteins with β strands by refining these parameters without introducing additional multibody energy terms. Of course, it might not be possible to optimize the potential for arbitrarily many proteins without introducing additional interaction terms. However, it is important to optimize the parameters as much as possible before introducing functional modifications because this will give us better insight into the limitations of a given potential and the types of additional interaction terms that are necessary for improvement.

Indeed, we observe that it is possible to refine the UNRES potential with three proteins 1bdd, 1fsd, and crambin (1ejg) *without* introducing additional multibody terms, where proteins 1fsd and 1ejg contain β strands. First, the parameters are optimized separately for these proteins, and an optimal parameter set for each of them is obtained. The potentials with optimized parameter sets yield global minimum energy conformations (GMECs) with root-mean-square deviations (RMSDs) of 1.7, 2.5, and 2.6 Å from the native structures for 1bdd, 1fsd, and 1ejg, respectively. Finally, the parameters are refined for the training set consisting of these three proteins, and a parameter set is obtained that correctly describe the energetics of these proteins simultaneously. The potential with the optimized parameter set yields GMECs with RMSDs of 1.8, 2.5, and 2.6 Å, respectively.

II. Methods

A. General Protocol. A brief description of our procedure is as follows. To check the performance of a potential energy function for a given set of parameters, one has to sample nativelike as well as non-native conformations. Non-native conformations can be obtained by an unrestricted conformational search that we call global CSA. Nativelike conformations are obtained by a restricted search that we call local CSA. In the local CSA, only the conformations whose RMSDs from the native structure are below a preset cutoff value are sampled.

Because a potential can be considered to describe the nature correctly if nativelike structures have lower energies than the non-native ones, the optimization criterion is given in terms of the energy gap, which is the difference between the lowest energy of the nativelike conformations and that of the non-native ones. We define the energy gap to be negative when the lowest energy of the native conformations is lower than that of non-native ones. We modify the parameters so that the energy gaps of the proteins in the training set decrease. The changes in the energy gaps are estimated by the linear approximation of the potential in terms of the parameters (See section D.). Because the positions of the local energy minima are shifted because of the parameter modification, it is necessary to reminimize their energies with the new parameters. We also search the conformational space with the newly obtained parameters to find new low-lying local energy minima. Together with the energy-reminimized conformations, these constitute a structural database that will be used for subsequent refinement of the parameters. We iterate these steps until all of the energy gaps become negative for proteins in the training set. A detailed explanation for each step of our procedure is given below.

B. Potential Energy Function. We use the UNRES force field,[24−28] where a polypeptide chain is represented by a sequence of α-carbon (C^α) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p) located in the middle between the consecutive C^αs. All of the virtual bond lengths are fixed: the $C^\alpha-C^\alpha$ distance is taken as 3.8 Å, and the $C^\alpha-SC$ distances are given for each amino acid type. The energy of the chain is given by

$$E = \sum_{i<j} U_{\mathrm{SCSC}}(i,j) + w_{\mathrm{SCp}} \sum_{i\neq j} U_{\mathrm{SCp}}(i,j) + w_{\mathrm{pp}} \sum_{i<j-1} U_{\mathrm{pp}}(i,j) + w_{\mathrm{b}} \sum_i U_{\mathrm{b}}(i) + w_{\mathrm{tor}} \sum_i U_{\mathrm{tor}}(i) + w_{\mathrm{rot}} \sum_i U_{\mathrm{rot}}(i) + w_{\mathrm{dis}} U_{\mathrm{dis}} + w^{(4)}_{\mathrm{el-loc}} \sum_{i<j} U^{(4)}_{\mathrm{el-loc}}(i,j) \quad (1)$$

where the w's are the relative weights that were refined in earlier works.[23,31,34] As described in detail in the Appendix, U_{SCSC}, U_{SCp}, U_{pp}, U_{tor}, and $U^{(4)}_{\mathrm{el-loc}}$ can be further decomposed into linear combinations of smaller parts, whose coefficients are refined in this work. Therefore, we may fix the values of w_{SCp}, w_{pp}, w_{tor}, and $w^{(4)}_{\mathrm{el-loc}}$ without loss of generality. We set them equal to unity for simplicity. Here, $U_{\mathrm{SCSC}}(i, j)$ represents the mean free energy of the hydrophobic (hydrophilic) interaction between the side chains of residues i and j, which is expressed by the Lennard-Jones potential $U_{\mathrm{SCp}}(i, j)$ that corresponds to the excluded-volume interaction between the side chain of residue i and the peptide group of residue j and the potential $U_{\mathrm{pp}}(i, j)$ that accounts for the electrostatic interaction between the peptide groups of residues i and j. The terms $U_{\mathrm{tor}}(i)$, $U_{\mathrm{b}}(i)$, and $U_{\mathrm{rot}}(i)$ denote the short-range interactions corresponding to the energies of virtual dihedral angle torsions, virtual angle bending, and side-chain rotamers, respectively. U_{dis} denotes the energy term that forces two cysteine residues to form a disulfide bridge. Finally, the four-body interaction term $U^{(4)}_{\mathrm{el-loc}}$ results from the cumulant expansion of the restricted free energy of the polypeptide chain. In contrast to the earlier works[31,34] where additional multibody terms were introduced,[35] $U^{(4)}_{\mathrm{el-loc}}$ is the only multibody term used in this work. Detailed forms of these terms are given in the Appendix. As discussed there, the total number of linear parameters that we adjust is 709. The functional form of eq 1, as well as the initial parameter set we use, is the one used in the CASP3 exercise.[29,30]

C. Global and Local CSA. In the protein-folding problem, the energy surface contains an astronomical number of local energy minima. The larger a protein is, the more likely it is

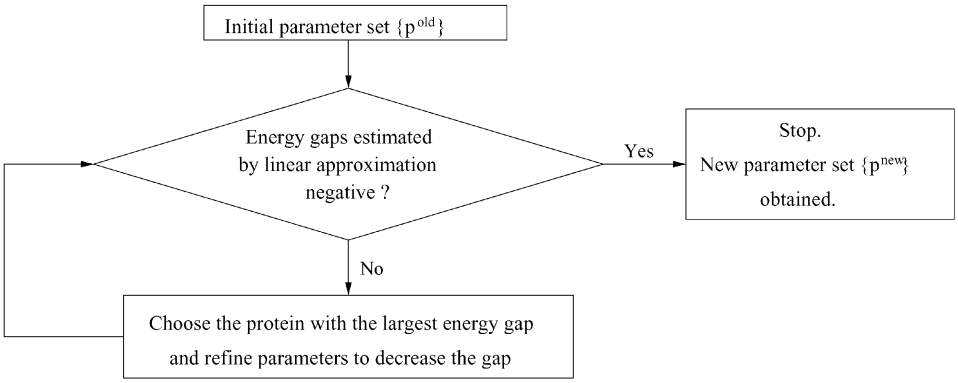

Figure 1. Flowchart for the part of the algorithm where the parameters are refined using a linear approximation for the energy gaps.

that there exist many local energy minima that correspond to very different structures. In general, it is not sufficient to consider only the lowest-energy conformation as a possible candidate for the native structure. Because the force-field parameters contain inevitable errors, one should take account of many distinct low-energy conformations. Therefore, it is necessary to search the whole conformational space.

