

CafeMol: A Coarse-Grained Biomolecular Simulator for Simulating Proteins at Work


Hiroo Kanzaki,[†] Nobuyasu Koga,[†] Naoto Hori,[†] Ryo Kanada,[†] Wenfei Li,^{†,‡} Kei-ichi Okazaki,[§] Xin-Qiu Yao,[†] and Shoji Takada^{*,†,||}

[†]Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan

[‡]Department of Physics, Nanjing University, Nanjing 210093, China

[§]Department of Physics, School of Advanced Science and Engineering, Waseda University, Tokyo 169-8555, Japan

^{||}Advanced Center for Computing and Communication, RIKEN, Wako Saitama 351-0198, Japan

 Supporting Information

ABSTRACT: For simulating proteins at work in millisecond time scale or longer, we develop a coarse-grained (CG) molecular dynamics (MD) method and software, CafeMol. At the resolution of one-particle-per-residue, CafeMol equips four structure-based protein models: (1) the off-lattice Go model, (2) the atomic interaction based CG model for native state and folding dynamics, (3) the multiple-basin model for conformational change dynamics, and (4) the elastic network model for quasiharmonic fluctuations around the native structure. Ligands can be treated either explicitly or implicitly. For mimicking functional motions of proteins driven by some external force, CafeMol has various and flexible means to “switch” the energy functions that induce active motions of the proteins. CafeMol can do parallel computation with modest sized PC clusters. We describe CafeMol methods and illustrate it with several examples, such as rotary motions of F₁-ATPase and drug exports from a transporter. The CafeMol source code is available at www.cafemol.org.

1. INTRODUCTION

Proteins work with their characteristic sequences, structures, and dynamics. For example, enzymatic activity relies on well-designed structural arrangement of several key residues at catalytic sites, and the enzymatic activity is often allosterically regulated by structural change dynamics upon binding to their regulatory molecules. Thus, interplay among sequences, structures, dynamics, and functions is the focus of our studies. Unfortunately, however, no single experiment is powerful to simultaneously address all these aspects. Structural biology methods, such as X-ray crystallography, are the most powerful for addressing sequence–structure relation in high spatial resolution, but their primacy is on static structure, and the dynamic information is limited. A broad range of biochemical experiments is the most useful for addressing overall functions and their relations to sequences. They, however, do not provide direct evidence in structure and dynamics. Single-molecule observations and laser chemistry experiments are the most direct to see protein motions and dynamics, but their spatial resolution is unavoidably low.

In this context, molecular dynamics (MD) simulations are candidates to fill the gap among these and other various experiments providing much of the time-dependent structural information. The conventional fully atomistic MD simulations^{1–4} have high spatial and temporal resolution, but currently they suffer from the time-range problem: Time scales reachable by these MD simulations are typically on the order of microseconds, which is orders of magnitude shorter than the typical time scales of most biological processes. Thus, as a complementary approach, coarse-grained (CG) molecular models have been

used for simulating much longer time scales of biomolecular systems. Naturally, CG MDs are popular in studying long time behaviors,^{5–10} such as folding of proteins,^{11–13} work of protein machines,^{10,14–18} and lipid membrane self-assembly and morphology change.^{19–22} We have recently been developing such CG models of proteins and applying them to various protein systems.^{13–15,18,23–25} In doing so, we developed a software, CafeMol. This paper presents the CafeMol methods and the software for simulating proteins at work with various CG models.

2. METHODS

2.1. CG Strategy. Coarse-graining is not a unique procedure at all, and each CG model is, to some extent, based on the developers' own perspectives. CafeMol is based on the energy landscape theory developed in protein folding study,^{26,27} although the range of its applicability is not limited to the folding. Proteins have evolved to have ability of folding to their own native structures. At the native structures of proteins, overall, most side-chains are extremely well-packed in their cores. Whenever one finds a charge in the core, it is either paired with its counter charge or it is functionally essential. Thus, except for functional reasons, the interactions at the native structures are highly consistent, as pointed out many years ago by Go.²⁸ Proteins gained through evolution, the foldability by minimizing the frustration at their native structures, which was termed

Received: February 13, 2011

Published: May 10, 2011

“principle of the minimum frustration” by Bryngelson and Wolynes.²⁶ The effective energy takes the minimal value at the native, and as the conformation deviates from the native, the effective energy, on average, increases, which leads to an overall funnel-like energy landscape, first coined by Onuchic et al.²⁹ Completeness of the side-chains packing at the native structure may resemble to the high-density packing in crystal. On the other hand, the denatured state is characterized with low-level side-chain packing and larger fluctuation, and thus it resembles fluid.

