Two Perturbations for Geometry Optimization of Off-Lattice Bead Protein Models

Hiroshi Takeuchi*

Division of Chemistry, Graduate School of Science, Hokkaido University, Sapporo 060-0810, Japan

Corresponding author phone: +81-11-706-3533; Fax: +81-11-706-3501; e-mail: takehi@sci.hokudai.ac.jp

Abstract: Referring to the optimization algorithm previously developed for atomic clusters, the present author develops an efficient method for geometry optimization of a coarse-grained protein model expressed with two kinds of beads (hydrophilic and hydrophobic ones). In the method, two types of geometrical perturbations, center-directed bead move and one bead rotation, are used to explore new configurations and local optimizations are performed after the perturbations. The center-directed bead move is used for hydrophobic beads and the one bead rotation is performed for both hydrophobic and hydrophilic beads. The optimization method was applied to protein models consisting of 13, 20, 21, and 34 beads. The present method produced the global minima of the 13-, 21-, and 34-bead models reported in the literature and updated the lowest energies of the protein models with 20 beads. These results indicate that the present method is efficient for searching for optimal structures of proteins.

Keywords: global optimization · protein model · potential energy

1 Introduction

Since chemical and physical properties of proteins depend on the structures, the determination of them is important for elucidating their functions. Proteins are of great interest and a lot of the structural investigations have been reported. The experimental methods of X-ray diffraction and NMR spectroscopy are powerful tools. However, the shortcomings of them (for example, limitations on crystallization and sizes of molecules treated by the spectroscopy) restrict molecules. Theoretically structures of molecules of biological interest have been investigated employing molecular dynamics and Monte Carlo simulations. These simulations have been used to examine time evolution and thermodynamic properties of the molecule under investigation. The results of the simulation depend on the initial condition of the system because of the limitations on the time scale and acceptable moves. The calculation may fail to locate the lowest-energy and low-lying configurations if high barriers between local minima on the potential energy surface are present. In the case, geometry optimization of the molecule would be useful to examine the reliability of the geometry obtained by the simulation. The geometry optimization is also necessary in structural investigations of proteins without sufficient structural information. The present study is aimed at developing an efficient geometry optimization method for proteins. The calculations of all-atom protein models are time-consuming and thereby one of the simplest protein models, the three-dimensional off-lattice model described with two kinds of beads (Figure 1),[1−26] is used in the present study.

The model consists of hydrophobic and hydrophilic beads labeled by A and B. This was first proposed by Stillinger et. al.[18] as a two-dimensional model and then extended to a three-dimensional one.[1-13,15,17,22,23] Because of the simplification of the three-dimensional model, side chains and hydrogen bonds existing in proteins are not taken into account. However, the potential energy surface of the model is still rugged and the search for the global minimum is difficult.[1-13,22,23] It should be noted that the model captures an important feature of proteins, i.e., formation of a single hydrophobic core.[1-8,10,12] For two dimensional off-lattice models with the same bead sequences as the three dimensional ones, [2-4,8,9,11,15,19,20,24,25] however, more than one hydrophobic cores are formed and hydrophilic beads isolate the cores. Hence the formation of the single hydrophobic core depends on the dimension of the model. Efficient geometry optimization methods for the three-dimensional models are promising for realistic protein models.

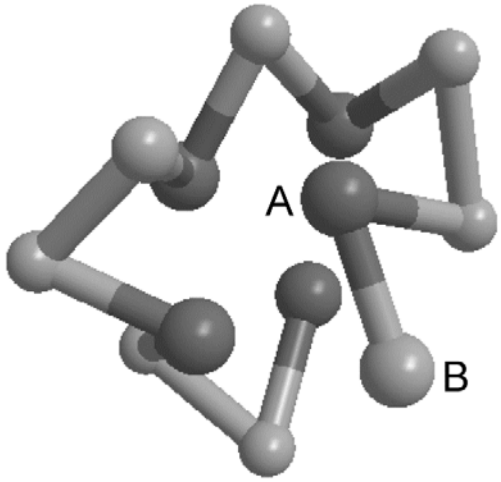

Figure 1. An off-lattice protein model with hydrophobic (A) and hydrophilic (B) beads.

The number of beads in the model is less than 90;[1-13,15,17,22,23] the bead sequences are described in the next section. The models with 13, 21, and 34 beads were investigated with different methods[2-13,22,23] and the lowest energy of each of them was obtained with at least two methods. Hence these models could be used as benchmark problems for geometry optimization. The geometries of the 20-bead models with 6 different sequences[1,5,10] were also calculated with a few methods. For the model with 55 beads,[2-9,11,13,15] no consensus on the global minimum is found although several methods are applied to it. Only the study of Zhang and Ma[13] reports the lowest energy of the 89-bead protein. Accordingly the global minima of the models with 55 and 89 beads are not clarified.

There are a lot of issues treating geometry optimizations. Structural predictions of clusters are included in them. As known in the studies on clusters,[27,28] a lot of stable configurations are possible for a cluster. Consequently it is difficult to search for its global minimum and the situation is similar to that for proteins. The present author previously developed the efficient method to optimize geometries of Lennard-Jones (LJ) atomic clusters.[29–31] The algorithm is based on specific geometrical perturbations and subsequent local optimizations. The calculations[29] reproduced the global minima of LJ homoclusters with 10 – 561 atoms reported previously and located new minima for 6 clusters. Heteroclusters with 2, 3, and 4 kinds of atoms were also investigated.[30,31] For the heteroclusters, the atomic compositions make the structure prediction complicated. The complicated cases were successfully optimized with the method: a lot of new global minima of the heteroclusters with up to 50 atoms were located. Accordingly the method could be applied to other global optimization problems. As a target of great interest, the present author selected structure prediction of proteins.

The purpose of the present study is to propose a new method for geometry optimization of three-dimensional off-lattice bead models. For the purpose, specific geometrical perturbations are developed for the models referring to the method for atomic clusters.[29] The proposed method is applied to the models with 13 to 55 beads. The lowest energies of the models are compared with the data reported in the literature to examine the efficiency of the method. The results are analyzed to elucidate structural features of the models and to examine the ability of the perturbations to lower the potential energies. Improvements of the method are finally discussed.