It has been shown that this multiple minima problem can be overcome by an efficient search algorithm such as the CSA method. In this work, extensive conformational searches are carried out by global and local CSA methods.[19−22] The CSA method can be considered to be a genetic algorithm that enforces broad sampling in its early stages and gradually allows the conformational search to be focused into a narrow conformational space in its later stages. As a consequence, many low-energy local minima including the GMEC of the benchmark protein can be identified for a given parameter set. Unless the parameters are properly optimized, these conformations can be quite different from the native structure. Therefore, in this case, we may consider the global CSA to be the sampling of the non-native conformations. However, the nativelike conformations are sampled by the local CSA search[23]. The local CSA is the restricted search where only the conformations whose C^{α} RMSD values are within a fixed cutoff, R_c, of the native conformation are sampled. Also, to find these nativelike structures, the initial conformations are prepared with the native backbone coordinates, whose energy is subsequently minimized.[44] The value of R_c should be large enough to sample representative nativelike conformations and at the same time small enough to eliminate non-native conformations.

D. Linear Approximation and Parameter Refinement. Once the energies of the non-native and nativelike conformations for all proteins in the training set are obtained, the parameters are modified as follows. We select the protein with the largest energy gap and change the parameters so that this energy gap decreases. The parameters are changed by small amounts at each step so the energy with the new parameters can be estimated by the linear approximation

$$E^{\text{new}} \approx E^{\text{old}} + \sum_i (p_i^{\text{new}} - p_i^{old}) \frac{\partial E^{\text{old}}}{\partial p_i} \tag{2}$$

where the p_i^{old} and p_i^{new} terms represent the parameters before and after modification, respectively. The parameter dependence of the position of the local minimum can be neglected in the linear approximation because the derivative in the conformational space vanishes at a local minimum.[23] In general, the derivative $\partial E/\partial p_i$ is a function of the parameters, but for linear parameters, it is just a constant that is independent of the parameters. In this work, we adjust only the linear parameters for simplicity, the total number of them being 709 for the UNRES potential. The details can be found in the Appendix. Therefore, the energy function can be written as

$$E = \sum_i p_i e_i \tag{3}$$

where the e_i terms are the coefficients that are independent of p_i. The change in the energy gap is estimated as

$$\begin{aligned}\Delta E_{\text{gap}} &= E_{\text{gap}}(\{p_j^{\text{new}}\}) - E_{\text{gap}}(\{p_j^{\text{old}}\}) \\ &= (E^{(\text{lowest N})}(\{p_j^{\text{new}}\}) - E^{(\text{lowest NN})}(\{p_j^{\text{new}}\})) - \\ &\qquad (E^{(\text{lowest N})}(\{p_j^{\text{old}}\}) - E^{(\text{lowest NN})}(\{p_j^{\text{old}}\})) \\ &= \sum_i [e_i(\text{lowest N}) - [e_i(\text{lowest NN})](p_i^{\text{new}} - p_i^{\text{old}})\end{aligned} \tag{4}$$

where E and e are evaluated for the lowest-energy nativelike (N) and non-native (NN) conformations. We fix the magnitude of the parameter change $\delta p_i \equiv p_i^{\text{new}} - p_i^{\text{old}}$ to be a certain fraction a of p_i^{old}. We use $a = 0.01$ in this study. The sign of δp_i is chosen to decrease the energy gap:

$$\delta p_i = -a p_i^{\text{old}} \,\text{sign}[e_i(\text{lowest N}) - e_i(\text{lowest NN})] \tag{5}$$

We repeat this procedure of selecting the protein with the largest energy gap and modifying the parameters until all the energy gaps estimated by eq 4 become negative for proteins in the training set. The flowchart for this part of the algorithm is shown in Figure 1.

E. Reminimization and New Conformational Search. Because the procedure of the previous section was based on the linear approximation of eq 4 and the number of conformations in the structural database is limited, we now have to evaluate the true energy gap using the newly obtained parameters. The breakdown of the parameter refinement may come from two sources. First, the conformations corresponding to the local minima of the potential for the original set of parameters are no longer necessarily those for the new parameter set. For this reason, we reminimize the energy of these conformations with the new parameters. Second, the local minima obtained using the CSA method with the original parameter set are only a tiny fraction of the whole set of local minima. After we changed the parameters, some of the local minima that were not considered because of their relatively high energies can now have low energies for the new parameter set. It is even possible that entirely new low-energy local minima appear. Therefore,

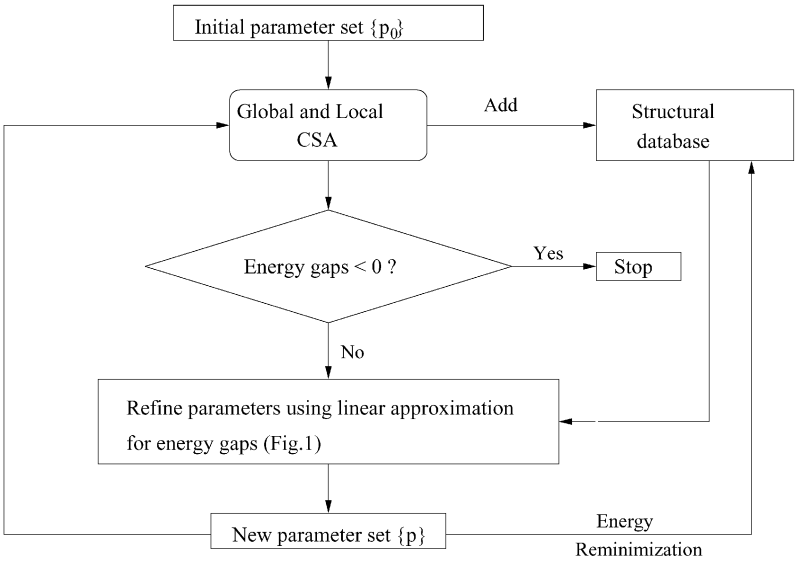

Figure 2. Flowchart for the whole protocol of the iterative parameter refinement. The parameter refinement step using a linear approximation for the energy gap corresponds to the flowchart of Figure 1.

these new minima are taken into account by performing subsequent CSA searches with the newly obtained parameter set.

F. Update of the Structural Database and Iterative Refinement of Parameters. The low-lying local energy minima found in the new conformational searches are added to the energy-reminimized conformations to form a structural database of local energy minima. The conformations in the database are used to obtain the energy gaps, and if their values are not satisfactory, these conformations are used for the new round of parameter refinement. As the procedure of (CSA → parameter refinement → energy reminimization) is repeated, the number of conformations in the structural database increases. As an example, the energy−RMSD plot of an energy-reminimized structural database for the protein 1ejg is shown in Figure 5c. The flowchart of the whole procedure is illustrated in Figure 2. This iterative procedure is continued until the energy gaps become negative for all proteins.

G. Choice of RMSD Cutoff. It is important to choose the RMSD cutoff judiciously for each protein in the training set in order to carry out the whole procedure efficiently. This cutoff is the criterion for distinguishing the nativelike and non-native conformations. The cutoff is necessary in two places in the procedure. It is used first in the local CSA where conformations with RMSDs below a preset cutoff value are sampled and second in the parameter refinement step where the conformations in the structural database are divided into nativelike and non-native families. In general, these two cutoff values can be different from each other. In addition, a separate cutoff value can be used for each iteration. In this work, we check the distribution of the RMSD versus the energy of the conformations by visual inspection and cluster them into nativelike and non-native families to determine the appropriate value of the RMSD cutoff for each protein. The values of the RMSD cutoffs used are given in Tables 1 and 2.

