NEW CONFORMATIONAL SEARCH METHOD USING GENETIC ALGORITHM AND KNOT THEORY FOR PROTEINS

Y. SAKAE
Department of Physics, Nagoya University,
Nagoya, Aichi 464-8602, Japan
E-mail: sakae@tb.phys.nagoya-u.ac.jp

T. HIROYASU
Department of Biomedical Information, Doshisha University,
Kyotanabe, Kyoto 610-0394, Japan
E-mail: tomo@is.doshisha.ac.jp

M. MIKI
Department of Intelligent Information Engineering and Sciences, Doshisha University,
Kyotanabe, Kyoto 610-0394, Japan
E-mail: mmiki@mail.doshisha.ac.jp

Y. OKAMOTO
Department of Physics, Nagoya University,
Nagoya, Aichi 464-8602, Japan
and
Structural Biology Research Center, Nagoya University,
Nagoya, Aichi 464-8602, Japan
E-mail: okamoto@phys.nagoya-u.ac.jp

We have proposed a parallel simulated annealing using genetic crossover as one of powerful conformational search methods, in order to find the global minimum energy structures for protein systems. The simulated annealing using genetic crossover method, which incorporates the attractive features of the simulated annealing and the genetic algorithm, is useful for finding a minimum potential energy conformation of protein systems. However, when we perform simulations by using this method, we often find obviously unnatural stable conformations, which have "knots" of a string of an amino-acid sequence. Therefore, we combined knot theory with our simulated annealing using genetic crossover method in order to avoid the knot conformations from the conformational search space. We applied this improved method to protein G, which has 56 amino acids. As the result, we could perform the simulations, which avoid knot conformations.

Keywords: Molecular Simulation; Simulated Annealing; Protein Folding; Genetic Algorithm; Knot Theory

1. Introduction

Computational simulations of biomolecular systems such as proteins and DNA are performed using molecular simulation techniques such as Monte Carlo (MC) and molecular dynamics (MD) methods. However, as the biomolecular system has a large number of degrees of freedom associated with a lot of atoms and is characterized by many local minima separated by high energy barriers, it is not yet possible to perform enough conformational searches in this

extremely high dimensional space, and efficient sampling techniques are required.

In order to solve this problem, various sampling and optimization methods for conformations of biomolecules have been proposed such as generalized-ensemble algorithms.[1] Simulated annealing[2] and genetic algorithm[3,4] have been recognized by researchers as powerful tools for difficult optimization problems. Simulated annealing mimics an annealing process in which the temperature of a system is lowered very slowly from a sufficiently high initial temperature to "freezing" temperature. This method has been applied to molecular simulations of biomolecules[5–11] (and also to various other research fields). The genetic algorithm mimics the process of natural evolution and has been applied to various research fields and is one of well-known techniques. The genetic algorithm uses the optimization procedures of natural gene-based evolution, that is, mutation, crossover, and replication. For a certain optimization problems, this algorithm has been found to be an excellent strategy to find global minima. The conformational search or optimization approaches for biomolecules using the genetic algorithm have also been performed.[12–16]

We proposed a new conformational search method, in which a simulated annealing simulation is combined with genetic algorithm, namely, *parallel simulated annealing using genetic crossover* (PSA/GAc),[17] and applied it to be the search of the global-minimum-energy structures for protein systems.[18,19] Here, the genetic crossover is one of the operations of genetic algorithm. The conformational search using simulated annealing is based on local conformational updates. On the other hand, the genetic algorithm is based on global conformational updates. Our method incorporates these two attractive features of the simulated annealing and the genetic crossover. In our previous work, in order to examine the effectiveness of our method, we compared our method with those of the conventional simulated annealing molecular dynamics simulations using an α-helical miniprotein, namely, Trp-cage.[20]

However, in the case of the conformational search of a protein constructed by a certain length of amino acids, we often found the lower energy conformation in spite of the completely different structure in comparison with the native structure. The structure has a very compact fold as if there is a knot in the string of the amino-acid sequence. For example, in Fig. 1, two conformations, namely, a native structure and a stable conformation obtained from the simulation by using our method, of a protein, which is the B1 domain in the immunoglobulin G (IgG) binding domains of protein G[21] and has 56 amino residues are shown. For the stable conformation obtained from the simulation, there is one knot in the string of the protein. Knot conformations of some proteins are already found by experiments of X-ray crystallography.[22,23] However, the knotted chains in the knot conformations obtained from the simulations by using our method have obviously different length from those of the experimental results. Although the length of the knotted chains known by experiments is at least 35–45 amino residues, the length of the knotted chains of the conformations obtained from the simulations is about 15 amino residues. Namely, the knot conformations obtained from the simulations are unnatural. As the reasons of getting the unnatural knot conformations, it is thought to be causally related to using the inaccurate force field and/or the unusual simulation technique in comparison with the conventional MC or MD. As far as we know, knotted conformations were never found with other conformational sampling methods, which suggests the powerfulness of our

conformational search method.