2 Method

2.1 Protein Model

The three-dimensional off-lattice model contains only two types of beads, A and B, which correspond to hydrophobic and hydrophilic residues, respectively. Since the model has the unit length for all bonds, independent structural parameters of the N-bead protein are $N – 2$ bond angles θ and $N – 3$ dihedral angles ϕ as shown in Figure 2. The present study adopts two kinds of potentials developed by Irbäck, Peterson, Potthast, and Sommelius[1] (abbreviated to the IPS potential) and by Stillinger, Head-Gordon, and Hirshfeld[18] (the SHH potential). The IPS potential energy of the N-bead protein is given by

$$E = \sum_{i=1}^{N-2} \vec{u}_{i,i+1} \cdot \vec{u}_{i+1,i+2} - \frac{1}{2} \sum_{i=1}^{N-3} \vec{u}_{i,i+1} \cdot \vec{u}_{i+2,i+3} + 4 \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \varepsilon_{\alpha\beta} \left(r_{ij}^{-12} - r_{ij}^{-6} \right) \quad (1)$$

Here $\vec{u}_{ij}$ means a unit vector directed from the *i*th bead to the *j*th bead and r_{ij} represents the distance between beads *i* (type α) and *j* (type β). The value of the coefficient $\varepsilon_{\alpha\beta}$ is 1 for A…A pairs and 0.5 for A…B and B…B pairs. When the first and second terms of the right-side of Eq. (1) are expressed by θ and ϕ, Eq. (1) is presented as follows:

$$E = -\sum_{i=1}^{N-2}\cos\theta_i - \frac{1}{2}\sum_{i=1}^{N-3}\left(\cos\theta_i\cos\theta_{i+1} - \sin\theta_i\sin\theta_{i+1}\cos\phi_i\right) + 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}\varepsilon_{\alpha\beta}\left(r_{ij}^{-12} - r_{ij}^{-6}\right) \qquad (2)$$

The potential energy of the 3-bead protein XXX (X = A or B) is presented as a function of $r_{1,3}$:

$$E_3 = -\left(1 - \frac{r_{1,3}^2}{2}\right) + 4\varepsilon_{\alpha\beta}\left(r_{1,3}^{-12} - r_{1,3}^{-6}\right) \qquad (3)$$

The above equation shows that the bond angles of 66 and 67° take the potential energy minima for $\varepsilon_{\alpha\beta} =$ 1 and 0.5, respectively; the corresponding minimum value $E_{\min}$ is −1.38 and −0.89, respectively. Similarly the 4-bead protein XXXX has the following potential energy if two bond angles take the same value θ:

$$E_4 = -2\cos\theta - \frac{1}{2}(\cos^2\theta - \sin^2\theta\cos\phi) + 4\varepsilon_{\alpha\beta}\left(r_{1,4}^{-12} - r_{1,4}^{-6}\right) \qquad (4)$$

where $r_{1,4}^2 = -2\sin^2\theta\cos\phi + 2\cos^2\theta - 4\cos\theta + 3$. Assuming that $\theta = 67°$ and $\varepsilon_{\alpha\beta} = 1$, numerical evaluation of the equation shows that E_4 takes the minimum values of −1.74 and −1.38 for the gauche and trans configurations ($\phi = 75$, 180°). The corresponding values are −1.24 and −1.33 for $\varepsilon_{\alpha\beta} = 0.5$. Hence AXXA favors the gauche configuration and BXXB and AXXB slightly prefer the trans one.

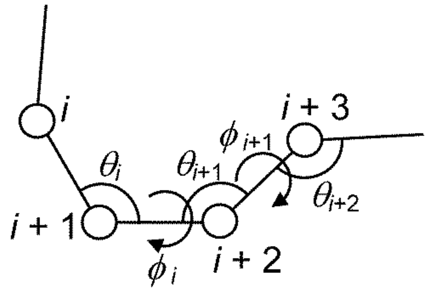

Figure 2. Definition of bond angles θ_i and torsional angles ϕ_i in the off-lattice protein model.

The potential function of the SHH model[18] is represented by

$$E = \frac{1}{4}\sum_{i=1}^{N-2}(1-\vec{u}_{i,i+1}\cdot\vec{u}_{i+1,i+2}) + 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}\left(r_{ij}^{-12} - C_{\alpha\beta}r_{ij}^{-6}\right) \tag{5}$$

The equation includes no torsional term, in contrast with the IPS expression. The value of the coefficient $C_{\alpha\beta}$ is 1 for A…A pairs, 0.5 for B…B pairs, and −0.5 for A…B pairs. For the 3-bead model of AXA, the bond angle of the most stable geometry is 69° and the energy is −0.66. The structures of AXB and BXB take the lowest potential energies at $\theta = 180°$. The model of AAAA with $\theta = 69°$ takes one stable conformation with $\phi = 70°$ and $E = -0.31$.

The values of $\varepsilon_{\alpha\beta}$ and $C_{\alpha\beta}$ in the IPS and SHH potentials indicate that in both models, the A…A van der Waals interaction has the lower energy at the equilibrium distance than the A…B and B…B interactions. Hence the A beads tend to aggregate to form hydrophobic cores.

The bead sequences of proteins used in the present study are listed in Table 1. The methods and potentials used in the previous studies are summarized in Table 2. The sequences of six 20-bead proteins are taken from the study of Irbäck et al.[1] and the sequences S13, S21, and S34 are generated according to the Fibonacci sequences.[2–9, 11–13] The lowest energies obtained in the literature are listed in Tables 3 and 4. For the bead models expressed with the IPS and SHH potentials, the previous studies[4,5,8,10,12] report that the folding to the lowest-energy conformation occurs and that the folding behaviors of them are qualitatively comparable with those of real proteins. Hence the development of the optimization method for the bead protein models would be useful for all-atom ones.

Table 1. Bead sequences of proteins, the number of optimization cycles $N_{\rm opt}$, and the number of successful cycles $N_{\rm s}$ for the potentials SHH and IPS.

Protein	Sequence	SHH		IPS	
		$N_{\rm opt}$	$N_{\rm s}$	$N_{\rm opt}$	$N_{\rm s}$
S13	$AB_2AB_2ABAB_2AB$	2000	1599	2000	625
S20.1	$BA_6BA_4BA_2BA_2B_2$	10000	34	10000	51
S20.2	$BA_2BA_4BABA_2BA_5B$	10000	10	10000	46
S20.3	$A_4B_2A_4BA_2BA_3B_2A$	10000	12	10000	239
S20.4	$A_4BA_2BABA_2B_2A_3BA_2$	10000	94	10000	88
S20.5	$BA_2B_2A_3B_3ABABA_2BAB$	10000	565	10000	87
S20.6	$A_3B_2AB_2ABAB_2ABABABA$	10000	101	10000	210
S21	$BABAB_2ABAB_2AB_2ABAB_2AB$	2000	70	2000	78
S34	$AB_2AB_2ABAB_2AB_2ABAB_2ABAB_2AB_2ABAB_2AB$	20000	4	100000	1

Table 2. Previous investigations on simulations for the sequences listed in Table 1.

