In Silico Chaperonin-Like Cycle Helps Folding of Proteins for Structure Prediction

Tadaomi Furuta,*†¶ Yoshimi Fujitsuka,* George Chikenji,*‡ and Shoji Takada*§¶

*Department of Chemistry, Faculty of Science, and Graduate School of Science and Technology, Kobe University, Nada, Kobe, Japan; †Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo, Tokyo, Japan; ‡Department of Computational Science and Engineering, Graduate School of Engineering, Nagoya University, Chigusa, Nagoya, Japan; §Department of Biophysics, Graduate School of Science, Kyoto University, Sakyo, Kyoto, Japan; and ¶Core Research for Evolutionary Science and Technology, Japan Science and Technology Agency, Kawaguchi, Japan

ABSTRACT Currently, one of the most serious problems in protein-folding simulations for de novo structure prediction is conformational sampling of medium-to-large proteins. In vivo, folding of these proteins is mediated by molecular chaperones. Inspired by the functions of chaperonins, we designed a simple chaperonin-like simulation protocol within the framework of the standard fragment assembly method: in our protocol, the strength of the hydrophobic interaction is periodically modulated to help the protein escape from misfolded structures. We tested this protocol for 38 proteins and found that, using a certain defined criterion of success, our method could successfully predict the native structures of 14 targets, whereas only those of 10 targets were successfully predicted using the standard protocol. In particular, for non-α-helical proteins, our method yielded significantly better predictions than the standard approach. This chaperonin-inspired protocol that enhanced de novo structure prediction using folding simulations may, in turn, provide new insights into the working principles underlying the chaperonin system.

INTRODUCTION

It has long been established that in protein-folding studies, many small proteins can spontaneously fold to their native structures, while most medium-to-large proteins cannot fold spontaneously at protein concentrations comparable to those in cellular environments. Indeed, it is well known that in vivo folding requires assistance from many types of molecular chaperones. The most widely studied molecular chaperone that assists the folding of substrate proteins is the bacterial chaperonin GroEL/ES system (1,2).

The GroEL/ES chaperonin complex is cylindrical, and its substrate protein is threaded into the GroEL/ES cage and then released from it (1,3–7). The substrate protein can attain folding to the native structure through this ATP-hydrolysis-driven cycle. Although the detailed mechanism of how GroEL/ES assists substrate folding still remains to be clarified, there are three possible scenarios, which are not mutually exclusive. First, as a result of the caging of some fraction of denatured substrate proteins in the cell, the aggregation of the proteins is suppressed, which is called the Anfinsen cage effect (8–10). Second, after a denatured substrate protein is threaded into the cage, GroEL/ES mechanically unfolds misfolded substrate proteins and then provides an environment for refolding, which increases the probability of successful folding (11–13). Third, a substrate protein placed in a relatively small confined space is more stable than that in a bulk environment and the confinement accelerates the protein folding, and thus the confinement itself is actively involved in substrate folding (14–18).

Here, we attempted to ask if we could utilize any of these mechanisms of chaperonin-assisted folding for de novo protein structure prediction via quasifolding simulations. The current status of de novo structure prediction may be best indicated by the results of the world-wide blind tests on predicting protein structures, i.e., the critical assessment of techniques for protein structure predictions (CASPs). In recent CASPs (19,21–23) (note that some information about CASPs is available at the web site: http://predictioncenter.gc.ucdavis.edu/), the most successful de novo prediction method in practice was the fragment assembly (FA) method developed by Baker and many others (24–29), in which the target tertiary structure is constructed by assembling short-length structures (fragments) taken from a structure database. Using the FA method with the coarse-grained energy function SimFold, which we developed in-house, we participated in CASPs, i.e., as Rokko (a human prediction group) and as Rokky (a server prediction group). In CASP6, both of these groups made top-level predictions in the new fold category, according to the assessment by B. K. Lee of the National Institutes of Health (19). Although the accuracy of prediction strongly depends on the particular protein, the folding of small proteins can often be predicted with reasonable accuracy and reliability, but in most cases, that of medium-to-large proteins cannot. Thus, the most serious bottleneck to accurate structure prediction for medium-to-large proteins is probably conformational sampling. Since this is the same step for which real proteins in vivo require the assistance of molecular chaperones, it may be possible to find better methods of sampling based on insights from the mechanisms of molecular chaperones.

Submitted July 3, 2007, and accepted for publication November 16, 2007.

Address reprint requests to Shoji Takada, Tel.: 81-75-753-4220; E-mail: takada@biophys.kyoto-u.ac.jp.

Editor: Ron Elber.

0006-3495/08/04/2558/08 $2.00

doi: 10.1529/biophysj.107.115261

Among the three scenarios described above, conformational sampling is related to the second and the third, and in this study, we tried to utilize the second mechanism to attain better conformational sampling in silico.

The mechanism by which GroEL/ES mechanically unfolds and refolds substrate proteins can be summarized as follows (11–13): A misfolded substrate protein that has some hydrophobic surfaces is first bound to the hydrophobic region at the rim of the GroEL cylinder. After the binding of an ATP and the capping of GroEL by GroES, the GroEL cylinder extends axially, which induces the bound substrate to mechanically unfold. Once the substrate has unfolded from a misfolded state, it has a chance to refold to the native state. GroEL releases the substrate to the outside of the cylinder at a certain rate. This mechanical unfolding-refolding is repeated until the substrate reaches the native state. Here, the role of the chaperonin is to provide the substrate multiple chances to refold from extended structures.

