Journal of the Korean Physical Society, Vol. 44, No. 3, March 2004, pp. 594∼598

Optimization of Potential-Energy Parameters for Folding of Several Proteins

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

Seung-Yeon Kim and Jooyoung Lee*
School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722

(Received 8 October 2003)

We introduce a novel approach to the study of the folding of proteins whose native structures are already known. We use an off-lattice atomistic potential energy. The parameters of the potential energy are simultaneously optimized for several proteins. The low-lying local-energy minima for these proteins are found by conformational space annealing. The parameters are modified in such a way that the native-like conformations are energetically more favored than the others. After the parameter optimization, one set of the parameters is obtained for the proteins. We then investigate Monte Carlo dynamics of these proteins by using this optimized potential energy. Our work is distinguished from earlier work in the literature, where folding was achieved by using simplified models such as lattice models. We apply our method to four proteins: betanova, 1fsd, 1vii, and 1bdd, and observe that at appropriate temperatures they fold into their native structure, starting from various non-native states. In all cases, rapid collapse is followed by a subsequent folding process, that takes place on a longer timescale. We also observe that for all proteins at low temperatures, the probability distributions of various quantities such as RMSD depend on initial conformations, showing their glassy behavior. At higher temperatures, this non-ergodic glassy behavior disappears. The results provide new insights into the folding mechanism, which is controlled not only by thermodynamic factors but also by kinetic factors. The way a protein folds into its native structure is also determined by the convergence point of early folding trajectories, which cannot be obtained from the free-energy surface.

PACS numbers: 05.20.Dd, 05.70.−a, 87.14.Ee, 87.15.−v
Keywords: Protein folding, Folding mechanism, Potential-parameter optimization, Monte Carlo dynamics

I. INTRODUCTION

The understanding of protein folding, that is, folding of a protein from its amino-acid sequence into a unique three-dimensional structure (native structure) is a long-standing challenge in modern biophysics. The native structure and folding pathways are indispensable for understanding the function and biological role of the protein [1–3]. Computer simulations [4–10] have been carried out to study the folding mechanism. However, simulation of protein-folding processes by using an atomistic model is a very difficult task. The difficulties come from two sources.

First, there are inherent inaccuracies in the potential-energy functions which describe the energetics of proteins. Potential-energy functions are generally parameterized 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.

Second, even if the global minimum-energy conformation is native-like, this does not guarantee that a protein will fold into its native structure on a reasonable timescale, because direct folding simulation by using an all-atom potential requires astronomical amounts of CPU time. Currently, typical simulation times are only about a few nanoseconds. An extensive folding simulation has been carried out for the 36-residue 1vii, where 1-μs molecular-dynamics simulation with an all-atom potential has been performed, producing only candidates

*E-mail: jlee@kias.re.kr; Fax: +82-2-958-3786 for folding intermediates [4]. For this reason, direct folding simulations have been mainly focused on simple models, such as lattice models [5,6], models where only native interactions are included (Go-type models) [7,8], and a model with discrete energy terms whose parameters are optimized *separately* for each protein [9]. Alternative indirect approaches have also been proposed, including unfolding simulations [8,10] starting from the folded state of a protein. However, it is not obvious whether the folding is the reverse of the unfolding [6,8]. Moreover, to the best of our knowledge, no one has yet succeeded in folding more than one protein into their native states by using a *single* potential, even with the use of simplified models [9].

In this work, we introduce a novel method of folding several proteins simultaneously. This method uses a *single* atomistic continuous potential which includes *all pairwise* (native and non-native) interactions, and yet allows us to carry out *folding* simulations starting from non-native conformations. We use an off-lattice potential energy function, the united-residue (UNRES) [11, 12] potential energy. The parameters of the UNRES potential energy are *simultaneously* optimized for several proteins. The low-lying local-energy minima for these proteins are found by the conformational space annealing (CSA) method [13]. The parameters are modified in such a way that the native-like conformations are energetically more favored than the others. After the parameter optimization, *one* set of the parameters is obtained for the proteins. The optimized UNRES potential energy is applied to the study of folding processes of these proteins.

