Structure optimization by conformational space annealing in an off-lattice protein model

Seung-Yeon Kim,[1] Sang Bub Lee,[1,2] and Jooyoung Lee[1]
[1]*School of Computational Sciences, Korea Institute for Advanced Study, 207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-722, Korea*
[2]*Department of Physics, Kyungpook National University, Taegu 702-701, Korea*
(Received 17 February 2005; published 26 July 2005)

The optimization results by conformational space annealing are presented for an off-lattice protein model consisting of hydrophobic and hydrophilic residues in Fibonacci sequences. The ground-state energies found are lower than those reported in the literature. In addition, the ground-state conformations in three dimensions exhibit the important aspect of forming a single hydrophobic core in real proteins. The energy landscape for the population of local minima is also investigated.

DOI: 10.1103/PhysRevE.72.011916 PACS number(s): 87.15.By, 87.10.+e, 05.10.−a

I. INTRODUCTION

As functions of proteins are known to be associated with their three-dimensional structures, understanding the protein structures is crucial in biological sciences. Unfortunately, the full three-dimensional structure is known only for far less than 1% of the proteins whose amino acid sequences are realized. Therefore, predicting the structure of a protein from its amino acid sequence is one of the most challenging problems in computational biology. Based on the minimum free-energy theory [1], many authors have developed various techniques to search for the structures of proteins, such as simulated annealing [2], genetic algorithm [3], simulated tempering [4,5], parallel tempering [6,7], multicanonical sampling [8,9], and conformational space annealing [10,11], to mention a few.

The native structure of a protein is associated with the structure of the global minimum of the free energy consisting of the intramolecular interaction among protein atoms and the intermolecular interaction between the protein and surrounding solvent molecules [1]. Since solving such a problem is too difficult for realistic protein models, often highly simplified but yet nontrivial model proteins are studied instead. The HP model by Dill *et al.* [12] is one such model, in which sequences consisting of two types of amino acids, the hydrophobic (A) and the hydrophilic (B) monomers, are configured as self-avoiding walks on square and simple cubic lattices. Stillinger and co-workers [13,14] studied a similar *off-lattice* AB protein model in two dimensions. Center-doped sequences and Fibonacci sequences of A and B monomers were studied using a potential including bending energy and Lennard-Jones energy. *Optimal* structures were sampled by the high-temperature Monte Carlo (HTMC) algorithm followed by subsequent local energy minimization by quasi-Newton and conjugate gradient routines [15,16]. Irbäck *et al.* have also considered the AB model but with random sequences to study the thermodynamic properties of folding to native states, using simulated tempering and multicanonical ensemble techniques [17–19]. Khokhlov and Khalatur used a similar model to design proteinlike copolymers consisting of A and B monomers, but with only the hydrophobic interactions [20].

Among these, only limited works were devoted to a search for the optimal structures with lowest energies and, even for the known minimum-energy conformations, no rigorous proof exists that they are indeed the global minimum, not one of many metastable local minima. Therefore, even for these simple model proteins, searching for the global minimum-energy conformations is still nontrivial, mainly due to the huge number of local minima in the energy landscape. Recently, Hsu, Mehra, and Grassberger [21] studied an off-lattice AB model in two and three dimensions using the improved pruned enriched Rosenbluth method with importance sampling (referred to as nPERMis) [22]. Fibonacci sequences of A and B monomers were studied. Low-energy conformations obtained in this approach are subsequently refined by applying the conjugate gradient descent method to obtain the final minimum-energy state. They have obtained putative optimal structures with energies lower than those reported earlier in Ref. [14]. Liang [23] have studied the same model by the annealing contour Monte Carlo (ACMC) algorithm and have reported the structures with even lower energies in two dimensions. (His work in three dimensions uses a different energy function from the work of Ref. [21].) More recently, Bachmann, Arkin, and Janke employed an elaborate technique of energy landscape paving minimizer (ELP), together with multicanonical sampling, and obtained even better results [24]. It is thus still not clear if the reported "optimal" structures are indeed the global minima in the seemingly complicated energy landscape.

