Design of a Protein Potential Energy Landscape by Parameter Optimization

Julian Lee,[†,‡,§] Seung-Yeon Kim,[§] and Jooyoung Lee*[,§]

Department of Bioinformatics and Life Sciences and Bioinformatics and Molecular Design Technology Innovation Center, Soongsil University, Seoul 156-743, Korea, and School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea

Received: October 10, 2003; In Final Form: December 10, 2003

We propose an automated protocol for designing the energy landscape of a protein energy function by optimizing its parameters. The parameters are optimized so that not only the global minimum-energy conformation becomes nativelike but also the conformations distinct from the native structure have higher energies than those close to the native structure. We classify low-energy conformations into three groups: supernative, nativelike, and non-native. The supernative conformations have all backbone dihedral angles fixed to their native values, and only their side chains are minimized with respect to energy. However, the nativelike and non-native conformations all correspond to the local minima of the energy function. These conformations are ranked according to their root-mean-square deviation (rmsd) of backbone coordinates from the native structure, and a fixed number of conformations with the smallest rmsd values are selected as nativelike conformations, whereas the rest are considered to be non-native conformations. We define two energy gaps $E_{\mathrm{gap}}^{(1)}$ and $E_{\mathrm{gap}}^{(2)}$. The energy gap $E_{\mathrm{gap}}^{(1)}$ ($E_{\mathrm{gap}}^{(2)}$) is the energy difference between the lowest energy of the non-native conformations and the highest energy of the nativelike (supernative) conformations. The parameters are modified to decrease both $E_{\mathrm{gap}}^{(1)}$ and $E_{\mathrm{gap}}^{(2)}$. In addition, the non-native conformations with larger rmsd values are made to have higher energies relative to those with smaller rmsd values. We successfully apply our protocol to the parameter optimization of the UNRES potential energy using the training set of betanova, 1fsd, the 36-residue subdomain of chicken villin headpiece (PDB ID 1vii), and the 10−55 residue fragment of staphylococcal protein A (PDB ID 1bdd). The new protocol of the parameter optimization shows better performance than earlier methods where only the difference between the lowest energies of nativelike and non-native conformations was adjusted without considering various nativelike degrees of the conformations. We also perform jackknife tests on other proteins not included in the training set and obtain promising results. The results suggest that the parameters we obtained using the training set of the four proteins are transferable to other proteins to some extent.

I. Introduction

The prediction of the 3D structure and the folding pathway of a protein solely from its amino acid sequence is one of the most challenging problems in biophysical chemistry. There are two major approaches to the protein structure prediction: so-called knowledge-based methods and energy-based methods. The knowledge-based methods,[1−4] which include comparative modeling and fold recognition, use the statistical relationship between the sequences and their 3D structures in the Protein Data Bank (PDB), without a deep understanding of the interactions governing the protein folding. Therefore, although these methods can be very powerful for predicting the structure of a protein sequence that has a certain degree of similarity to those in the PDB, they cannot provide a fundamental understanding of the protein-folding mechanism.

However, the energy-based methods,[5−11] which are also called the physics-based methods, are based on the thermodynamic hypothesis that proteins adopt native structures that minimize their free energies.[12] Understanding the fundamental principles of protein folding by these methods will lead not only to a successful structure prediction, especially for proteins having no similar sequences in the PDB, but also to a clarification of the protein-folding mechanism.

However, there have been several major obstacles to the successful application of energy-based methods to the protein-folding problem. First, there are inherent inaccuracies in potential energy functions that describe the energetics of proteins. Second, even if the global minimum-energy conformation is nativelike, this does not guarantee that a protein will fold into its native structure on a reasonable timescale unless the energy landscape is properly designed, as summarized in the Levinthal paradox.

Physics-based potentials are generally parametrized from quantum mechanical calculations and experimental data on model systems. 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 can be successfully applied to the protein-folding problem.

[*] Corresponding author. E-mail: jlee@kias.re.kr.
[†] Department of Bioinformatics and Life Sciences, Soongsil University.
[‡] Bioinformatics and Molecular Design Technology Innovation Center, Soongsil University.
[§] Korea Institute for Advanced Study.

10.1021/jp037076c CCC: $27.50 © 2004 American Chemical Society
Published on Web 03/17/2004

An iterative procedure that systematically refines the parameters of a given potential energy function was recently proposed[13] and successfully applied to the parameter optimization[13−16] of an UNRES potential energy.[17−19] The method exploits the high efficiency of the conformational space annealing (CSA) method[20−24] in finding distinct low-energy conformations. For a given set of proteins whose low-lying local minimum-energy conformations for a given energy function is found by the CSA method, one modifies the parameter set so that nativelike conformations of these proteins have lower energies than non-native ones. The method consists of the following steps:

(1) Low-lying local minimum-energy conformations are searched with no constraints, which is called the global CSA search. For many proteins, the conformations resulting from the global CSA are non-native conformations for parameters that are not yet optimized.

(2) Nativelike conformations are searched by the local CSA search, where low-lying local minimum-energy conformations, among those whose root-mean-square deviation (rmsd) of the backbone C^{α} coordinates from the native structure is below a given cutoff value $R_{\text{cut}}^{(1)}$, are sampled.

(3) The nativelike and non-native conformations from steps 1 and 2 are added to the structural database of each protein.

(4) Among the conformations in the structural database, those with rmsd values below a given cutoff value $R_{\text{cut}}^{(2)}$ are defined as nativelike conformations, whereas the rest are defined as non-native ones. The parameters are optimized in such a way as to minimize the energy gaps

$$E_{\text{gap}} = E_{\text{min}}^{\text{N}} - E_{\text{min}}^{\text{NN}} \tag{1}$$

for *all* proteins in the training set, where $E_{\text{min}}^{\text{N}}$ ($E_{\text{min}}^{\text{NN}}$) is the minimum energy among the energies of the nativelike (non-native) conformations in the structural database.

(5) After the parameters are modified in step 4, the conformations in the structural database are not local minimum-energy conformations anymore. Therefore, it is necessary to reminimize these conformations using the potential energy with the new parameters.

(6) In general, with the new parameters, there may exist many additional low-lying local minima of the potential energy, which are absent in the structural database. Therefore, it is necessary to go back to step 1 and perform CSA searches with the new parameters. These steps are repeated until the performance of the optimized parameters is satisfactory (i.e., the global CSA search finds nativelike conformations with reasonably small values of rmsd from their corresponding native structures). Because the size of the structural database of local minimum-energy conformations grows after each iteration, the efficiency of the parameter optimization increases as the algorithm proceeds.

It would be desirable to include many proteins in the training set that represent many structural classes of proteins. The optimization method was successfully applied to the parameter optimization of the UNRES potential for a training set consisting of three proteins of structural classes α and α/β [15] without introducing additional multibody terms.[11,14,16,25] However, it was still difficult to optimize the parameters of the UNRES potential for a training set containing β proteins.

In this work, we propose a new protocol where the parameters are modified so as to make conformations with larger rmsd values have higher energy values relative to those with smaller rmsd values. This goal is achieved by using the following modified energy,

$$E_{\text{modified}} = E + 0.3 \text{ rmsd} \tag{2}$$

when calculating the energy gaps. The numerical value of 0.3 is chosen empirically as an optimal value by inspecting the performance of the parameter-optimization protocol for one protein—betanova—although performances with other values such as 0.1 or 0.5 are also reasonably good for this protein.