II. PARAMETER OPTIMIZATION OF A POTENTIAL ENERGY

We propose a procedure (Fig. 1) where the parameters of an empirical potential-energy function are modified so as to make conformations with larger values of C^{α}-RMSD (root-mean-square deviation) from the native structure having higher values of energy relative to those with smaller values of RMSD. This goal is achieved by using the following modified energy:

$$E_{\text{modified}} = E + 0.3 \text{ RMSD}, \qquad (1)$$

when calculating the energy gaps (see subsection II. 3), where the numerical value of the coefficient 0.3 is an arbitrarily chosen value. We also tried the values of 0.1 and 0.5, with similar results to those in the case of 0.3. In the current method, we introduce *super-native* 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. Among the conformations with non-zero RMSD, 50 (an arbitrary number) conformations with the lowest RMSD values are selected as *native-like* conformations, and the rest are considered as *non-native* conformations. The super-native conformations would furnish low-lying local minima with small RMSD values after the reminimization procedure with new optimized parameters. Generally, the RMSD values of the native-like conformations become smaller as the iteration of the parameter optimization continues.

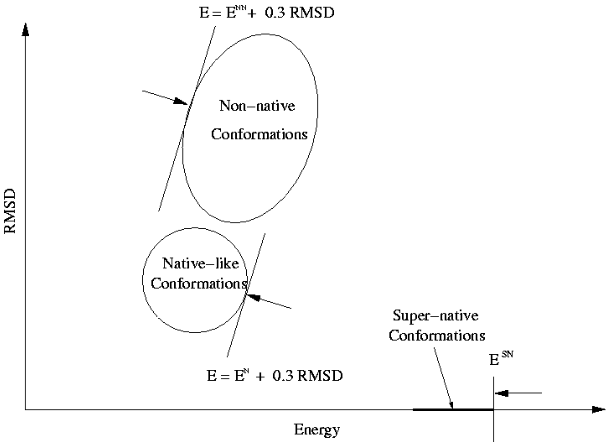

Fig. 1. Schematic showing the potential-parameter optimization. The minimum modified energy of the non-native family, and the maximum modified energies of native-like and super-native families, define the energy gaps. The arrows indicate the direction of the optimization.

1. Potential-energy Function

In the UNRES potential, a protein 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 located in the middle between the consecutive C^{α}'s (Fig. 2). All the virtual bond lengths are fixed: the C^{α}-C^{α} distance is taken as 3.8 Å, and C^{α}-SC distances are given for each amino acid type. There are two backbone angles and two SC angles per residue (no SC for glycines). The energy of a protein is given by [11,12]

$$E = U_{\text{dis}} + \sum_{i<j} \left[U^{(4)}_{\text{el-loc}}(i,j) + U_{\text{ss}}(i,j) \right] + \sum_{i \neq j} U_{\text{sp}}(i,j) + \sum_{i<j-1} U_{\text{pp}}(i,j) + \sum_{i} \left[U_{\text{b}}(i) + U_{\text{t}}(i) + U_{\text{r}}(i) \right]. \quad (2)$$

Here, U_{dis} denotes the energy term which forces two cysteine residues to form a disulfide bridge. The four-body interaction term $U^{(4)}_{\text{el-loc}}$ results from the cumulant expansion of the restricted free energy of the protein. $U_{\text{ss}}(i,j)$ represents the mean free energy of the hydrophobic (hydrophilic) interaction between the side-chains of residues i and j, which is expressed by Lennard-Jones potential, $U_{\text{sp}}(i,j)$ corresponding to the excluded-volume interaction between the side-chain of residue i and the peptide group of residue j, and the potential $U_{\text{pp}}(i,j)$

