Protein three-dimensional structure generation with an empirical hydrophobic penalty function

Kazunori Toma

Computer Science Department, Asahi Chemical Industry Co. Ltd., Shizuoka, Japan

Given current computational environments, it is worthwhile to establish amino acid residue-level functions which approximate protein folds quite well. Such functions must be the interim steps toward protein three-dimensional structure prediction. I have shown that an empirical hydrophobic penalty function of protein, derived from the number of residues in a sphere around each residue, could be utilized to distinguish the correctly folded structure from the incorrect ones. In order to assess the predictive power of the penalty function, I had generated conformations by randomly changing main chain dihedral angles, and applied the penalty function to them. If only a local region was allowed to change its conformation, nativelike structures could be generated within a reasonable computational time. In global simulations, however, a considerable number of nonnative conformations, which gave as small a penalty value as that of the native protein, were found. Although some of the conformations were compact and globular, they were quite different from the native structure in that they lacked most of the secondary structures. This result shows that the penalty function alone cannot define the native structure, and that substructure information may help the penalty function to reach the correctly folded structure.

Keywords: protein structure prediction, α-carbon model, off-lattice model, hydrophobic penalty function, residues in a sphere

INTRODUCTION

Prediction of the protein tertiary structure from its amino acid sequence alone is one of the unsolved fundamental problems in molecular biology. While practical demands for such a method are always increasing, the problem has remained to be answered for about three decades. The crucial obstacles at present are the computational time limit, which does not allow us to elaborate the vast conformational space, and the lack of a proper potential function to evaluate the protein fold.[1,2]

Even today, the computational time limit makes atomic detailed empirical potential energy functions impractical in predicting protein structures. To surmount this limitation, several simplified models of protein have been proposed.[3] Among them, lattice models are rather successful.[4–8]

More than a decade ago, Gō and his coworkers,[4] and more recently, Dill and his coworkers[5] used square or cubic lattice models to reveal some of the fundamental aspects of protein folding. Their models, however, are too simple to be incorporated in the protein structure prediction scheme.

More realistic lattice models were introduced by Skolnick and his coworkers to study the folding of various types of proteins.[6] They successfully refolded proteins from extended structures, simply because their models were designed so as to fold proteins. Thus, their models must be classified as a kind of folding potential, and cannot be modified to the predictive one with ease.

Covell and Jernigan,[7] and Hinds and Levitt[8] employed lattice models in a more predictive way. They exhaustively searched compact folding conformations on the lattice model and assessed the resultant structures with a packing or contact potential function. They claimed that nativelike structures could be found in a small portion of conformations which gave good energy values.

Although those successes of lattice models are promising, it is necessary to go beyond them, if a more realistic protein structure prediction is pursued. Since it was pointed out that lattice models might favor the formation of secondary structure,[9] we may encounter difficult problems once we go into a continuous model, which should be an interim step toward a precise atomic model.

Because such off-lattice models are simpler in spirit, they were first reported more than a decade ago,[10–12] and they have been recently revisited, as computers have become more powerful.[13] The most fundamental problem of this approach is that there is no proper potential function known. The problem is so profound that the best result obtained by this type of approach did not come closer to the native structure than 4.0-Å root-mean-square displacement (RMSD).[13] This level of accomplishment simply means that the resultant structure is somewhat globular.[14]

In the previous paper, I proposed a new empirical penalty

Color Plates for this article are on page 234.

Address reprint requests to Dr. Toma at the Computer Science Department, Asahi Chemical Industry Co. Ltd., 2-1, Samejima, Fuji, Shizuoka, Japan.
Received 8 December 1992; revised 29 January 1993; accepted 2 February 1993

function (RIS-PF; *RIS* stands for *residues in a sphere*), using the number of amino acid residues in a sphere around each residue.[15] Although Nishikawa and Ooi had used the idea to predict the radial distribution of amino acids from the sequence information,[16,17] my proposal was the first attempt to apply it directly in the three-dimensional context. The RIS-PF was shown to be able to distinguish the native fold from incorrect ones.

As an extension of the former study, and also as an attempt to use RIS-PF in protein structure prediction, I would like to report here a combinational approach of the penalty function with an off-lattice model.

METHODS

Penalty function

Though described in the former paper,[15] some of the important points are briefly summarized in the following. The RIS of a sphere size is calculated as the number of residues in the sphere of a defined radius around a given residue. The position of each residue was represented by the Cα coordinate. All residues other than the central one were counted. Radii from 6 Å to 14 Å, with a 1-Å increment, were used to generate the RIS values for each sphere size. The standard RIS value for each amino acid was defined as the average of real values of the 27 data set protein structures taken from the Protein Data Bank (PDB).[18] The standard deviation (sd) for each standard value was also calculated.

The penalty value for each radius size is defined by the following formula.

$$\text{Penalty Value} = \sum_{i}^{N} (|\text{RIS}_i^{\text{cal}} - \text{RIS}_{aa}^{\text{stand}}|/\text{sd}_{aa})$$

where RIS^{cal} denotes the RIS of a generated structure, and $\text{RIS}^{\text{stand}}$ does that of the standard value for amino acids. The native structure gave a smaller penalty value than incorrect ones did.

Structure generation

For the trial coordinate generation, the main chain model of proteins was adopted. Although only Cα coordinates were used in the RIS calculations, it must be noted that the coordinates of N and C were also calculated. In the local simulations, in which only a limited number of residues were allowed to change their conformations, the crystal structure was used without adjusting bond lengths and bond angles, while the standard ALA residue coordinates and the C-N-Cα bond angle of ECEPP[19] were employed in the global simulations, in which all residues could alter their conformations. From the consideration of protein size dependence of RIS-PF,[15] human lysozyme,[20] which consists of 130 residues and is out of the data set, was taken from PDB as a model protein.