Inspired by this scheme, we designed a multiple-unfolding protocol for substrate proteins using FA simulations with SimFold (30). Somewhat similar ideas can be found in the literature (31–33). We then tested our protocol for 38 small-to-medium proteins and compared its performance with that of the standard protocol. We found that the chaperonin-inspired protocol yielded better sampling results than the standard protocol, i.e., with a certain defined criterion (that is, at least one of five predicted structures was within 6.5 Å of the native structure), the chaperonin-based simulation successfully predicted the native structures of 14 out of 38 proteins studied, whereas the standard protocol only predicted the structures of 10 proteins successfully.

MATERIALS AND METHODS

SimFold is an empirical energy function of proteins with a coarse-grained representation of the chain. The backbone of a peptide in SimFold is explicitly represented, while each side chain is simplified as a sphere located at the center of mass of the side-chain atoms. The solvent molecules are not explicitly included, but the effect of the solvent is taken into account through the effective energy term of the protein interactions. Many parameters in the energy function terms are statistically optimized based on the native-structure information of a training set of protein structures. Our optimized total energy potential (V_{tot}) has the form (34,30)

$$V_{\text{tot}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{torsion}} + V_{\text{vdW}} + V_{\text{HP}} + V_{\text{HB}} + V_{\text{pairwise}} + V_{\text{Rot}} + V_{\text{Rg}}, \quad (1)$$

where V_{bond} stands for the main-chain 1-2 bond (r) potential, V_{angle} for the main-chain 1-2–3 angle (θ) potential, and V_{torsion} for the main-chain 1-2-3-4 torsional (ϕ,ψ,ω) potential. Here, V_{vdW} stands for the van der Waals potential, V_{HP} for the hydrophobic interaction potential, and V_{HB} for the hydrogen-bond potential. Furthermore, V_{pairwise} stands for the pairwise potential, V_{Rot} for the rotamer potential, and V_{Rg} for the radius of gyration potential (set to zero in the default parameter set).

Using this default SimFold force field, we carried out de novo structure prediction by fragment assembly simulated annealing (FASA) in a three-step procedure: Step 1 involved fragment preparation, Step 2 involved FASA using the energy function Simfold, and Step 3 involved a model selection via a clustering analysis. These steps were executed as follows:

Step 1. We first prepared 300 fragments each for every possible nine-residue segment that could be aligned along the amino-acid chain for each target by using the publicly available version of Rosetta software (Rosetta fragment selection Ver. 1.1 (35)), which we called the fragment library.

Step 2. We then conducted FASA using this fragment library, performing Metropolis Monte Carlo (MC) judgment at each step with the SimFold force field at the annealing temperature for that step, which was linearly decreased from 1000 K to 240 K over a period of 500,000 steps. We generated 800 sampling structures for each of the targets and extracted 400 lower-energy sampling structures that were filtered based on the SimFold energy.

Step 3. Finally, we applied hierarchical clustering analysis based on pairwise root mean-square deviations (RMSDs) for these 400 structures by searching for a cutoff length of RMSD such that the size of the best cluster was ~10% of the number of all structures analyzed, i.e., 400.

We then selected the clustering centers of the top five clusters as the prediction models. This was our basic procedure for predicting structures. After the prediction, we calculated the RMSDs of the five model structures compared to the native structures for all targets.

A chaperonin-like cycle was implemented by using the following oscillatory hydrophobic potential term in SimFold, which was utilized in the Metropolis MC judgment in Step 2 above,

$$V_{\text{HP}} = \begin{cases} cV_{\text{HP}} & (t \mod(p) < \tau) \\ V_{\text{HP}} & (t \mod(p) \geq \tau) \end{cases}, \quad (2)$$

where c stands for the coefficient multiplying the default hydrophobic potential, t stands for the time, p stands for the period, and τ stands for the interval that the chaperonin-like-cycle effect continues in the annealing steps. This chaperonin-like cycle was continued until the 350,000th step and the nonreduced hydrophobic interaction was used for the period of the final 150,000 steps so that a correct structure was not broken during this low-temperature period. We termed this chaperonin-like-cycle sampling ''SimFold-CC'' and termed default sampling without this cycle ''SimFold.''

RESULTS AND DISCUSSION

Chaperonin-based folding simulation

We designed protocols that imitated chaperonin functions in quasifolding simulations using fragment assembly simulated annealing (FASA; see Materials and Methods). It is known that, in standard FASA simulations, once a protein finds a compact structure, it rarely undergoes global conformational changes thereafter, because such conformational changes would cause steric collisions. Thus, when the first such compact structure attained is far from the native structure, it is very difficult for the sampling structure to reach the native structure at the end of the simulation. During the simulated annealing runs with the SimFold energy function, we periodically applied additional energies which forced the protein structure to extend.

First, we tried a harmonic potential in the radius of gyration $V_{\text{Rg}} = c(R_{\text{g}} - R_{\text{g}}^0)^2$, which was applied periodically during a simulation, where R_{g} is the radius of gyration of the simulated protein and R_{g}^0 is the estimated radius of gyration of the native structure, which is expressed as $R_{\text{g}}^0 = 2.96N^{1/3} - 0.84$, where N is the chain length (30). Since we found that this

method had severe problems after tuning the parameters and testing the performance, we did not choose it. The first cause of its poor performance was that we did not know the radius of gyration of the native structure before prediction, and thus R_g^0 could be a poor estimate for some proteins. In such cases, this additional potential severely disrupts conformational sampling. The second cause was that the potential V_{Rg} forces the protein to extend independent of the goodness of the structure. Therefore, we often observed that the applied potential destroyed not only the misfolded structures but also the correctly folded ones. Thus, the final structures were often very poor.