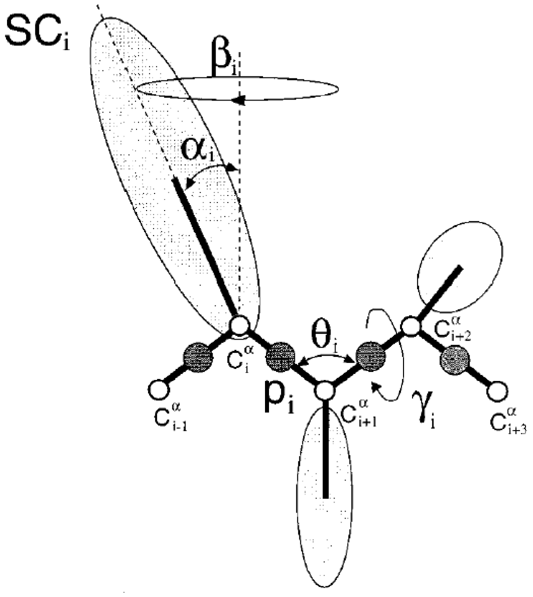

Fig. 2. United-residue representation of a protein. The interaction sites are side-chain ellipsoids of different sizes (SC) and peptide-bond centers (p) indicated by shaded circles, whereas the α-carbon atoms (small empty circles) are introduced to define the backbone-local interaction sites and to assist in defining the geometry. The virtual C$^\alpha$-C$^\alpha$ 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, with a variable bond angle, α_i, formed by SC$_i$ and the bisector of the angle defined by C$^\alpha_{i-1}$, C$^\alpha_i$ and C$^\alpha_{i+1}$, and with a variable dihedral angle β_i of counterclockwise rotation about the bisector, starting from the right side of the C$^\alpha_{i-1}$, C$^\alpha_i$, C$^\alpha_{i+1}$ frame.

accounts for the electrostatic interaction between the peptide groups of residues i and j. The terms $U_\mathrm{b}(i)$, $U_\mathrm{t}(i)$ and $U_\mathrm{r}(i)$ denote the short-range interactions corresponding to the energies of virtual angle bending, virtual dihedral angle torsions, and side-chain rotamers, respectively. The total number of linear parameters which we adjust is 715 [12].

2. Global and Local CSA

In order to check the performance of a potential-energy function for a given set of parameters, one has to sample super-native, native-like, and non-native conformations for each protein in the training set. For this purpose, we perform two types of conformational search, the local and global CSA searches. The local CSA search is defined as the restricted search for the super-native conformations in the space of the side-chain angles. The other conformations are obtained from an unrestricted conformational search which 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.

3. Parameter Refinement

Since a potential can be considered to describe the native state correctly if native-like structures have lower energies than the non-native ones, the parameters are optimized to minimize the energy gaps $E^{(1)}_\mathrm{gap}$ and $E^{(2)}_\mathrm{gap}$,

$$E^{(1)}_\mathrm{gap} = E^\mathrm{N} - E^\mathrm{NN}, \quad E^{(2)}_\mathrm{gap} = E^\mathrm{SN} - E^\mathrm{NN}, \qquad (3)$$

for each protein in the training set, where E^N and E^SN are the highest energies of the native-like and super-native conformations, respectively, and E^NN is the lowest energy of the non-native conformations (Fig. 1). The energies are the modified ones that are weighted with the RMSD values of the conformations as in Eq. (1). Weighting the energy with the RMSD value has the effect of *pushing harder* the high RMSD conformations compared to the ones with lower RMSD values. The RMSD value is easy to calculate, and consequently it becomes easier to automate the procedure. The parameter optimization is carried out by minimizing the energy gaps $E^{(1)}_\mathrm{gap}$ and $E^{(2)}_\mathrm{gap}$ of each protein in turn, while imposing the constraints that all the other energy gaps, including those from the other proteins, do not increase. The energy gaps are evaluated using linear approximation.