Coarse-graining is relatively easy for the fluid-like denatured state because we primarily need to approximate the statistical average over the ensemble. Conversely, the native state is more difficult to be described by CG models because of the high level of side-chain packing, which is a very specific and a non-self-averaging property. If the side-chain architecture is lost by coarse-graining, then very surely we lose those specific interactions, to some extent.³⁰ Thus purely physicochemical coarse-graining is not effective for approximating the native state. Instead, an evolutionary perspective in the minimal frustration principle can be used as a guiding principle for coarse-graining. Namely, we assume, as an extreme, that all the interactions found at the native structure are attractive. Simultaneously, we require that the protein can take nearly random coil at sufficiently high temperature. These two requirements led to the so-called Go model, first developed in the lattice representation of proteins.³¹ However, of course, protein dynamics near the native is not well approximated by the lattice representation but is well-approximated by a quasiharmonic potential, such as the elastic network model (ENM).^{32,33} The ENM model is good only near the native. A model that is similar to the ENM near the native structure and simultaneously that shares the concept of the lattice Go model was developed and called the off-lattice Go model or the perfect funnel model.¹² The off-lattice Go model represents quasiharmonic fluctuations near the native structure and simultaneously realizes the perfect funnel energy landscape in global conformational space. CafeMol employs the off-lattice Go model developed by Clementi, Nyemyer, and Onuchic^{12,13} and its derivative as a basic CG model of proteins.

For applying CG MD to protein functional dynamics, however, the standard Go model is not sufficient because we often need to simulate conformational change of proteins, which is beyond the range of standard Go model. To simulate conformational change, many extensions of the Go model to multiple-basin cases have been proposed.^{18,23,34,35} CafeMol equips Okazaki et al.’s version of the multiple-basin potential,²³ which is an important feature of CafeMol since the multiple-basin model simulation is not easily fit with standard MD codes.

We note that the Go model and its derivative, as well as ENM, explicitly depend on the native structure. Thus, these models are often called the structure-based model or the native-centric model. The “structure-based MD” is quite different in concept from the conventional fully atomistic MD because the latter uses a physicochemically derived force field and thus the Hamiltonian does not explicitly depend on the native structure. Structure-based model, by design, says that the native structure is the most stable state, and this design implicitly and indirectly takes into account chemical interactions in a very crude sense. In addition, CafeMol has a new model that combines the structure-based Go model and the fully atomistic force field by using a multiscale protocol, which we term the atomic

interaction based CG (AICG) model.³⁶ In the AICG model, we can take into account chemical feature of interactions observed at the native state, without significant increase in computational time.

By CG MD, we can easily simulate protein dynamics in time scales comparable to milliseconds or longer, but the long-time simulation alone is not sufficient to simulate “proteins at work”. In cells, many proteins work as “machines”. For the machines to work lively, some free energy source is necessary. Many protein machines use chemical energy as the free energy source, such as energy from ATP hydrolysis and that by ion passage through membrane. These chemical events cannot be well represented by CG models, and thus we need to mimic them in some ways. For the purpose, we proposed to “switch” the energy function at a certain time.¹⁴ By switching, we put some energy into the protein systems, and proteins start to “work” as machines. Switching MD has some similarity, in concept, to the Brownian ratchet model studied in modeling molecular motors.³⁷ One of the key advantages of CafeMol, in comparison with other MD packages, is to equip various means to conduct dynamic “switching” simulations, which try to mimic roles of the energy source given to the system and turn on the active motion of the machines. Simulations of molecular motors, such as F₁-ATPase,¹⁴ AAA+ motor,²⁴ kinesin (Kanada et al unpublished), and a multidrug transporter²⁵ are examples of these functions.