Author(s)	Method(s)	Potential and sequence	Ref.
Irbäck, Peterson, Potthast, and Sommelius	Monte Carlo (MC) + quenching	IPS (S20.1 to 20.6)	1
Hsu, Mehra, and Grassberger	Pruned-enriched-Rosenbluth (PER) + energy minimization	SHH (S13, 21, 34)	2
Liang	Annealing contour Monte Carlo (ACMC)	IPS (S13, 21, 34)	3
Kim, Lee, and Lee	Conformational space annealing (CSA)	SHH, IPS (S13, 21, 34)	4
Bachmann, Arkin, and Janke	Energy landscape paving (ELP) and multicanonical Monte Carlo (MMC)	SHH, IPS (S13, 21, 34, 20.1 to 20.6)	5
Huang and Liu	Heuristic conjugate gradient (HCG) and pruned-enriched-Rosenbluth (PER) + energy minimization	SHH (S13, 21, 34)	6
Chen and Huang	Heuristic algorithm (HA)	SHH (S13, 21, 34)	7
Kim, Straub, and Keyes	Statistical temperature annealing (STA), statistical temperature	SHH, IPS (S13, 21, 34)	8

	molecular dynamics (STMD), and force-biased multicanonical molecular dynamics (FBMMD)		
Zhang and Ma	Molecular dynamics with Wang-Landau sampling (MDWL)	SHH, IPS (S34)	9
Schnabel, Bachmann, and Janke	Multicanonical Monte Carlo (MMC)	SHH (S20.1, 20.3 and 20.4)	10
Lee, Joo, Kim, and Lee	Conformational space annealing (CSA)	SHH, IPS (S13, 21, 34)	11
Arkin	Energy landscape paving (ELP)	IPS (S13, 21, 34)	12
Zhang and Ma	Temperature space random walk (TSRW)	IPS (S34)	13
Kalegari and Lopes	Differential evolution (DE)	IPS (S34)	22
Parpinelli, Benítez, Cordeiro, and Lopes	Particle swarm optimization (PSO)	IPS(S13, 21, 34)	23

Table 3. The lowest potential energies obtained for the sequences S20.1 – S20.6.

Method	S20.1	S20.2	S20.3	S20.4	S20.5	S20.6
	SHH					
MMC[10]	–33.8236		–33.5838	–34.4892		
MMC[5]	–33.766	–33.920	–33.582	–34.496	–19.647	–19.322
ELP[5]	–33.810	–33.926	–33.578	–34.498	–19.653	–19.326
This work	–33.8433	–33.9449	–33.6097	–34.5263	–19.6614	–19.3468
	IPS					
MC[1]	–58.3			–57.2		
MMC[5]	–58.306	–58.880	–59.293	–59.068	–51.525	–53.359
ELP[5]	–58.317	–58.914	–59.338	–59.079	–51.566	–53.417
This work	–58.3246	–58.9202	–59.3476	–59.0913	–51.5721	–53.4226

Table 4. The lowest potential energies obtained for the sequences, S13, S21, and S34.

Method	SHH			IPS		
	S13	S21	S34	S13	S21	S34
PER[2]	–4.9616	–11.5238	–21.5678			
ACMC[3]				–26.5066	–51.7175	–94.0431

CSA[4]	−4.9746	−12.3266	−25.5113	−26.4714	−52.7865	−97.7321
MMC[5]	−4.967	−12.296	−25.321	−26.496	−52.915	−97.273
ELP[5]	−4.967	−12.316	−25.476	−26.498	−52.917	−97.261
PER[6]	−4.9717	−11.9834	−24.0224			
HCG[6]	−4.9744	−12.2553	−24.8149			
HA[7]	−4.9746	−12.0617	−23.0441			
STA[8]	−4.9746	−12.3266	−25.5113	−26.5066	−52.9339	−97.6171
STMD[8]	−4.9667	−12.3176	−25.4932	−26.5052	−52.9100	−97.5570
FBMMD[8]				−26.4354	−52.7040	−97.3281
MDWL[9]						−98.3571
CSA[11]	−4.9746	−12.3266	−25.5113	−26.5066	−52.9339	−98.3571
ELP[12]				−26.39	−52.67	−97.32
TSRW[13]						−98.3571
DE[22]				−26.507	−50.3613	−92.0962
PSO[23]				−24.888	−46.611	−80.409
This work	−4.9746	−12.3266	−25.5113	−26.5066	−52.9339	−98.3571

2.2 Geometry Optimization

2.2.1 Algorithm

The present optimization method was developed from the algorithm for atomic LJ clusters.[29] In the algorithm, cluster geometries are optimized with two geometrical perturbations followed by local optimizations. These perturbations are made by interior and surface operators (see Figure 3). The interior operator moves the highest-energy atom or group to the interior of the geometry and the surface one moves it to the most stable positions on the surface. Similar geometrical perturbations are carried out for geometries of proteins; as shown in Figure 3, the highest-energy bead is moved to the interior of the protein (center-directed bead move) and to a position in the vicinity of it keeping the two rigid bonds (one bead rotation). These moves are exactly extension of the interior and surface operators and no information on the global minima previously obtained for the protein models is used for the development. The details of the geometrical perturbations are described in the next subsection. The local optimization is essential to enhance efficiency of optimization methods as shown by Wales et al.[32] in the study using the basin-hopping algorithm.

The flowchart for the developed method is shown in Figure 4. The optimization method adopts a monotonic descent algorithm (energy lowering is acceptable) and a lot of initial geometries are used for exploration of the potential energy surface. The values of the bond angles and dihedral angles of initial geometries are randomly generated. These values are locally optimized with a limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) quasi-Newton method.[33] The local optimization terminates when the following condition is satisfied for the IPS potential:

$$\sqrt{\sum_{i=1}^{2N-5} g_i^2 \Big/ (2N-5)} \le 10^{-3} \tag{6}$$

Here g_i means the gradient of E with respect to the ith independent structural parameter. For the SHH potential, the condition is modified by changing the right-side value from 10^{-3} to 10^{-2} because of the differences between the values of g_i for the SHH and IPS potential functions. The optimal potential energies of the proteins used in the present study take the accuracy of the order of 1×10^{-5} for both the potentials. In the L-BFGS method, the approximations of inverse Hessian matrix are derived from successive corrections.[34] Koslover and Wales[35] show that the number of the successive corrections significantly affects convergence speed. According to the preliminary calculations on the S21 protein, the number of the successive corrections was set to be 200 and 300 for the IPS and SHH potentials, respectively.

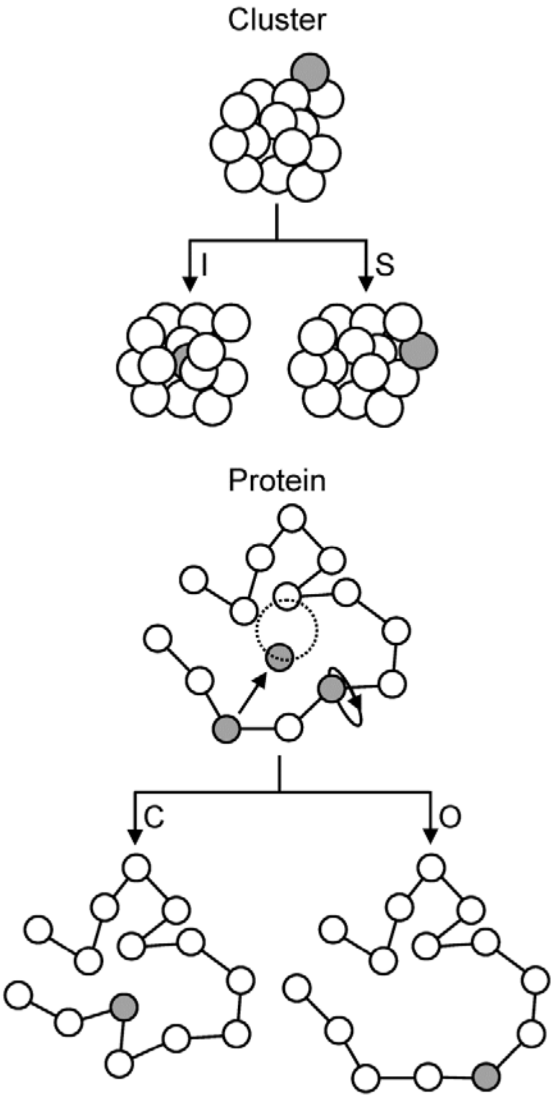

Figure 3. Geometrical perturbations for a cluster and a protein model. New geometries of a cluster are generated by moving the highest-energy atom to the interior (I) and surface (S). For a protein, center-directed bead move (C) and one bead rotation (O) operators perturb the geometry.