4. Reminimization

Since we optimize the parameters by using the linear approximation, 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 so for the new parameter set. For this reason, we reminimize the energy of these conformations with the new parameters. Since the super-native conformations are not local minimum-energy conformations, even with the original parameters, the reminimization of these conformations with the new parameters would furnish low-lying local minima with small values of RMSD. Second, the local minima obtained by using the CSA method with the original parameter set may constitute only a small fraction of the low-lying local minima. After the change of the parameters, some of the local minima, which were not considered due to 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 with the newly obtained parameter set.

5. Iterative Refinement

The low-lying local energy minima found in the new conformational searches are added into 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. This iterative procedure is continued until sufficiently good native-like conformations are found from the global CSA search.

6. Four Proteins in the Training Set

We apply our method to a training set consisting of four proteins. They are betanova (20 residues, three-stranded β-sheet), 1fsd (28 residues, one β-hairpin and one α-helix), 1vii (36 residues, three-helix bundle) and 1bdd (46 residues, three-helix bundle). They represent structural classes of small proteins. Fifty conformations were sampled in each CSA search; 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, and the smallest values of RMSD found from the CSA search are 5.1, 3.6, 4.9 and 4.0 Å. After the 28-th 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 found are 1.6, 1.7, 1.6 and 1.6 Å. The RMSD values become even smaller after the 40-th 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. We observe a linear slope of 0.3 in the energy vs. RMSD plot for the low-lying states.

We have performed conformational searches for proteins not contained in the training set, which are usually called jackknife tests [14]. We find that the performance of the optimized parameters is reasonably good, and 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 to some extent transferable to other proteins.

III. FOLDING OF SEVERAL PROTEINS

We apply the UNRES potential energy with the optimized parameters after the 40-th iteration to the study on folding *dynamics* of proteins betanova, 1fsd, 1vii and 1bdd, by using Monte Carlo dynamics. In Monte Carlo simulation, the values of angles of a protein are perturbed one at a time, typically about 15°, and the backbone angles are chosen three times more frequently than SC angles. The perturbed conformation is accepted according to the change in the potential energy, following the Metropolis rule. Since only small angle changes are allowed one at a time, the resulting Monte Carlo dynamics can be viewed as equivalent to the real dynamics.

At a fixed temperature, at least ten independent simulations starting from various non-native states of a protein were performed with up to 10^9 Monte Carlo steps (MCS) for each run. Since we conducted the simulations for more than 10 different temperatures, the total number of *long-time* runs for a protein was more than 100. During simulation, the values of RMSD from the native structure and the radius of gyration (R_g) were calculated by using C^α coordinates. The lowest RMSD values from folding simulations are 0.78 Å, 1.07 Å, 1.58 Å and 2.07 Å for betanova, 1fsd, 1vii and 1bdd, respectively. The fractions of the native contacts (Q) were also measured during simulations, where Q is calculated from the native structure. A native contact is defined to exist when two C^α's separated by more than two residues in sequence are placed within 7.0 Å. Distributions of RMSD, Q and R_g are also accumulated during the whole simulations. To investigate the early folding trajectories in detail, we also performed *short-time* simulations of 10^5 MCS. We divided the 10^5 MCS into 19 intervals (ten 10^3 MCS and subsequently nine 10^4 MCS), and took the average over the conformations in each interval. These averages were again averaged over 100 independent simulations at a fixed temperature, starting from random conformations. The same procedure was applied to 100 independent short-time simulations, starting from a fully extended conformation.

We observe [15] that for all proteins at low temperatures, the probability distributions of various quantities such as RMSD depend on initial conformations, showing their glassy behavior. At higher temperatures, this non-ergodic glassy behavior disappears. We also observe [15] that all proteins fold into their native-like conformations at appropriate temperatures. In all cases, rapid collapse is followed by a subsequent folding process that takes place on a longer timescale.