III. Results

We consider a training set consisting of three proteins. They are the 10−55 fragment of the B domain of staphylococcal protein A (1bdd), 1fsd, and crambin (1ejg), which are 46, 28, and 46 residues long, respectively. We first refine the parameters for these proteins separately to check whether our protocol for the iterative parameter refinement is feasible.

TABLE 1: Values of RMSD Cutoffs Used for Local CSA Searches and Parameter Refinements for the Case of Separate Optimizations (Å)[a]

iteration	1bdd	1fsd	1ejg
0	(no local search)[b]	3.0	3.0
0 → 1	3.0	3.5	4.0
1	2.5	3.0	2.5
1 → 2	2.2	2.8, 3.0[c]	3.0
2	2.5	3.0	2.5
2 → 3	2.2	2.8	2.5
3	2.0	3.0	2.5
3 → 4	2.0	2.7	2.5
4	2.0	3.0	2.5
4 → 5	1.8	2.6	2.5
5	2.0	3.0	2.4
5 → 6	1.8	2.8	2.5
6	2.0	3.0	(no local search)[b]
6 → 7	1.8	2.6	2.5
7	2.0	3.0	(no local search)[b]
7 → 8	1.8	2.6	2.5
8	2.0	3.0	(no local search)[b]
8 → 9	1.8	2.6	2.5
9	2.0	3.0	(no local search)[b]
9 → 10	1.8	2.6	2.6
10	2.0	3.0	2.5
10 → 11	1.8	2.6	2.6
11	2.0	3.0	(no local search)[b]
11 → 12		2.6	2.6
12		3.0	(no local search)[b]
12 → 13			2.6
13			2.5

[a] The integer i denotes the ith iteration of the CSA search, and $i \rightarrow i + 1$ denotes the parameter refinement step from the ith to $i + 1$ iteration. [b] The local CSA was not carried out because the global CSA was enough to find nativelike conformations. [c] These values of the RMSD cutoff are used sequentially during the parameter refinement.

A. Separate Parameter Refinement for Each Protein. This is the simplest case of having one protein in the training set. The first example is 1bdd, which was the target protein of the previous study of the weight optimization.[23] In that work, a negative energy gap was found after six iterations, and the GMEC that has a 2.2-Å RMSD deviation from the native structure was obtained. We start with the same initial parameters that were used in CASP3.[23,29,30]

TABLE 2: Values of the RMSD Cutoff Used for Local CSA Searches and Parameter Refinements for the Case of Simultaneous Optimizations (Å)

iteration	1bdd	1fsd	1ejg
0 → 1[a]	1.8	2.6	2.6
1	2.5	3.0	2.5
1 → 2	1.8	2.6	2.6
2	2.0	3.0	2.5
2 → 3	1.7, 1.8[b]	2.5, 2.6[b]	2.55, 2.6[b]
3	2.0	3.0	2.5
3 → 4	1.9, 1.8 [b]	2.6, 2.5, 2.6[b]	2.6, 2.5, 2.6[b]
4	2.0	3.0	2.5
4 → 5	1.8	2.55	2.6
5	2.0	3.0	2.5
5 → 6	1.9	2.6	2.6
6	2.0	3.0	2.5
6 → 7	1.9	2.6	2.6
7	2.0	3.0	3.0

[a] The initial conformational search is not necessary because we use the structural databases accumulated from the separate optimizations of three proteins. [b] These values of the RMSD cutoff are used sequentially during the parameter refinement.

In the CSA sampling with the original parameter set, the GMEC has a RMSD of 3.8 Å from the native structure,[15] as shown in Figure 3a. We set $R_c = 3.0$ Å, adjust the parameters according to it, and proceed to the next iteration. A negative energy gap is found after three iterations of parameter refinement with $R_c = 2.2$ Å, and the global CSA search yields the global minimum at 2.2 Å. Furthermore, from the local CSA run, we find conformations with RMSDs lower than 2.0 Å (Figure 3b). Therefore, we repeat the procedure with lower values of R_c. After 11 iterations, a GMEC with RMSD = 1.7 Å is obtained (Figure 3c). We have proceeded with an even lower value of R_c. However, the energy gap does not improve after the 11th iteration. Therefore, we take the result from the 11th iteration as the final optimized parameter set for this protein. The results are shown in Figure 3c. Similar procedures are repeated for 1fsd and 1ejg to obtain optimized parameter sets that yield GMECs with RMSD values of 2.5 and 2.6 Å, respectively. Details are shown in Tables 1 and 2 and Figures 3, 4, and 5.

B. Simultaneous Parameter Refinement for Three Proteins. Again, the initial parameter set is the one used in CASP3.[29,30] Because a large number of conformations were already accumulated in the structural databases during the separate parameter optimizations for three proteins, we use them to start the iterative procedure of simultaneous parameter refinement for the three proteins.

By choosing appropriate values of the RMSD cutoff for each iteration, we obtain an optimized parameter set after six iterations, yielding GMECs with RMSD values of 1.8, 2.5, and 2.6 Å for 1bdd, 1fsd, and 1ejg, respectively. The energies and RMSDs of the conformations obtained with the optimized parameters are plotted in Figure 6, and the C^α trace of the GMEC conformations are shown in Figure 7 along with the native conformations. The numerical values of the optimized parameters are provided in the Supporting Information. As an example of the changes of interaction terms due to the parameter optimization, the torsional energies between residues that are neither glycine nor proline, with the optimized and original parameter set, are plotted in Figure 8.

IV. Jackknife Test

It should be noted that the purpose of the present work is not to provide a potential that is transferable to all proteins but to develop a methodology for optimizing potential parameters

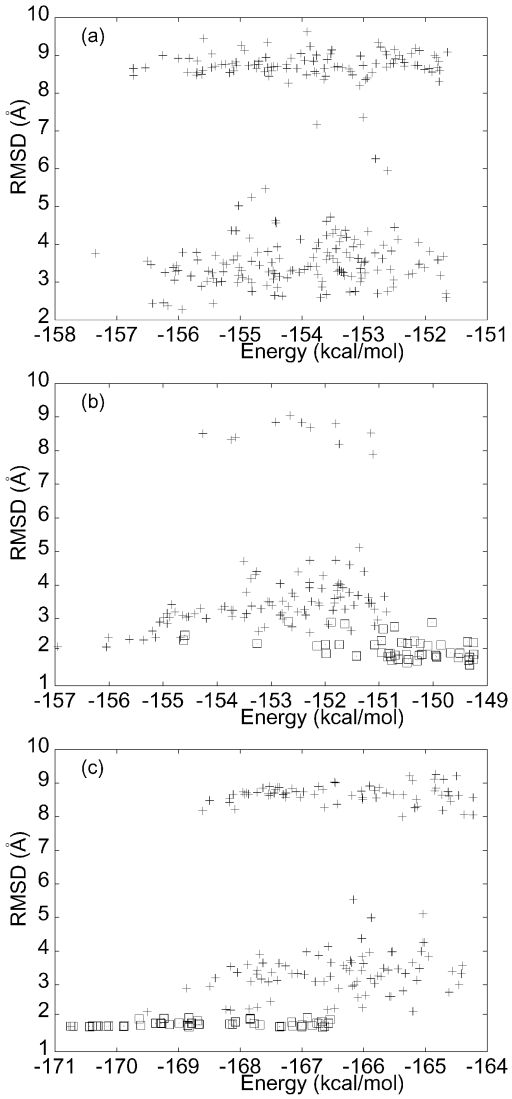

Figure 3. Plots of the UNRES energy and C^α RMSD (from the native structure) for conformations of 1bdd obtained from CSA searches. (a) Result with the initial parameter set. We observe that the GMEC has an RMSD of 3.8 Å and that there are nativelike conformations with RMSDs of less than 2.4 Å. (b) Result after three iterations. In all of the figures in this paper, the plus signs denote the conformations from the global CSA, and the squares denote those from the local CSA. We observe that the GMEC has an RMSD of 2.2 Å and that there are nativelike conformations with RMSDs of less than 2.0 Å. (c) Final result after 11 iterations. We observe that the GMEC has an RMSD of 1.7 Å.