After the local optimization of the initial geometry, it is modified with center-directed bead move operator for the highest-energy A bead; the calculation of the energy of each bead is described later. The geometry perturbed with the operator is optimized with the quasi-Newton method. If the potential energy of the model is improved with the local optimization, the geometry is updated. When the potential energy of the model is not improved during the last twenty perturbations, the one bead rotation operator is carried out. Twenty perturbations gave the best efficiency compared with five, ten, forty, sixty, and eighty perturbations.

The center-directed bead move is not carried out for B beads. As mentioned in the subsection 2.1, A beads tend to aggregate in the interior of the model. Consequently B beads prefer exterior positions of the protein. This tendency is contrary to the center-directed bead move operator.

The one bead rotation operator selects the A bead with the highest potential energy (the number of the moved beads *m* is initially set to be 1) and moves it as shown in Figure 3. The local optimization is carried out for the perturbed geometry. If the energy of the protein is not improved by moving the highest-energy bead, the second highest-energy bead and the third highest-energy bead are separately moved to perturb the geometry of the protein. Then the number of the moved beads increases up to 4 one by one if no energy lowering occurs after the local optimization. When energy lowering is observed, the geometry is updated and the number of the moved beads is initialized. If the operator performed for $m = 4$ does not improve the energy, the optimization algorithm proceeds to the one bead rotation operator for B beads. The above mentioned process is repeated for them.

The termination condition is that the lowest energy obtained before the center-directed bead move operator is equal to that obtained after the one bead rotation operator. This means that the update of the geometry is not performed with both the center-directed bead moves and one bead rotations.

Local optimization is necessary as explained by the two following factors. First, the center-directed bead move generates a high-energy configuration with high probability because of short distances between the moved bead and its surrounding beads. Consequently the move rarely lowers the energy of the protein. However, the local optimization significantly improves it, leading to a probability for energy lowering compared with the original energy. This is also the case for one bead rotation. Second, local optimization transforms the potential energy surface to less rugged one, indicating that it makes the search for the global minimum easier.[32,36] The combined use of local optimization and efficient sampling on potential energy surface is performed in the previous studies.[1,6-8,11,20]

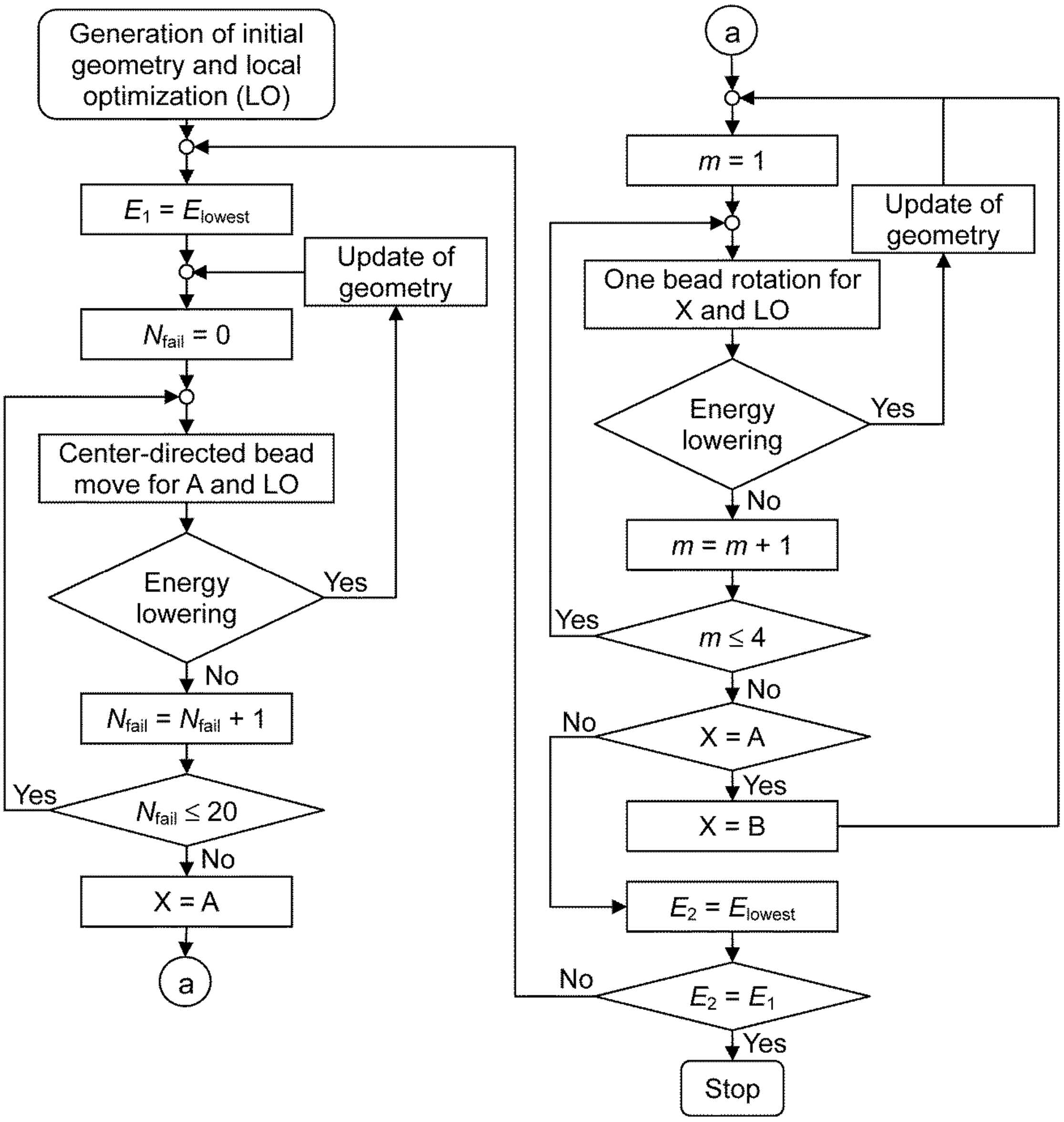

Figure 4. Flowchart of a cycle of the optimization algorithm for the off-lattice bead model.