We then noticed that the misfolded conformations or misfolded parts of the chain may be less stable than correctly folded conformations or correctly folded parts. Therefore, we reasoned that by reducing the interaction energy, we might be able to preferentially destroy the misfolded parts while preserving the correctly folded parts. To this end, we decided to periodically modulate the strength of the hydrophobic interaction, which is the major energy term that makes the chain compact in real proteins and in SimFold energy. Here, we should note that the mechanical extension of substrates by a chaperonin and by modulation of the hydrophobicity are not identical, but as can be seen from Fig. 1 *b*, Fig. 2 *c*, Fig. 3 *b*, and Fig. 4 *c* (explained below), reduction of the hydrophobicity tends to release incorrectly formed contacts and extend the structure, with an effect on folding similar to mechanical unfolding by a chaperonin.

Using a set of six mainly β-rich test proteins (PDB: 2gb1, 1csp, 1csk, 1ten, 1tza, and 1ey0), we tuned parameters in the chaperonin-based simulation protocol (see Materials and Methods). Based on the results of test simulations, we chose a period (p) of 10,000 Monte Carlo (MC) steps in which the first interval (τ) of 10 MC steps has reduced hydrophobic energy. The reduction coefficient (c) in the hydrophobic term was briefly optimized to 0.25 (*Param*2 in Table 1). With this parameter set, we could successfully predict the native structure of three small proteins under the criteria that at least one of the five cluster centers had a root mean-square deviation (RMSD) from the native structure of <6.5 Å. We adopted this parameter set as the best chaperonin-like cycle method (we called it ''SimFold-CC,'' since it uses SimFold force field with a chaperonin-like cycle, and we called the default SimFold simply ''SimFold''). Even though this parameter set was the best of those we tested, it failed to predict three out of six targets; two of the unsuccessful targets (1tza and 1ey0) had chain length longer than 100, suggesting that such long structures were still difficult to predict.

A typical MC trajectory with the chaperonin-like cycle protocol is illustrated in Figs. 1 and 2 for the conformational sampling of protein G (PDB: 2gb1). Protein G has the fol-

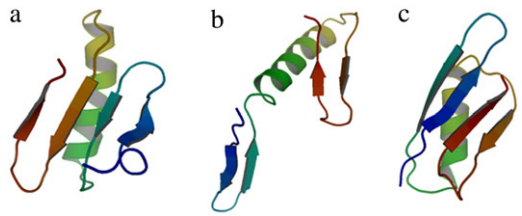

FIGURE 1 (*a*) Snapshot at 214,000th step: RMSD is 7.41 Å, R_g is 11.14 (local minimum), and β2 and β3 create an incorrect parallel β-sheet. (*b*) Snapshot at 222,000th step: RMSD is 12.48 Å, R_g is 14.83 (local maximum), and there is no distant β-pair. (*c*) Snapshot at 243,500th step: RMSD is 1.81 Å, R_g is 10.75, and β1 and β4 create correct parallel β-sheet. Figures were prepared with PyMol (43).

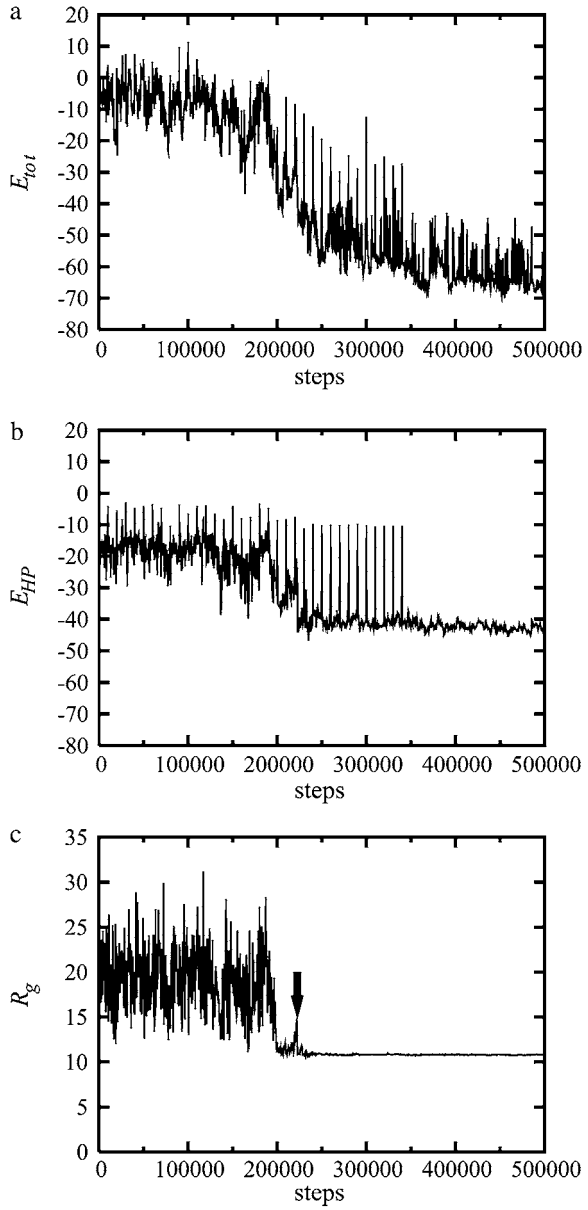

FIGURE 2 Time series of (*a*) the total energy, (*b*) the hydrophobic energy, and (*c*) the radius of gyration for FASA simulation trajectory with chaperonin-based protocol for protein G. The arrow in Fig. 2 *c* indicates a small peak at the 222,000th step, which corresponds to the mechanically unfolded structure (shown in Fig. 1 *b*).