of a *given* potential. Applying this method to develop a transferable potential is beyond the scope of this paper, and a much larger training set and even additional interaction terms might be necessary in order to achieve it. In fact, it is quite nontrivial to check whether such a procedure is possible at all. However, we performed conformational searches for proteins not included in the training set, which is called a jackknife test, and found some interesting features.

It should be noted that a mere comparison of low-energy conformations found from the optimized parameters with the native structure is not very meaningful. Rather, we should check if the low-energy conformation from the new parameters is closer to the native structure in comparison to those from the *original* parameters. We considered the 1−32 segment of the 36-residue protein 1bba. This protein contains a C-terminal α helix with an N-terminal extended strand parallel to the helix.

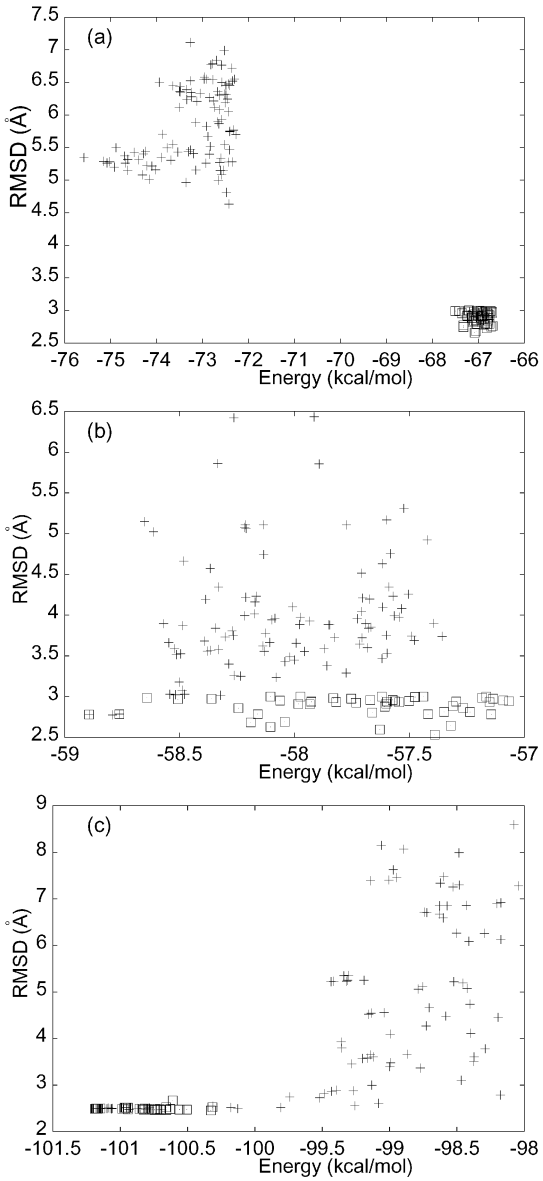

Figure 4. Plots of the UNRES energy and C^{α} RMSD (from the native structure) for conformations of 1fsd obtained from global and local CSA searches. (a) Result with the initial parameter set. We observe that the GMEC has an RMSD of 5.4 Å and that there are nativelike conformations with RMSDs of less than 2.8 Å. (b) Result after three iterations. We observe that the GMEC has an RMSD of 2.8 Å and that there are nativelike conformations with RMSDs of less than 2.6 Å. (c) Final result after 12 iterations. We observe that the GMEC has an RMSD of 2.5 Å.

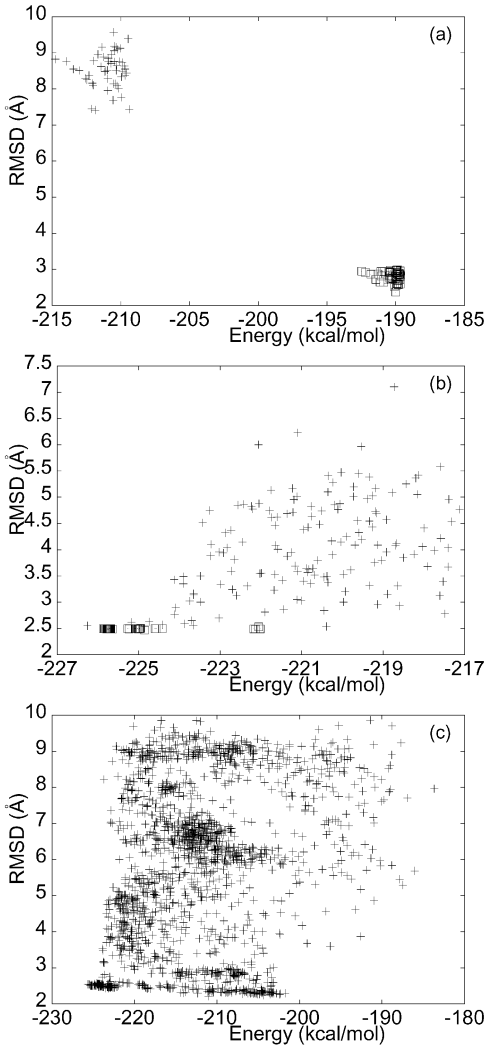

Figure 5. Plots of the UNRES energy and C^{α} RMSD (from the native structure) for conformations of 1ejg obtained from global and local CSA searches and the conformations in the structural database. (a) Result with the initial parameter set. We observe that the GMEC has an RMSD of 8.8 Å and that there are nativelike conformations with RMSDs of less than 2.5 Å. (b) Final result after 13 iterations. We observe that the GMEC has an RMSD of 2.6 Å. (c) Plot of the conformations in the structural bank whose energies are reminimized with the optimized parameter set.

The NMR structure of the protein is shown in Figure 9a. Using CSA, 200 low-energy conformations are sampled for both the original parameters and the optimized parameters, respectively. The lowest RMSD values are 6.2 and 5.8 Å for the original and optimized parameters, respectively, whereas the GMECs' RMSD values are 7.6 and 7.9 Å, respectively. Although the RMSD values are rather large, interesting qualitative differences in the secondary structures of the sampled conformations are observed, which is difficult to recognize from the RMSD values alone. The lowest RMSD conformation and GMEC for the optimized parameters are shown in Figure 9b and c, and those for the original parameters are shown in Figure 9d and e. We observe that, for conformations from the optimized parameters, the α helices extend to the end of the C-terminal, which is in good agreement with the native structure. However, the corresponding α helices are incomplete near the C-terminal for the conformations obtained with the original parameters. In addition, the extended strand at the N-terminal is reproduced better with the optimized parameters. We find that these are the prevailing features of all 400 conformations we have sampled.