Therefore, we propose the improved conformational search method, which can avoid the unnatural knot conformations. In order to check whether a knot conformation or not, we use the Kauffman polynomial[24] in the mathematical theory of knots. If a trial conformation generated by the crossover operation in the simulation has knots, the conformation is rejected regardless of the value of the potential energy. In this paper, we performed the conformational search of protein G by using the improved PSA/GAc and the conventional one in order to examine the simulation results of the improved method.

In section 2 the details of our conformational search method and its improved one are given. In section 3 the results of applications of the folding simulations of protein G are presented. Section 4 is devoted to conclusions.

2. Method

2.1. *Parallel simulated annealing molecular dynamics using genetic crossover*

Let M be the total number of *individuals*. In parallel simulated annealing using genetic crossover (PSA/GAc), a crossover operation is carried out in a fixed interval of a certain time steps of the M parallel conventional simulated annealing simulations. The entire process of the general formalism of parallel simulated annealing using genetic crossover[17–19] is illustrated in Fig 2 (in the schematic illustration there, we have $M = 6$). In parallel simulated annealing molecular dynamics using genetic crossover (PSAMD/GAc), M conventional simulated annealing molecular dynamics simulations (instead of Monte Carlo simulations) are performed in parallel. Although we employed a genetic one-point crossover in our previous study,[17–19] we can employ various kinds of genetic crossover operations such as one-point crossover, two-point crossover, etc. In this study, we employed the genetic two-point crossover, and we refer to the entire method as PSAMD/GAc2. The crossover operation in this method exchanges a part of dihedral angles between two conformations of a protein.

In the two-point crossover operation, the following procedure is carried out (see Fig. 3) :

(1) $M/2$ pairs of conformations are selected from "parental" group randomly.

(2) Consecutive amino acids of length n residues in the amino-acid sequence of the conformation are selected randomly for each pair of selected conformations.

(3) All dihedral angles (in backbone and side chains) in the selected n amino acids are exchanged between the selected pairs of conformations.

Note that the length n of consecutive amino-acid residues is in general different for each pair of selected conformations. Motivated by the fragment assembly method,[25] we take n to be an integer ranging from 2 to 10. In this procedure, we obtain two new "child" conformations. After that, we have to select two superior "chromosomes" (conformations) from the total

of four conformations (two parental conformations and two new child conformations). We perform the energy minimizations for these four conformations by a standard method such as Newton-Raphson method and conjugate gradient method. We then select two lower-energy conformations based on the four minimized energy values. Finally, using the selected two energy-minimized conformations, the parallel simulated annealing simulations continue.

In our previous works, we did not perform the energy minimization after the genetic crossover operation. However, the conformations generated by the genetic crossover operation often have unusually high potential energy, because the genetic crossover operation brings about a large global change of conformations. This leads to very low acceptance ratio of child conformations. Therefore, in this study, we perform the energy minimization after the genetic crossover operation in order to avoid this difficulty of low acceptance ratio. Because the conformational change by the energy minimization is very small (in the example of a mini-protein presented below, the root-mean-square deviations of C_α atoms between before and after energy minimizations was only about 0.45 Å on the average), we believe that this energy minimization does not affect the nature of the new conformational generation of the crossover operations.

2.2. *PSAMD/GAc with knot theory*

In this paper, in order to check whether a trial conformation, which generated by the genetic crossover operation, has knots or not, we use the Kauffman polynomial[24] in knot theory.

2.2.1. *Calculation of knot invariants*

To characterize the topological properties of knots and links of strings algebraically, polynomials can be used. These polynomials are knot invariants, which have been discovered and constructed, and have been proposed several polynomials. The Kauffman polynomial $F(L; a, x)$ is one of them and is a two-variable (a and x) invariant.

$$F(L; a, x) = a^{-t(\tilde{L})}\Lambda(|\tilde{L}|; a, x). \tag{1}$$

Here, L is a link. A knot is an embedding of a single circle into three-dimensional space, while a link is an embedding of a collection of circles. The sign of $|\ |$ means unoriented knots, and the tilde ˜ means a link represented by a link diagram. $\Lambda(|\tilde{L}|; a, x)$ is defined by the conditions and the Skein relation, which is recursion relations relating the invariants of knots, in Fig. 4(b). Four knots $|L_+|$, $|L_-|$, $|L_\infty|$, and $|L_{-\infty}|$ in Fig. 4(b) correspond to the line configurations $+$, $-$, ∞, and $-\infty$ in Fig. 4(a), respectively. $t(\tilde{L})$ is the sum of the signs of all the crossings. If a knot is unknot (trivial knot), the knot invariant estimated by the Kauffman polynomial is equal to 1 ($F = 1$), if it is other knots, the knot invariant is a polynomial except 1 ($F \neq 1$). Namely, by estimating the knot invariant, we can determine whether a conformation has knots or not.

2.2.2. *Estimation of knotting properties of a protein*

We need to construct a knot diagram from a protein conformation in order to obtain the knot invariant. At first, the coordinate points of C_α atoms in a protein conformation are projected

on X-Y Cartesian coordinate space. These points are connected from N-terminal to C-terminal by lines. If two lines intersect, the crossing point is defined as a crossing on the knot diagram, and the sign of the crossing is determined by the relation of Z Cartesian coordinates of the crossing point on the two lines. After that, the two points of the first and last C_α atoms are connected by as few crossings as possible. We use a collection of these lines connected by the points of C_α atoms as a knot diagram.

2.2.3. *Flow of PSAMD/GAc with knot theory*

A chart of the PSAMD/GAc simulation process with knot theory is shown in Fig. 5. In our improved simulation, the calculation of the knot invariant for a conformation is performed after the process of crossover operations. In the conventional PSAMD/GAc simulation process, we select two lower-energy conformations based on the four minimized energy values of four conformations (two parental conformations and two new child conformations). On the other hand, in the improved process, if both two new child conformations generated by genetic crossover operations do not have knots, the simulation process is the same as the conventional one. If one of two new child conformations has knots, we select two lower-energy conformations based on the three minimized energy values of three conformations (two parental and one child conformations) except one child knot conformation. If both two conformations have knots, we do not perform the procedure of the selection, namely, two parental conformations are selected, and after that, the simulation continues.

3. Results and Discussion

We applied our improved method to the protein G (PDB code: 1PGA).[21] Protein G from *Streptococcus* also binds human immunoglobulin G (IgG). This protein consists of a series of small binding domains separated by linkers and a cell-wall anchor near the C-terminus. Two (in some strains, three) of the domains bind IgG. The IgG-binding domains of protein G are identified as B1, B2, etc., numbering from the N-terminus of the native protein G molecule. We used the B1 domain which consists of a four-stranded β-sheet and an α-helix, and was engineered for production as a 56 residue protein with N-terminal methionine (this position was threonine in the wild type) (see Fig. 1(a)).

We incorporated PSAMD/GAc2 by modifying the TINKER program package[26] modified by us. The unit time step was set to 2.0 fs, and all bonds to hydrogen atoms at ideal bond lengths were constrained by RATTLE method.[27] Each simulation was carried out for 2.0 nsec (hence, it consisted of 1,000,000 MD steps) with 32 individuals ($M = 32$) and repeated 5 times. The temperature during MD simulations was controlled by Berendsen method.[28] For each run the temperature was decreased exponentially from 1000 K to 200 K. As for the conformational energy calculations, we used the AMBER ff96 force field.[29] As for solvent effects, we used the GB/SA model[30,31] included in the TINKER program package.[26] These folding simulations were started from a fully extended conformation and different sets of randomly generated initial velocities (for repetition of 5 times). The genetic crossover operations in PSAMD/GAc2 simulation were performed 1000 times at the fixed interval of 1000 MD steps. Moreover, we incorporated the calculation program of knot invariants by the Kauffman

polynomial to the PSAMD/GAc2 program based on a program of Ochiai *et al.*[32] After the genetic crossover operation, the energy minimization by the quasi-newton method (L-BFGS)[33] included in TINKER was performed. Additionally, we performed the conventional simulated annealing molecular dynamics simulations for comparison. In order to balance the computational cost, we performed 160 simulation runs of 2 nsec in length ($32 \times 5 = 160$). The other simulation conditions were the same (except for with or without crossover operations).

We remark on the dependence of the frequency of knotted conformation creation on the force fields. We found that three out of five simulations with OPLS-AA/L and one out of five simulations with CHARMM22 created knotted conformations, while five out of five simulations with AMBER ff96 found knotted conformations. Because AMBER ff96 gave the most number of knotted conformations, we present the results of our knot-avoiding method with AMBER ff96 below.

In Fig. 6, the lowest-energy final minimized conformations obtained from the normal PSAMD/GAc2, and the improved PSAMD/GAc2 with knot theory are shown. As these results, all the conformations obtained from the normal PSAMD/GAc2 have unnaturally knot conformations. On the other hand, the conformations obtained from the improved PSAMD/GAc2 with knot theory have the stable conformations without knots.

In Fig. 7, the minimized potential energy of the final 160 conformations obtained from the normal PSAMD/GAc2, the improved PSAMD/GAc2 with knot theory, and the conventional simulated annealing is shown. As a reference, the value for the native conformation is also shown. Here (and in Fig. 8 below), the "native conformation" means the conformation that was obtained as follows. A canonical MD simulation of 100 psec at a low temperature (200 K) with the initial conformation being the native PDB conformation was first performed. The final conformation was then energy-minimized. The heavy-atom RMSD of this "native conformation" from the PDB coordinates was 1.4 Å. In comparison with the conventional simulated annealing method, the potential energy is obviously lower in both the normal PSAMD/GAc2 and the improved PSAMD/GAc2 with knot theory as a whole. The lowest energy and the average energy obtained from the normal PSAMD/GAc2 are −2322.7 kcal/mol and −2306.6 kcal/mol, respectively. Those obtained from the improved PSAMD/GAc2 with knot theory are −2310.4 kcal/mol and −2297.6 kcal/mol, respectively. On the other hand, those obtained from the conventional simulated annealing method are −2277.3 kcal/mol and −2237.9 kcal/mol. The differences of the energy values between the normal and improved PSAMD/GAc2 are 12.3 kcal/mol and 9.0 kcal/mol. The differences of the energy values between the normal PSAMD/GAc2 and the conventional simulated annealing are 45.4 kcal/mol and 68.7 kcal/mol. As these results, the conformations obtained from both the normal and improved PSAMD/GAc2 are more stable than those of the conventional simulated annealing. Namely, by incorporating the crossover operation into the simulated annealing method, we can obtain more stable structures than the conventional simulated annealing method. Additionally, the conformations obtained from the normal PSAMD/GAc2 are slightly more stable than the improved PSAMD/GAc2 with knot theory. This result shows that the more unnatural conformations with knots are more stable than the conformations without knots. We suppose that one of the reasons is the inaccuracy of the force field for the simulations.

In Fig. 8, the radius of gyration of the final minimized conformations obtained from the normal PSAMD/GAc2 and the improved PSAMD/GAc2 with knot theory, and the conventional simulated annealing are shown. These results obviously illustrate that the final conformations obtained from both the normal PSAMD/GAc2 and the improved PSAMD/GAc2 with knot theory become more compact conformations in comparison of those of the conventional simulated annealing on the whole. Namely, the improved method as well as the normal method can search compact conformations.

4. Conclusions

In this article, for the parallel simulated annealing using genetic crossover (PSA/GAc), we proposed the improved method, which can avoid the unnatural knot conformations. In order to check whether a conformation has knots or not, we used the Kauffman polynomial in the mathematical theory of knots and links and incorporated the check function to PSA/GAc.

As a test simulation, we applied this improved conformational search method to the protein G. We succeeded in performing the simulations which avoided unnatural knot conformations and could obtain stable conformations as well as the normal PSAMD/GAc2, in comparison with the conventional simulated annealing. Additionally, the knot conformations obtained from the normal PSAMD/GAc2 were slightly more stable than the unknoted conformations obtained from the improved PSAMD/GAc2 with knot theory. One of the supposable reasons is inaccuracy of the force field for the simulations. Therefore, in a future work we are going to perform the conformational search by PSAMD/GAc2 with the force field optimized by our optimization methods.[34,35]

Once all these preparations are successfully made, we will be ready to apply the present method to multi-scale modelling of biosystems.

Acknowledgments

The computations were performed on SuperNOVA of Miki Laboratory at Doshisha University and the computers at the Research Center for Computational Science, Institute for Molecular Science. This work was supported, in part, by the Grants-in-Aid for the Academic Frontier Project, "Intelligent Information Science", for Scientific Research on Innovative Areas ("Fluctuations and Biological Functions"), and for the Next Generation Super Computing Project, Nanoscience Program from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

1. A. Mitsutake, Y. Sugita and Y. Okamoto, *Biopolymers (Pept. Sci.)* **60**, 96 (2001).
2. S. Kirkpatrick, C. D. Gelatt Jr. and M. P. Vecchi, *Science* **220**, 671 (1983).
3. J. H. Holland, *Adaptation in Natural and Artificial Systems* (The University of Michigan Press, Ann Arbor, 1975).
4. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, 1989).
5. M. Nilges, G. M. Clore and A. M. Gronenborn, *FEBS Lett.* **229**, 317 (1988).

6. A. T. Brünger, *J. Mol. Biol.* **203**, 803 (1988).
7. A. T. Brünger, M. Karplus and G. A. Petsko, *Acta Cryst.* **A45**, 50 (1989).
8. S. R. Wilson, W. Cui, J. W. Moskowitz and K. E. Schmidt, *Tetrahedron Lett.* **29**, 4373 (1988).
9. H. Kawai, T. Kikuchi and Y. Okamoto, *Protein Eng.* **3**, 85 (1989).
10. C. Wilson and S. Doniach, *Proteins* **6**, 193 (1989).
11. Y. Okamoto, M. Fukugita, T. Nakazawa and H. Kawai, *Protein Eng.* **4**, 639 (1991).
12. T. Dandekar and P. Argos, *Protein Eng.* **5**, 637 (1992).
13. S. Sun, *Protein Science* **2**, 762 (1993).
14. R. Unger and J. Moult, *J. Mol. Biol.* **231**, 75 (1993).
15. A. A. Rabow and H. A. Scheraga, *Protein Science* **5**, 1800 (1996).
16. J. Lee, H. A. Scheraga and S. Rackovsky, *J. Comput. Chem.* **18**, 1222 (1997).
17. T. Hiroyasu, M. Miki and M. Ogura, *Proceedings of the 44th Institute of Systems* , 113 (2000).
18. T. Hiroyasu, M. Miki, M. Ogura and Y. Okamoto, *J. IPS Japan* **43**, 70 (2002), in Japanese.
19. T. Hiroyasu, M. Miki, M. Ogura, K. Aoi, T. Yoshida and Y. Okamoto, *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003)* , 117 (2003).
20. J. W. Neidigh, R. M. Fesinmeyer and N. H. Andersen, *Nature Struct. Biol.* **9**, 425 (2002).
21. T. Gallagher, P. Alexander, P. Bryan and G. L. Gilliland, *Biochemistry* **33**, 4721 (1994).
22. K. Lim, H. Zhang, A. Tempczyk, W. Krajewski, N. Bonander, J. Toedt, A. Howard, E. Eisenstein and O. Herzberg, *Proteins* **51**, 56 (2003).
23. O. Nureki, M. Shirouzu, K. Hashimoto, R. Ishitani, T. Terada, M. Tamakoshi, T. Oshima, M. Chijimatsu, K. Takio, D. G. Vassylyev, T. Shibata, Y. Inoue, S. Kuramitsu and S. Yokoyama, *Acta Cryst.* **D58**, 1129 (2002).
24. L. H. Kauffman, *Trans. Am. Math. Soc.* **318**, 417 (1990).
25. K. T. Simons, C. Kooperberg, E. Huang and D. Baker, *J. Mol. Biol.* **268**, 209 (1997).
26. Tinker program package software available at http://dasher.wustl.edu/tinker/.
27. H. C. Andersen, *J. Comput. Phys.* **52**, 24 (1983).
28. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.* **81**, 3684 (1984).
29. P. A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot and A. Pohorille, *Computer simulations of biological systems* (Escom, Netherlands, 1997), ch. The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data, pp. 83–96.
30. W. C. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson, *J. Am. Chem. Soc.* **112**, 6127 (1990).
31. D. Qiu, P. S. Shenkin, F. P. Hollinger and W. C. Still, *J. Phys. Chem. A* **101**, 3005 (1990).
32. M. Ochiai, S. Yamada and E. Toyoda, *Computer Aided Knot Theory* (Makino Shoten, 1996). in Japanese.
33. J. Nocedal, *Math. Comp.* **35**, 773 (1980).
34. Y. Sakae and Y. Okamoto, *Chem. Phys. Lett.* **382**, 626 (2003).
35. Y. Sakae and Y. Okamoto, *Mol. Sim.* **36**, 159 (2010).

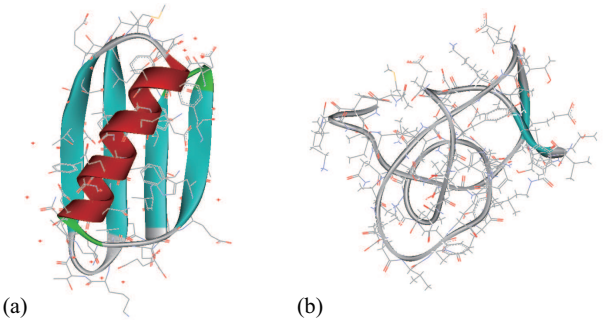


Fig. 1. The structure of Protein G. (a) is the native structure (PDB ID: 1PGA). (b) is the final conformation obtained from PSAMD/GAc2.

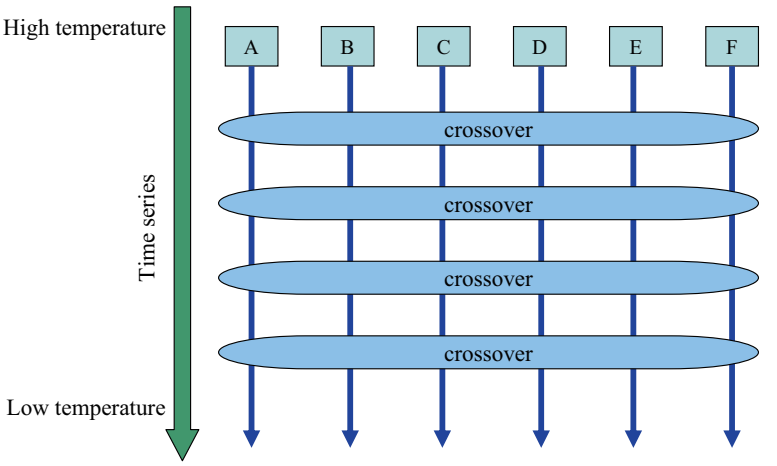


Fig. 2. Schematic process of the parallel simulated annealing using genetic crossover. In this method, the crossover operation, which is shown in Fig. 3, is performed during parallel simulated annealing simulations.

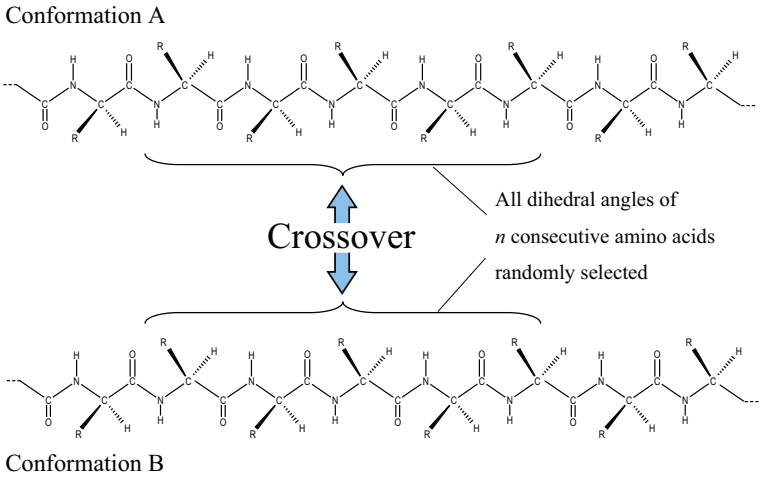
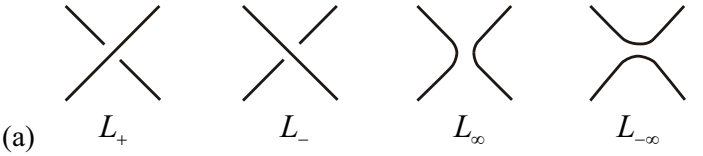


Fig. 3. Schematic process of the two-point crossover operation. In this process, all dihedral angles (in backbone and side chains) within the randomly selected n consecutive amino acids are exchanged between a pair of conformations. Motivated by the fragment assembly method,[25] we take n to be an integer ranging from 2 to 10.



(b)

(1) $\Lambda(O;a,x)=1$

(2) $\Lambda(|\,L_+\,|;a,x)+\Lambda(|\,L_-\,|;a,x)=x\{\Lambda(|\,L_\infty\,|;a,x)+\Lambda(|\,L_{-\infty}\,|;a,x)\}$

(3) $\Lambda($  $;a,x)=a\Lambda($  $;a,x)$

$\Lambda($  $;a,x)=a^{-1}\Lambda($  $;a,x)$

Fig. 4. The four line configurations (a), the Skein relation (2) of (b), and the conditions (1,3) of (b) defined by the Kauffman polynomial.

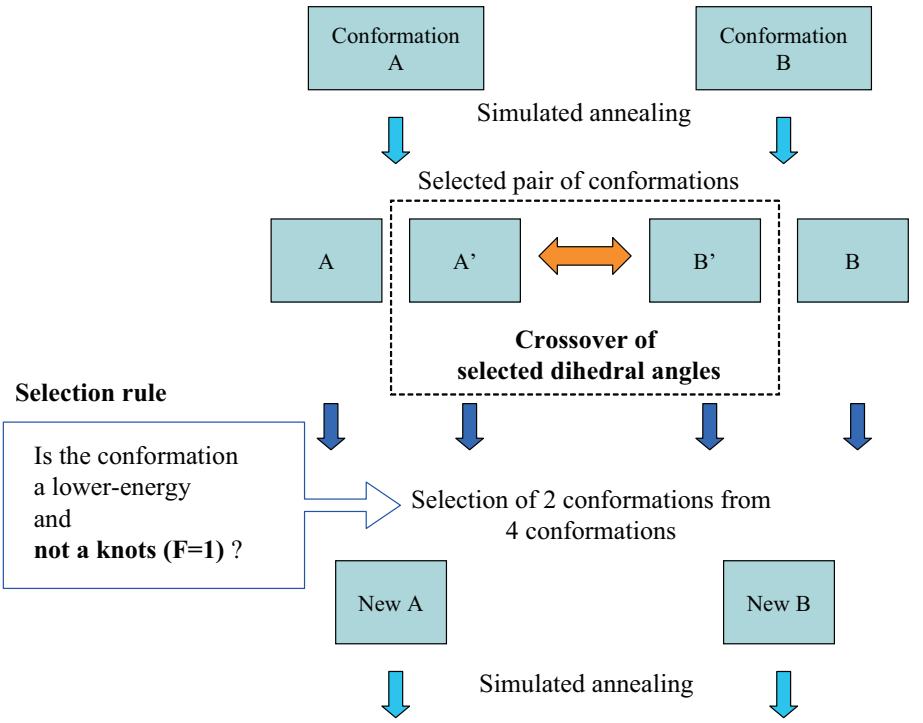


Fig. 5. Schematic process of the simulated annealing using genetic crossover with knot theory. In this simulation, the two unknotted conformations are selected by using the Kauffman polynomial after genetic crossover operations.

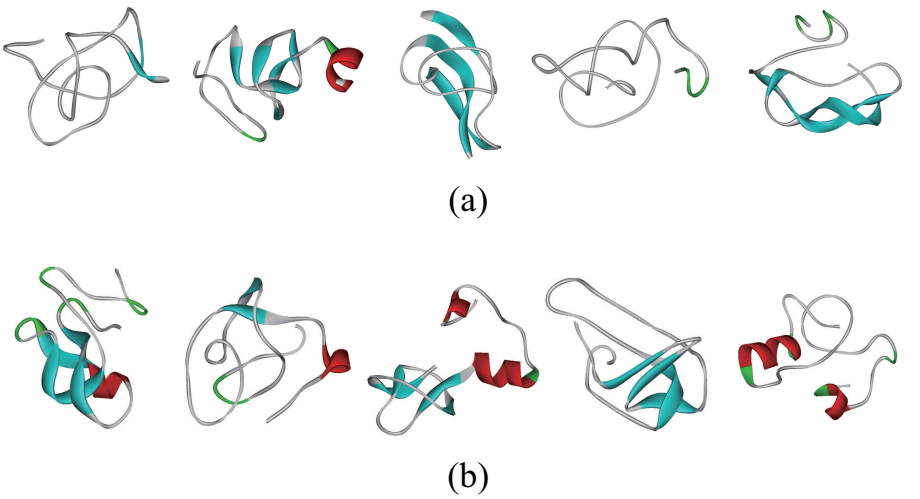


Fig. 6. The final conformations obtained from PSAMD/GAc2. (a) shows the conformations obtained from the normal method, and (b) shows the conformations obtained from the improved method with knot theory. The simulations were performed five times for both cases.

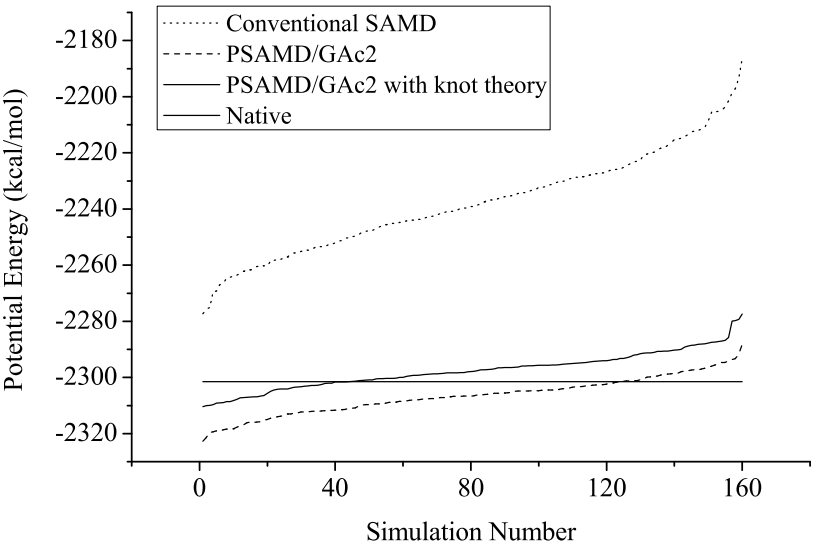


Fig. 7. Comparison of the minimized potential energy of the final conformations obtained from the conventional simulated annealing MD simulation (dotted line), the normal PSAMD/GAc2 (broken line), the improved PSAMD/GAc2 with knot theory (normal line). The value for the native structure is also shown (normal horizontal line).

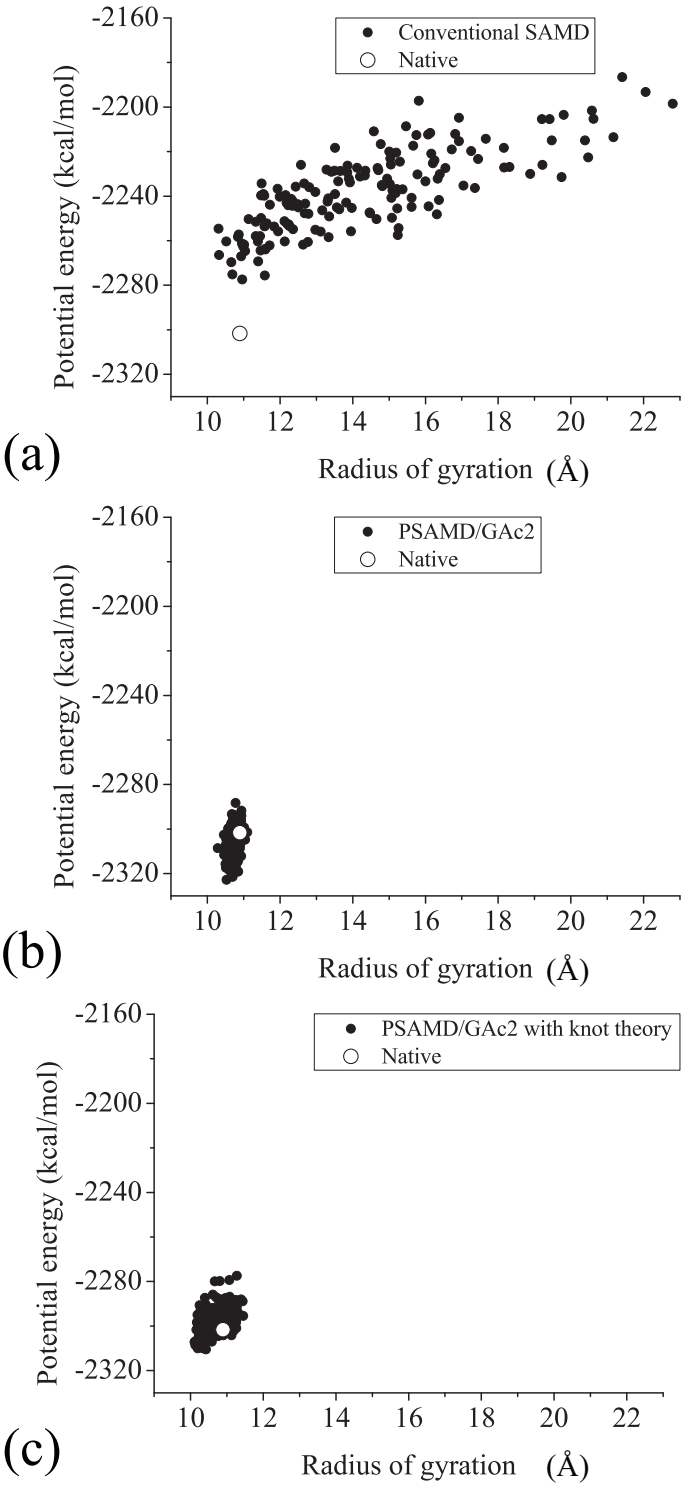


Fig. 8. Radius of gyration of the final minimized conformations obtained from the conventional simulated annealing (a), the normal PSAMD/GAc2 (b), the improved PSAMD/GAc2 with knot theory (c). The radius of gyration was caluclated with respect to all atoms. The value for the native structure is also shown (open circle).