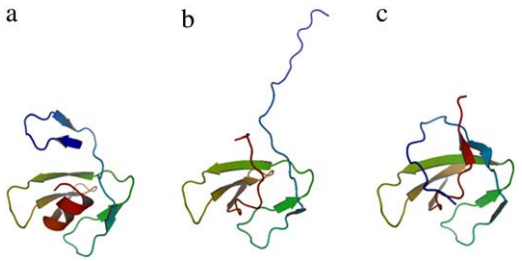

FIGURE 3 (*a*) Snapshot at 307,500th step: RMSD is 4.48 Å, R_g is 11.30 (local minimum), and there is a nonnative β-hairpin at the N-terminal and an α-helix at the C-terminal. (*b*) Snapshot at 345,500th step: RMSD is 3.15 Å, R_g is 15.65 (local maximum), the β-hairpin is broken, and the α-helix is melted. (*c*) Snapshot at 450,000th step: RMSD is 2.88 Å, R_g is 11.00, and the folding is correct. Figures were prepared with PyMol (43).

lowing secondary structure elements from the N- to the C-terminal direction: β1-β2-α1-β3-β4, where β is a β-strand and α is an α-helix. Among them, the β1-β2 and β3-β4 pairs create β hairpins, and the β1-β4 pair creates an anti-parallel β sheet in the native structure. Fig. 1 *a* shows a snapshot of a misfolded structure at the 214,000th step, which created an incorrect parallel β-sheet between the β2- and β3-strands. This misfolded structure was then broken at the 222,000th step (shown in Fig. 1 *b*). Finally, a correct parallel β-sheet between the β1- and β4-strands was constructed at the 243,500th step (Fig. 1 *c*). The time series of the total energy (E_{tot}), the hydrophobic energy (E_{HP}), and the radius of gyration (R_g) for this trajectory are shown in Fig. 2. The time series of the total energy for FASA with SimFold-CC is depicted in Fig. 2 *a*. Chaperonin-like-cycle effects that continued until the 350,000th step appear as spikes in the time series of the hydrophobic energy (shown in Fig. 2 *b*); these periodic spikes represent reductions in the hydrophobic interaction. These two figures represent features of the FASA with SimFold-CC. The time series of the radius of gyration R_g showed a small peak at the 222,000th step, at which R_g reached 15 Å (indicated by the *arrow*). This peak corresponds to the structure in Fig. 1 *b*, which was a mechanically unfolded structure derived from the previous more compact one in Fig. 1 *a*. Then, R_g decreased again until the protein reached the correct folding (shown in Fig. 1 *c*). After escaping from the misfolded state, the unscaled hydrophobic energy increased from a local minimum value of -38.3 kcal/mol at the 204,000th step to a value of ~ -20 kcal/mol at the 221,500th step. The temperature was $\sim$700 K around these steps, and thus the Boltzmann factor for this excitation was roughly $\exp(-14) < 10^{-6}$, which was too small for escape to occur. With the reduction coefficient (*c*) of 0.25, the Boltzmann factor became $\sim\exp(-3.5) = 0.03$, enabling escape to occur. These kinds of escape from misfolded structures occurred in the conformational sampling of FASA with SimFold-CC.

Another interesting case of escaping from a misfold trap was observed in the conformational sampling for the Src SH3 domain (PDB: 1csk) (shown in Fig. 3). The Src SH3 domain

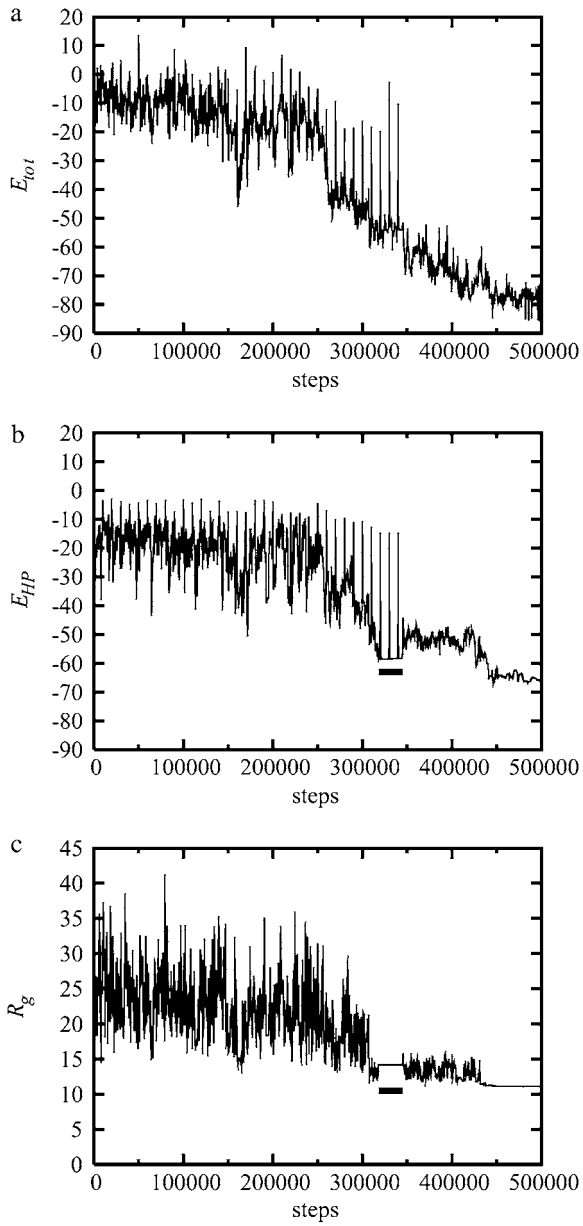

FIGURE 4 Time series of (*a*) the total energy, (*b*) the hydrophobic energy, and (*c*) the radius of gyration for FASA simulation trajectory with chaperonin-based protocol for Src SH3 domain. Bars in Fig. 4, *b* and *c*, indicate a flat basin, which corresponds to a stable misfolded structure (shown in Fig. 3 *a*).