In this paper, the AB model with Fibonacci sequences of A and B atoms is investigated in two and three dimensions using the conformational space annealing (CSA) method [10,11,25–29]. First, we find new candidates for global minima with energies lower than reported in the literature. Second, based on more complete conformational analysis on the energy landscape of the AB model, we find that the model in two dimensions does not provide low-energy structures containing single hydrophobic cores often observed in real proteins. On the other hand, the model in three dimensions provides the lowest-energy structures containing single hydrophobic cores in contrast with existing results.

II. MODELS AND METHODS

The AB model consists of hydrophobic A monomers and hydrophilic B monomers. The subclass of Fibonacci sequences is defined recursively by

$$\Lambda_0 = \mathrm{A}, \quad \Lambda_1 = \mathrm{B}, \quad \Lambda_i = \Lambda_{i-2} * \Lambda_{i-1}, \tag{1}$$

where the asterisk denotes the concatenation operator. The present work will focus on the cases of $6 \leqslant i \leqslant 9$. For $i \leqslant 5$, chains are quite short and their optimal conformations can be obtained in a trivial fashion [14]. The sequences of chains are denoted as S_{13}, S_{21}, S_{34}, and S_{55} for, respectively, $i=6$, 7, 8, and 9, where the subscripts indicate the sizes of chains.

In two dimensions, we consider the energy functions consisting of the bending energy and the van der Waals interaction energy, given as

$$E_2 = \sum_{i=1}^{N-2} \frac{1}{4}(1+\cos\theta_{i,i+1}) + 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N} [r_{ij}^{-12} - C_2(\sigma_i,\sigma_j) r_{ij}^{-6}], \tag{2}$$

where $\theta_{i,j}$ is the angle between the ith and jth bonds and r_{ij} is the distance between monomers i and j. The constant $C_2(\sigma_i,\sigma_j)$ is +1, $+\frac{1}{2}$, and $-\frac{1}{2}$ for, respectively, AA, BB, and AB pairs, thus yielding a strong attraction between AAs, a weak attraction for BBs, and a weak repulsion between ABs. All bond lengths are fixed to unity. In three dimensions, on the other hand, previous works considered two different energy functions depending on the authors. Recent work of Ref. [21] considered the energy function given in Eq. (2), neglecting the torsional energies. Since it is known that, among the two bonded interaction terms the bending energy is the most crucial [19], the torsional energy may be neglected for simplicity. However, there are other works [19], including the more recent work of Ref. [23], that considered the torsional energy implicitly, with the energy function given as

$$E_3 = \sum_{i=1}^{N-2} \cos\theta_{i,i+1} - \frac{1}{2}\sum_{i=1}^{N-3} \cos\theta_{i,i+2} + 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N} C_3(\sigma_i,\sigma_j) \times (r_{ij}^{-12} - r_{ij}^{-6}), \tag{3}$$

where $C_3(\sigma_i,\sigma_j)$ is +1 for AA pairs and $+\frac{1}{2}$ for BB and AB pairs. Therefore, all nonbonded interactions are attractive but AA interactions carry the highest weight. In the present work, we consider both energy functions. We refer to the model with the energy function in Eq. (2) as model I, and that with Eq. (3) as model II. It should be noted that the energy function of model I is identical to that of Ref. [21], while that of model II is equivalent to that of Ref. [23].

Our global optimization is the CSA method developed by Lee, Scheraga, and Rackovsky [10], which unifies the essential ingredients of the simulated annealing [30], genetic algorithm [31], and Monte Carlo with minimization [32]. We realize that there are other optimization algorithms; however, we find that CSA is simpler and more efficient for relatively short chains. For the details of the CSA method, the readers should consult with the authors' published papers [10,11,25–29].

In our CSA method, 50 random initial conformations are generated to divide the phase space, and an additional 50 conformations are added whenever the search space is enlarged. When generating initial conformations in three dimensions, the inversion symmetry in dihedral angles ($\{\phi\} \rightarrow \{-\phi\}$) is taken into account. Consequently, the first dihedral angle ϕ_1 is selected randomly in the range of $[0,\pi]$, while the rest $N-3$ angles are sampled in the whole range of $[0,2\pi]$.

TABLE I. The lowest-energies of the two-dimensional AB model by conformational space annealing (CSA), in comparison with those by the high-temperature Monte Carlo method (HTMC), the improved pruned-enriched-Rosenbluth method with importance sampling (nPERMis), and the annealing contour Monte Carlo method (ACMC).