We also performed jackknife tests on other proteins, whose results are not shown here. For some α/β proteins, notably 1L4V, the resulting conformations have similar qualitative features to those described above (i.e., the optimized parameters perform better in assigning secondary structures). However, the results for pure α or β proteins are not as conclusive.

V. Conclusions and Discussion

We applied the general protocol for the force-field parameter optimization of Lee et al.[23] to the UNRES potential used in CASP3.[29,30] We optimized the parameters separately for the 10−55 fragment of staphylococcal protein A (1bdd), 1fsd, and crambin (1ejg) and obtained an optimal parameter set for each of them, giving GMECs with RMSD values of 1.7, 2.5, and

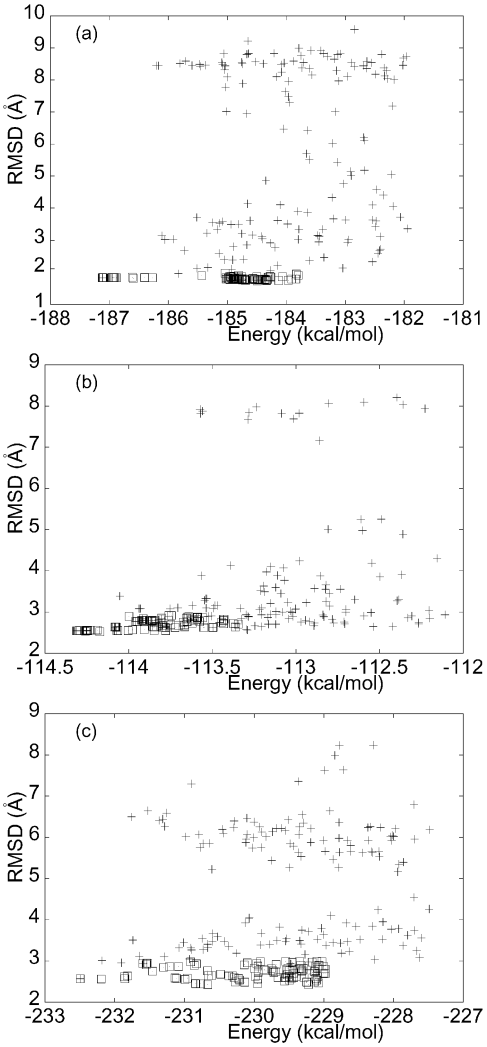

Figure 6. Plots of the UNRES energy and Cα RMSD (from the native structure) for conformations of the three proteins obtained from global and local CSA searches with the parameter set that is optimized for these proteins simultaneously. (a) Plot for 1bdd. We observe that the GMEC has an RMSD of 1.8 Å. (b) Plot for 1fsd. The GMEC has an RMSD of 2.5 Å. (c) Plot for 1ejg. The GMEC has an RMSD of 2.6 Å.

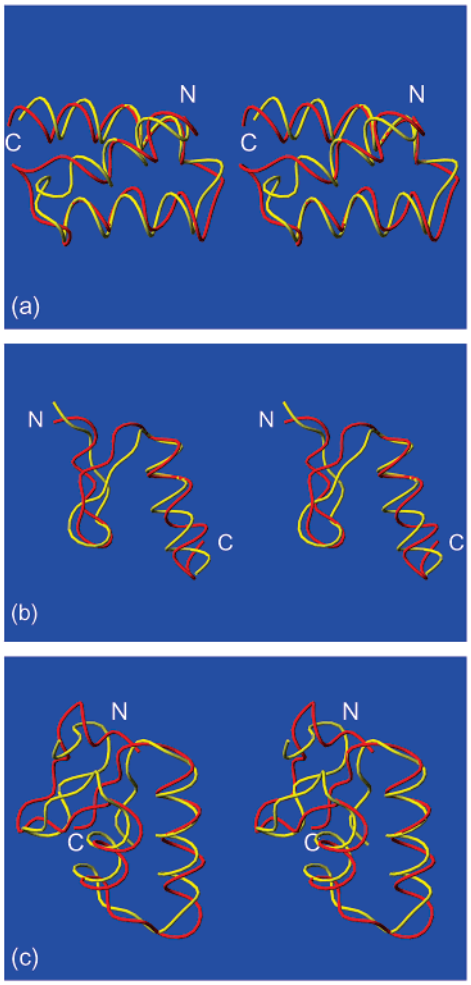

Figure 7. (a) Cα trace of 1bdd. The native structure is shown in red, and the GMEC found with the optimized parameters is shown in yellow. The RMSD is 1.8 Å. (b) Cα trace of 1fsd. The native structure is shown in red, and the GMEC found with the optimized parameters is shown in yellow. The RMSD is 2.5 Å. (c) Cα trace of 1ejg. The native structure is shown in red, and the GMEC found with the optimized parameters is shown in yellow. The RMSD is 2.6 Å.

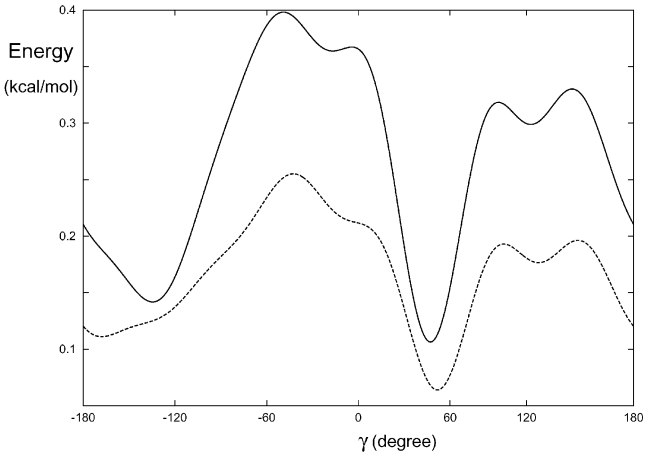

Figure 8. Torsional potential between the residues that are neither glycine nor proline as a function of the torsional angle γ. The solid (dashed) line is obtained with the optimized (original) parameters.

2.6 Å, respectively. We also obtained a parameter set that correctly describes the energetics of these three proteins simultaneously. This optimized parameter set yielded GMECs with RMSD values of 1.8, 2.5, and 2.6 Å for 1bdd, 1fsd, and 1ejg, respectively.

In contrast to the earlier works[23,31] where only the relative weights were optimized, we refined all 709 linear parameters of the UNRES potential. This enabled us to optimize the UNRES potential of eq 1 without introducing additional multibody energy terms. In particular, we demonstrated for the first time that the energetics of proteins containing β strands can be correctly described using the energy terms in eq 1 only.

It would be interesting to determine how many proteins can be energetically well-described using a given force field. This should provide a good measure of the efficacy of existing force fields. Once the parameters for a potential are successfully refined for the proteins in a given training set, we should perform a jackknife test on proteins not included in the training set. If this test is successful, we may confidently use this potential to predict the unknown structure of a given amino acid sequence.