is classified into the all-β protein class, and has an SH3-like barrel fold. Fig. 3 *a* shows a snapshot of a misfolded structure containing a clear stable α-helix at the C-terminal and a weak β-hairpin at the N-terminal that does not exist in the native structure. This weak β-hairpin was easily broken within a few steps, but the stable α-helix structure lasted until the 340,000th step. These features appear in the time series of the hydrophobic energy and the radius of gyration, where a flat basin clearly exists between the 300,000th and 350,000th steps (shown in Fig. 4, *b* and *c*, respectively). After the final

TABLE 1 Optimization of chaperonin-like cycle parameters

Serial	PDB	SCOP	ACO	Naa	Default	Param1	Param2	Param3	Param4
1	2gb1	$\alpha+\beta$	9.6	56	**2.2**	**3.3**	**2.2**	**1.9**	**3.4**
2	1csp	All β	11.1	67	6.7	7.1	**5.3**	7.2	6.9
3	1csk	All β	11.1	71	7.1	6.8	**4.6**	**5.8**	**5.8**
4	1ten	All β	15.35	90	8.3	**6.4**	8.2	7.1	9.5
5	1tza	All β	17.58	134	15.4	15.7	15.3	18.7	14.2
6	1ey0	All β	13.71	149	13.4	15.3	14.7	14.4	13.4

SCOP stands for SCOP classification (39) (SCOP web site: http://scop.mrc-lmb.cam.ac.uk/scop/); *ACO* stands for the absolute contact order, which is a measure of the topological complexity of the native state and is related to the folding rate (41) (PERL script for calculating the contact order is available at the web site: http://depts.washington.edu/bakerpg/contact_order/contactOrder.pl.). *Naa* stands for the number of amino acids in each target. *Default* stands for the RMSD (Å) of the best cluster center of the top five clusters from all native structures using the default SimFold, i.e., $c = 1$. *Param* (1–4) stands for the RMSD (Å) of the best cluster center of the top five clusters from all native structures using the SimFold-CC (chaperonin-like cycle) with the coefficients of the hydrophobic interaction. Param1, $c = 0.0$; Param2, $c = 0.25$; Param3, $c = 0.5$; and Param4, $c = 0.75$. For all the simulations in this table, we used the parameters $p = 10{,}000$ and $\tau = 10$. The bold values stand for successful prediction based on the criterion that RMSD was <6.5 Å. The proteins in this table were sorted based on the N_{aa}.

effect of the chaperonin-like cycle at the 340,000th step, this α-helix structure was gradually broken (shown in Fig. 3 *b*). The global structure then became correctly folded at the final stage (shown in Fig. 3 *c*). Along the trajectory, after escaping from the misfolded state, the unscaled hydrophobic energy increased from -58 kcal/mol to -44 kcal/mol at a temperature of 500 K, which gave a Boltzmann factor of $\sim\exp(-14) < 10^{-6}$, which was identical to that in the case of protein G. It seems that a reduction coefficient (c) of 0.25 effectively enabled escape from the misfolded state when the unscaled Boltzmann factor was too small, i.e., $\sim 10^{-6}$. As we saw in the above two cases, a reduction coefficient of 0.25 worked effectively in this framework, so we adopted this parameter set. However, in general, the energy barrier for escaping from a misfolded state would depend on the misfolded structure and also on the protein studied. Thus, in some cases, a protein-dependent optimal reduction coefficient might be needed. How strongly the barrier for escaping from the misfolded state depends on the protein would be more difficult to assess, and this remains quite an interesting issue for future studies.

TABLE 2 Predictions for 38 proteins

					SimFold		SimFold-CC	
Serial	PDB	SCOP	ACO	Naa	Clus	Best	Clus	Best
1	1utg	All α	5.1	70	7.4	4.2	8.5	4.9
2	1rpo	All α	5.8	61	**2.4**	1.8	**2.8**	1.9
3	2reb	$\alpha+\beta$	6.3	60	7.9	5.4	**6.4**	4.7
4	1orc	All α	6.5	64	**3.3**	2.7	**3.1**	2.5
5	1a8o	All α	6.6	70	**4.4**	3.2	**3.8**	3.2
6	1ail	All α	6.8	70	8.2	4.5	7.1	4.6
7	1jhf	All α	6.8	71	**5.0**	3.5	**6.2**	3.3
8	1r69	All α	6.9	63	**2.1**	1.7	**1.8**	1.6
9	1ig5	All α	7.5	75	**3.2**	2.6	**2.4**	2.4
10	1l0h	All α	7.6	76	7.7	4.9	8.9	7.2
11	1dol	$\alpha+\beta$	7.8	71	11.2	6.4	11.0	7.8
12	1vcc	All β	8.2	77	8.5	6.0	**5.5**	5.1
13	1lfb	All α	8.3	77	**5.1**	4.0	**5.8**	3.9
14	1tif	$\alpha+\beta$	8.4	76	7.9	5.0	7.1	5.0
15	1cei	All α	9	85	9.5	6.9	11.4	6.7
16	1pgx	$\alpha+\beta$	9.1	70	8.7	5.8	8.6	5.9
17	1a68	$\alpha+\beta$	9.3	87	7.9	5.8	**6.4**	6.4
18	1beo	All α	9.4	98	10.9	8.0	11.8	8.5
19	4pti	Small	10	58	8.3	4.8	8.4	5.9
20	1ay7	α/β	10.1	89	7.7	6.7	7.1	6.6
21	1vqh	All β	10.5	86	11.4	8.5	11.3	9.0
22	1hb6	All α	10.6	86	8.2	5.0	**3.8**	3.8
23	1cun	All α	10.6	109	**6.1**	2.4	9.2	3.8
24	1vif	All β	10.7	60	9.9	6.3	11.0	6.8
25	1aba	α/β	10.9	87	10.1	6.1	8.6	5.5
26	1csp	All β	11.1	67	**5.3**	4.3	**5.7**	3.2
27	1nb5	$\alpha+\beta$	11.2	98	8.8	8.0	9.1	8.8
28	1bov	All β	11.3	69	10.0	5.2	9.9	5.0
29	1nxb	Small	11.3	62	8.8	4.9	**6.0**	6.0
30	1msi	All β	11.5	66	11.0	7.0	10.1	8.0
31	1ffg	α/β	11.6	68	8.7	7.2	8.8	7.1
32	1tuc	All β	11.7	61	8.4	6.3	8.1	5.9
33	1ctf	$\alpha+\beta$	12.2	68	**6.0**	4.1	**6.0**	4.8
34	1hoe	All β	14.9	74	11.5	7.9	12.1	8.9
35	1who	All β	15.1	97	13.4	8.0	11.4	9.0
36	1bdo	All β	16.3	80	13.4	9.6	11.3	9.5
37	1ris	$\alpha+\beta$	17.9	97	10.6	6.6	6.5	6.4
38	2acy	$\alpha+\beta$	19.3	98	9.0	8.2	9.8	7.6