	HTMC	nPERMis	ACMC	This work
S_{13}	−3.2235	−3.2939	−3.2941	−3.2941
S_{21}	−5.2881	−6.1976	−6.1979	−6.1980
S_{34}	−8.9749	−10.7001	−10.8060	−10.8060
S_{55}	−14.4089	−18.5154	−18.7407	−18.9110

III. RESULTS AND DISCUSSIONS

The full sequences of chains corresponding to Λ_6, Λ_7, Λ_8, and Λ_9 are given, respectively, as

$$S_{13} = \mathrm{ABBABBABABBAB},$$

$$S_{21} = \mathrm{BABABBAB\ ABBABBABABBAB},$$

$$S_{34} = \mathrm{ABBABBABABBAB\ BABABBAB} \times \mathrm{ABBABBABABBAB},$$

and

$$S_{55} = \mathrm{BABABBAB\ ABBABBABABBAB} \times \mathrm{ABBABBABABBAB\ BABABBAB} \times \mathrm{ABBABBABABBAB}.$$

Table I shows the results of the lowest energy in two dimensions, along with the values reported in the literature by HTMC [14], nPERMis [22], and ACMC [23] for comparison. Our results for S_{13}, S_{21}, and S_{34} are identical to the ACMC results but slightly better than the nPERMis results. This indicates that the results obtained by ACMC and CSA may indeed correspond to the ground-state energies. However, for S_{55}, our result for the lowest-energy conformation is approximately 1% lower in energy compared to that by ACMC, demonstrating that the best conformation reported by ACMC corresponds to a metastable local minimum.

Figure 1 depicts the lowest-energy conformations in two dimensions. The conformations for S_{13} and S_{21} are identical to those by nPERMis and ACMC (see, e.g., Fig. 1 of Refs. [21,23]), as can be expected from their almost identical ground-state energies in Table I. This implies that all three methods find ground-state conformations reasonably well for these short chains. For S_{34}, on the other hand, the lowest-energy conformation by CSA is identical to that by ACMC but is different from that by nPERMis. For S_{55}, the best conformation obtained by CSA differs from that by ACMC,

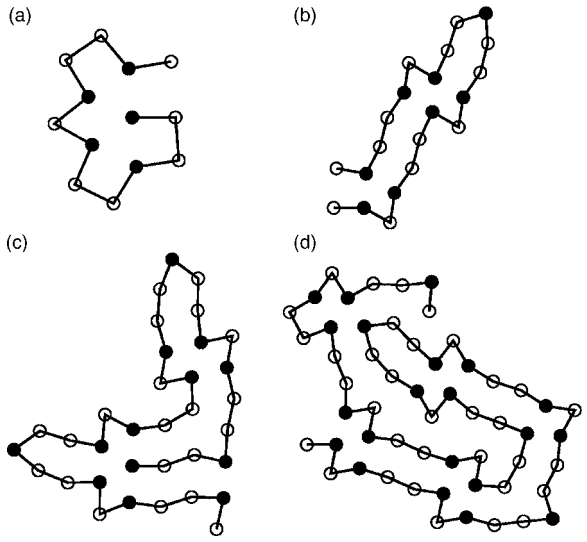

FIG. 1. The lowest-energy conformations of the two-dimensional AB model for (a) S_{13}, (b) S_{21}, (c) S_{34}, and (d) S_{55}. The filled symbols represent hydrophobic A monomers and the open symbols hydrophilic B monomers.

as the energy difference between them is significantly large.

Table II shows the lowest energies of our model I and model II in three dimensions, along with the results by nPERMis, ACMC, and ELP. Our results for model I are better than those of the nPERMis for all cases, with the energy difference increasing gradually for longer chains, and are also slightly better than those by ELP except for S_{55}. (For S_{55}, we believe that we would obtain similar results if we scanned a wider search space.) It should be noted that all of our results are consistent with the corresponding results by ELP within less than 0.2%. Results for model II are also better than ACMC results except for S_{13}. For S_{13}, our result is smaller by 0.13% than that of ACMC. (We carried out intensive calculations; however, we were not able to reach the energy by ACMC.) On the other hand, for other cases, our results are better by 2% for S_{21} and up to as much as 11% for S_{55}. This indicates that the conformations reported earlier in Ref. [23] for model II are not for the ground states, but are for metastable local minima. Again, our results are consistent with the corresponding results by ELP, but our results are better for longer chains. (Note that our result for S_{34} is better by about 5% than that by ELP.)

Figure 2 displays the lowest-energy conformations of model I. It should be noted that the conformation for S_{55} obtained by nPERMis contains two hydrophobic cores, while

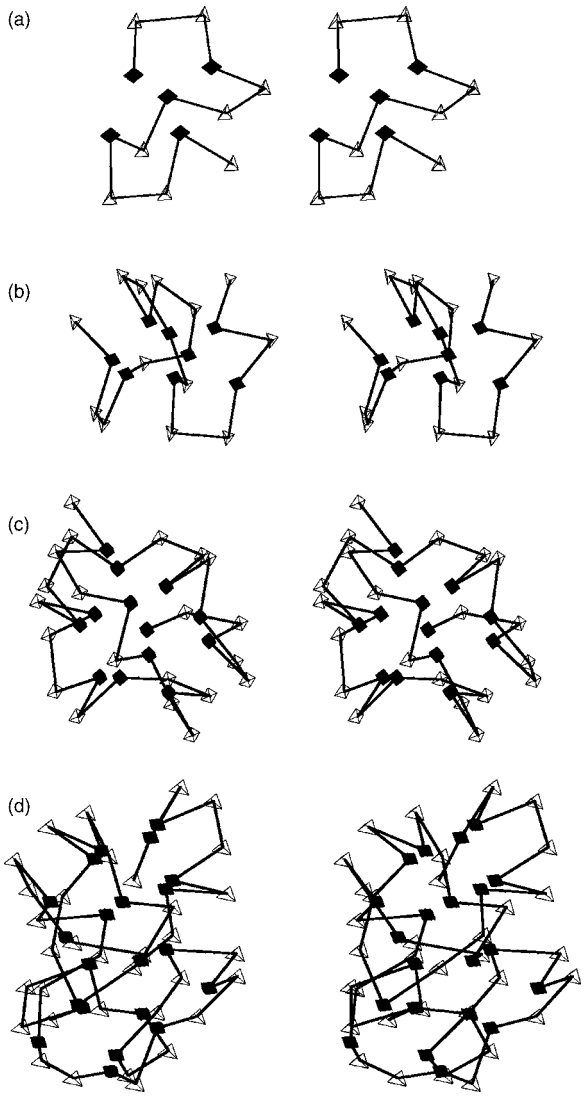

FIG. 2. The stereographic views of the lowest-energy conformations of model I for the three-dimensional AB model for (a) S_{13}, (b) S_{21}, (c) S_{34}, and (d) S_{55}. The filled cubes represent hydrophobic A monomers and the open tetrahedrons hydrophilic B monomers.

our results form a single hydrophobic core, in contrast to the two-dimensional cases, where several clusters of three to five hydrophobic residues were observed. This suggests that the AB model in three dimensions with Fibonacci sequences displays the important feature of forming a single hydrophobic core as observed in real proteins. Figure 3 shows the ground-state conformations of model II. It is clear that our result for

TABLE II. The lowest energies of model I and model II for the three-dimensional AB model by CSA, in comparison with those by nPERMis, ELP, and ACMC, respectively.

	nPERMis	ELP	Model I	ACMC	ELP	Model II
S_{13}	−4.9616	−4.967	−4.9746	−26.5066	−26.498	−26.4714
S_{21}	−11.5238	−12.316	−12.3266	−51.7575	−52.917	−52.7865
S_{34}	−21.5678	−25.476	−25.5113	−94.0431	−92.746	−97.7321
S_{55}	−32.8843	−42.428	−42.3418	−154.5050	−172.696	−173.9803

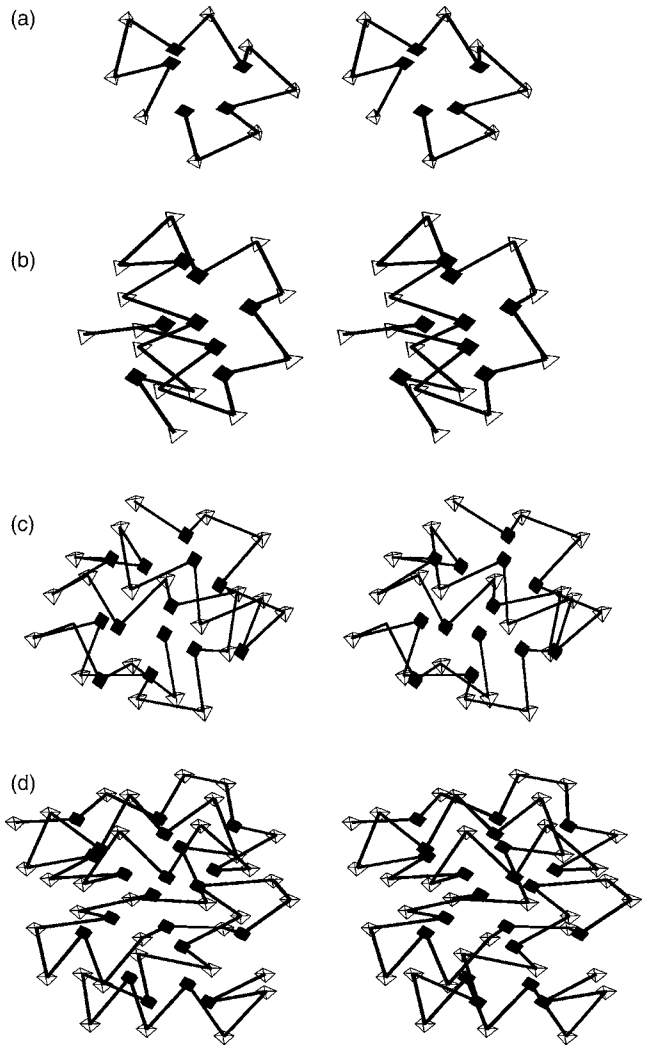

FIG. 3. The stereographic views of the lowest-energy conformations of model II for the three-dimensional AB model for (a) S_{13}, (b) S_{21}, (c) S_{34}, and (d) S_{55}.

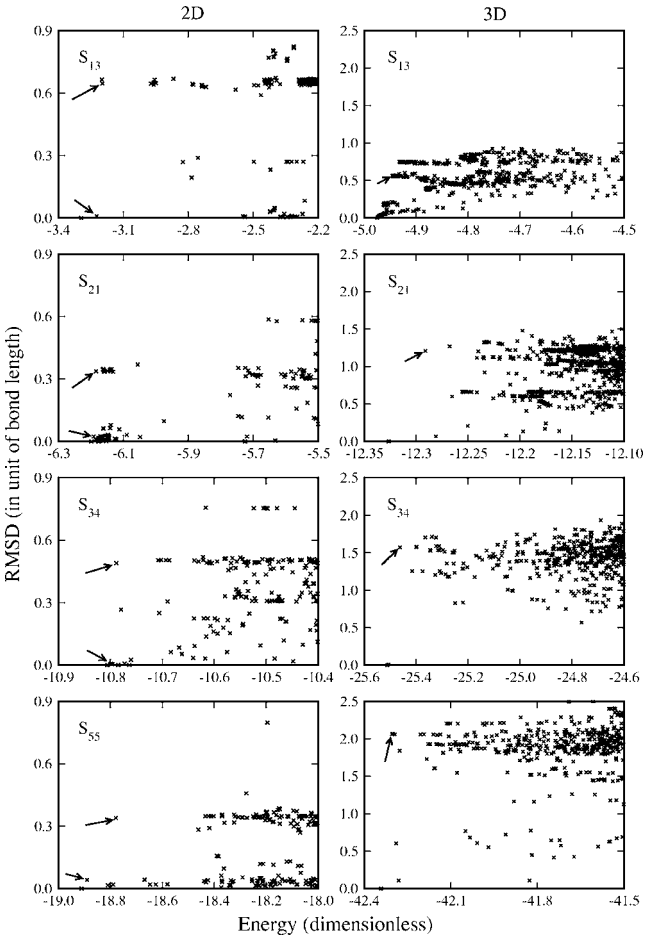

FIG. 4. The distribution of RMSD vs energy for low-lying local minimum-energy conformations for, from top to bottom, S_{13}, S_{21}, S_{34}, and S_{55}. The RMSD is measured from the lowest-energy conformation listed in Tables I and II. The left plots are for two dimensions and the right plots for three dimensions. The conformations indicated by arrows are shown in Figs. 5 and 6.