2.2.2 Geometrical Perturbations

The geometrical perturbation operators move a bead or a few beads with the highest potential energy to different positions. For the IPS model, the potential energy of the *i*th bead $E(i)$ is defined by

$$E(i) = \vec{r}_{i-2,i-1} \cdot \vec{r}_{i-1,i} + \vec{r}_{i,i+1} \cdot \vec{r}_{i+1,i+2} - \frac{1}{2}\left(\vec{r}_{i-3,i-2} \cdot \vec{r}_{i-1,i} + \vec{r}_{i,i+1} \cdot \vec{r}_{i+2,i+3}\right) + \sum_{j \neq i,i\pm 1} V(i,j) \tag{7}$$

For $i = 1, 2, 3, N - 2, N - 1, N$, one or two terms in Eq. (7) are omitted since subscripts are not positive or exceed N. Eq. (7) satisfies the following equation:

$$E = \frac{1}{2}\sum_{i=1}^{N} E(i) \tag{8}$$

Hence the potential energy of each of the beads equivalently contributes (with the same weight) to the total energy of the protein. The bead with the highest potential energy is selected referring to Eq. (7). On the other hand, m beads ($m \neq 1$) with the highest potential energy are selected as follows: (i) For all combinations of m beads (numbering of m beads is represented by $k_1, k_2, \ldots, k_m$), calculate the contribution $E(k_1, k_2, \ldots, k_m)$ of m beads to the potential energy of the protein. For example, 2 beads i and j have the potential energy:

$$E(i,j) = \begin{cases} E(i) + E(j) - V(i,j) + \vec{r}_{j,j+1} \cdot \vec{r}_{j+1,j+2} & \text{if } j = i + 2 \qquad (9\text{a}) \\ E(i) + E(j) - V(i,j) + \frac{1}{2}\vec{r}_{j,j+1} \cdot \vec{r}_{j+2,j+3} & \text{if } j = i + 3 \qquad (9\text{b}) \\ E(i) + E(j) - V(i,j) & \text{otherwise} \qquad (9\text{c}) \end{cases}$$

The third and fourth terms of the right side are necessary since the terms of opposite sign (e.g. $V(i,j) - \vec{r}_{j,j+1} \cdot \vec{r}_{j+1,j+2}$ for eq. (9a)) are duplicately included in the sum of the first and second terms. The energy of m beads can be evaluated following the above manners. (ii) Select the m beads with the highest potential energy from all the combinations. The number m is a predetermined integer as described before. Eqs. (7) and (9) are modified for the SHH potential since the potential function is different from the IPS one.

The center-directed bead move operator changes the position of the highest-energy A bead (kth bead) to the surface of a sphere. It takes the radius of 0.5 and the center coincident with the center of mass of the protein (see Figure 3). The center of mass of the protein is calculated assuming the unit mass for all beads. A few beads may be moved together with the moved bead to keep the unit bond length (Figure 5). For the jth bead ($j = k + 1$), the new position is given according to the position of the ith bead ($i = k + 2$) by the following equation:

$$\vec{r}_{k,j}^{\,new} = \begin{cases} \vec{r}_{k,j} / \left|\vec{r}_{k,j}\right| & \text{if } \left|\vec{r}_{k,i}\right| \geq 2 \qquad (10\text{a}) \\ \dfrac{1}{2}\vec{r}_{k,i} + \sqrt{1 - \dfrac{\left|\vec{r}_{k,i}\right|^2}{4}}\,\vec{u}_{k,i}^{\perp} & \text{otherwise} \qquad (10\text{b}) \end{cases}$$

Here $\vec{r}_{k,j}$ is the vector directed from the position of the moved A bead to the original positions of the *j*th bead and $\vec{u}_{k,i}^{\perp}$ is an arbitrary unit vector perpendicular to $\vec{r}_{k,i}$. The above equation is repeatedly used until all bonds take the unit length.

The one bead rotation is shown in the middle drawing in Figure 5b. The moved bead corresponds to the *j*th bead in the drawing and the moved position is calculated according to Eq. (10b). This operator is performed for both A and B beads as shown in the algorithm in Figure 4.

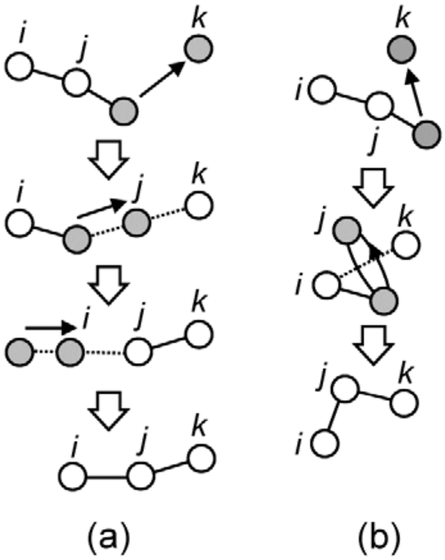

Figure 5. New positions of the bead (*j*th bead) next to the moved bead (*k*th bead). Two types of moves are shown: (a) Case with $\left|\vec{r}_{k,i}\right| \geq 2$. The *i*th bead is also moved; (b) case with $\left|\vec{r}_{k,i}\right| < 2$.

3 Results and Discussion

For each protein, a lot of optimization cycles were performed with different initial geometries. The number of initial geometries N_{opt} is listed in Table 1 together with the number of the optimization cycles which locate the lowest-energy configuration (N_{s}). The lowest energies obtained in the present study are compared with the results of the previous studies as shown in Tables 3 and 4. Since the article of Irbäck et al.[1] did not give numerical data for the lowest energy, approximate values for S20.1 and S20.4 were obtained from the figures in the article.

The successful rate t (= N_s/N_{opt}) is a useful parameter to evaluate the efficancy of the optimization algorithm. The rates for the SHH potential do not significantly differ from those for the IPS one. The effect of the number of beads on the rates is found as follows: t(S34) < t(S20) < t(S21) < t(S13). This is reasonable since an increasing number of the beads mainly extend the searching space on the potential energy surface.

For the S34 model with the IPS and SHH potentials, the number of geometries calculated in a cycle of the algorithm shown in Figure 4 (the number of calculated configurations per initial geometry, N_{conf}) is 150 and 170, respectively (CPU time is 1 – 2 min.). This is reduced to 100 for the S13 model. These geometries are optimized with the quasi-Newton method. The most difficult case in the present study (the S34 model with the IPS potential) required 1.5×10^7 local optimizations to search for the global minimum once.

The comparison of the parameters to evaluate the efficiency, t and N_{conf}, with the corresponding values of the previous studies is impossible since the information is not available from the literature. Hence only the lowest energies of the models are discussed below.

3.1 The Lowest Energies of the Model Proteins

For the S13, S21, and S34 models, the lowest energies are reproduced at least twice in the previous studies as shown in Table 4. Hence these values would correspond to the true global minima. The results for the S34 proteins indicate that the molecular dynamics with Wang-Landau sampling,[9] the conformational space annealing,[11] and the temperature space random walk algorithm[13] are the most efficient.