SCOP, *ACO*, and *Naa* are the same as in Table 1. *SimFold Clus* stands for the RMSD (Å) of the best cluster center of the top five clusters from all native structures using the default SimFold, i.e., $c = 1$. *SimFold-CC Clus* stands for the RMSD (Å) of the best cluster center of the top five clusters from all native structures using SimFold with the chaperonin-like cycle, i.e., with parameter set $c = 0.25$, $p = 10{,}000$, and $\tau = 10$. *SimFold Best* and *SimFold-CC Best* stand for the best RMSD in the conformational sampling for all targets. The bold values stand for successful prediction based on the criterion that RMSD was <6.5 Å, and the underlined values indicate that at least one structure with RMSD of <6.5 Å was sampled in the conformational sampling. Proteins in this table were sorted based on the ACO.

Benchmark test

Next, using the tuned protocol of SimFold-CC, we carried out a benchmark test on the method and compared its performance with that of a standard protocol for FASA (Table 2). For this purpose, we used a set of 38 small-to-medium protein domains that covered various topologies, i.e., all α, all β, α/β, and $\alpha+\beta$ domains. This set was the one we had previously used (36) and was a subset of the benchmark test used by Simons et al. (27). We created five models for prediction, as described in Materials and Methods, and if at least one of the five had an RMSD from the native structure of <6.5Å, the prediction was considered successful. The cutoff of 6.5 Å is more or less a standard value for de novo structure prediction of proteins that have multiple difficulties (37), and one out of five chosen models is what CASP has been using for prediction. Note that some structures with an RMSD of <6.5 Å have a different topology from that of the native

structure, especially for β-proteins, as was demonstrated by Fujitsuka et al. (36). As listed in Table 2, SimFold-CC could successfully predict the structures of 14 out of 38 proteins using this criterion, which is more than the 10 proteins whose structures were successfully predicted using the standard SimFold method. This comparison is visually depicted in Fig. 5, where, for each protein, we have plotted the best RMSDs obtained by SimFold-CC on the x axis and that obtained by the standard method on the y axis. Both methods could successfully predict 9 out of 38 targets (1rpo, 1orc, 1a8o, 1jhf, 1r69, 1ig5, 1lfb, 1csp, and 1ctf), which are indicated by the solid circles in the figure. For the five targets depicted by the solid triangles in Fig. 5 (2reb, 1vcc, 1a68, 1hb6, and 1nxb), only SimFold-CC was successful in predicting the structure. For the one target (1cun) depicted by the solid square in Fig. 5, only SimFold was successful in predicting the structure. Neither method was successful for the other 23 targets (*asterisks* in the figure), which were, on average, larger proteins and thus were inherently difficult to predict. We found that the chaperonin-based method performed better especially for proteins with β-sheets, for which conformational sampling was more difficult on average. We conducted statistical tests on a head-to-head comparison by using RMSDs, and found that SimFold-CC was significantly better than the standard protocol ($p = 0.0045 < 0.05$) for a set of 24 non-α-helical proteins. The same test using the GDT_TS scores yielded $p = 0.043 < 0.05$ (38). However, the difference between the two methods for α-helical proteins was not statistically significant. Perhaps conformational sampling is not a bottleneck for folding of α-helical proteins of these sizes.

Some successful prediction models of different fold classes superimposed on their native structures are shown in Fig. 6. These successful predictions for proteins in different fold classes suggest that the conformational sampling capability of the SimFold-CC method is robust. We previously carried out a benchmark test using the publicly available version of Rosetta (ab initio Ver. 1.2) (35) and an identical setup, and found that the structures of 12 out of 38 proteins could successfully be predicted using the same criterion (36).