The parameters of an UNRES energy function were also successfully optimized by Liwo et al.[16] using energy gaps that depend on the nativelikeness of the conformations. However, whereas these authors used the secondary structure contents for the criterion of nativelikeness of a conformation, we use rmsd. The idea of correlating rmsd to energy was also used in optimizing a contact potential[26] for the study of fold recognition, where the distance rmsd was used instead of the coordinate-based rmsd used in this work.

The new method is more natural than previous methods,[13−15] where non-native conformations were treated equally regardless of their rmsd values. It also turns out that the new method is much more efficient than the previous ones and allows us now to optimize the parameters for a training set containing a β protein. Additional new features are introduced in the current method to overcome several major drawbacks of the previous methods, as discussed below.

First, in previous methods,[13−15] arbitrarily chosen values of rmsd cutoffs $R_{\text{cut}}^{(1)}$ and $R_{\text{cut}}^{(2)}$ were used as the criteria for separating nativelike conformations from non-native ones, which were set at each iteration by inspecting the distribution of rmsd values of conformations. This rather arbitrary procedure made it difficult to automate the optimization procedure. Moreover, for some proteins, the value of $R_{\text{cut}}^{(1)}$ had to be taken as a large number to have a nonzero number of nativelike conformations, in which case the local CSA search is not meaningful. This can happen for a protein where the initial parameter set is so bad that there exist no local minimum-energy conformations that are nativelike. This problem is solved in the current method by introducing what we call the supernative conformations, whose backbone angles are fixed to the values of the native structure and only side-chain angles are minimized with respect to the energy. In the current method, the local CSA search is defined as the restricted search for supernative conformations in the space of the side-chain angles. Because the C^{α} rmsd values for the supernative conformations are zero by definition, an arbitrary cutoff value $R_{\text{cut}}^{(1)}$ is no longer necessary. Also, the supernative conformations can be found for any parameter set. Although supernative conformations are unstable with respect to energy, minimizing the energy gap between their highest energy and the lowest energy of non-native conformations stabilizes their energies. Therefore, because of the reminimization procedure with new optimized parameters, the supernative conformations would furnish low-lying local minima with small rmsd values that accumulate as the iteration proceeds. This makes the current method more efficient than the earlier methods, where it was difficult to optimize the parameters unless local minimum-energy conformations with small values of rmsd exist with the initial parameters. In addition to the supernative conformations, we define nativelike conformations as the 50 conformations with the smallest rmsd values in the structural database. Although 50 is an arbitrary number, it can be kept as a fixed number, and again the cutoff value $R_{\text{cut}}^{(2)}$ is set automatically. Also, because the size of the structural database grows after each iteration of the parameter optimization, the total number of local minimum-energy conformations becomes as large as several thousand, and the 50 nativelike conformations

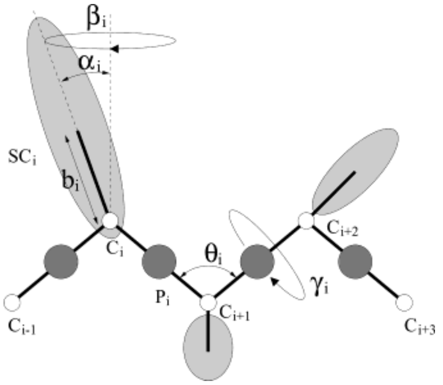


Figure 1. United-residue representation of a protein. The interaction sites are side-chain ellipsoids of different sizes (SC) and peptide-bond centers (p) indicated by shaded circles, whereas the C^α atoms (small empty circles) are introduced to define the backbone−local interaction sites and to assist in defining the geometry. The virtual $C^\alpha-C^a$ bonds have a fixed length of 3.8 Å, corresponding to a trans peptide group; the virtual bond (θ) and dihedral (γ) angles are variable. Each side chain is attached to the corresponding α-carbon with a different but fixed bond length, b_i, and variable bond angle, α_i, formed by SC_i and the bisector of the angle defined by C^α_{i-1}, C^α_i, and C^α_{i+1} and with a variable dihedral angle β_i of counterclockwise rotation about the bisector, starting from the right side of the C^α_{i-1}, C^α_i, C^α_{i+1} frame.

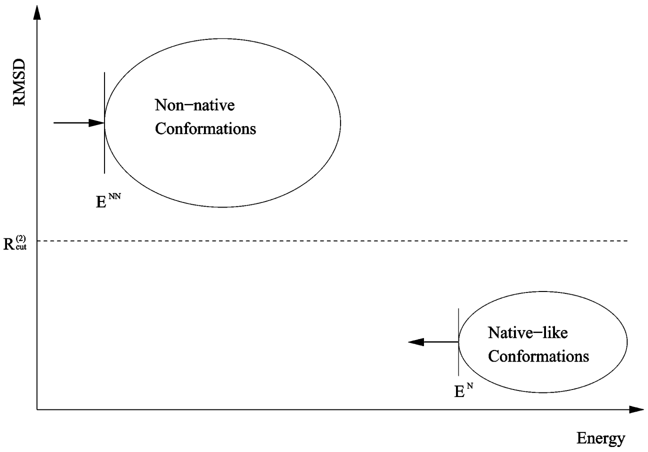


Figure 2. Schematic of the old method in terms of the energy and rmsd. The conformations in the structural database are divided into nativelike and non-native conformations with an arbitrary rmsd cutoff $R^{(2)}_{\text{cut}}$. The minimum energies of these two families define the energy gap. (See the text.) The arrows indicate the direction of the optimization.

comprise only about 1% of them at the final stage of the parameter optimization. Therefore, we expect that the performance of our procedure is not sensitive to the precise number of nativelike conformations. As mentioned above, the low-lying local minima with small rmsd values can be provided from the reminimization procedure of the supernative conformations. Generally, the rmsd values of these nativelike conformations become smaller as the iteration of the parameter optimization continues. Both supernative and nativelike conformations are used to calculate the energy gaps.

Second, we introduce linear programming to perform the parameter optimization systematically on the basis of linear approximation. This allows one to optimize the parameters so that the energy gap of a protein is minimized while imposing the constraint that the other energy gaps, including those of the other proteins, do not increase if they are positive and do not become positive if they are negative. This is in contrast to the optimization method of earlier work, where the protein with the largest energy gap was selected in turn, and the energy gap of that protein was reduced without imposing any constraint on the energy gaps of the other proteins in the set. Because linear programming has the effect of simultaneously decreasing the energy gaps of all of the proteins in the training set, it is especially powerful when there are many proteins in the training set.

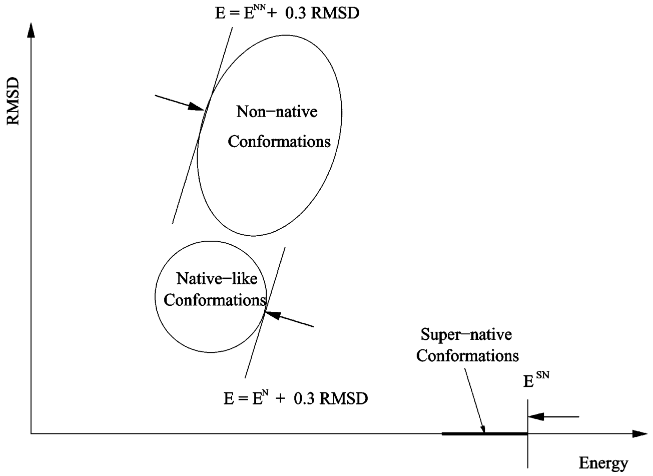


Figure 3. Schematic of the new method in terms of the energy and rmsd. The energy along the vertical axis is the one without the 0.3 rmsd term. Among the conformations with nonzero rmsd's, 50 conformations with the smallest rmsd values are selected as the nativelike conformations, and the rest are considered to be the non-native conformations. Supernative conformations are those with zero rmsd. The supernative conformations furnish the candidates for the low-lying nativelike local minima after the reminimization procedure with new optimized parameters. The lowest modified energy of the non-native family and the highest modified energies of nativelike and supernative families define the energy gaps. (See the text.) The arrows indicate the direction of the optimization.

TABLE 1: rmsd Values of the GMECs Found from the Global CSA Search Using the Initial Parameters along with the Optimized Parameters after the 28th and 40th Iterations (Å)[a]

protein	0th	28th	40th
betanova (β: 20 aa)	6.6 (5.1)	4.1 (1.6)	1.5 (1.5)
1fsd (α/β: 28 aa)	5.6 (3.6)	1.9 (1.7)	1.7 (1.3)
1vii (α: 36 aa)	6.3 (4.9)	2.7 (1.6)	1.7 (1.2)
1bdd (α: 46 aa)	9.6 (4.0)	3.1 (1.6)	1.9 (1.7)
1bbg (α/β: 40 aa)	8.7 (6.3)	7.9 (5.3)	7.3 (5.9)
1ccn (α/β: 46 aa)	7.7 (6.4)	9.5 (7.0)	6.5 (6.0)
1hnr (α/β: 47 aa)	9.9 (5.1)	9.7 (5.9)	9.2 (5.2)
1kbs (β: 60 aa)	11.2 (9.5)	11.3 (10.0)	10.1 (7.6)
1neb (β: 60 aa)	10.9 (9.3)	11.3 (8.8)	9.6 (9.1)
1bba (α: 36 aa)	8.9 (8.1)	8.1 (6.8)	12.0 (10.7)
1idy (α: 54 aa)	11.9 (6.6)	11.6 (7.4)	7.5 (6.2)
1prb (α: 53 aa)	10.2 (7.0)	11.1 (5.4)	7.1 (5.1)
1pru (α: 56 aa)	8.4 (7.1)	11.3 (6.4)	8.4 (7.6)
1zdb (α: 38 aa)	7.7 (6.3)	7.6 (4.9)	5.0 (4.5)

[a] The numbers inside the parentheses are the smallest values of rmsd found. The structural class and the chain length of each protein are also shown inside the parentheses next to the protein name. The energies are not displayed because their numerical values have no physical meaning because of the fact that the overall scale of the linear parameters is not fixed in our protocol.

Third, we use the highest energies of the nativelike and supernative conformations to calculate the energy gaps. This is in contrast to earlier work in which the lowest energy of the nativelike conformations was used instead. The old procedures may result in lowering only the nativelike conformation with the lowest energy. The current procedure lowers the energies of all nativelike conformations and is more efficient.

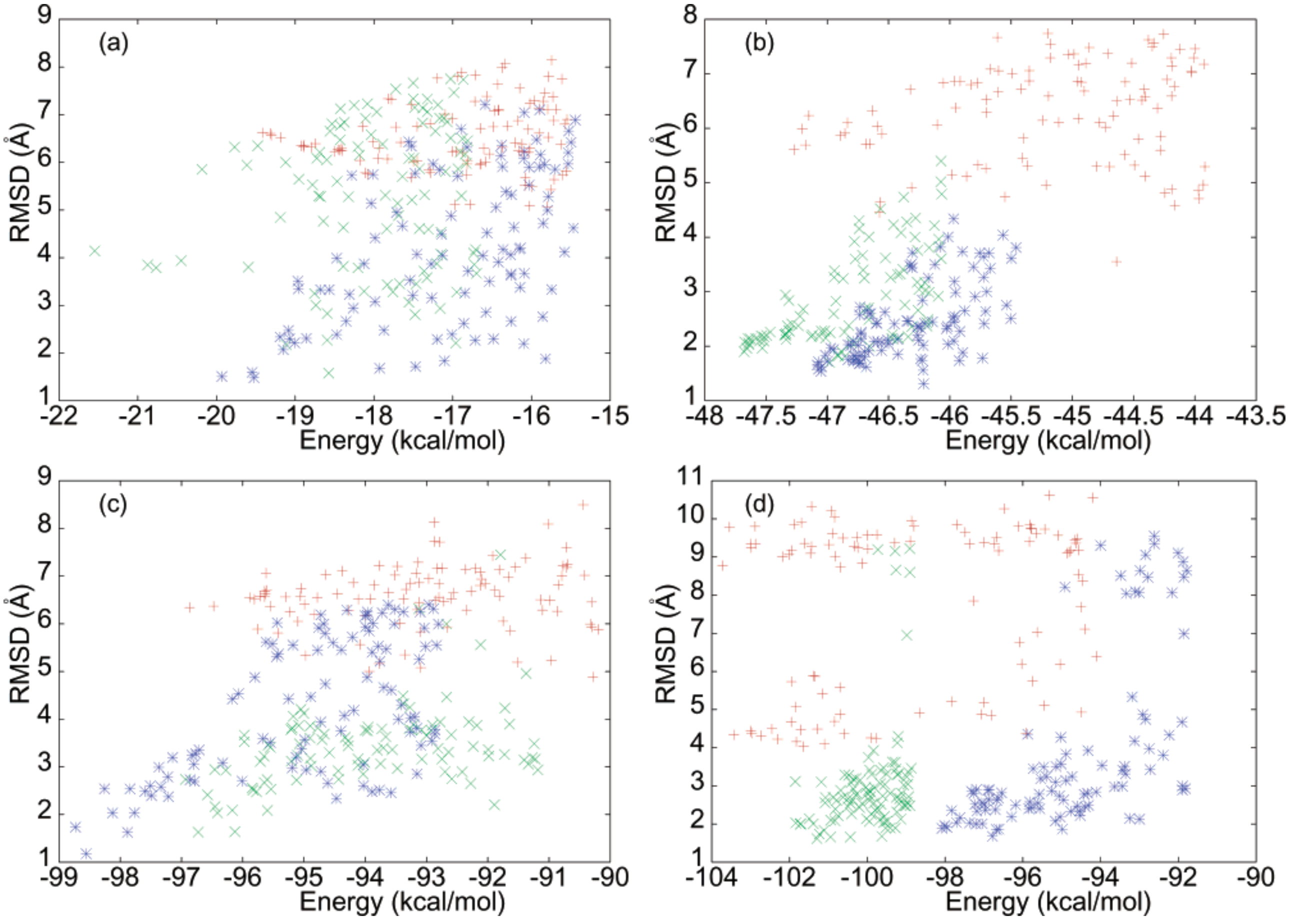


Figure 4. Plots of the UNRES energy and C$^\alpha$ rmsd (from the native structure) of four proteins obtained from a global CSA search using the initial and refined parameters. The red, green, and blue crosses denote the results obtained using the parameters before the optimization, after the 28th iteration, and after the 40th iteration, respectively. The results are shown for (a) betanova, (b) 1fsd, (c) 1vii, and (d) 1bdd.

In this work, we successfully apply this method to the optimization of linear parameters in the UNRES potential energy for a training set consisting of betanova, 1fsd, the 36-residue subdomain of chicken villin headpiece (PDB ID 1vii), and the 10−55 residue fragment of staphylococcal protein A (PDB ID 1bdd). We obtain the global minimum-energy conformations (GMECs) of these proteins with rmsd values of 1.5, 1.7, 1.7, and 1.9 Å, respectively. The proteins in the training set are β (betanova), α/β (1fsd), and α (1vii and 1bdd) proteins, which cover representative structural classes of small proteins in nature. The basic form of the UNRES potential that we use, where the only multibody term is the four-body term, is the one that was used for the successful prediction of unknown structures of proteins in CASP3.[7,10,27] With the optimized parameters, we have performed jackknife tests on various proteins not included in the training set, and we find promising results.

II. Methods

A. Potential Energy Function. We use the UNRES force field,[17−19] where a polypeptide chain is represented by a sequence of α-carbon (C$^\alpha$) atoms linked by virtual bonds with attached united side-chains (SC) and united peptide groups (p) located in the middle between consecutive C$^\alpha$'s (Figure 1). All of the virtual bond lengths are fixed: the C$^\alpha$−C$^\alpha$ distance is taken to be 3.8 Å, and the C$^\alpha$−SC distance is given for each amino acid type. The energy of the chain is given by

$$E = \sum_{i<j} U_{\mathrm{SCSC}}(i,j) + \sum_{i\neq j} U_{\mathrm{SCp}}(i,j) + \sum_{i<j-1} U_{\mathrm{pp}}(i,j) + \sum_i U_{\mathrm{b}}(i) + \sum_i U_{\mathrm{tor}}(i) + \sum_i U_{\mathrm{rot}}(i) + U_{\mathrm{dis}} + \sum_{i<j} U^{(4)}_{\mathrm{el-loc}}(i,j) \qquad (3)$$

As described in detail in the Appendix of ref 15, 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. 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, $U_{\mathrm{SCp}}(i, j)$ corresponds to the excluded-volume interaction between the side chain of residue i and the peptide group of residue j, and the potential $U_{\mathrm{pp}}(i, j)$ 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. The functional form, as well as the initial parameter set that we

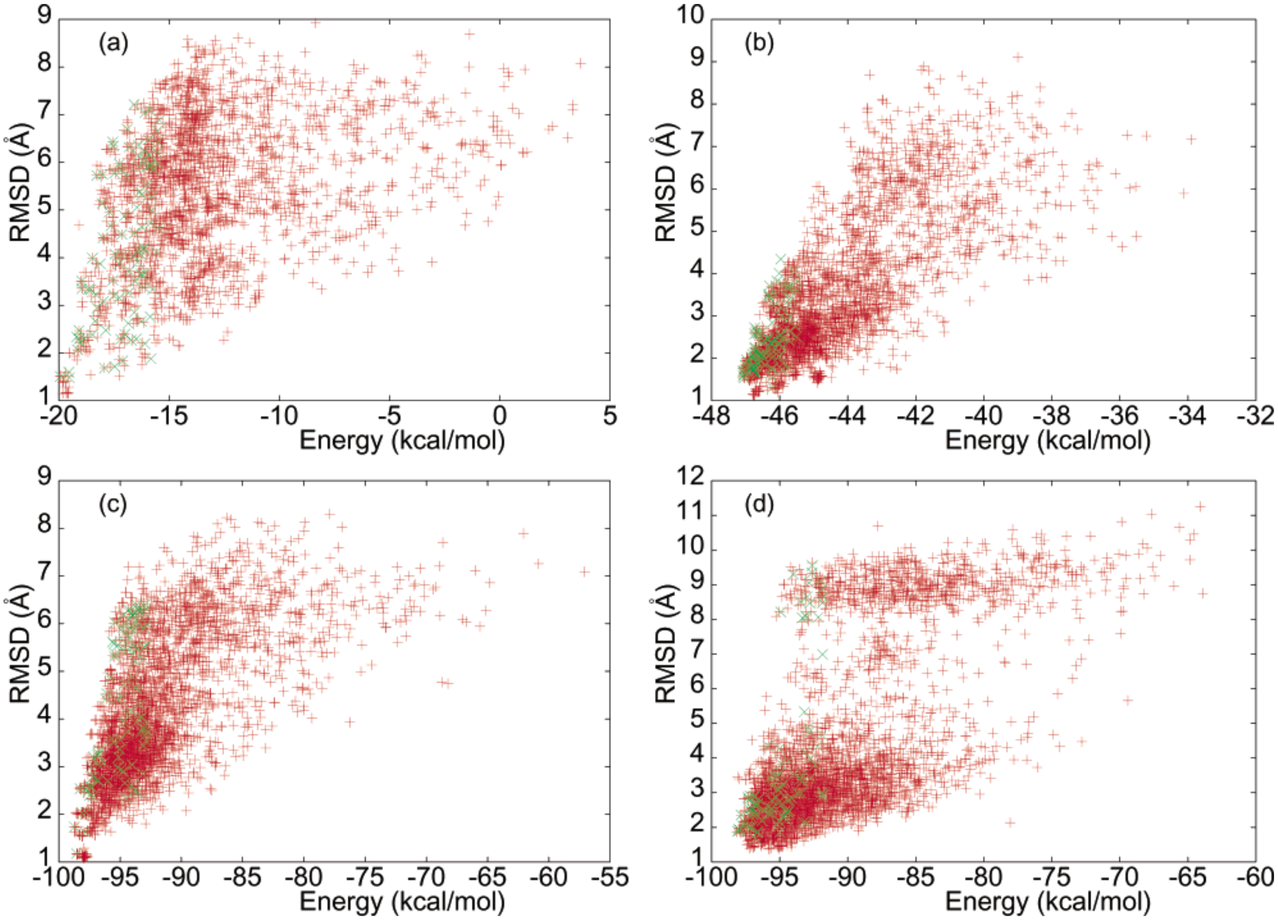


Figure 5. Plots of the UNRES energy and C$^\alpha$ rmsd (from the native structure) of local minimum-energy conformations in the structural database of four proteins accumulated after the 40th iteration of parameter optimization (red) and the new conformations obtained from the global CSA using these parameters (green). The results are shown for (a) betanova, (b) 1fsd, (c) 1vii, and (d) 1bdd.

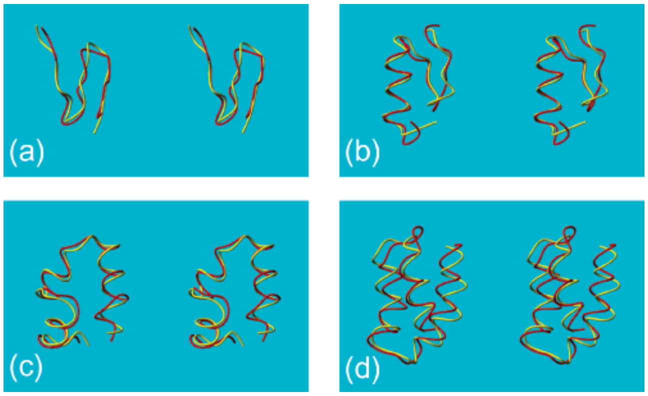


Figure 6. C$^\alpha$ trace of GMEC found with the optimized parameters shown together with the native structure for each of the four proteins in the training set. The native structure is shown in red, and the GMEC is shown in yellow. The GMECs are shown for (a) betanova, with a rmsd value of 1.51 Å, (b) 1fsd, with a rmsd value of 1.65 Å, (c) 1vii, with a rmsd value of 1.73 Å, and (d) 1bdd, with a rmsd value of 1.89 Å. Plots were prepared with the program MOLMOL.[37]

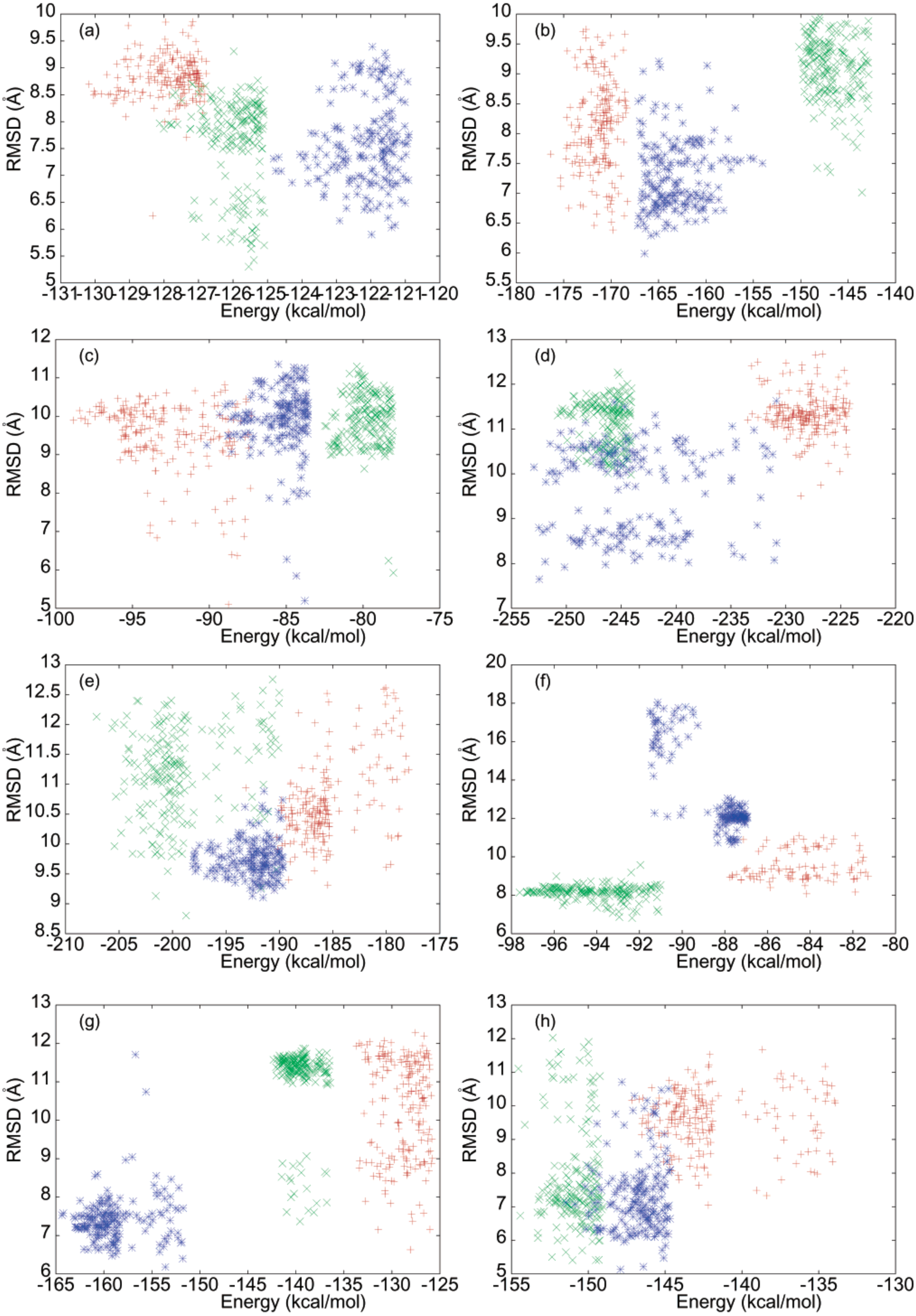
use, is the one used in the CASP3 exercise.[7,27] The total number of linear parameters that we adjust is 715.[28]

B. Global and Local CSA. To check the performance of a potential energy function for a given set of parameters, one has to sample supernative, nativelike, and non-native conformations for each protein in the training set. For this purpose, we perform two types of conformational searches: local and global CSA. In the local CSA, the backbone angles of the conformations are fixed to the values of the native conformations, and only the side-chain angles are minimized with respect to the energy. We call the resulting conformations the supernative. The other conformations are obtained from an unrestricted conformational search that we call global CSA. The conformations obtained from the local and global searches are added to the structural database of local minimum-energy conformations for each protein.

C. Parameter Refinement Using Linear Programming. The changes in energy gaps are estimated by the linear approximation of the potential energy in terms of the parameters. Among the conformations with nonzero rmsd values in the structural database, 50 (an arbitrary number) conformations with the smallest rmsd values are selected as the nativelike conformations, and the rest are considered to be the non-native ones. Because a potential can be considered to describe the nature correctly if nativelike structures have lower energies than non-native ones, the parameters are optimized to minimize the energy gaps $E_{\rm gap}^{(1)}$ and $E_{\rm gap}^{(2)}$,

$$E_{\rm gap}^{(1)} = E^{\rm N} - E^{\rm NN}$$
$$E_{\rm gap}^{(2)} = E^{\rm SN} - E^{\rm NN} \qquad (4)$$

for each protein in the training set, where $E^{\rm N}$ and $E^{\rm SN}$ are the highest energies of the nativelike and supernative conformations, respectively, and $E^{\rm NN}$ is the lowest energy of the non-native conformations. The energies are the modified ones that are weighted with the rmsd values of the conformations as in eq 2. Weighting the energies with the rmsd values makes the large rmsd conformations have high energies compared to the ones


(a)
RMSD (Å)
10
9.5
9
8.5
8
7.5
7
6.5
6
5.5
5
-131
-130
-129
-128
-127
-126
-125
-124
-123
-122
-121
-120
Energy (kcal/mol)
(b)
RMSD (Å)
10
9.5
9
8.5
8
7.5
7
6.5
6
5.5
-180
-175
-170
-165
-160
-155
-150
-145
-140
Energy (kcal/mol)
(c)
RMSD (Å)
12
11
10
9
8
7
6
5
-100
-95
-90
-85
-80
-75
Energy (kcal/mol)
(d)
RMSD (Å)
13
12
11
10
9
8
7
-255
-250
-245
-240
-235
-230
-225
-220
Energy (kcal/mol)
(e)
RMSD (Å)
13
12.5
12
11.5
11
10.5
10
9.5
9
8.5
-210
-205
-200
-195
-190
-185
-180
-175
Energy (kcal/mol)
(f)
RMSD (Å)
20
18
16
14
12
10
8
6
-98
-96
-94
-92
-90
-88
-86
-84
-82
-80
Energy (kcal/mol)
(g)
RMSD (Å)
13
12
11
10
9
8
7
6
-165
-160
-155
-150
-145
-140
-135
-130
-125
Energy (kcal/mol)
(h)
RMSD (Å)
13
12
11
10
9
8
7
6
5
-155
-150
-145
-140
-135
-130
Energy (kcal/mol)

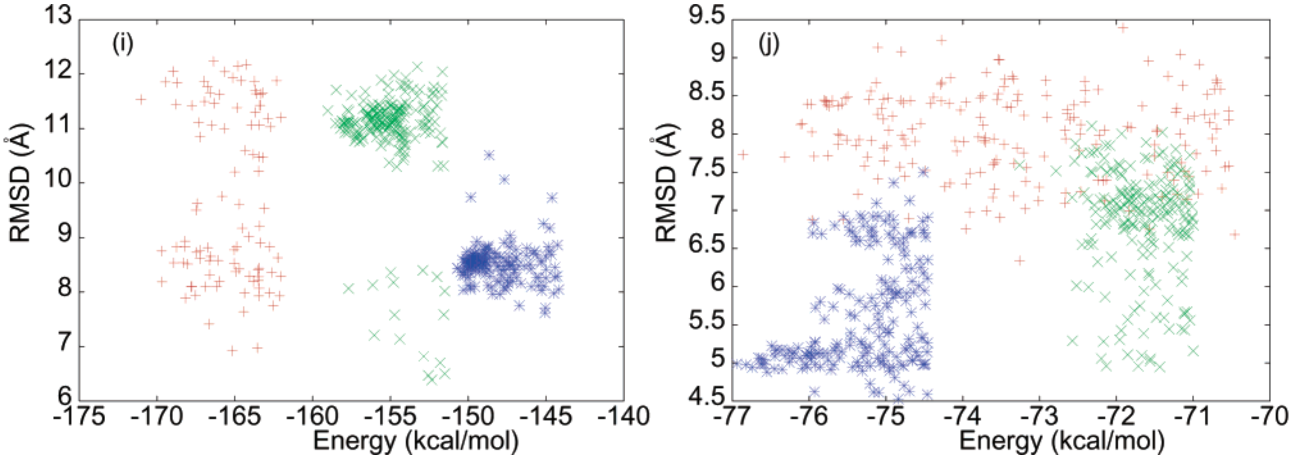


Figure 7. Results of the jackknife test. Plots of the UNRES energy versus C$^{\alpha}$ rmsd (from the native structure) of the low-lying local-energy-minimum conformations. Conformations obtained from the global CSA using initial parameters and parameters obtained after the 28th and 40th iterations of optimization are shown in red, green, and blue, respectively. The results are shown for (a) 1bbg, (b) 1ccn, (c) 1hnr, (d) 1kbs, (e) 1neb, (f) 1bba, (g) 1idy, (h) 1prb, (i) 1pru, and (j) 1zdb.

with small rmsd values. This idea is somewhat similar to the hierarchical optimization method proposed in ref 16, where the secondary structure contents were used as the criterion for ranking the nativeness of the conformations. In this work, we simply use the rmsd values. The rmsd value is easier to calculate; consequently, it becomes easier to automate the procedure. Parameter optimization is carried out by minimizing the energy gaps $E^{(1)}_{\rm gap}$ and $E^{(2)}_{\rm gap}$ of each protein in turn, while imposing the constraints that all of the other energy gaps, including those from the other proteins, do not increase.

In this work, we adjust only the linear parameters for simplicity, the total number of them being 715 for the UNRES potential. Therefore, the energy of a local minimum-energy conformation can be written as

$$E = \sum_j p_j e_j(x_{\rm min}) \tag{5}$$

where the e_j's are the energy components evaluated with the coordinates $x_{\rm min}$ of a local minimum-energy conformation. Because the positions of local minima also depend on the parameters, the full parameter dependence of the energy gaps is nonlinear. However, if the parameters are changed by small amounts, then the energy with the new parameters can be estimated by the linear approximation

$$E^{\rm new} \approx E^{\rm old} + \sum_i (p_i^{\rm new} - p_i^{\rm old}) e_i(x_{\rm min}) \tag{6}$$

where the $p_i^{\rm old}$ and $p_i^{\rm new}$ terms represent the parameters before and after the 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.[13] The additional 0.3 rmsd term of eq 2 vanishes in these expressions for the same reason. The changes of the energy gaps are estimated to be

$$\begin{aligned}\Delta E^{(1)}_{\rm gap} &= E^{(1)}_{\rm gap}(\{p_j^{\rm new}\}) - E^{(1)}_{\rm gap}(\{p_j^{\rm old}\}) \\ &= (E^{\rm N}(\{p_j^{\rm new}\}) - E^{\rm NN}(\{p_j^{\rm new}\})) - \\ &\qquad (E^{\rm N}(\{p_j^{\rm old}\}) - E^{\rm NN}(\{p_j^{\rm old}\})) \\ &= \sum_j (e_j^{\rm N} - e_j^{\rm NN})(p_j^{\rm new} - p_j^{\rm old}) \end{aligned} \tag{7}$$

$$\begin{aligned}\Delta E^{(2)}_{\rm gap} &= E^{(2)}_{\rm gap}(\{p_j^{\rm new}\}) - E^{(2)}_{\rm gap}(\{p_j^{\rm old}\}) \\ &= (E^{\rm SN}(\{p_j^{\rm new}\}) - E^{\rm NN}(\{p_j^{\rm new}\})) - \\ &\qquad (E^{\rm SN}(\{p_j^{\rm old}\}) - E^{\rm NN}(\{p_j^{\rm old}\})) \\ &= \sum_j (e_j^{\rm SN} - e_j^{\rm NN})(p_j^{\rm new} - p_j^{\rm old}) \end{aligned} \tag{8}$$

The magnitude of the parameter change $\delta p_j \equiv p_j^{\rm new} - p_j^{\rm old}$ is bounded by a certain fraction ϵ of $p_j^{\rm old}$. We use $\epsilon = 0.01$ in this study. First, the vector δp_j is chosen within the bound to decrease the energy gap $\Delta E^{(1)}_{\rm gap}$ of the selected protein as much as possible while imposing the constraint that any positive values among $E^{(2)}_{\rm gap}$ and the energy gaps of the other proteins do not increase and negative values do not become positive. Denoting the energy gaps of the kth protein as $E^{(p=1,2)}_{\rm gap}(k)$ and assuming that the ith protein is selected for the decrease in the energy gap, this problem can be phrased as follows:

Minimize

$$\Delta E^{(1)}_{\rm gap}(i) = \sum_j (e_j^{\rm N}(i) - e_j^{\rm NN}(i))(p_j^{\rm new} - p_j^{\rm old}) \tag{9}$$

with constraints

$$|\delta p_i| \le \epsilon \tag{10}$$

$$\Delta E^{(2)}_{\rm gap}(i) = \sum_j (e_j^{\rm SN}(i) - e_j^{\rm NN}(i))(p_j^{\rm new} - p_j^{\rm old}) \le \left\{\begin{array}{ll} 0 & \text{if } E^{(2)}_{\rm gap}(i) > 0 \\ -E^{(2)}_{\rm gap}(i) & \text{otherwise} \end{array}\right\} \tag{11}$$

$$\Delta E^{(p=1,2)}_{\rm gap}(k \neq i) = \sum_j (e_j^{\rm (S)N}(k) - e_j^{\rm NN}(k))(p_j^{\rm new} - p_j^{\rm old}) \le \left\{\begin{array}{ll} 0 & \text{if } E^{(p)}_{\rm gap}(k) > 0 \\ -E^{(p)}_{\rm gap}(k) & \text{otherwise} \end{array}\right\} \tag{12}$$

This is a global optimization problem where the linear parameters p_j are the variables. The object function to minimize and the constraints are all linear in p_j. This type of optimization problem is called linear programming. It can be solved exactly, and many algorithms have been developed to solve the linear programming problem. We use the primal-dual method with supernodal Cholesky factorization[29] in this work, which finds an accurate answer with reasonable computational cost.

After minimizing $\Delta E^{(1)}_{\text{gap}}(i)$, we solve the same form of linear programming problem where $\Delta E^{(2)}_{\text{gap}}(i)$ is now the objective function and the other energy-gap changes become constrained. Then we select another protein and repeat this procedure (300 times in this work) of minimizing $\Delta E^{(1)}_{\text{gap}}$ and $\Delta E^{(2)}_{\text{gap}}$ in turn.

It should be noted that we do not put any constraints on the overall scale of the parameters. Because the energies are proportional to the overall parameter scale and because this scale changes freely during the optimization process, the overall energy scale is not determined in our protocol. Therefore, the numerical value of the energy has no physical meaning in our work.

The current parameter-optimization procedure is different from the one used in earlier work,[13−15] where the optimization was performed without using supernative conformations and energy was not weighted with the rmsd value. The earlier procedure and the current one are shown schematically in Figures 2 and 3, respectively, in terms of the energy and rmsd.

D. Reminimization and New Conformational Search. Because the procedure in the previous section was based on linear approximation equations (eqs 7 and 8), we now have to evaluate the true energy gaps using the newly obtained parameters. The breakdown of the linear approximation 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 valid for the new parameter set. For this reason, we reminimize the energy of these conformations with the new parameters. Because supernative conformations are not local minimum-energy conformations, even with the original parameters, the unconstrained reminimization of these conformations with the new parameters may furnish low-lying local minima with small values of rmsd. Second, the local minima obtained using the CSA method with the original parameter set may constitute only a small fraction of low-lying local minima. After the modification of 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 distinct low-energy local minima appear. Therefore, these new minima are taken into account by performing subsequent CSA searches (see section B) with the newly obtained parameter set.

E. 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, which 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.[15] This iterative procedure is continued until sufficiently good nativelike conformations are found from the global CSA search.

III. Results

A. Four Proteins in the Training Set. We apply our protocol to a training set consisting of four proteins. They are the designed protein betanova, 1fsd, the 36-residue subdomain of chicken villin headpiece (HP36 or 1vii), and the 10−55 fragment of the B domain of staphylococcal protein A (1bdd), which are 20, 28, 36, and 46 residues long, respectively. The protein betanova is a β protein, 1fsd is an α/β protein, and the rest are α proteins that represent structural classes of small proteins. The initial parameter set is the one used in CASP3.[7,27]

Fifty conformations were sampled in each CSA search, and the global minimum-energy conformations (GMECs) found with the initial parameters have rmsd values of 6.6, 5.6, 6.3, and 9.5 Å, respectively. The smallest values of rmsd found from the CSA search are 5.1, 3.6, 4.9, and 4.0 Å. After the 28th iteration of the parameter refinement, the conformations with smaller values of rmsd are found from the global CSA search. The GMECs have rmsd values of 4.1, 1.9, 2.7, and 3.1 Å, and the smallest values of rmsd that are found are 1.6, 1.7, 1.6, and 1.6 Å. The rmsd values become even smaller after the 40th iteration, with rmsds of GMECs being 1.5, 1.7, 1.7, and 1.9 Å and the smallest values of rmsd being 1.5, 1.3, 1.2, and 1.7 Å. The rmsds of the GMECs and the smallest rmsds for these parameters are shown in the Table 1. Their energies are not displayed because their numerical values have no physical meaning, which is due to the fact that the overall scale of the linear parameters is not fixed in our protocol. The results of the global search with the initial and optimized parameter set for the four proteins are also plotted in different colors in terms of energy and rmsd in Figure 4. The local minimum-energy conformations accumulated in the structural databases after the 40th iteration of the parameter refinement are shown in Figure 5 along with the global CSA search results. The C^{α} traces of the GMECs of the four proteins found using the parameters obtained after the 40th iteration of optimization are shown in Figure 6 along with the native conformations.

We also observe a linear slope of 0.3 in the energy versus rmsd plot for the low-lying states. It turns out that the energy landscape designed in this way ensures good foldability. In fact, we observe successful folding of all four proteins into their native states in the direct Monte Carlo folding simulation with the UNRES potential, using the parameters after the 40th iteration of the refinement.[30]

B. Jackknife Tests. We have performed conformational searches for proteins not contained in the training set, which are usually called jackknife tests. We selected proteins of various structural classes, composed of no more than 60 amino acids residues. We find that the performance of the optimized parameters is reasonably good and that the optimized parameter set provides better performance compared to the results from the initial parameter set. This implies that the optimized parameters are not overfitted to the four proteins in the training set but are transferable to other proteins to some extent.

We have considered proteins 1bbg, 1ccn, 1hnr, 1kbs, 1neb, 1bba, 1idy, 1prb, 1pru, and 1zdb, with the number of amino acid residues being 40, 46, 47, 60, 60, 36, 54, 53, 56, and 38, respectively. 1bbg, 1ccn, and 1hnr are α/β proteins, 1kbs and 1neb are β proteins, and the rest are α proteins. The rmsd values of the GMECs and the best structures found with the initial and optimized parameter sets are summarized in Table 1. The results of the global CSA search with these parameters are shown in Figure 7 in terms of energy and rmsd. We find that the results for protein 1zdb are particularly notable. In Figure 8, the C^{α} traces of the GMECs found with the initial and optimized parameters are shown together with the native structure. We see that although this protein is not included in the training set the GMEC becomes more and more similar to the native structure as the parameter-optimization procedure is continued. These results suggest that the parameters we obtained from the training set of the four proteins perform better than