Before tackling these more challenging problems, there are several points in our protocol that should be improved. First, in the step of parameter refinement, we decreased the largest among the energy gaps of the proteins without any restriction and repeated this procedure. However, this can become quite inefficient as the number of proteins in the training set increases.

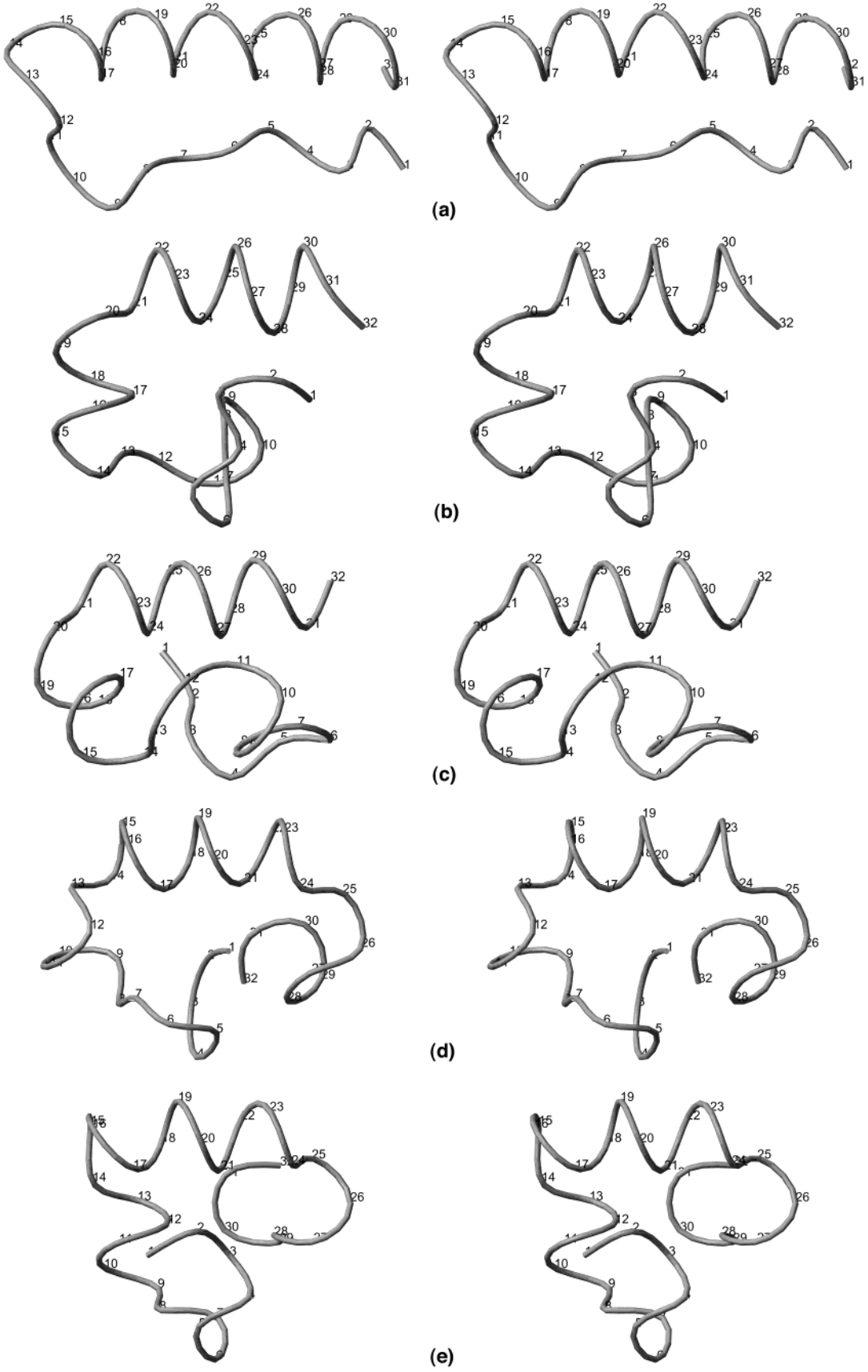

Figure 9. C^{α} trace of the 1−32 segment of the protein 1bba. (a) Native structure. Residues 15−32 form an α helix, and there is an extended strand consisting of residues 1−12. (b) Conformation with the lowest RMSD found with the optimized parameters. The RMSD is 5.8 Å. We observe that the α helix at the C-terminal is partially formed, consisting of residues 15−19 and 21−32. (c) GMEC with the optimized parameters. The RMSD is 7.9 Å. Again, the α helix is correctly formed except at residue 20. (d) Lowest RMSD conformation found with the original parameters. The RMSD is 6.2 Å. The position of the α helix is shifted to 9−12, 14−24, and 27−29. (e) GMEC with the original parameters, with an RMSD of 7.6 Å. Again, we find that α helices are formed at the wrong positions: 10−13 and 15−24.

We need to implement a constrained optimization where we require that the energy gaps of the other proteins in the training set do not increase while that of a given protein decreases. Second, the value of the RMSD cutoff at each iteration was determined from visual inspection. It would be better if one could devise a natural criterion for choosing RMSD cutoffs. Third, in principle, one can also refine nonlinear parameters, which was not carried out in this work. Finally, although we considered only the UNRES potential for parameter optimization in this work, it is straightforward to apply the procedure to other potentials such as ECEPP, AMBER, and CHARMM[36−41] with various solvation terms. All of these points are left for future study.

Acknowledgment. We thank Seung-Yeon Kim and Keehyoung Joo for useful discussions and help in fulfilling this work. This work was carried out on our own Linux PC cluster of 76 AMD processors.

Appendix A: UNRES Potential and Its Linear Parameters

The united residue (UNRES) potential is given by the expression[24−28]

$$E = \sum_{i<j} U_{\mathrm{SCSC}}(i,j) + w_{\mathrm{SCp}} \sum_{i\neq j} U_{\mathrm{SCp}}(i,j) + w_{\mathrm{pp}} \sum_{i<j-1} U_{\mathrm{pp}}(i,j) + w_{\mathrm{b}} \sum_{i} U_{\mathrm{b}}(i) + w_{\mathrm{tor}} \sum_{i} U_{\mathrm{tor}}(i) + w_{\mathrm{rot}} \sum_{i} U_{\mathrm{rot}}(i) + w_{\mathrm{dis}} U_{\mathrm{dis}} + w^{(4)}_{\mathrm{el-loc}} \sum_{i<j} U^{(4)}_{\mathrm{el-loc}}(i,j) \quad \text{(A1)}$$

1. Side-Chain Interactions. The interactions between the side chains are given by the Lennard-Jones-type potential:

$$U_{\mathrm{SCSC}}(i,j) = \frac{a_{\mathrm{SCSC}}(t_i,t_j)}{r_{ij}^{12}} + \frac{b_{\mathrm{SCSC}}(t_i,t_j)}{r_{ij}^{6}} \quad \text{(A2)}$$

$t_i = 1, \ldots, 20$ is the amino acid type of the ith residue. The linear parameters we optimize in this work are a_{SCSC} and b_{SCSC}, which comprise a total of $\frac{(20)(21)}{2} \times 2 = 420$ parameters.