CONCLUSIONS

We developed a simple and effective sampling method for prediction of the structures of proteins via mimicking the physical nature of the chaperonin cycle. Our novel method outperformed a standard method with the same energy function in a benchmark test, i.e., it successfully predicted the structures of 14 out of 38 target proteins using a certain criterion, whereas only the structures of 10 proteins were correctly predicted using the standard method. A statistical test suggested that this novel approach made significantly better

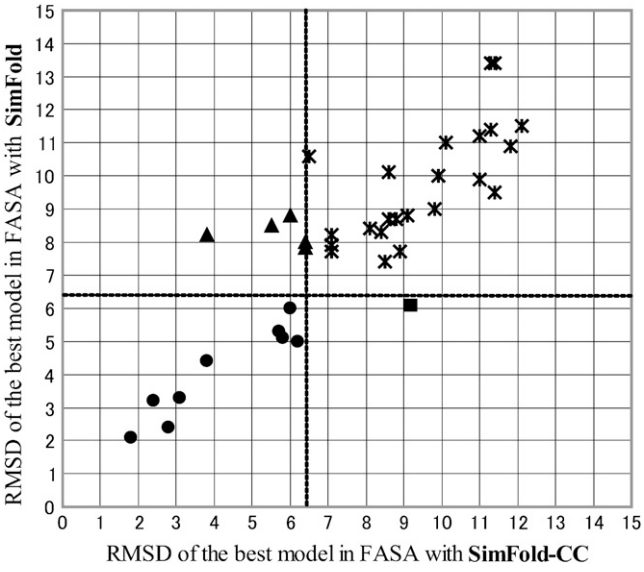

FIGURE 5 Comparison between RMSDs of the best models in SimFold-CC versus those in SimFold; each point corresponds to a protein in the test set. The horizontal axis shows the RMSD (Å) of the best model out of the five prediction models with SimFold-CC and the vertical axis shows that obtained using the standard SimFold method. Dashed lines indicate RMSD equal to 6.5 Å, i.e., the criterion of success used here. Solid circles correspond to proteins for which both methods succeeded. Only SimFold-CC succeeded for proteins indicated by solid triangles, while standard SimFold, but not SimFold-CC, succeeded for the protein indicated by a solid square. Asterisks correspond to proteins for which neither method succeeded.

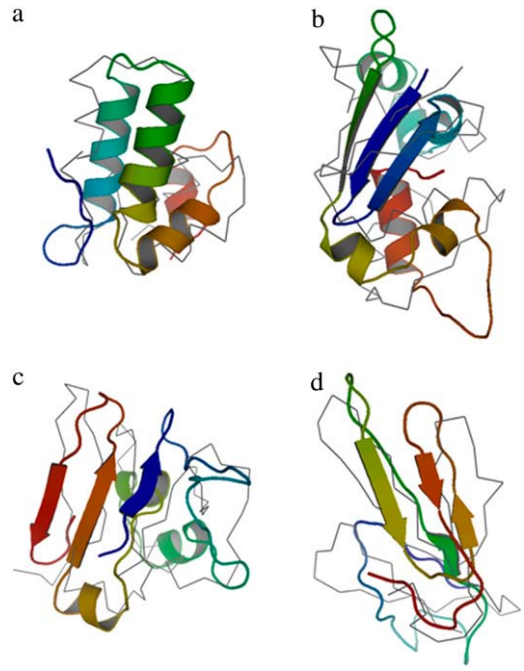

FIGURE 6 Some successful prediction models obtained using SimFold-CC superimposed on native structures (*a*) PDB# 1a8o; RMSD = 3.8 Å; SCOP: All α. (*b*) PDB# 1a68; RMSD = 6.4 Å; SCOP: $\alpha+\beta$. (*c*) PDB# 1vcc; RMSD = 5.5 Å; SCOP: All β. (*d*) PDB# 1nxb; RMSD = 6.0 Å; SCOP: small. Thick colored lines stand for backbones of native structures. Figures were prepared with PyMol (43).

predictions than the standard one for non-α-helical targets. The simple idea of a chaperonin-like cycle is quite a general and promising concept for enhancing sampling, which is not limited to fragment assembly simulations (31–33). Moreover, time-dependent modulations of interactions other than changes of the hydrophobicity are also useful (32). Theoretically, this approach can easily be extended to the multicanonical-ensemble (29) or the replica exchange (31) method, for which even better performance is expected. Findings about the efficacy of the method, in turn, may provide new insights into the power of mechanical unfolding as the working principle underlying the chaperonin system.

This work was partly supported by a Grant-in-Aid for Scientific Research in the Priority Area of ''Chemistry of Biological Processes Created by Water and Biomolecules'' from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and by the Global Center of Excellence Program ''Formation of a Strategic Base for Biodiversity and Evolutionary Research: from Genome to Ecosystem'' from Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