The folding mechanism suggested in this study is as follows: There are two aspects of folding dynamics, (i) non-equilibrium kinetic properties and (ii) equilibrium thermodynamic properties (Fig. 3). The non-equilibrium kinetic properties, relevant to the early folding trajectories (fast process), can be examined only by direct folding simulations. The free energy surface, an equilibrium thermodynamic property, dictates the way an initially collapsed state completes its folding (slow process). The way a protein folds into its native structure, i.e., either horizontally or diagonally in the (Q, R_g) plane, is determined by the position of (Q, R_g), where early folding trajectories converge, relative to the native state. It appears that the slow folding process of α-proteins such as 1vii and 1bdd occurs in a diagonal fashion, as compared

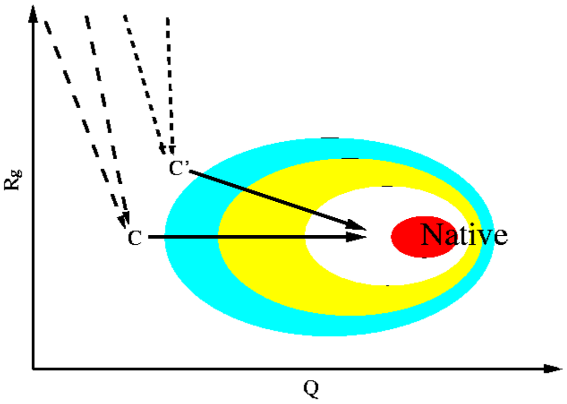

Fig. 3. Schematic of the folding trajectories in the (Q, R_g) plane. The contour represents the free-energy surface, which is an equilibrium property. Even for proteins with identical free energy landscape, the early folding trajectories (dashed lines) may converge into different points (C or C′). The solid lines represent the later part of the folding trajectories dictated by the free-energy landscape. The position of the convergence point of a protein is determined by its kinetic properties. This information can be obtained only by direct *folding* simulations.

to that of proteins (for example, betanova and 1fsd) containing β-strands.

IV. CONCLUSION

We have proposed a general method for potential-parameter optimization, and applied it to the UNRES potential. We optimized the 715 linear parameters so that they correctly describe the energetics of several proteins simultaneously. By using the optimized potential energy, we successfully carried out direct folding simulations of more than one protein. The results provide new insights into the folding mechanism.

ACKNOWLEDGMENTS

This work was supported by grants No. R01-2003-000-11595-0 (Jooyoung Lee) and No. R01-2003-000-10199-0 (Julian Lee) from the Basic Research Program of the Korea Science & Engineering Foundation.

REFERENCES

[1] C.-M. Ghim and J.-M. Park, J. Korean Phys. Soc. **40**, 1077 (2002).

[2] C. E. Lee, K. W. Lee, C. H. Lee and D. K. Oh, J. Korean Phys. Soc. **42**, 1182 (2003).

[3] J. Kim, Y. Park, B. Kahng and H. Y. Lee, J. Korean Phys. Soc. **42**, 162 (2003).

[4] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).

[5] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990); A. Sali, E. Shakhnovich and M. Karplus, Nature (London) **369**, 248 (1994); J. N. Onuchic, Z. Luthey-Schulten and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[6] A. R. Dinner and M. Karplus, J. Mol. Biol. **292**, 403 (1999).

[7] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).

[8] J.-L. Shea and C. L. Brooks, Annu. Rev. Phys. Chem. **52**, 499 (2001).

[9] E. Kussel, J. Shimada and E. Shakhnovich, Proc. Natl. Acad. Sci. USA **99**, 5343 (2002).

[10] E. M. Boczko and C. L. Brooks, Science **269**, 393 (1995).

[11] A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky and H. A. Scheraga, J. Comput. Chem. **18**, 849 (1997).

[12] J. Lee, K. Park and J. Lee, J. Phys. Chem. B **106**, 11647 (2002).

[13] J. Lee, H. A. Scheraga and S. Rackovsky, J. Comput. Chem. **18**, 1222 (1997).

[14] J. Lee, S.-Y. Kim and J. Lee, J. Phys. Chem. B, in press.

[15] S.-Y. Kim, J. Lee and J. Lee, J. Chem. Phys., in press.