2. Peptide−Peptide Interaction. The peptide−peptide interaction is decomposed into a Lennard-Jones-type interaction and an electrostatic interaction

$$U_{\mathrm{pp}}(i,j) = U_{\mathrm{LJ}}(i,j) + U_{\mathrm{es}}(i,j) \quad \text{(A3)}$$

with the Lennard-Jones-type interaction

$$U_{\mathrm{LJ}}(i,j) = \frac{a_{\mathrm{pp}}(I_i,I_j)}{r_{ij}^{12}} + \frac{b_{\mathrm{pp}}(I_i,I_j)}{r_{ij}^{6}} \quad \text{(A4)}$$

and the electrostatic interaction

$$U_{\mathrm{es}}(i,j) = \frac{a_{\mathrm{el}}(I_i,I_j)}{r_{ij}^{6}}[4 + (\cos\alpha_{ij} - 3\cos\beta_{ij}\cos\gamma_{ij})^2 - 3(\cos^2\beta_{ij} + \cos^2\gamma_{ij})] + \frac{b_{\mathrm{el}}(I_i,I_j)}{r_{ij}^{3}}(\cos\alpha_{ij} - 3\cos\beta_{ij}\cos\gamma_{ij}) \quad \text{(A5)}$$

where $\cos\alpha_{ij} = (n_i \cdot n_j)$, $\cos\beta_{ij} = (n_i \cdot r_{ji})$, $\cos\gamma_{ij} = (n_j \cdot r_{ji})$, and n_i is the vector along the ith peptide. The integer I_i denotes the type of peptide group i. There are only two types, proline and nonproline, with $I_i = 1, 2$. The term with $j = i + 1$ is not included in the summation of eq A1 because it can be absorbed into the local energy terms such as the bending energy. The linear parameters we adjust for this interaction are a_{pp}, b_{pp}, a_{el}, and b_{el}, which comprise a total of $4 \times \frac{(2)(3)}{2} = 12$ parameters.

3. Side-Chain Peptide Interaction. The side-chain peptide interaction is given by a Lennard-Jones-type interaction:

$$U_{\mathrm{SCp}}(i,j) = \frac{a_{\mathrm{SCp}}(I_i,t_j)}{r_{ij}^{12}} + \frac{b_{\mathrm{SCp}}(I_i,t_j)}{r_{ij}^{6}} \quad \text{(A6)}$$

Again, $j = i \pm 1$ is not included in the summation of eq A1 because it can be absorbed into the local rotamer energy. Therefore, the nearest neighbors that can contribute to this interaction are $j = i \pm 2$. If we use the same parameter values for these residues, they dominate this interaction, and we get unphysical results. Therefore, one usually uses smaller parameter values for these residues to avoid problems. This may seem to be an ad hoc treatment, but conceptually one may justify it by noting that for the residues close in sequence the quantum effect may become important, which modifies the classical interaction parameters. Therefore, we define additional peptide groups $I_i = 3, 4$ for nonproline and proline with $j = i \pm 2$. The linear parameters to be refined are a_{SCp} and b_{SCp}, which comprise a total of $(2)(20)(4) = 160$ parameters.

4. Disulfide Bridge Energy. To form disulfide bridges between cysteins, the following energy term is introduced:

$$U_{\mathrm{dis}} = w_{\mathrm{dis}} \sum_{i'} \frac{1}{2}(D(h_1(i'), h_2(i')) - D_0)^2 \quad \text{(A7)}$$

where i' and $h_1(i')$, $h_2(i')$ label the disulfide bridges and the residue numbers forming that bridge, respectively. The overall weight w_{dis} is the only linear parameter to be refined for this term.

5. Torsional Energy. The twist of the virtual bond between the $i - 2$ and $i - 1$ residues defines the torsion angle γ_i. Therefore, the torsion energy for γ_i depends on the amino acid types of these residues, which we denote by the integers J_{i-2} and J_{i-1}. There are three types of amino acid residues for this interaction—glycine, proline, and the rest—with $J_i = 1$, 3, and 2, respectively. For the torsion energy between two prolines, that is, when $J_{i-2} = J_{i-1} = 3$, we have

$$U_{\mathrm{tors}}(i) = \sum_{j=1}^{3} (v_1(j+1,3,3)\cos(j\gamma_i) + v_2(j+1,3,3)\sin(j\gamma_i) + |v_1(j+1,3,3)| + |v_2(j+1,3,3)|) + \begin{cases} v_1(1,3,3)\dfrac{1+\cos 3\gamma_i}{1-\cos 3\gamma_i} & \text{for } -\dfrac{\pi}{3} < \gamma_i < \pi \\ 0 & \text{for } -\pi \le \gamma_i \le -\dfrac{\pi}{3} \end{cases} \quad \text{(A8)}$$

and

$$U_{\text{tors}}(i) = \sum_{j=1}^{6}(v_1(j, J_{i-2}, J_{i-1}) \cos(j\gamma_i) + v_2(j, J_{i-2}, J_{i-1}) \sin(j\gamma_i) + |v_1(j, J_{i-2}, J_{i-1})| + |v_2(j, J_{i-2}, J_{i-1})|)$$

otherwise. The linear parameters to be refined are $v_1(j, k, l)$ and $v_2(j, k, l)$ with $j = 1, ..., 6$ and $k, l = 1, 2, 3$, which comprise a total of 108 parameters.

6. Local Side-Chain Energy. This energy is the negative log of a probability distribution, which is given by the sum of Gaussian peaks:

$$U_{\text{rot}}(i) = -\log\left[\sum_{j=1}^{n(t_i)} \exp\left(b_{j,t_i} - \frac{1}{2}\vec{z}_{j,i} G_{j,i} \vec{z}_{j,t_i}\right)\right]$$

where $\vec{z}_{j,i} \equiv \vec{x}_i - \vec{c}_{j,t_i}$, $n(t_i)$ is the number of Gaussian peaks in the distribution that is the function of the amino acid type of the ith residue t_i, and $\vec{x}_i = (\cot \theta_i, \alpha_i, \beta_i)$ for $k = 1, 2, 3$, with θ_i, α_i, and β_i being the bending angle and the polar angles of the side chain, respectively.[24−28] The values of the nonlinear parameters b_{j,t_i}, c_{j,t_i}, and $G_{j,i}$ are fixed in this work to those values used in CASP3,[17,29,30] so the only linear parameter to be refined for this term is its overall weight w_{rot}.

7. Bending Energy. The form is similar to the local side-chain energy except that there are two Gaussian peaks for all amino acids types. We have

$$U_{\text{b}}(i) = -\log\left[\exp\left(-\frac{(\theta_i - \theta_{\text{c}})^2}{2\sigma_{\text{c}}(\theta_{\text{c}})^2}\right) + k(\theta_{\text{c}}) \exp\left(-\frac{(\theta - \theta_0(t_i))^2}{\sigma_0(t_i)^2}\right)\right]$$

where

$$\theta_{\text{c}} = a_1(t_i) \cos(\gamma_i) + a_2(t_i) \sin(\gamma_i) + b_1(t_i) \cos(\gamma_{i+1}) + b_2(t_i) \sin(\gamma_{i+1})$$

$$\sigma_{\text{c}}(\theta_{\text{c}})^{-1} = 2(p_3(t_i)\theta_{\text{c}}^3 + p_2(t_i)\theta_{\text{c}}^2 + p_1(t_i)\theta_{\text{c}} + p_0(t_i))^2 + 2s_0(t_i)$$

$$k(\theta_{\text{c}}) = \exp\left(g_1(t_i) - \frac{(\theta_{\text{c}} - g_2(t_i))^2}{2g_3(t_i)^2}\right)$$

and again the values of the nonlinear parameters $a_j(t_i)$, $b_j(t_i)$, $p_j(t_i)$, $s_0(t_i)$, and $g_j(t_i)$ are fixed to those used in CASP3,[17,29,30] so the overall weight w_{b} is the only linear parameter to be optimized for this term.

8. Multibody Term. If the ith residue is in contact with the jth residue, then this term contributes if the $i + 1$ residue is also in contact with the $j + 1$ or $j - 1$ residue. In this case, the energy reads

$$U^{(4)}_{\text{el-loc}}(i, j) = -p_{u(i,j),u(i+1,k)} f_{i,j} f_{i+1,k}[C_+ E_{\text{es}(+)}(i, j) E_{\text{es}(+)}(i + 1, k) + C_- E_{\text{es}(-)}(j, i) E_{\text{es}(-)}(i + 1, k)] \quad \text{(A9)}$$

where $k = j + 1$ or $j - 1$. $f_{i,j}$ is the contact function that is 1 when the distance between the residues is less than a given cutoff, 0 when they are far away from the cutoff, and a smooth function in the intermediate region. $C_\pm$ are fixed numbers that are independent of the residue type, and

$$E_{\text{es}(\pm)}(i, j) = [4(1 + \cos \alpha_{ij}) \pm (\cos \alpha_{ij} - 3 \cos \beta_{ij} \cos \gamma_{ij})^2 - 3(\cos \beta_{ij} \pm \cos \gamma_{ij})^2)]^{1/2}$$

where α_{ij}, β_{ij}, and γ_{ij} are the same as in the peptide−peptide interaction. The integer $u(i, j) = 1, 2, 3$ when the (i, j) pair is (nonproline, nonproline), (nonproline, proline), and (proline, proline), respectively. We see that $p_{u(i,j),u(i+1,k)}$ comprises a total of six parameters because it is symmetric under the exchange of two indices.

9. Total Number of Linear Parameters. Therefore, the total number of linear parameters we adjust is

$$420 \text{ (SC−SC)} + 12 \text{ (p−p)} + 160 \text{ (SC−p)} + 108 \text{ (torsion)} + 6 \text{ (multibody)} + 1 \text{ (bending)} + 1 \text{ (local side chain)} + 1 \text{ (disulfide bridge)} = 709$$

Supporting Information Available: Numerical values of optimized parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Lewis, P. N.; Scheraga, H. A. *Arch. Biochem. Biophys.* **1971**, *144*, 584.
(2) Warme, P. K.; Momany, F. A.; Rumball, S. V.; Tuttle, R. W.; Scheraga, H. A. *Biochemistry* **1974**, *13*, 768.
(3) Jones, T. A.; Thirup, S. *EMBO J.* **1986**, *5*, 819.
(4) Clark, D. A.; Shirazi, J.; Rawlings, C. J. *Protein Eng.* **1991**, *4*, 751.
(5) Rooman, M. J.; Wodak, S. J. *Biochemistry* **1992**, *31*, 10239.
(6) Johnson, M. S.; Overington, J. P.; Blundell, T. L. *J. Mol. Biol.* **1993**, *231*, 735.
(7) Sipple, M. J. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 473.
(8) Fischer, D.; Rice, D.; Bowie, J. U.; Eisenberg, D. *FASEB J.* **1996**, *10*, 126.
(9) Skolnick, J.; Kolínski, A.; Ortiz, A. R. *J. Biol.* **1997**, *265*, 217.
(10) Bystroff, C.; Baker, D. *J. Mol. Biol.* **1998**, *281*, 565.
(11) Scheraga, H. A. *Int. J. Quantum Chem.* **1992**, *42*, 1529.
(12) Scheraga, H. A. *Biophys. Chem.* **1996**, *59*, 329.
(13) Vásquez, M.; Némethy, G.; Scheraga, H. A. *Chem. Rev.* **1994**, *94*, 2183.
(14) Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proteins: Struct., Funct., Genet.* **1999**, *suppl. 3*, 204.
(15) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025.
(16) Liwo, A.; Lee, J.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5482.
(17) Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Saunders, J. A.; Gibson, K. D.; Scheraga, H. A. *Int. J. Quantum Chem.* **2000**, *77*, 90.
(18) Anfinsen, C. B. *Science (Washington, D.C.)* **1973**, *181*, 223.
(19) Lee, J.; Scheraga, H. A.; Rackovsky, S. *J. Comput. Chem.* **1997**, *18*, 1222.
(20) Lee, J.; Scheraga, H. A.; Rackovsky, S. *Biopolymers* **1998**, *46*, 103.
(21) Lee, J.; Scheraga, H. A.; Rackovsky, S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025.
(22) Lee, J.; Liwo, A.; Scheraga, H. A. *Int. J. Quantum Chem.* **1999**, *75*, 255.
(23) Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7291.
(24) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1697.
(25) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1715.
(26) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849.

(27) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874.

(28) Liwo, A.; Kaźmierkiewicz, R.; Czaplewski, C.; Groth, M.; Oldziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259.

(29) Lee, J.; Liwo, A.; Ripoll, D.; Pillardy, J.; Scheraga, H. A. *Proteins: Struct., Funct., Genet.* **1999**, *suppl. 3*, 204.

(30) *Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction;* Asilomar Conference Center, Asilomar, U.S.A., December 13−17, 1998; http://predictioncenter.llnl.gov/casp3/Casp3.html.

(31) Pillardy, J.; Czaplewski, C.; Liwo, A.; Wedemeyer, W. J.; Lee, J.; Ripoll, D.; Arlukowicz, P.; Oldziej, S.; Arnautova, Y. A.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7299.

(32) Pillardy, J.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D.; Kaźmierkiewicz, R.; Oldziej, S.; Wedemeyer, W. J.; Gibson, K. D.; Arnautova, Y. A.; Saunders, J.; Ye, Y.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2329.

(33) *Fourth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction;* Asilomar Conference Center, Asilomar, U.S.A., December 3−7, 2000; http://predictioncenter.llnl.gov/casp4/Casp4.html.

(34) Liwo, A.; Arlukowicz, P.; Czaplewski, C.; Oldziej, S.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1937.

(35) Liwo, A.; Czaplewski, C.; Pillardy, J.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Chem. Phys.* **2001**, *115*, 2323.

(36) Momany, F. A.; Mcguire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* **1975**, *79*, 2361.

(37) Némethyi, G.; Pottle, M. S.; Scheraga, H. A. *J. Phys. Chem.* **1983**, *87*, 1883.

(38) Sipple, M. J.; Némethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1984**, *88*, 6231.

(39) Némethyi, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. *J. Phys. Chem.* **1992**, *96*, 6472.

(40) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1.

(41) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

(42) More recently, Liwo et al.[34] used two training sets consisting of one protein 1fsd and one protein 1igd to optimize the UNRES potential, with a different optimization criterion.

(43) This form of the UNRES potential was used in CASP4[32,33]

(44) To be precise, these initial conformations, which we call first bank,[19−22] are allowed to have RMSDs greater than R_c. However, only new conformations with RMSDs less than R_c are allowed to update the bank.