One simulation step consists of three parts. At first, one residue is randomly selected from the flexible part, or from the whole sequence, to which a conformational change should be applied.

Then, new dihedral angles, ϕ and ψ, are assigned to the residue. Each ω is set to 180°. The notation of dihedral angles follows the IUPAC-IUB nomenclature.[21] The ϕ–ψ pair is taken so as to be roughly in the allowed region of the Ramachandran preference,[22] and PRO and GLY are treated separately from other residues, as shown in Figure 1. The squares on the Ramachandran map were equally weighted.

The resultant new conformation is examined using distances between Cα atoms, first with the residue clash criterion. If some residues come closer than a defined length (typically 3.0 Å), the structure is abandoned, and the simulation goes back to the first residue selection part. If a starting conformation itself has some structural overlaps already, this part is modified so that a new conformation with a smaller number of clashes than the previous structure is taken. Because RIS values of small radii are incorporated in the next part of the simulation, the clash criterion may seem unnecessary. If residue clashes are ignored, however, folded structures can become nonsensical flat ones.

The new structure, which goes through the clash stage, then proceeds to the RIS-PF part. The RIS penalty values of nine different radii and their average are calculated. If more than a certain number of them (typically five) give better values than the previous structure, the new conformation is kept for further simulation steps. Even if a structure does not satisfy this condition, it is adopted with a certain probability to escape from local minima. A random number is generated, and if the number falls into a given range (typically 1.0%), the structure survives for a further simulation step. After this part, the simulation goes back to the first part.

Since there are several adjustable parameters in the simulation, some of them are changed and examined, as discussed in the following sections.

Characterization of results

Other than RIS-PF itself, generated conformations were characterized by RMSD from the native structure after two structures were optimally superimposed.

$$D = \frac{\sqrt{\sum_{i}^{N} (r_i^{\text{cal}} - r_i^{\text{nat}})^2}}{N}$$

where N is the total number of residues, r^{cal} is the Cα coordinate of the generated structure and r^{nat} is that of the native one.

Since several authors reported the goodness of their results by the RMSD of the following definition, this D' was also calculated.

$$D' = \sqrt{\frac{\sum_{i}^{N} \sum_{j \neq i}^{N} (d_{ij}^{\text{cal}} - d_{ij}^{\text{nat}})^2}{N(N-1)}}$$

where d_{ij}^{cal} denotes the distance between ith and jth residues of the simulated structure, and d_{ij}^{nat} does that of the native one. Although D and D' were quite different in the absolute values, their behaviors were almost the same, as can be seen in the results. Thus, D' is shown in the figures only for a reference.

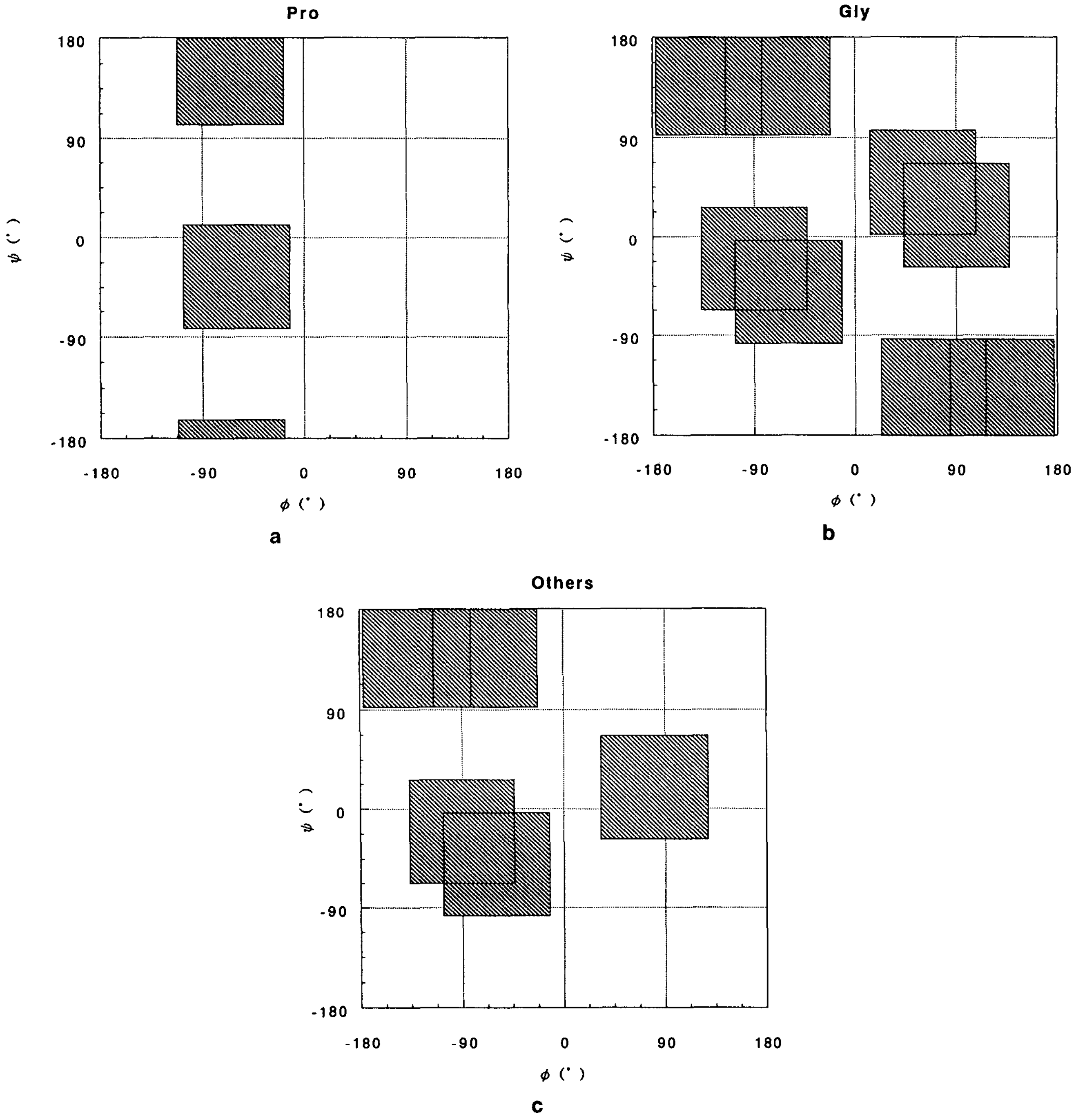

Figure 1. Allowed regions in the φ and ψ map: for PRO (a), for GLY (b), and for other residues (c).

Radii of gyration were also calculated to see the compactness of the simulated structures.

$$R = \frac{\sum_{i}^{N} \sqrt{(r_i - r_0)^2}}{N}$$

where r_0 is the coordinate of the center of mass. Only Cα coordinates were used without any weights.

Most of the calculations were performed on either a SiliconGraphics 4D/35 or a Hewlett Packard 9000/750 workstation, and tube models were displayed by using the MOLGRAPH software.[23]

RESULTS

Local simulation

The simplest procedure would be to make only one residue flexible. Starting from the native structure, the φ and ψ dihedral angles of one selected residue were changed. Since

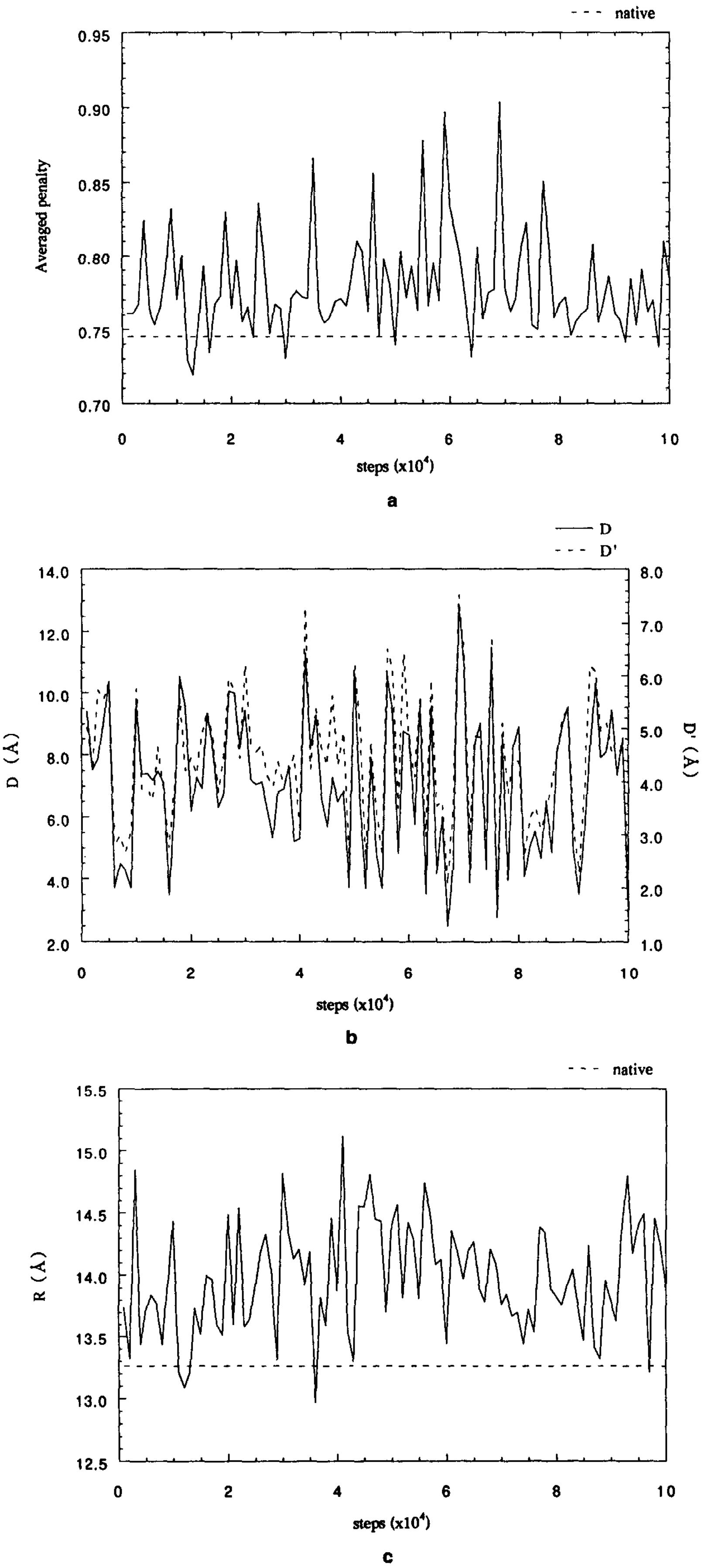

Figure 2. Local simulation of ten consecutive residues shown in trajectories: of averaged RIS penalty value (a), of RMSD (b), and of radius of gyration (c). Each point corresponds to the structure giving the lowest penalty over 1000 steps of the simulation.

the simulation was allowed to take randomly some conformations which became worse in penalty values, generated conformations sometimes deviated from the native structure. Most of the structures which gave good penalty values, however, were fairly close to the native one in this type of simulation.

The situation was the same even if the flexible part was extended to two or three consecutive residues. Within a short computational time, native-like structures were frequently observed. Then, the simulation went on to the consecutive 10 residues. Although this is superficially a simple extension of the trivial simulations, it has some practical meanings. If the domain structures of a protein were fairly well predicted, a short junction of less than 10 residues would be enough to link them. Thus, the local simulation procedure could be utilized as a step in protein structure prediction.

Several simulation conditions were also examined at this stage, because a typical 10^5 step simulation finished in a reasonable computational time. The clash radius was changed from 2.0 Å to 3.5 Å, with 0.5-Å increments. While there was no clear difference in the performance if 3.0 Å or 3.5 Å was chosen, smaller radii gave poorer results. The problem of how often bad conformations are allowed to survive must be considered together with the number of RIS values of different radii that should become better. A ratio of 1% gave reasonable access to both nonnative and nativelike structures, if the RIS number was set to 5. For the same RIS number, a bad conformation ratio of 5% made conformations deviate away from the native structure. As this criterion is related to the simulation temperature of the Monte Carlo methods, this result seemed reasonable.

Simulation runs first started from the native structure. The clash radius was set to 3.0 Å, the RIS number to 5, and the bad conformation allowance to 1%. During a 10^5 step run several conformations came close to the native structure. The flexible part was changed so as to examine the positional dependence. Although there were small positional differences, all trials gave similar, successful results, indicating that RIS-PF could guide the conformation to nativelike structures.

Because these results could be biased by the initial structure of the native one, simulations were also started from structures in which the flexible part was extended. The result of a typical run that allowed residues 100 to 109 to move is shown in Figure 2. In this trajectory, each point is the representative of 10^3 steps of the simulation, which gave the best average RIS penalty value during the period. It is easy to discern that nativelike structures were generated several times during this simulation, and that the structure sometimes drifted from the globular nativelike structures.

For visual help, the native structure, the initial structure, and the best RMSD structure are shown together in the Color Plates 1 and 2, in which the flexible part is highlighted. In this particular case, structural overlaps can be seen in the starting structure. The best RMSD structure looks very much like the native structure, although the flexible part itself takes a different conformation.

One of the proposed structure prediction schemes is to build up secondary structures into the tertiary one.[24] Therefore, more realistic local simulations would make several regions flexible. The result of one such simulation, in which three parts were moved, is shown in Figure 3. The three parts were residues 24–27, 55–58, and 98–101, corresponding to turn structures, and the simulation condition was the same as in the above simulation. One of the generated conformations could come as close as 6.5 Å to the native structure, indicating that the present method can be fairly practical. If the number of flexible parts was increased, however, the computational time soon became a real problem.

Global simulation

Since the results of the local simulations were promising, global simulations were also investigated in the hope that some nativelike conformations could be reached within a limited computational time. Starting from the extended structure, 2×10^5 steps of simulation were pursued with the typical condition. Although several cases of adjustable parameters were examined, there was no clear progress in the results.

Figure 4 shows the result of a 2×10^5 step simulation. From every 10^3 steps one conformation which gave the best average RIS penalty value was saved. From Figure 4b, it is obvious that no nativelike fold appeared in the simulation run. A further important point, however, is that there were conformations which gave similar RIS penalty values and compactness to the native structure, as can be seen in Figures 4a and 4c. This result means that the present formulation of RIS-PF is not powerful enough to specify only nativelike structures as the best ones.

To investigate the results further, sampled conformations were investigated using molecular graphics. Two representative conformations which gave relatively good RIS penalty values are shown in Color Plate 3. When compared with the native structure of Color Plate 1, the lack of regular secondary structures is easily discernible.

Then a simulation run from the native structure was performed to see whether the same result as that from the extended structure was obtained. The condition was the same as that of the former simulation, except that the native structure was used as a starting conformation. The results of the simulation shown in Figure 5 show no difference from Figure 4 other than the very first part, where the native structure collapsed rapidly.

Since the probability that allows a bad conformation should correspond to the temperature, the 1% condition may be deemed as a high-temperature simulation, so that the denaturing proceeded along the simulation. If 0% was taken as the allowance ratio and the penalty values of all radii had to become smaller—meaning no bad conformation was allowed—the deviation from the native structure was very small, giving a validation of the treatment.

DISCUSSION

Some of the advantages of the present coordinate generation method over the lattice models must first be summarized. Since the N and C atom coordinates are implicitly calculated in each simulation step, the model can easily be extended to an atomic detailed model. Also, since the dihedral angles are generated based on the Ramachandran-like preferences, the model incorporates handedness in a very natural way,

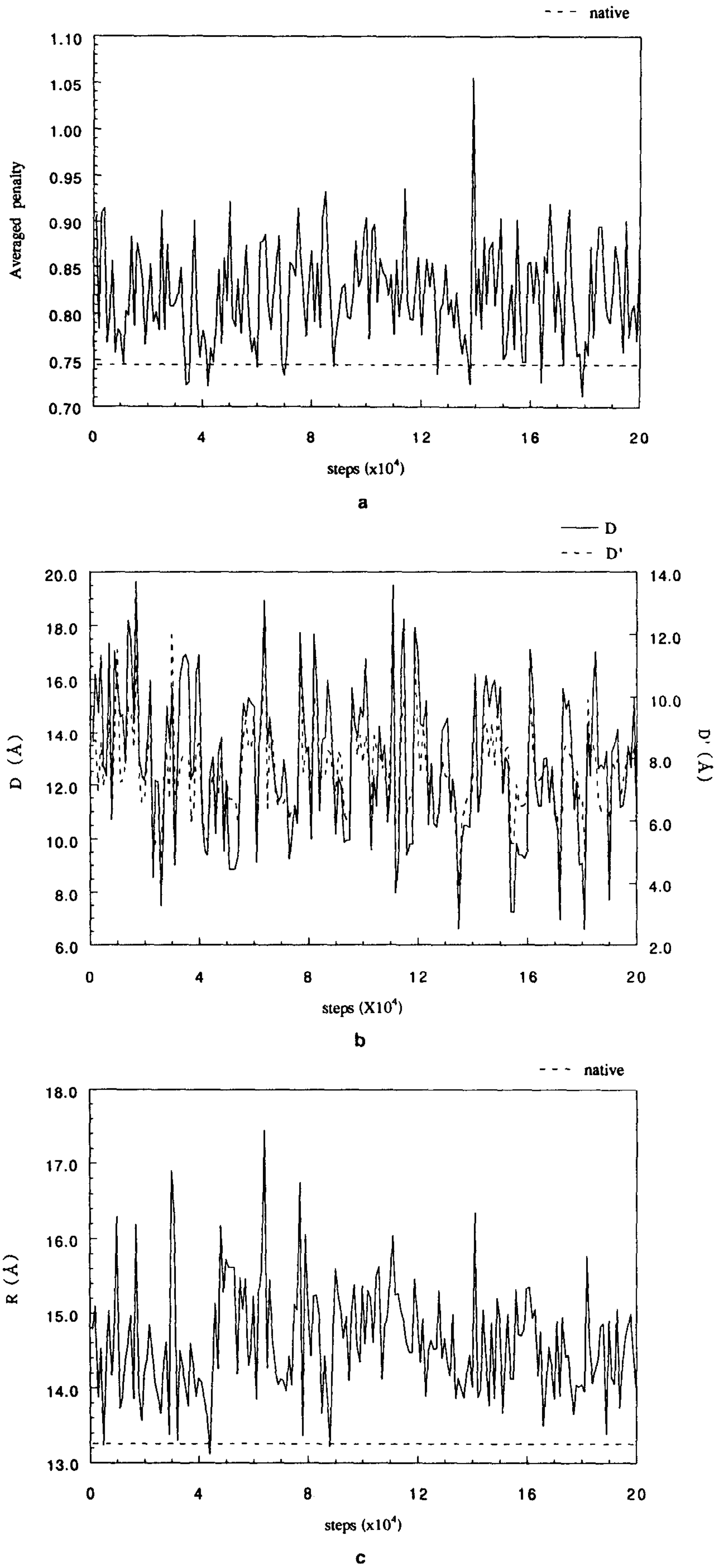

Figure 3. Local simulation of three turns shown in trajectories: of averaged RIS penalty value (a), of RMSD (b), and of radius of gyration (c).

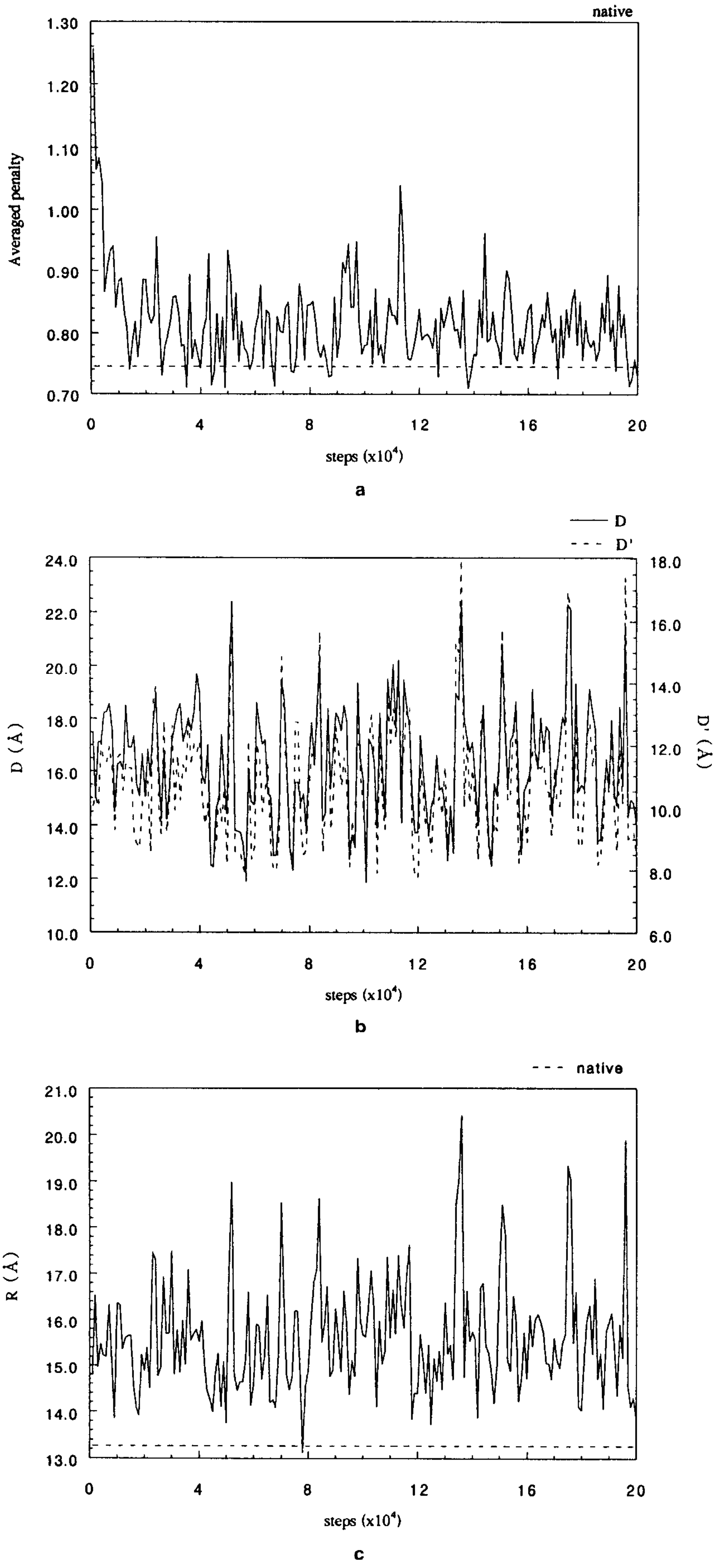

Figure 4. Global simulation starting from the extended conformation shown in trajectories: of averaged RIS penalty value (a), of RMSD (b), and of radius of gyration (c).

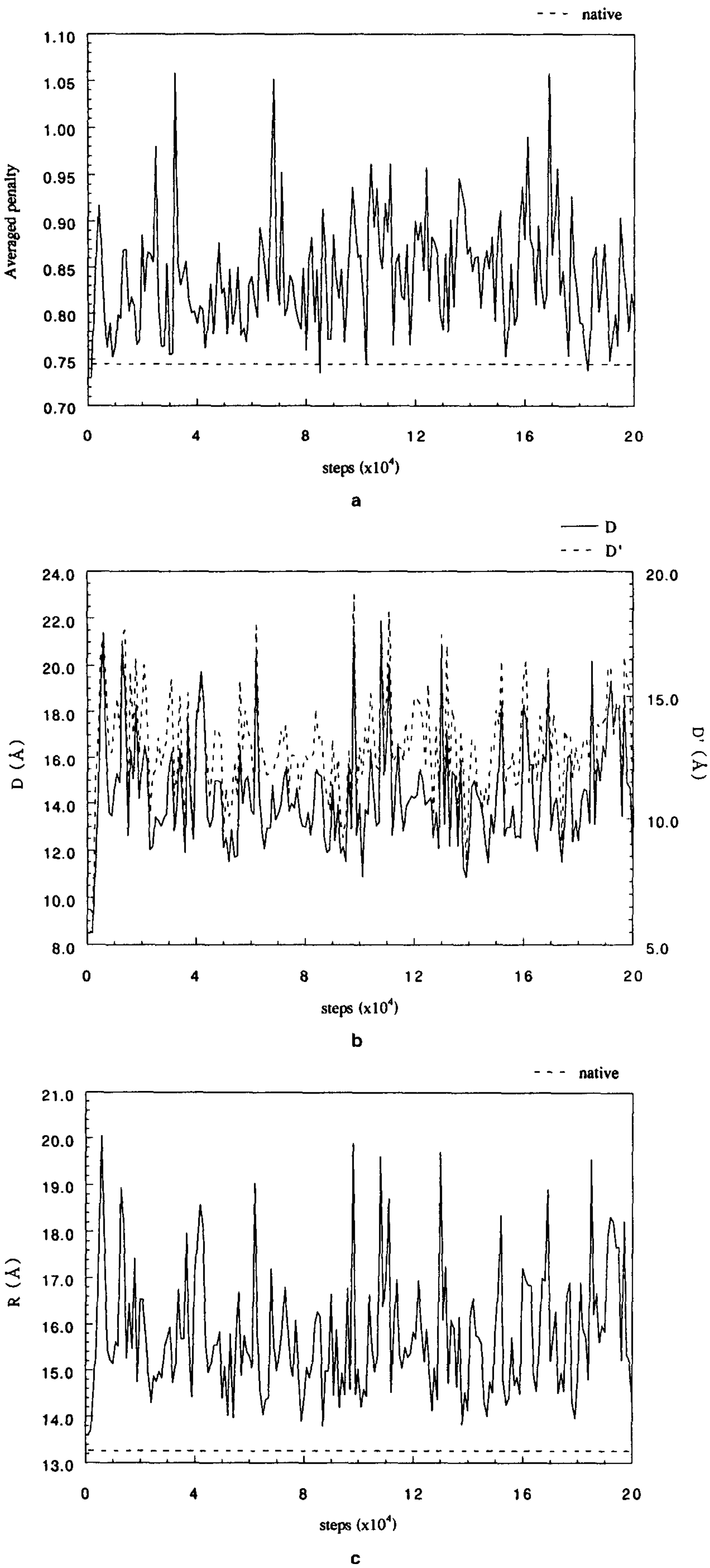

Figure 5. Global simulation starting from the native structure shown in trajectories: of averaged RIS penalty value (a), of RMSD (b), and of radius of gyration (c).

and the mirror image fold is discriminated. Those points must be overcome when we go from lattice models to atomic models.

With the molecular graphics program ALPHA,[25] handmade partially unfolded structures had been constructed, and were assessed with RIS-PF in the previous paper.[15] Except for domain structures, RIS-PF could distinguish the native structure. From that result, successful refolding in the local simulations could be foreseen. The result may be biased by the partial structures which were kept intact from the native structure, since those parts must bear much information about the native packing. Even with such aids, the flexible part in the simulation could not come close to the native conformation, which indicated the limitation of the RIS-PF method. In must be noted, however, that the reconstruction of nativelike structures lends some credit to RIS-PF in the real use. Since a 10^5 step simulation took about 5 hours on a Hewlett Packard workstation, it is hopeful that RIS-PF can be used in the packing procedure of well-predicted partial structures, like an additional step from secondary structure predictions.[24]

Although the local simulation could be pursued within a reasonable time, it was clear that a simple extension of it must confront the computational time limit. Thus, the global simulation was rather an assessment of RIS-PF itself than a protein three-dimensional structure prediction. The global simulations from both the extended and the native structures showed that nonnative structures could give as small RIS penalty values and radii of gyration as the native structure, revealing some significant problems of RIS-PF.

Although one of the important advantages of RIS-PF over the conventional pairwise potentials is the consideration of different radii in a discrete way, the present treatment does not fully utilize this advantage. As pointed out by Behe et al., it is unrealistic to model a pairwise distribution in a simple function.[26] Nishikawa and Ooi had shown that radius of 14 Å was better than 8 Å to predict radial distribution of amino acid residues from a sequence,[17] which was the first notion that RIS of different radii behaved differently. From the present simulations like the one shown in Figure 6, it was clearly demonstrated that RIS of small radius and of large one behaved differently, while RIS of similar radii showed similar tendency. This result indicates that the results of the simulations were a compromise between shorter and longer range interactions. Incorporation of this balance more explicitly into the simulation and also the refinement of other adjustable parameters must be investigated in future studies.

A more significant problem is the shape of the penalty function. Covell and Jernigan,[7] and more recently, Hinds and Levitt[8] showed that a simple contact energy criterion proposed by Miyazawa and Jernigan[27] could not specify nativelike structures among about 10^3 to 10^5 structures. Although the formulation is different, the RIS approach is similar in using the contact number as a main factor of the folding, especially a RIS value of 8 Å. Wako and Kubota used a RIS value 14 Å in their characterization of the structures resulting from the distance-constraint approach for the structure prediction.[28] They showed that the RIS values were comparable with their function, which means a RIS value 14 Å is as bad as the distance constraint. Those results seem to arise from the same problem. From the resultant structures shown in Color Plate 3, it was obvious that simulated structures lacked the ordinary secondary structure elements. This result can be understood to come from the fact that RIS-PF and other potentials are spherical in nature, so that they cannot distinguish the cases shown in Figure 7, where thick lines represent local structures.

Another important problem is the computational time. Because Chan and Dill[28] claimed that packing was the major factor in determining secondary structures, there was a little hope that some of the compact structures given by the

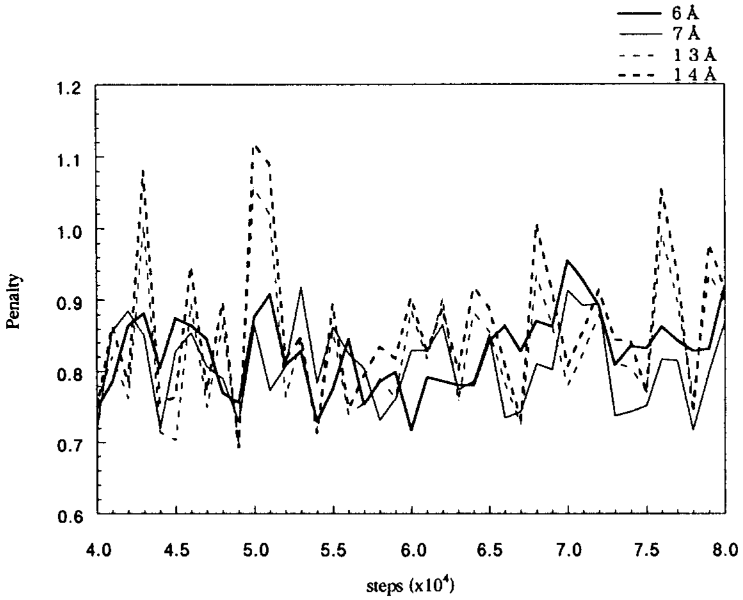

Figure 6. Close-up of a trajectory to show RIS-PF behaviors of various radii: 6 Å (thick line), 7 Å (thin line), 13 Å (thin dotted line), and 14 Å (thick dotted line).

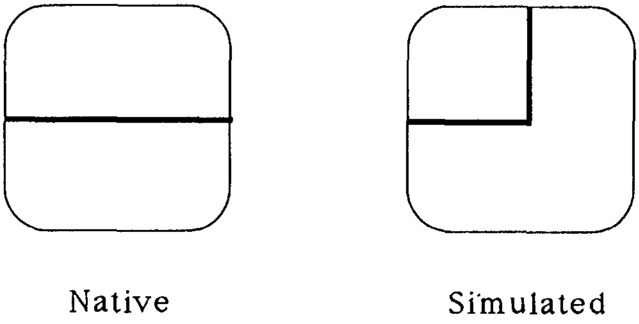

Figure 7. Fundamental problem of spherical penalty functions.

present simulation might contain secondary structures automatically. This expectation was simply wrong, as discussed already. On the other hand, Gregoret and Cohen pointed out that secondary structure formation can be biased by the lattice treatment itself.[9] Because the global assessment of RIS-PF is far from perfect at present, I should refrain from jumping to a conclusion. It seems, however, that the compactness alone does not promise a computationally practical scheme of the protein structure prediction. It is interesting to point out that even a very successful folding model needs a preference of local secondary structure for a simulation to finish in a reasonable computational time.[6]

Because the computational time depends on the vastness of the conformational space to be searched, the structure generation method cannot be discussed separately from the development of the penalty function. Since my intention is to make an off-lattice model, the most practical alternative would be the database type approach that I proposed in the previous paper.[15] Some studies which dealt with similar database searches in the prediction have appeared since then.[30,31] Because the database method limits the search space and introduces secondary structure information, the approach is now considered as the next step of the present study, together with the refinement of RIS-PF itself.

ACKNOWLEDGMENTS

The author thanks Dr. N. Gō for his helpful comments, and Ms. T. Kamata for her technical assistance.

REFERENCES

1 Gō, N. Theoretical studies of protein folding. *Ann. Rev. Biophys. Bioeng.* 1983, **12,** 183–210

2 Jaenicke, R. Protein folding: Local structures, domains, subunits, and assemblies. *Biochem.* 1991, **30,** 3147–3161

3 Troyer, J.M. and Cohen, F.E. Simplified models for understanding and predicting protein structure. *Rev. Comput. Chem.* 1991, **2,** 57–80

4 Gō, N., Abe, H., Mizuno, H., and Taketomi, H. In: *Protein Folding.* (N. Jaenicke, Ed.) Elsevier, Amsterdam (1980) pp. 167–181, and references therein

5 Yue, K. and Dill, K.A. Inverse protein folding problem: Designing polymer sequences. *Proc. Natl. Acad. Sci. USA.* 1992, **89,** 4163–4167, and references therein

6 Godzik, A.G., Skolnick, J., and Kolinski, A. Simulations of the folding pathway of triose phosphate isomerase-type α/β barrel proteins. *Proc. Natl. Acad. Sci. USA.* 1992, **89,** 2629–2633, and references therein

7 Covell, D.G. and Jernigan, R.L. Conformations of folded proteins in restricted space. *Biochem.* 1990, **29,** 3287–3294

8 Hinds, D.A. and Levitt, M. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA.* 1992, **89,** 2536–2540

9 Gregoret, L.M. and Cohen, F.E. Protein folding: Effect of packing density on chain conformation. *J. Mol. Biol.* 1991, **219,** 109–122

10 Levitt, M. and Warshel, A. Computer simulation of protein folding. *Nature.* 1975, **253,** 694–698

11 Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 1976, **104,** 59–107

12 Kuntz, I.D., Crippen, G.M., Kollman, P.A., and Kimelman, D. Calculation of protein tertiary structure. *J. Mol. Biol.* 1976, **106,** 983–994

13 Wilson, C. and Doniach, S. A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins.* 1989, **6,** 193–209

14 Hagler, A.T. and Honig, B. On the formation of protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. USA.* 1978, **75,** 554–558

15 Toma, K. Number of residues in a sphere around a certain residue can be used as a hydrophobic penalty function of proteins. *J. Mol. Graphics.* 1991, **9,** 78–84

16 Nishikawa, K. and Ooi, T. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int. J. Peptide Protein Res.* 1980, **16,** 19–32

17 Nishikawa, K. and Ooi, T. Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J. Biochem.* 1986, **100,** 1043–1047

18 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542

19 Scheraga, H.A. and Paine, G.H. Conformational energy calculations on polypeptides and proteins: Use of a statistical mechanical procedure for evaluating structure and properties. *Ann. N.Y. Acad. Sci.* 1986, **482,** 60–68

20 Artymiuk, P.J. and Blake, C.C.F. Refinement of human lysozyme at 1.5-Å resolution. *J. Mol. Biol.* 1981, **152,** 737–762

21 IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chain: tentative rules (1969). *J. Biol. Chem.* 1970, **245,** 6489–6497

22 MacArthur, M.W. and Thornton, J.M. Influence of proline residues on protein conformation. *J. Mol. Biol.* 1991, **218,** 397–412

23 MOLGRAPH is a molecular graphics software package licenced from Daikin Industries, Ltd. (Tokyo)

24 Saito, N., Shigaki, T., Kobayashi, Y., and Yamamoto, M. Mechanism of protein folding: I. General considerations and refolding of myoglobin. *Proteins.* 1988, **3,** 199–207

25 Toma, K. Simple protein model building tool. *J. Mol. Graphics*. 1987, **5,** 101–102

26 Behe, M.J., Lattman, E.E., and Rose, G.D. The protein-folding problem: The native fold determines packing, but does packing determine the native fold? *Proc. Natl. Acad. Sci. USA*. 1991, **88,** 4195–4199

27 Miyazawa, S. and Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasichemical approximation. *Macromol.* 1985, **18,** 534–552

28 Wako, H. and Kubota, Y. Distance-constraint approach to higher order structures of globular proteins with empirically determined distances between amino acid residues. *J. Protein Chem.* 1991, **10,** 233–243

29 Chan, H.S. and Dill, K.A. Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA*. 1990, **87,** 6388–6392

30 Crippen, G.M. Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochem.* 1991, **30,** 4232–4237

31 Jones, D.T., Taylor, W.R., and Thornton, J.M. A new approach to protein fold recognition. *Nature*. 1992, **358,** 86–89