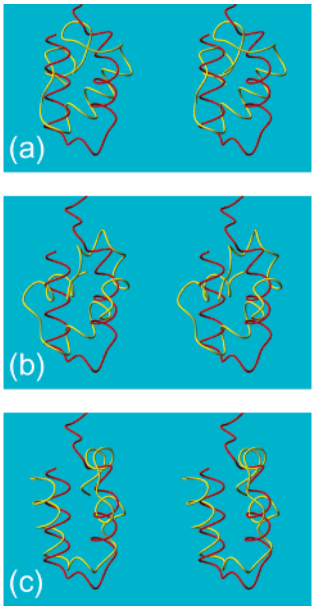


Figure 8. C^{α} traces of the GMECs of 1zdb for various parameter sets. The native structure is shown in red, and the GMECs found with the optimized parameters are shown in yellow. It should be noted that there are conformations with even smaller values of rmsd among those found from the CSA search. The results are shown with (a) initial parameters (rmsd of 7.7 Å), (b) parameters after the 28th iteration of optimization (rmsd of 7.6 Å), and (c) parameters after the 40th iteration of optimization (rmsd of 5.0 Å). We observe that the GMEC becomes more and more similar to the native structure as the parameter optimization continues, although this protein is not included in the training set. This strongly suggests that the optimized parameters are transferable to other proteins to some extent. The plots were prepared with the program MOLMOL.[37]

the initial parameter set and are transferable to other proteins to some extent.

IV. Discussion

We have proposed a general protocol for force field parameter optimization and landscape design and have applied it to the UNRES potential. We optimized the 715 linear parameters so that they correctly describe the energetics of four proteins simultaneously. This optimized parameter set yielded GMECs with rmsd values of 1.5, 1.7, 1.7, and 1.9 Å for betanova, 1fsd, 1vii, and 1bdd, respectively. In the process, we designed the energy landscape to have good foldability.[30] It seems that the current parameter-optimization method achieves this goal by constructing the protein-folding funnel,[31] which is believed to be an essential property of the protein energy functions in nature. 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.

Liwo et al.[16] also successfully carried out the parameter optimization of an UNRES potential utilizing the CSA method. Their protocol is similar to ours in that they use energy gaps that depend on the nativelikeness of conformations, but the secondary structure content was used for the criterion of nativelikeness instead of the rmsd. Also, there are several additional multibody terms in their UNRES potential, and the interaction between side chains is a Gay-Berne potential instead of the simpler Lennard-Jones form used in this work. They optimized their parameters separately for proteins 1fsd and 1igd. The parameters optimized using 1igd were tested on proteins betanova, 1fsd, and 1bdd to obtain impressive results.

In contrast to earlier protocols,[13,15] where the value of the rmsd cutoff was manually specified for each protein at each iteration to define nativelike conformations, we now have defined 50 conformations with the smallest rmsd values in the structural database for each protein as the nativelike conformations and have used supernative structures, which have zero rmsd values by definition, to furnish candidates for low-lying nativelike conformations with small values of the rmsd. This enabled us to automate the whole procedure using a shell script. However, there is still some arbitrariness in our protocol, such as choosing 50 nativelike conformations and giving the slope of 0.3 in eq 2. Although 0.3 is chosen as an optimal value by examining the performance of our protocol for one protein—betanova—this criterion is rather arbitrary because the overall performance of the protocol is not sensitive to the precise numerical value of the coefficient. In fact, one can obtain reasonably good results using other values such as 0.1 and 0.5. In addition, the optimal value may depend on the proteins that we use for the training. Therefore, one should consider a more systematic way of determining the value of the slope. Finally, it should be noted that although we have considered only the UNRES potential for parameter optimization in this work it is straightforward to apply the procedure to other potentials such as ECEPP,[32] AMBER,[33] and CHARMM[34] with various solvation terms.[35,36] All of these points are left for future study.

Acknowledgment. We thank Ki Hyung Joo and Il-Soo Kim for useful discussions and help in carrying out this work. This work was supported by grants R01-2003-000-11595-0 (J.L.) and R01-2003-000-10199-0 (J.L.) from the Basic Research Program of the Korea Science & Engineering Foundation. The calculations were carried out on the Linux PC cluster of 214 AMD processors at KIAS.

References and Notes

(1) Jones, T. A.; Kleywegt, G. J. *Proteins Struct. Funct. Genet. Suppl.* **1999**, *3*, 30.

(2) Murzin, A. G. *Proteins Struct. Funct. Genet. Suppl.* **1999**, *3*, 88.

(3) Tramontano, A.; Leplae R.; Morea V. *Proteins: Struct., Funct., Genet. Suppl.* **2001**, *5*, 22.

(4) Sippl, M. J.; Lackner, P.; Domingues, F. S.; Prlić, A.; Malik, R.; Andreeva, A.; Wiederstein, M. *Proteins: Struct., Funct., Genet., Suppl.* **2001**, *5*, 55.

(5) Orengo, C. A.; Bray, J. E.; Hubbard, T.; LoConte, L.; Sillitoe, I. *Proteins: Struct., Funct., Genet. Suppl.* **1999**, *3*, 149.

(6) Lesk, A. M.; LoConte L.; Hubbard, T. *Proteins: Struct., Funct., Genet. Suppl.* **2001**, *5*, 98.

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

(8) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025.

(9) Liwo, A.; Lee, J.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5482.

(10) 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.

(11) 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.

(12) Anfinsen, C. B. *Science* **1973**, *181*, 223.

(13) Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A. *J. Phys. Chem. B* **2001**, *105*, 7291.

(14) 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.

(15) Lee, J.; Park, K.; Lee, J. *J. Phys. Chem. B* **2002**, *106*, 11647.

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

(17) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849.

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

(19) 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.

(20) Lee, J.; Scheraga, H. A.; Rackovsky. S. *J. Comput. Chem.* **1997**, *18*, 1222.

(21) Lee, J.; Scheraga, H. A.; Rackovsky. S. *Biopolymers* **1998**, *46*, 103.

(22) Lee, J.; Scheraga, H. A. *Int. J. Quantum Chem.* **1999**, *75*, 255.

(23) Lee, J.; Lee, I. H.; Lee, J. *Phys. Rev. Lett.* **2003**, *91*, 080201.

(24) Kim S.-Y.; Lee S. J.; Lee J. *J. Chem. Phys.* **2003**, *119*, 10274.

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

(26) Mairov V. N.; Crippen G. M. *J. Mol. Biol.* **1992**, *227*, 876.

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

(28) The number of parameters is larger than the 709 used in ref 15 because now we allow different weights for the four-body terms for parallel and antiparallel strands. Also, whereas the side-chain peptide group interaction and 14 rescaling (see Appendix of ref 15) were performed in an asymmetric manner in earlier works, we do it symmetrically.

(29) Mészáros, C. A. *Comput. Math. Appl.* **1996**, *31* 49.

(30) Kim, S.-Y.; Lee, J.; Lee, J. *J. Chem. Phys.* **2004**, *120*, in press.

(31) Onuchic, J. N.; Luthy-Schulten, Z.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545.

(32) 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.

(33) 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.

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

(35) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086.

(36) Wesson, L.; Eisenberg, D. *Protein Sci.* **1992**, *1*, 227.

(37) Koradi, R.; Billeter, M.; Wuthrich, K. *J. Mol. Graphics* **1996**, *14*, 51.