1. Sigler, P. B., Z. Xu, H. S. Rye, S. G. Burston, W. A. Fenton, and A. L. Horwich. 1998. Structure and function in GroEL-mediated protein folding. *Annu. Rev. Biochem.* 67:581–608.
2. Hartl, F. U., and M. Hayer-Hartl. 2002. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*. 295:1852–1858.
3. Braig, K., Z. Otwinowski, R. Hegde, D. C. Boisvert, A. Joachimiak, A. L. Horwich, and P. B. Sigler. 1994. The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. *Nature*. 371:578–586.
4. Xu, Z., A. L. Horwich, and P. B. Sigler. 1997. The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. *Nature*. 388:741–750.
5. Weissman, J. S., Y. Kashi, W. A. Fenton, and A. L. Horwich. 1994. GroEL-mediated protein folding proceeds by multiple rounds of binding and release of nonnative forms. *Cell*. 78:693–702.
6. Weissman, J. S., C. M. Hohl, O. Kovalenko, Y. Kashi, S. Chen, K. Braig, H. R. Saibil, W. A. Fenton, and A. L. Horwich. 1995. Mechanism of GroEL action: productive release of polypeptide from a sequestered position under GroES. *Cell*. 83:577–587.
7. Rye, H. S., S. G. Burston, W. A. Fenton, J. M. Beechem, Z. Xu, P. B. Sigler, and A. L. Horwich. 1997. Distinct actions of *cis* and *trans* ATP within the double ring of the chaperonin GroEL. *Nature*. 388:792–798.
8. Weber, F., F. Keppel, C. Georgopoulos, M. K. Hayer-Hartl, and F. U. Hartl. 1998. The oligomeric structure of GroEL/GroES is required for biologically significant chaperonin function in protein folding. *Nat. Struct. Biol.* 5:977–985.
9. Eliss, R. J. 1994. Molecular chaperones. Opening and closing the Anfinsen cage. *Curr. Biol.* 4:633–635.
10. Saibil, H. R., D. Zheng, A. M. Roseman, A. S. Hunter, G. M. Watson, S. Chen, A. auf der Mauer, B. P. O'Hara, S. P. Wood, N. H. Mann, L. K. Barnett, and R. J. Ellis. 1993. ATP induces large quaternary rearrangements in a cage-like chaperonin structure. *Curr. Biol.* 3: 265–273.
11. Todd, M. J., G. H. Lorimer, and D. Thirumalai. 1996. Chaperonin-facilitated protein folding: optimization of rate and yield by an iterative annealing mechanism. *Proc. Natl. Acad. Sci. USA*. 93:4030–4035.
12. Betancourt, M. R., and D. Thirumalai. 1999. Exploring the kinetic requirements for enhancement of protein folding rates in the GroEL cavity. *J. Mol. Biol.* 287:627–644.
13. Shtilerman, M., G. H. Lorimer, and S. W. Englander. 1999. Chaperonin function: folding by forced unfolding. *Science*. 284:822–825.
14. Brinker, A., G. Pfeifer, M. J. Kerner, D. J. Naylor, F. U. Hartl, and M. Hayer-Hartl. 2001. Dual function of protein confinement in chaperonin-assisted protein folding. *Cell*. 107:223–233.
15. Zhou, H. X., and K. A. Dill. 2001. Stabilization of proteins in confined spaces. *Biochemistry*. 40:11289–11293.
16. Klimov, D. K., D. Newfield, and D. Thirumalai. 2002. Simulations of β-hairpin folding confined to spherical pores using distributed computing. *Proc. Natl. Acad. Sci. USA*. 99:8019–8024.
17. Takagi, F., N. Koga, and S. Takada. 2003. How protein thermodynamics and folding mechanisms are altered by the chaperonin cage: molecular simulations. *Proc. Natl. Acad. Sci. USA*. 100:11367–11372.
18. Tand, Y.-C., H.-C. Chang, A. Roeben, D. Wischnewski, N. Wischnewski, M. J. Kerner, F. U. Hartl, and M. Hayer-Hartl. 2006. Structural features of the GroEL-GroES nano-cage required for rapid folding of encapsulated protein. *Cell*. 125:903–914.
19. Vincent, J. J., C. H. Tai, B. K. Sathyanarayana, and B. Lee. 2005. Assessment of CASP6 predictions for new and nearly new fold targets. *Proteins*. 61:67–83.
20. Reference deleted in proof.
21. Aloy, P., A. Stark, C. Hadley, and R. B. Russell. 2003. Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins*. 53:436–456.
22. Moult, J., K. Fidelis, B. Rost, T. Hubbard, and A. Tramontano. 2005. Critical assessment of methods of protein structure prediction (CASP)–round 6. *Proteins*. 61:3–7.
23. Moult, J., K. Fidelis, A. Zemla, and T. Hubbard. 2003. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*. 53:334–339.
24. Simons, K. T., C. Kooperberg, E. Huang, and D. Baker. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–225.
25. Jones, D. T. 1997. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*. 29:185–191.
26. Bowie, J. U., and D. Eisenberg. 1994. An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. USA*. 91: 4436–4440.
27. Simons, K. T., C. Strauss, and D. Baker. 2001. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* 306:1191–1199.
28. Jones, D. T. 2001. Predicting novel protein folds by using FRAGFOLD. *Proteins*. 45:127–132.
29. Chikenji, G., Y. Fujitsuka, and S. Takada. 2003. A reversible fragment assembly method for de novo protein structure prediction. *J. Chem. Phys*. 119:6895–6903.
30. Fujitsuka, Y., S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes. 2004. Optimizing physical energy functions for protein folding. *Proteins*. 54:88–103.
31. Fukunishi, H., O. Watanabe, and S. Takada. 2002. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. *J. Chem. Phys*. 116: 9058–9067.
32. Fan, H., and A. E. Mark. 2004. Mimicking the action of folding chaperones in molecular dynamics simulations: application to the refinement of homology-based protein structures. *Protein Sci.* 13: 992–999.
33. Liu, P., X. Huang, R. Zhou, and B. J. Berne. 2006. Hydrophobic aided replica exchange: an efficient algorithm for protein folding in explicit solvent. *J. Phys. Chem. B*. 110:19018–19022.
34. Takada, S. 2001. Protein folding simulation with solvent-induced force field: folding pathway ensemble of three-helix-bundle proteins. *Proteins*. 42:85–98.
35. Bonneau, R., C. E. Strauss, C. A. Rohl, D. Chivian, P. Bradley, L. Malmström, T. Robertson, and D. Baker. 2002. De novo prediction of

three-dimensional structures for major protein families. *J. Mol. Biol.* 322:65–78.

36. Fujitsuka, Y., G. Chikenji, and S. Takada. 2006. SimFold energy function for de novo protein structure prediction: consensus with Rosetta. *Proteins*. 62:381–398.
37. Kihara, D., H. Lu, A. Kolinski, and J. Skolnick. 2001. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. USA*. 98: 10125–10130.
38. Zelma, A., C. Venclocas, J. Moult, and K. Fidelis. 2001. Processing and evaluation of predictions in CASP4. *Proteins*. 45:13–21.
39. Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540.
40. Reference deleted in proof.
41. Plaxco, K. W., K. T. Simons, and D. Baker. 1998. Contact order, transition state placement, and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–994.
42. Reference deleted in proof.
43. DeLano, W. L. 2002. The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA.