Journal of the Korean Physical Society, Vol. 46, No. 3, March 2005, pp. 707∼712

Protein Structure Prediction Based on Fragment Assembly and the β-Strand Pairing Energy Function

Julian Lee

Department of Bioinformatics and Life Sciences,
Computer Aided Molecular Design Research Center,
Bioinformatics and Molecular Design Technology Innovation Center, Soongsil University, Seoul 156-743

Seung-Yeon Kim and Jooyoung Lee*

School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722

(Received 12 October 2004)

We develop an improved version of PROFESY, a novel method for ab-initio prediction of protein tertiary structures based on fragment assembly and global optimization. In contrast to the primitive version presented earlier, where the hydrogen bond was defined only in terms of inter-atom distance, the angle dependence is now correctly incorporated. This new feature allows us to obtain low-energy conformations with a reasonable number of beta strands, in contrast to the earlier version in which the fraction of alpha helices was excessively large on average. In order to enhance the performance of the prediction method, we optimize the linear parameters of the energy function so that native-like conformations become energetically more favorable than non-native ones for proteins with known structures. We test the feasibility of the parameter optimization procedure by applying it to a training set consisting of two proteins of structural class $\alpha + \beta$: 1FSD and 1PQS. We use the resulting parameter set for the jackknife test on several proteins from various structural classes. The results are quite promising. In particular, for protein 2GB1, the predictions improve dramatically with the optimized parameter set compared to the original parameters, despite the fact that 2GB1 was *not included in the training set.* This result suggests that parameters trained for a relatively small number of proteins are transferrable to other proteins to some extent.

PACS numbers: 87.15.Aa, 87.15By, 87.15.-v
Keywords: Protein folding, Tertiary structure prediction, *Ab-initio* prediction, Fragment assembly, Global optimization, Parameter optimization

I. INTRODUCTION

The prediction of the unique tertiary (three - dimensional) structure of a protein from its amino-acid sequence alone is one of the most important and challenging problems in biological science today. Information on the tertiary structure of a protein is quite crucial in understanding the function and the biological role of the protein. Popular approaches to this problem include comparative modeling [1–5] and fold recognition [6–9], which are classified as knowledge-based methods [5, 10, 11]. These methods use statistical information on sequences and their three-dimensional structures in structural databases such as Protein Data Bank (PDB) to predict the unknown structure of a protein. Obviously, these methods can be used only when the amino-acid sequence of a target protein with unknown structure is related to those of one or more proteins with known structures.

On the other-hand, when homologous or weakly homologous sequences with known structures are not available, we turn to new-fold (or *ab-initio*) methods [5, 12–26], which include energy-based methods. Energy-based methods are based on the thermodynamic hypothesis [27] that the native structure of a protein corresponds to the global minimum of its free energy for its physiological environment. However, although much progress has been made in energy-based methods [13–16], successful prediction of protein structure solely from the potential energy function still remains as a challenging problem. For this reason, most recent new-fold prediction methods use information on known structures to some degree. One of the popular trends among such methods is to determine the tertiary structure of a target protein by assembling fragments generated from the PDB. The effect of short-range interactions are incorporated by using fragments from the PDB, and only long-range interaction terms are included in the energy function, which are minimized in order to find conformations with optimal tertiary pack-

*E-mail: jlee@kias.re.kr

ing [17–20].

As in pure energy-based methods, there are two crucial elements for successful application of the fragment-based method for structure prediction, which are an accurate energy (or score) function and a powerful optimization algorithm for finding low-energy conformations. In previous papers [25, 26], we introduced a fragment-based protein structure prediction method PROFESY (PROFile Enumerating SYstem), which utilized the fragment library obtained from the secondary structure prediction method PREDICT (PRofile Enumeration DICTionary) [28,29]. In contrast to earlier methods where only simple sampling algorithms, such as the simulated annealing method, were used for generating low-energy conformations, PROFESY applied a powerful global optimization algorithm, conformational space annealing (CSA) [30–34], for sampling low-energy conformations. However, the energy function used was rather primitive, and various parameters in energy terms were set by crude guesses. Although some promising results were obtained from the benchmark tests of PROFESY, which were believed to be mainly due to the high efficiency of the sampling method, it was necessary to construct a reasonably accurate energy function for successful application of PROFESY to protein structure prediction.

This problem was partly addressed in Ref. [26], where the effect of the solvent water, which was absent in the earliest version of PROFESY, was incorporated in an indirect manner. Also, parameter optimization was carried out using a training set consisting of three proteins. However, the hydrogen bonding term introduced in order to facilitate the β- strand pairing between extended fragments had a serious defect in that the hydrogen bond was assumed to exist between any hydrogen and oxygen whose distance was shorter than a certain cutoff, regardless of the hydrogen bonding angles. Furthermore, the hydrogen bond was allowed for any residues which were far away in the sequence, regardless of their secondary structures, which diminished the performance of this energy term in distinguishing correctly paired beta-strands. Therefore, with the old version of PROFESY, it was extremely difficult to obtain low-energy conformations with beta-strand pairings.

In this work, we remedy the situation by introducing a new form of hydrogen bond term with a correct angular dependence. We also allow a hydrogen bond only between extended fragments. Because the correct angular dependence of the hydrogen bonding is expected to remove any hydrogen bonding between alpha helices or coils, by allowing hydrogen bonds between extended fragments only, we can save much computational time.

To test the new form of the energy function, we optimize the relative weights of the energy terms by using training set consisting of two proteins, 1FSD and 1PQS, that belong to the structural class of $\alpha + \beta$ proteins. We tested the optimized parameters on several proteins which are not included in the training set and obtained promising results. In particular, for protein G (2GB1), the predictions with the optimized parameter set were dramatically improved compared to those with the original parameters, despite the fact that protein G was *not included in the training set.* The result suggests that parameters trained for a relatively small number of proteins are transferable to other proteins to some extent.

II. FRAGMENT ASSEMBLY

We first briefly describe the way one generates conformations by using fragment assemblies [25]. The fragment libraries used in PROFESY are constructed using the recently proposed secondary structure prediction method PREDICT [28,29]. For each residue of a query sequence, a window of size fifteen is considered, where the center of the window is located on the residue under consideration. The fragment library of this residue is the collection of twenty backbone structures of the corresponding twenty nearest patterns in the pattern database of PREDICT. After fragment libraries have been constructed for all residues of a query sequence, full-length chain conformations can be constructed by assembling the fragments in these libraries.

III. ENERGY FUNCTION

The energy function is given by

$$U = \sum_{i,j} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} \right) - w_h E_{\rm hb} + w_{MJ} E_{MJ}. \qquad (1)$$

The first term is the Lennard-Jones 6-12 van der Waals energy introduced to avoid steric clashes, where r_{ij} is the distance between the i-th and the j-th atoms. In order to incorporate possible quantum effects, we use separate values of the van der Waals interaction strength between the neighboring residues, which we call A' and B', which are not indicated in Eq. (1) for notational simplicity. In the second term, $N_{\rm hb}$ is the number of hydrogen bonds between residues, which is introduced to facilitate the β-strand pairing between extended fragments. Its functional form is given by

$$E_{\rm hb} = \sum_{i,j} f\left(\frac{r_{ij}^{(HO)} - r_h}{\Delta_h}\right) \frac{(\cos\theta_m - \cos\theta)}{\cos\theta_m + 1} \qquad (2)$$

when the angle formed by nitrogen, hydrogen, and oxygen, θ, is greater than $\theta_m = 146^\circ$, which is the minimum allowed value for this angle. This term is absent when $\theta < \theta_m$. Here, i, j are the residue indices, and $r_{ij}^{(HO)}$ is the distance between hydrogen and oxygen atoms in the

i-th and the j-th residues, respectively. The smoothed step function $f(x)$ is defined by

$$f(x) = \begin{cases} 1 & (x \le -1) \\ \frac{1}{2} - \frac{15}{16}x + \frac{5}{8}x^3 - \frac{3}{16}x^5 & (-1 < x < 1) \\ 0 & (x \ge 1). \end{cases} \quad (3)$$

We used this function instead of the sharp step function so that we can take the derivative with respect to the parameters, which is required for optimizing the parameters. The numerical coefficients in the polynomial are determined so that the function $f(x)$ and its derivative are continuous up to second order. The last term in Eq.(1) is a Miyazawa-Jernigan-type contact term [35, 36] between C_β atoms and is introduced to incorporate the solvation effect in an indirect manner. Its functional form is given by

$$E_{MJ} = \sum_{i,j} f(\frac{r_{ij} - r_c}{\Delta_c}), \quad (4)$$

where r_{ij} is the $C_\beta - C_\beta$ distance. The definition for the smoothed step function $f(x)$ is the same as above. We optimize a total of 10 parameters: A, B, A', B', w_h, w_{MJ}, r_h, r_c, Δ_h, and Δ_c.

IV. LOCAL MINIMIZATION

In order to apply the conformational space annealing (CSA) method to the fragment assembly of a protein tertiary structure prediction method such as PROFESY, one must define the concept of local minimization. In PROFESY, a conformation generated from a fragment assembly is locally minimized with respect to the energy by randomly selecting a residue and attempting to replace a part of the fifteen-residue-long fragment of the chain by another one in the corresponding library. If the new fragment can be inserted smoothly into the existing chain and if the new conformation is lower in energy than the existing one, the former replaces the latter. This process is continued either for 10 $N_{\rm seq}$ times, where $N_{\rm seq}$ is the length of the protein, or until the update attempts fail for $N_{\rm seq}$ consecutive times, whichever is encountered first, which completes the local minimization.

V. CONFORMATIONAL SAMPLING

The CSA [30–32] is a powerful global optimization algorithm that has played an integral role in the recent success of the energy-based method for protein structure prediction [13–16]. A population of local minimum-energy conformations are maintained in the CSA method, which is called the bank. The diversity of the bank is directly controlled in CSA by introducing a distance measure $D(A, B)$ between two conformations A and B and comparing it to a cutoff value, $D_{\rm cut}$. As the algorithm proceeds, $D_{\rm cut}$ is gradually reduced, playing the role of the temperature in simulated annealing. Hence, the name "conformational space annealing." The annealing of $D_{\rm cut}$ amounts to shifting the emphasis from the diversity of sampling at the early stage of the algorithm to obtaining low-energy conformations at later stages, enabling efficient sampling of low-lying local minimum-energy conformations.

In order to check the performance of a potential energy function for a given set of parameters, one has to sample native-like and non-native conformations for each protein in the training set. For this purpose, we perform two types of conformational search, local and global CSA searches. The global CSA search samples low-energy conformations with respect to the physical energy, Eq.(1), and with an unoptimized parameter set in the energy function, the resulting conformations are non-native. The native-like conformations are sampled by using a local CSA search, where the restraint function

$$F_{res} = \sum_{ij} |R_{ij} - R_{ij}^{(N)}| \quad (5)$$

is minimized instead of the physical energy function in Eq.(1). In this expression, the R_{ij} and $R_{ij}^{(N)}$ are the $C_\alpha - C_\alpha$ distances between the i-th and the j-the residues, for the generated conformation and the native conformation, respectively.

It should be noted that most of the native-like conformations obtained from the local CSA search are not local minimum of the physical energy function in Eq.(1). The conformations obtained from the global and the local CSA searches are added to the structural database of local minimum-energy conformations for each protein obtained from the earlier iteration of CSA searches (see section VII).

VI. PARAMETER REFINEMENT USING LINEAR PROGRAMMING

The procedure of parameter optimization, which is described in this section, is almost the same as that described in [26,37–39]. First, the changes in the energy gaps are estimated by using a linear approximation of the potential energy in terms of the parameters. Since a potential can be considered to describe nature correctly if a native-like structure has lower energy than non-native ones, the parameters are optimized to minimize the energy gap,

$$E_{\rm gap} = E^{\rm N} - E^{\rm NN}, \quad (6)$$

for each protein in the training set, where $E^{\rm N}$ and $E^{\rm NN}$ are the lowest energies of the native-like conformations and the non-native conformations, respectively. We add to the energy a term proportional to the root mean square deviation (RMSD) values of the conformations:

$$E = U + w_{\rm RMSD}\ {\rm RMSD}. \quad (7)$$

The additional term is introduced in order to make the conformations with large RMSD have high energies compared to ones with small RMSD values after the parameter optimization [39]. The numerical value of $w_{\rm RMSD}$ is determined at the initial stage of every iteration of the parameter optimization and depends on the energy scale of the conformations:

$$w_{\rm RMSD} = 0.5|E{\rm gap}|/{\rm RMSD}_0, \quad (8)$$

where $RMSD_0$ is the RMSD value of the conformation with the lowest energy.

The parameter optimization is carried out by minimizing the energy gap $E_{\rm gap}$ of each protein in turn while imposing the constraints that the energy gaps of the other proteins do not increase. Changing the parameters by small amounts, we can estimate the energy with the new parameters by using a linear approximation:

$$E(\mathbf{x_{min}}; p^{\rm new}) \approx E(\mathbf{x_{min}}; p^{\rm old}) + \sum_i (p_i^{\rm new} - p_i^{\rm old}) \frac{\partial E(\mathbf{x_{min}}; p^{\rm old})}{\partial p_i},$$

where the $p_i^{\rm old}$ and $p_i^{\rm new}$ 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 conformational space vanishes at a local minimum [37]. The additional term $w_{\rm RMSD}$ RMSD of Eq.(7) vanishes in these expressions for the same reason. 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 $\epsilon = 0.001$ for the linear parameters and $\epsilon = 0.00001$ for the nonlinear parameters in this study. The resulting optimization problem is a linear programming problem [39], which is solved by using the primal-dual method with supernodal Cholesky factorization [40]. We select each protein in the training set in turn and repeat this procedure (300 times in this work) of minimizing $\Delta E_{\rm gap}$.

VII. ITERATIVE REFINEMENT OF PARAMETERS

Since the change in the energy gaps after the parameter change was estimated using linear approximations, we now have to evaluate the true energy gaps by using the newly obtained parameter set. Therefore, we reminimize the conformations in the structural database with the new parameter set. We also perform the CSA search with the new parameters [38,39]. The low-lying local energy minima found in the new conformational searches are added to the 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 [38,39]. This iterative procedure is continued until sufficiently good native-like conformations are found from the CSA search.

VIII. RESULTS

We have applied our protocol to a training set consisting of two $\alpha + \beta$ proteins, Full Sequence Design 1 Of $\beta\beta\alpha$ Motif (PDB ID: 1FSD) and C-Terminal Opca Domain Of Ycdc24P (PDB ID: 1PQS). The initial parameter set for the hydrogen bond is chosen to be close to the physical values, $w_h = 23.3$ kJ/mol, $R_h = 3.1$ Å, $\Delta_h = 0.1$ Å, and those for the contact terms were taken from the work of Miyazawa and Jernigan [36]. The initial parameter set for the van der Waals interactions were determined more or less arbitrarily from previous experiences [25,26].

Fifty conformations were sampled in each CSA search, and the global minimum-energy conformations (GMECs) found with the initial parameters had RMSD values of 6.5 and 11.7 Å for 1FSD and 1PQS, respectively, and the smallest values of RMSD found from the CSA search were 4.4 and 8.0 Å, respectively. After the 6-th iteration of the parameter refinement, the GMECs resulting from the global CSA had RMSD values of 7.9 and 11.4, and the smallest values of RMSD found were 3.4 and 8.5 Å, respectively.

Indeed, the results above are not really impressive. In fact, the overall distributions of energy and RMSD show that the performance of the optimized parameters are more or less comparable to the original ones, and at this stage, the parameter optimization seems to be not so successful. However, we tested the performance of the parameters on several proteins *not included in the training set*, and surprisingly, we obtained promising results.

We performed tests on four proteins, Immunoglobulin-Binding B Domain of Staphylococcal Protein A (PDB ID: 1BDD), Thermostable Subdomain From Chicken Villin Headpiece (PDB ID: 1VII), H-Ns DNA-Binding Domain (PDB ID: 1HNR), and Immunoglobulin-Binding Domain of Streptococcal protein G (PDB ID: 2GB1). Two proteins 1BDD and 1VII belong to the structural class of α protein whereas the others are $\alpha + \beta$ proteins. For 1BDD, 1VII, 1HNR, and 2GB1, respectively, the global minimum-energy conformations (GMECs) found with the initial parameters had RMSD values of 8.2, 5.9, 7.5, and 9.1 Å, and the smallest values of RMSD found from the CSA search were 5.1, 4.5, 5.8 and 4.1 Å. Using the parameter set obtained after the 6-th iteration of the parameter refinement, the GMECs resulting from the global CSA had RMSD values of 6.1, 5.5, 10.3 and 2.4, and the smallest values of RMSD found were 4.6, 3.5, 6.9, and 1.3 Å, respectively. We find that the performance of the optimized parameters are either almost

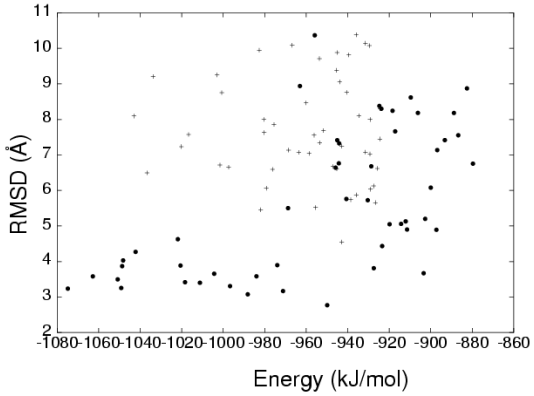

Fig. 1. Plots of the energy and RMSD values of low energy conformations obtained from CSA search using the initial and the refined parameters for the protein 2GB1. The crosses and the filled circles denote the results obtained using the parameters before the optimization and after the 6-th iteration, respectively.

the same as that of the original parameter set or only marginally improved for the three proteins 1BDD, 1VII, and 1HNR. However, for protein 2GB1, we see that the performance of the parameters is dramatically enhanced after the 6-th iteration of the optimization. This result can be more graphically shown Fig. 1, where the results of the global search with the initial and the optimized parameter sets for the protein 2GB1 are plotted in terms of the energy and the RMSD.

IX. DISCUSSION

In this work, we have improved PROFESY, a novel method for the prediction of protein tertiary structures based on fragment assembly, by introducing a new type of hydrogen bond energy term and systematically optimizing the energy parameters. The parameter optimization was performed by applying the general protocol for force-field parameter optimization and landscape design, which have been used previously only in the context of the pure energy-based method. Using this procedure, we optimized 10 parameters so that they correctly described the energetics of two selected proteins simultaneously. We found that the parameter optimization was not so successful for the proteins in the training set, because the performance of the parameters after the optimization procedure were more or less the same as that of the original ones. However, after applying the optimized parameters for four other proteins not included in the training set, we saw that the optimized parameter set was not worse than the original one, which itself is a nontrivial result. More interestingly, we found that the performance of the parameter set was dramatically enhanced for the protein 2GB1, despite the fact that it was *not included* in the training set. This result suggests that in the process of optimizing the parameters for the proteins in the training set, we obtained parameters which described the protein 2GB1 much better than the original parameters did, although their performance is comparable to that of the original ones for the proteins in the training set. The reason behind this curious fact remains a subject for the further investigation.

ACKNOWLEDGMENTS

This work was supported by grant 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] A. Fiser, R. K. G. Do and A. Sali, Protein Sci. **9**, 1753 (2000).
[2] A. Kolinski, M. R. Betancourt, D. Kihara, P. Rotkiewicz and J. Skolnick, Proteins **44**, 133 (2001).
[3] P. A. Bates, L. A. Kelley, R. M. MacCallum and M. J. E. Sternberg, Proteins Suppl. **5**, 39 (2001).
[4] C. Venclovas, Proteins Suppl. **5**, 47 (2001).
[5] D. Baker and A. Sali, Science **294**, 93 (2001).
[6] K. K. Koretke, R. B. Russell and A. N. Lupas, Proteins Suppl. **5**, 68 (2001).
[7] A. G. Murzin and A. Bateman, Proteins Suppl. **5**, 76 (2001).
[8] K. Karplus, R. Karchin, C. Barrett, S. Tu, M. Cline, M. Diekhans, L. Grate, J. Casper and R. Hughey, Proteins Suppl. **5**, 86 (2001).
[9] M. G. Williams, H. Shirai, J. Shi, H. G. Nagendra, J. Mueller, K. Mizuguchi, R. N. Miguel, S. C. Lovell, C. A. Innis, C. M. Deane, L. Chen, N. Campillo, D. F. Burke, T. L. Blundell and P. I. W. de Bakker, Proteins Suppl. **5**, 92 (2001).
[10] J. Moult, T. Hubbard, K. Fidelis and J. T. Pedersen, Proteins Suppl. **3**, 2 (1999).
[11] J. Moult, K. Fidelis, A. Zemla and T. Hubbard, Proteins Suppl. **5**, 2 (2001).
[12] A. M. Lesk, L. L. Conte and T. J. P. Hubbard, Proteins Suppl. **5**, 98 (2001).
[13] J. Lee, A. Liwo and H. A. Scheraga, Proc. Natl. Acad. Sci. **96**, 2025 (1999).
[14] J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy and H. A. Scheraga, Proteins Suppl. **3**, 204 (1999).
[15] J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, J. A. Saunders, K. D. Gibson and H. A. Scheraga, Int. J. Quant. Chem. **77**, 90 (2000).
[16] A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy and H. A. Scheraga, Proc. Natl. Acad. Sci. **96**, 5482 (1999).
[17] K. T. Simons, C. Kooperberg, E. Huang and D. Baker, J. Mol. Biol. **268**, 209 (1997).
[18] K. T. Simons, C. Strauss and D. Baker, J. Mol. Biol. **306**, 1191 (2001).

[19] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. M. Strauss and D. Baker, Proteins Suppl. **5**, 119 (2001).
[20] D. T. Jones, Proteins Suppl. **5**, 127 (2001).
[21] D. M. Standley, V. A. Eyrich, Y. An, D. L. Pincus, J. R. Gunn and R. A. Friesner, Proteins Suppl. **5**, 133 (2001).
[22] D. Xu, O. H. Crawford, P. F. LoCascio and Y. Xu, Proteins Suppl. **5**, 140 (2001).
[23] J. Skolnick, A. Kolinski, D. Kihara, M. Betancourt, P. Rotkiewicz and M. Boniecki, Proteins Suppl. **5**, 149 (2001).
[24] D. Kihara, H. Lu, A. Kolinski and J. Skolnick, Proc. Natl. Acad. Sci. **98**, 10125 (2001).
[25] J. Lee, S. -Y. Kim, K. Joo, I. Kim and J. Lee, Proteins **58**, 704 (2004).
[26] J. Lee, S. -Y. Kim and J. Lee, Biophysical Chemistry, in press.
[27] C. B. Anfinsen, Science **181**, 223 (1973).
[28] K. Joo, J. Lee, S. -Y. Kim, I. Kim, S. J. Lee and J. Lee, J. Korean Phys. Soc. **44**, 599 (2004).
[29] K. Joo, I. Kim, J. Lee, S. -Y. Kim, S. J. Lee and J. Lee, J. Korean Phys. Soc. **45**, 1441 (2004).
[30] J. Lee, H. A. Scheraga and S. Rackovsky, J. Comp. Chem. **18**, 1222 (1997).
[31] J. Lee, H. A. Scheraga and S. Rackovsky, Biopolymers **46**, 103 (1998).
[32] J. Lee and H. A. Scheraga, Int. J. Quant. Chem. **75**, 255 (1999).
[33] S. -Y. Kim, S. J. Lee and J. Lee, J. Chem. Phys. **119**, 10274 (2003).
[34] J. Lee, I. H. Lee and J. Lee, Phys. Rev. Lett. **91**, 0802011 (2003).
[35] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[36] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[37] J. Lee, D. R. Ripoll, C. Czaplewski, J. Pillardy, W. J. Wedemeyer and H. A. Scheraga, J. Phys. Chem. B **105**, 7291 (2001).
[38] J. Lee, K. Park and J. Lee, J. Phys. Chem. B **106**, 11647 (2002).
[39] J. Lee, S. -Y Kim and J. Lee, J. Phys. Chem. B **108**, 4525 (2004).
[40] C. A. Mészáros, Computers & Mathematics with Applications **31**, 49 (1996).