S_{13} is similar to that of ACMC due to almost identical energies. However, for other cases, the conformations are different from those by ACMC, due to considerably different energies. In all cases, the conformations form a single hydrophobic core, as for the cases of model I. It should be noted that the conformations of model II are more compact than those of model I for the same sequences, due mainly to the attractive long-range interactions among AB atoms.

We also sample the populations of distinct local minimum energy conformations and calculated corresponding RMSDs measured from the global minimum-energy conformation for each sequence. This enables us to investigate energy landscape. In order to describe dimensional differences in energy landscape for the same energy function, we present the results of only model I. Figure 4 shows the distributions of the RMSD versus energy for two and three dimensions. From the figure, we observe three features that describe the differences in the results between two and three dimensions. First, as the chain length increases, the RMSD gaps between the global minima and the conformations indicated by upper arrows decrease in two dimensions, while they increase in three dimensions. Second, in two dimensions, there are many low-energy conformations similar to the global minimum (i.e., those with small RMSD values), while in three dimensions there are not many of those in a relative sense. Third, in two dimensions, many local minima are located close to the global minima and low-energy conformations are divided into two groups (particularly for S_{21}), one with larger RMSD values and the other with smaller RMSD values, while in three dimensions, the global minimum is well separated from the rest of the local minima except for S_{13}. In three dimensions, the global minimum of S_{13} is not isolated in the plot but is close to other local minima. The local minima appear to be divided into several subgroups. It is interesting to observe that the conformations belonging to the lower two groups are similar to each other, and the lowest-energy conformation found by nPERMis belongs to the group with RMSD values close to 0.2. For other sequences, i.e., for S_{21}, S_{34}, and S_{55}, we observe that their global minima are well separated from the rest. We believe that this would make searching for the global minimum particularly difficult.

Low-energy conformations with relatively large RMSD values are also studied. Plotted in Fig. 5 are the conforma-

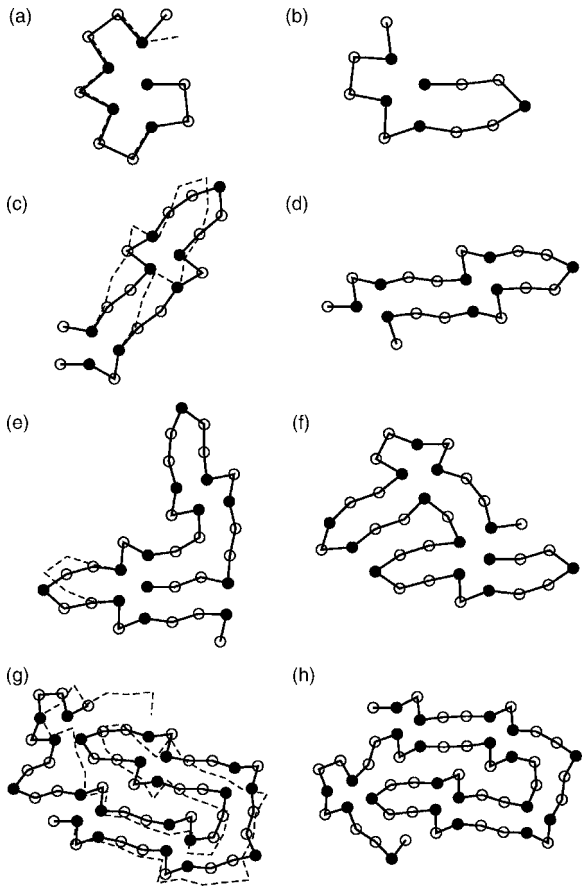

FIG. 5. The low-energy conformations of the two-dimensional AB model with smaller RMSD values (left) and with larger RMSD values (right). The sequences, the values of energy, and RMSD are (a) S_{13} $(-3.2235, 0.0081)$, (b) S_{13} $(-3.1990, 0.6651)$, (c) S_{21} $(-6.1911, 0.0234)$, (d) S_{21} $(-6.1838, 0.3368)$, (e) S_{34} $(-10.8058, 0.0001)$, (f) S_{34} $(-10.7883, 0.4905)$, (g) S_{55} $(-18.8894, 0.0426)$, and (h) S_{55} $(-18.7786, 0.3396)$. The dotted lines superimposed on the left plots represent the lowest-energy conformations.

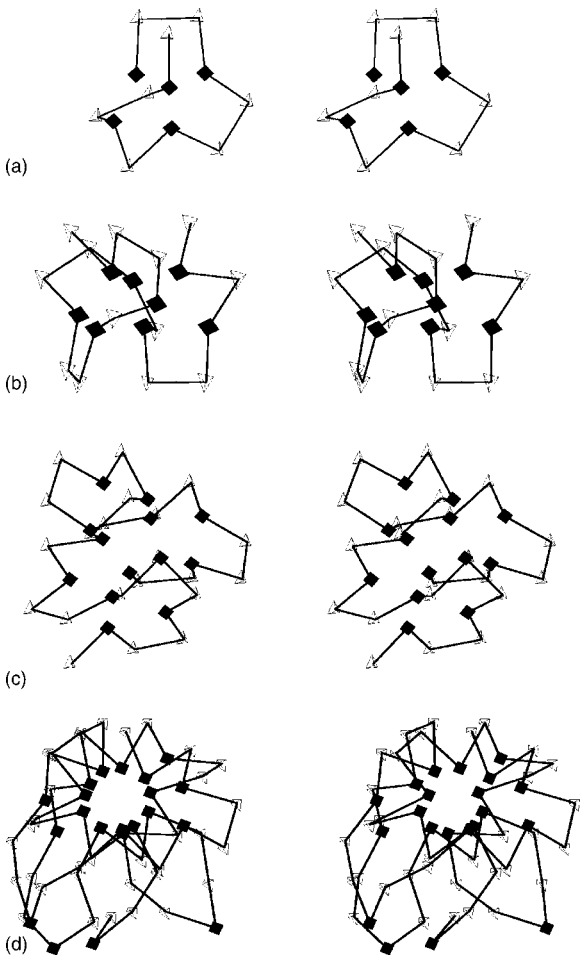

FIG. 6. Stereographic views of the low-energy conformations of model I for the three-dimensional AB model. The conformations are quite distinct from the lowest-energy conformations shown in Fig. 2. The sequences, the values of energy, and RMSD are (a) S_{13} $(-4.9457, 0.5649)$, (b) S_{21} $(-12.2912, 1.2087)$, (c) S_{34} $(-25.4614, 1.5718)$, and (d) S_{55} $(-42.3037, 2.0661)$.

tions in two dimensions; (a), (c), (e), and (g) are for the smaller values of RMSD corresponding to the lower arrows in Fig. 4, and (b), (d), (f), and (h) are for the larger values corresponding to the upper arrows. The conformations look quite different while their energies are close to each other. The source for such energetic similarity comes from the fact that, for each chain, the number and the size of the corresponding hydrophobic clusters are identical (except for S_{13}). It should be emphasized that conformations for S_{55} shown in Figs. 5(g) and 5(h) are of lower energies than those obtained earlier by other methods.

Low-energy conformations, which are quite distinct from the global minima in three dimensions, are plotted in Fig. 6. They are indicated by arrows in Fig. 4. For S_{13}, since the conformations in the two low-RMSD groups in Fig. 4 are similar as mentioned earlier, the minimum-energy conformation in the group of RMSD values close to 0.6 is shown instead. All conformations form single hydrophobic cores as for the global minimum-energy conformations. It should be noted that the energies of the conformations in the figure are much lower than the lowest energies obtained by nPERMis [21] for all cases except S_{13}.

IV. SUMMARY AND CONCLUDING REMARKS

An off-lattice model protein consisting of two types of monomers, hydrophobic (A) and hydrophilic (B) residues, with Fibonacci sequences is studied by conformational space annealing. It is found that some of our results for the ground-state energies are lower than the best values reported in the literature for both two and three dimensions (S_{55} in two dimensions and all cases except S_{55} for model I and S_{34} and S_{55} for model II in three dimensions). Consequently, some of the present results provide updated candidates for the global minima. In two dimensions, the conformations of the ground states form several clusters of three to five hydrophobic residues, whereas single hydrophobic cores are observed in three dimensions. Based on this, it is concluded that the AB model with Fibinacci sequences in three dimensions mimics the real protein reasonably well, while in two dimensions the model becomes unrealistic to describe proteins.

The distributions of the RMSD values from the global minima are also analyzed as a function of energy. It is found that in three dimensions there exist at least two distinct subgroups of local minima, one with larger RMSD values and the other with smaller RMSD values. In three dimensions, the lowest-energy conformations are generally located far away from the majority of conformations. This situation makes it particularly more difficult to search for the lowest-energy conformations. The lowest-energy conformations form single hydrophobic cores as observed in real proteins.

ACKNOWLEDGMENTS

J.L. and S.-Y.K. were supported by the Basic Research Program of the Korea Science and Engineering Foundation under Grant No. R01-2003-000-11595-0. S.B.L. was supported by the Korean Council for University Education, support for 2004 Domestic Faculty Exchange Program.

[1] C. B. Anfinsen, Science **181**, 223 (1973).
[2] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Science **220**, 671 (1983).
[3] R. S. Judson, M. E. Colvin, J. C. Meza, A. Huffer, and D. Gutierrez, Int. J. Quantum Chem. **44**, 277 (1992).
[4] A. P. Lyubartsev, A. A. Martinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, J. Comput. Phys. **96**, 1776 (1992).
[5] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[6] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[7] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[8] U. H. E. Hansmann and Y. Okamoto, J. Comput. Chem. **14**, 1333 (1993); Physica A **212**, 415 (1994).
[9] N. Nakajima, H. Nakamura, and A. Kidera, J. Chem. Phys. **101**, 817 (1997).
[10] J. Lee, H. A. Scheraga, and S. Rackovsky, J. Comput. Phys. **18**, 1222 (1997).
[11] J. Lee and H. A. Scheraga, Int. J. Quantum Chem. **75**, 255 (1999).
[12] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989); D. Shortle, H. S. Chan, and K. A. Dill, Protein Sci. **1**, 201 (1992).
[13] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993).
[14] F. H. Stillinger and T. Head-Gordon, Phys. Rev. E **52**, 2872 (1995).
[15] T. Head-Gordon and F. H. Stillinger, Phys. Rev. E **48**, 1502 (1993).
[16] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, *Numerical Recipes* (Cambridge University Press, New York, 1986).
[17] A. Irbäck and F. Potthast, J. Chem. Phys. **103**, 10298 (1995).
[18] A. Irbäck, C. Peterson, and F. Potthast, Phys. Rev. E **55**, 860 (1997).
[19] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius, J. Chem. Phys. **107**, 273 (1997).
[20] A. R. Khokhlov and P. G. Khalatur, Phys. Rev. Lett. **82**, 3456 (1999).
[21] H.-P. Hsu, V. Mehra, and P. Grassberger, Phys. Rev. E **68**, 037703 (2003).
[22] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003); Phys. Rev. E **68**, 021113 (2003).
[23] F. Liang, J. Chem. Phys. **120**, 6756 (2004).
[24] M. Bachmann, H. Arkin, and W. Janke, Phys. Rev. E **71**, 031906 (2005).
[25] J. Lee, A. Liwo, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **96**, 2025 (1999).
[26] J. Lee, I.-H. Lee, and J. Lee, Phys. Rev. Lett. **91**, 080201 (2003).
[27] S.-Y. Kim, S. J. Lee, and J. Lee, J. Chem. Phys. **119**, 10274 (2003).
[28] J. Lee, S.-Y. Kim, K. Joo, I. Kim, and J. Lee, Proteins **56**, 704 (2004).
[29] S.-Y. Kim, S. J. Lee, and J. Lee, J. Korean Phys. Soc. **44**, 589 (2004).
[30] P. von Laarhoven and E. H. L. Arts, *Simulated Annealing: Theory and Applications* (Kluwer, Dordrecht, 1992).
[31] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machines Learning* (Addison-Wesley, Reading, MA, 1989).
[32] Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).