The lowest energies of S13, S21, and S34 for the IPS potential, −26.5066, −52.9339, and −98.3571, were reproduced by the present method. The best results of S13, S21, and S34 previously obtained for the SHH potential are also in good agreement with the present ones. These results indicate that the present method is useful for calculating optimal geometries of the proteins.

The present method improves the lowest energies of the 20-bead models. Other efficient methods such as the conformational space annealing[11] are necessary to examine if the energies obtained in the present study correspond to the global minima.

The differences between the previous and present potential energies are considered to be partially due to artifact in the literature. For example, in the present study, the global-minimum energy of the S20.1 SHH protein is –33.8433 whereas the second lowest one is –33.8131. Hence the energy reported by Schnabel et al., [10] –33.8236, does not correspond to a local minimum on the potential energy surface.

Similar situations are found for many previous results. In the present study, the second lowest energies of the SHH models of S13, S20.1 to S20.6, S21, and S34 are –4.9627, –33.8131, –33.9364, –33.6058, –34.3658, –19.6182, –19.3441, –12.2912, and –25.4884, respectively. The corresponding values for the IPS potential are –25.9806, –58.1840, –58.7503, –58.7606, –58.6221, –51.5232, –53.3735, –52.4744, and –98.0757, respectively. If the energy in the previous study is lower than the above value and is not equal to the lowest energy listed in Tables 3 and 4, it must be artifact.

The energies of the S13, S21, and S34 IPS models reported in the literature[37] are not listed in Table 4 since the present author could not reproduce them using the geometries given in the supplementary material. This is because there are discrepancies in the implementation of the IPS potential.

3.2 Structures

Although the structures of the IPS and SHH models are clarified, the features of them are not discussed in detail. It is important to examine the features since the information may make optimization methods faster.

The results in the present study show that the single hydrophobic core is formed in each of the lowest-energy configurations of the proteins (Figures S1 and S2, and Table S1 in Supporting Information), in agreement with the results reported by Kim et al.,[4] Kim et al.,[8] and Lee et al.[11] To compare geometries of 200 low-lying configurations of each protein, the principle moments of inertia of them were calculated. Figure 6 shows the principle moments of inertia of the configurations of the S21 protein. The configurations take asymmetric shapes because of $I_A \neq I_B \neq I_C$. The values of the principle moments of inertia for the IPS potential are smaller than the corresponding values for the SHH one. Hence the structures for the IPS potential are more compact than those for the SHH one. This is the case for the other proteins (Figure S3 in Supporting Information). As a feature of the SHH potential, the AXB and BXB bond angles tend to take 180° whereas the corresponding value for the IPS one is 67°. This feature affects the bond angles of 200 low-lying configurations (see Figure 7); the bond angles of the SHH protein are usually larger than those of the IPS one. This results in the compact structures of the IPS protein.

Figure 8 shows the torsional angles of 200 low-lying configurations of S21. The IPS potential restricts possible configurations compared with the SHH one. This is due to the lack of torsional terms in the SHH potential. The lack may lead to the broad torsional distribution. The above discussion on the bond and torsional angles is valid for the other proteins (Figures S4 and S5 in Supporting Information).

The broad distribution for the SHH model (Figure 8) indicates that the barrier to the internal rotation is low in energy. Hence the potential energy surface of the SHH model is less rugged than that of the IPS model, in agreement with the suggestion in the previous study.[11] This may result in the high successful rates for the SHH model compared with the corresponding rates for the IPS one. However, this is the case for only the S13, S20.5 and S34 models as shown in Table 1. For the other models, the successful rates would be considerably affected by separation of the deepest basin from the other basins[4] and the width of the deepest basin.[11]

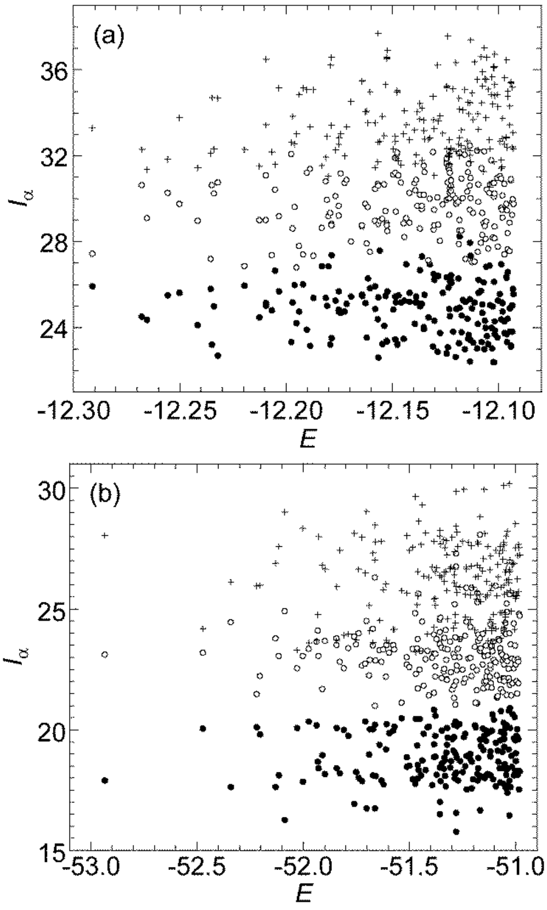

Figure 6. The principle moments of inertia of 200 low-lying configurations of the S21 protein: (a) SHH potential and (b) IPS potential. The values of I_A, I_B and I_C are plotted with symbols of closed circles, open circles, and pluses, respectively.

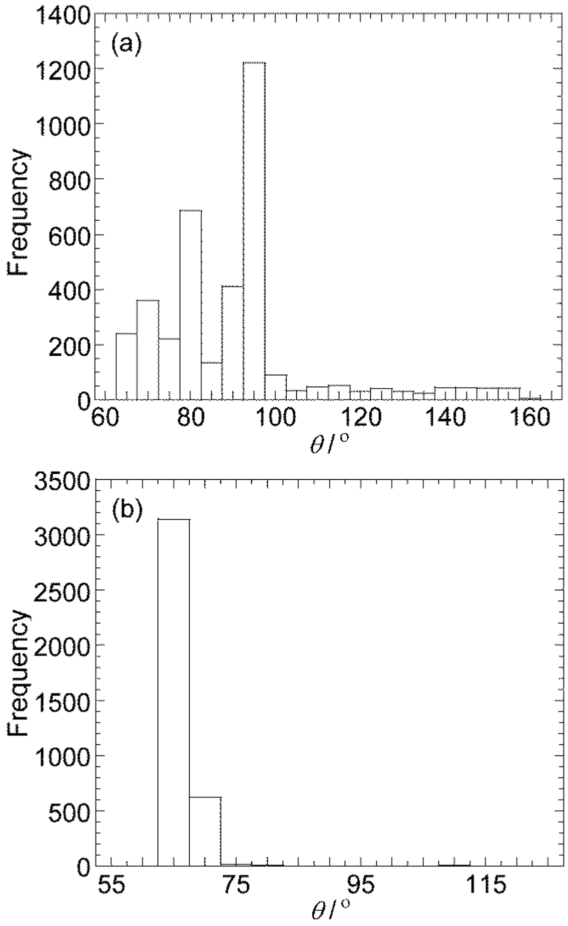

Figure 7. The bond angles of 200 low-lying configurations of the S21 protein: (a) SHH potential and (b) IPS potential.

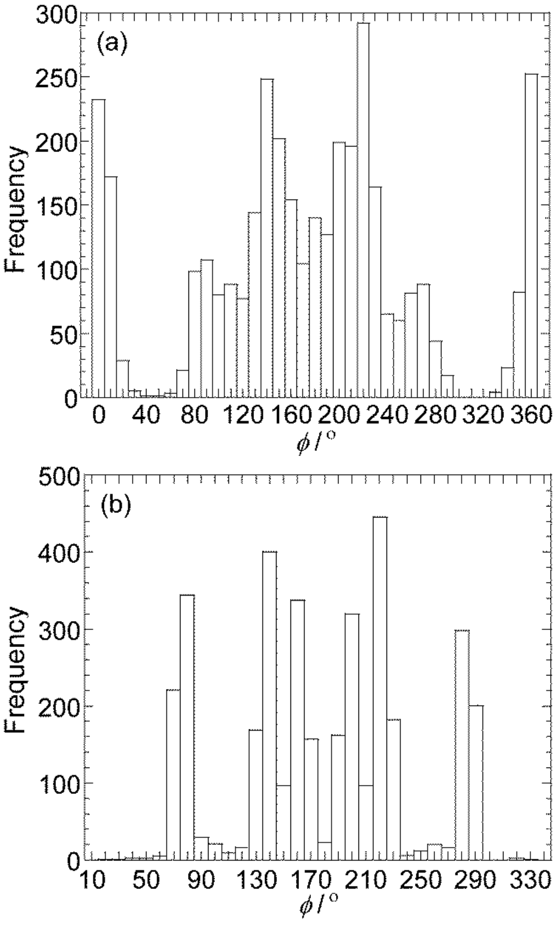

Figure 8. The torsional angles of 200 low-lying configurations of the S21 protein: (a) SHH potential and (b) IPS potential.

3.3 Efficiency of the Geometrical Perturbations

The efficiency of the optimization method is determined by the ability of the center-directed bead move and one bead rotation operators to lower potential energies. Table 5 lists the potential energies lowered by the operators; the data were obtained from the results of all calculations. The center-directed bead move operator considerably reduces the energies of the proteins with the IPS and SHH potentials. It is shown using the correlation between compactness of structures and energies that the collapse of the model proceeds with decreasing energy.[8] Hence the center-directed bead move operator exclusively induces the collapsed structure with the hydrophobic core. The resulting structure is locally modified with the one bead rotation operator. These optimization steps are similar to the folding process of the collapse model. [38]

Table 5. Energy lowering of the SHH and IPS models induced by the center-directed bead move (C) and one bead rotation (O) perturbations. The values averaged over all optimization calculations are listed.

	S13	S20.1	S20.2	S20.3	S20.4	S20.5	S20.6	S21	S34
					SHH				
C	−2.95	−4.80	−5.41	−10.24	−9.37	−9.47	−10.02	−7.23	−12.93
O	−0.04	−0.18	−0.17	−0.16	−0.16	−0.20	−0.16	−0.09	−0.56
					IPS				
C	−4.20	−7.39	−7.58	−8.18	−7.98	−7.21	−7.72	−7.60	−13.52
O	−0.21	−0.90	−0.78	−0.72	−0.62	−0.66	−0.60	−0.60	−1.30

As discussed above, the center-directed bead move perturbation is an efficient strategy for geometry optimization. The efficiency originates from the following factors: (1) As described before, the A beads prefer to occupy the core of the protein. The core formation of the hydrophobic beads is enhanced with the operator; (2) The energy of the hydrophobic beads near the center of the protein contributes to the total energy more significantly than that of the hydrophilic beads in the surface. This suggests that perturbations of the interior part of the geometry give possibilities for considerably lowering the potential energy; (3) The perturbation induced by the move of an A bead diffuses to its near beads through the structural change shown in Figure 5a and the local optimization. This would lead to improvement of the whole geometry of the protein.

Since the one bead rotation operator generates geometries which cannot be produced by the center-directed bead move one, it works as a supplementary perturbation for the center-directed bead move one. Hence the combined use of the two perturbations enhances the efficiency of the present optimization method.

The results of the LJ clusters[29] show that the interior operator as shown in Figure 3 significantly improves the potential energies. The study on the ternary and quaternary LJ clusters[31] suggests that the atom-type conversion for interior atoms is efficient for improving the energies and structures. According to these results as well as the present results, it is considered that perturbations of the interior combined with local optimizations are an excellent optimization algorithm commonly used for aggregations of atoms independently of the existence of bonds.

3.4 Application to Larger Protein Models

Some studies[2-9,11,13,15] treat larger proteins of 55 and 89 beads with Fibonacci sequences as described below. Global optimizations of these models are so difficult that no consensus on the global minimum of each protein can be obtained.

The present method was used for geometry optimization of the 55-bead protein. The sequence is expressed by the S21 sequence combined with the S34 one. The lowest energy of the IPS protein was obtained to be −174.9241 from 200000 optimization cycles. The energy of the protein is lower than those given in the literature (−154.505,[3] −173.9803,[4] −172.696,[5] −174.5681[8], −157.112[22], and −115.758[23]) but is higher than the values of −176.6913[11] and −178.1339.[9,13] For the SHH potential, 100000 optimization cycles located the configuration with the lowest energy of −43.3658. The value is lower than the values reported previously, −32.8843[2], −42.3418[4], −42.428[5], −42.5195[6], −38.1977[7], −42.5781,[8] and −42.5781.[15] However, the literature[9, 11] gives the energies of −44.8765 and −44.7983. Since the calculation of the 55-bead protein was time-consuming (it took a few months in this work), no more geometry optimization could be carried out. Execution of more calculations might yield energies comparable to the putative lowest energies in the literature.[9,11,13] The global minimum energy for the 89-bead protein is reported to be −311.6134.[13] It is one of the most challenging problems which must be solved with geometry optimization methods.

To obtain the global minimum of the 55-bead protein, a powerful multi-core computer would be useful. Different approaches can be derived by modifying initial geometry generator and geometrical perturbations. Since a lot of initial configurations are randomly generated, many optimization steps are needed for energy lowering of them. The present method aided with the strategy to generate promising initial configurations[7] is a candidate of modifications. The geometrical perturbations are unique compared with geometrical changes in the other optimization methods.[1-26] The combined use of the perturbations and other algorithms may increase the efficiency of the optimization method. For example, a hybrid algorithm of the geometrical perturbations and the crossover operator which is used in one of the most efficient methods, the conformational space annealing,[11] could perform geometry optimization better than the single algorithm.

Optimization methods are divided into two groups, unbiased and biased ones. The above modifications are based on unbiased algorithms. Biased algorithms usually search for global minima quicker than unbiased ones since the former restricts search space on the potential energy surfaces. The incorporation of the geometrical features obtained in the section 3.2 (constraints for bond and torsional angles) into the method may generate promising geometries at initial and perturbation steps in the algorithm.

The present method would be applicable to small proteins with realistic potentials. If the above mentioned improvements are performed, the application would be extended to larger protein models. In the future studies, more realistic models[39-42] will be useful for evaluating the improved method. For the models, a new optimization method could be developed using the secondary structures predicted with accuracy of approximately 80%.[43,44]

4 Conclusions

The present study is aimed at developing an efficient optimization algorithm for proteins. The off-lattice model described with two kinds of beads was used to examine the efficiency of the optimization method. Two types of geometrical perturbations were developed referring to the algorithm proposed for atomic clusters. The center-directed hydrophobic bead move operator proposed in the present study is crucial for the geometry optimization of the protein model since it significantly lowers the potential energy. Efficiency of the interior geometrical perturbations is found for the Lennard-Jones clusters. Hence the center-directed particle move followed by local optimization would be efficient for any systems. Incorporation of the secondary structure prediction and/or other optimization strategies into the present algorithm would be a future direction of the optimization algorithm for proteins with realistic potentials.

Acknowledgement

The author thanks the anonymous reviewers for helpful discussion and comments.

References

[1] A. Irbäck, C. Peterson, F. Potthast, O. Sommelius, *J. Chem. Phys.* **1997**, *107*, 273–282.

[2] H. -P. Hsu, V. Mehra, P. Grassberger, *Phys. Rev. E* **2003**, *68*, 037703.

[3] F. Liang, *J. Chem. Phys.* **2004**, *120*, 6756–6763.

[4] S. -Y. Kim, S. B. Lee, J. Lee, *Phys. Rev. E* **2005**, *72*, 011916.

[5] M. Bachmann, H. Arkin, W. Janke, *Phys. Rev. E* **2005**, *71*, 031906.

[6] W. Huang, J. Liu, *Biopolymers* **2006**, *82*, 93–98.

[7] M. Chen, W. Huang, *J. Zhejiang Univ. SCIENCE B* **2006**, *7*, 7–12.

[8] J. Kim, J. E. Straub, T. Keyes, *Phys. Rev. E* **2007**, *76*, 011913.

[9] C. Zhang, J. Ma, *Phys. Rev. E* **2007**, *76*, 036708.

[10] S. Schnabel, M. Bachmann, W. Janke, *Phys. Rev. Lett.* **2007**, *98*, 048103.

[11] J. Lee, K. Joo, S. -Y. Kim, J. Lee, *J. Comput. Chem.* **2008**, *29*, 2479–2484.

[12] H. Arkin, *Phys. Rev. E* **2008**, *78*, 041914.

[13] C. Zhang, J. Ma, *J. Chem. Phys.* **2009**, *130*, 194112.

[14] A. Irbäck, C. Peterson, F. Potthast, *Phys. Rev. E* **1997**, *55*, 860–867.

[15] J. Kim, J. E. Straub, T. Keyes, *Phys. Rev. Lett.* **2006**, *97*, 050601.

[16] J. Kim, T. Keyes, J. E. Straub, *Phys. Rev. E* **2009**, *79*, 030902(R).

[17] T. Dash, P. K. Sahu, *J. Comput. Chem.* **2015**, *36*, 1060−1068.

[18] F. H. Stillinger, T. Head-Gordon, C. Hirshfeld, *Phys. Rev. E* **1993**, *48*, 1469−1477.

[19] F. H. Stillinger, T. Head-Gordon, *Phys. Rev. E* **1995**, *52*, 2872−2877.

[20] J. Liu, S. Xue, D. Chen, H. Geng, Z. Liu, *J. Biol. Phys.* **2009**, *35*, 245−253.

[21] A. Irbäck, F. Potthast, *J. Chem. Phys.* **1995**, *103*, 10298−10305.

[22] D. H. Kalegari, H. S. Lopes, Proc. 2013 IEEE Symposium on Differential Evolution (SDE), IEEE, 2013, 143–150.

[23] R. S. Parpinelli, C. M. V. Benítez, J. Cordeiro, H. S. Lopes, *J. Mult.-Valued Logic & Soft Computing*, **2014**, *22*, 267–286.

[24] J. Liu, Y. Sun, G. Li, B. Song, W. Huang, *Comput. Biol. Chem.* **2013**, *47*, 142–148.

[25] B. Li, Y. Li, L. Gong, *Eng. Appl. Artif. Intell.* **2014**, *27*, 70–79.

[26] R. S. Parpinelli, H. S. Lopes, Proc. 2013 Brazilian Conference on Intelligent Systems, IEEE, 2013, 64–69.

[27] B. Hartke, *WIREs Comput. Mol. Sci.* **2011**, *1*, 879–887.

[28] J. M. C. Marques, F. B. Pereira, *J. Mol. Liq.* **2015**, *210*, 51−63.

[29] H. Takeuchi, *J. Chem. Inf. Model.* **2006**, *46*, 2066–2070.

[30] H. Takeuchi, *Comput. Theoret. Chem.* **2014**, *1050*, 68−73.

[31] H. Takeuchi, *Chem. Phys.* **2015**, *457*, 106−113.

[32] D. J. Wales, J. P. K. Doye, *J. Phys. Chem. A* **1997**, *101*, 5111−5116.

[33] D. C. Liu, J. Nocedal, *J. Math. Prog.* **1989**, *45*, 503−528.

[34] E. K. P. Chong, S. H. Źak, An introduction to optimization, John Wiley & Sons: New Jersey, 2008; Chapter 11.

[35] E. F. Koslover, D. J. Wales, *J. Chem. Phys.* **2007**, *127*, 234105.

[36] M. C. Prentiss, D. J. Wales, P. G. Wolynes, *J. Chem. Phys.* **2008**, *128*, 225106.

[37] I. Kolossváry, K. J. Bowers, arXiv: 1205.4705 [physics.comp-ph], Cornell University Library.

[38] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, H. S. Chan, *Protein Science* **1995**, 561–602.

[39] J. D. Honeycutt, D. Thirumalai, *Proc. Natl. Acad. Sci. USA*, **1990**, *87*, 3526–3529.

[40] S. Brown, N. J. Fawzi. T. Head-Gordon, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 10712−10717.

[41] A. Irbäck, F. Sjunnesson, S. Wallin, *Proc. Natl. Acad. Sci. USA*, **2000**, *97*, 13614−13618.

[42] J. Maupetit, P. Tuffery, P. Derreumaux, *Proteins* **2007**, *69*, 394–408.

[43] D. W. A. Buchan, S. M. Ward, A. E. Lobley, T. C. O. Nugent, K. Bryson, D. T. Jones, *Nucleic Acids Res*. **2010**, *38*, W563−W568.

[44] A. Drozdetskiy, C. Cole, J. Procter, G. J. Barton, *Nucleic Acids Res*. **2015**, *43*, W389−W394.