The CafeMol employs a simplified representation of proteins where one particle is assigned to each amino acid most often placed at C_α atoms (can be at C_β or at the geometric center of amino acids, though). Four models are included: (1) the off-lattice Go model,¹² (2) the AICG model,³⁶ which is a chemically tuned extension to the Go model in which residue–pairwise interactions are modeled by their atom-based interaction derived from an all-atom force field, (3) the multiple-basin model,²³ which is a minimal model that represents energy landscape with more than one basins, and (4) the ENM.³² The Go model and the AICG model can simulate global and local folding/unfolding of proteins as well as native-state fluctuations. The multiple-basin model can be used to simulate conformational change dynamics. As such, the model assumes knowledge of (at-least) two structures corresponding to end-points of the conformational change. The ENM is to represent quasiharmonic fluctuation around the native structure. In addition, one can optionally turn on ligand models, generic electrostatic interactions, and empirical forms of hydrophobic interactions. The CafeMol implements various simulation protocols: (1) The constant temperature MD by Newtonian dynamics with Berendsen thermostat³⁸ and by Langevin dynamics. Based on the constant temperature MD, many higher order protocols are available. (2) One can automatically estimate the folding transition temperature by iterative folding/unfolding simulations in different temperatures. (3) Simulated annealing simulations can be done. (4) The one- and two-dimensional replica exchange MD can be run. (5) During MD simulations, one can switch (suddenly change) the native structure information. This is a simple way to mimic proteins at work driven by some external forces, such as ATP hydrolysis free energy or proton motive force. (6) During MD simulations, one can switch the relative stability in the multiple-basin model. This is another and somewhat more sophisticated way to mimic activation by some driving force. The protocols 5 and 6 are the major advantages of CafeMol over other methods and packages.

It may be convenient to compare CafeMol software with other publicly available means for CG MDs of proteins. Standard and some extended Go models can be run by GROMACS with the SMOG tool³⁹ and by Charmm with the Go-model-builder tool.⁴⁰ They are convenient since they can run on common MD packages, which are familiar to many people, although some extensions, such as multiple-basin potentials, may be more difficult because of the framework given by the packages. The ENM and its extensions are simpler models and thus can be run conveniently by many ways including a software RedMD,¹⁰ another software REACH,⁴¹ and web servers iGNM/oGNM/oANM.⁴² These tools, by themselves, cannot handle large-scale deviations from the native structure that include complete or partial unfolding. More flexible ways to realize CG MD may be to use general tools, such as a CG-builder of VMD/NAMD packages^{3,43} and some modules of LAMMPS.⁴⁴ They are flexible and thus may require more adaptation to particular applications. No publicly available method, other than CafeMol, can run conformational change dynamics between two (or more than two) given structures, to our knowledge. Also, internal framework for switching the potential is a unique feature of CafeMol.

CafeMol is a standalone software written in Fortran90 with MPI and C preprocessing directives. Thus, it can run in virtually any computer that has a Fortran90 compiler. The CafeMol source code, together with the parameter set and sample input files, is freely available to anyone at the web page, <http://www.cafemol.org>. A summary on the structure of the code and used file formats are given in Supporting Information.

2.2. Models and Energy Functions. *2.2.1. Off-Lattice Go Model.* CafeMol implements the off-lattice Go model developed by Clementi, Nyemyer, and Onuchic.¹² For a protein with the number of amino acids n_{aa} , the Go model potential $V_{Go}(\mathbf{R}|\mathbf{R}_0)$ is defined by

$$V_{Go}(\mathbf{R}|\mathbf{R}_0) = \sum_i K_b(b_i - b_{i,0})^2 + \sum_i K_\theta(\theta_i - \theta_{i,0})^2 + \sum_i \{K_\phi^{(1)}[1 - \cos(\phi_i - \phi_{i,0})] + K_\phi^{(3)}[1 - \cos 3(\phi_i - \phi_{i,0})]\} + \sum_{i < j - 3}^{\text{nat contact}} \varepsilon_{go} \left[5 \left(\frac{r_{ij0}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij0}}{r_{ij}} \right)^{10} \right] + \sum_{i < j - 3}^{\text{non-native}} \varepsilon_{ev} \left(\frac{d}{r_{ij}} \right)^{12} \quad (1)$$

Here, \mathbf{R} is the Cartesian coordinates of the simulated protein as a $3n_{aa}$ -dimensional vector, b_i is the i -th virtual bond length defined as $|\mathbf{r}_{i+1} - \mathbf{r}_i|$ ($1 \leq i \leq n_{aa} - 1$), where \mathbf{r}_i stands for the Cartesian coordinate of the i -th amino acid (= CG particle), θ_i is the i -th bond angle between two consecutive virtual bond vectors, $\mathbf{r}_{i+1} - \mathbf{r}_i$ and $\mathbf{r}_{i+2} - \mathbf{r}_{i+1}$ ($1 \leq i \leq n_{aa} - 2$), ϕ_i is the i -th dihedral angle around the $i + 1$ -th virtual bond $\mathbf{r}_{i+2} - \mathbf{r}_{i+1}$ ($1 \leq i \leq n_{aa} - 3$), and r_{ij} is the distance between i -th and j -th amino acids. All parameters with the subscript 0 are the constants which have the values of the corresponding variables at the native structure \mathbf{R}_0 , which corresponds to the structure at the bottom of the folding funnel. K_b , K_θ , $K_\phi^{(1)}$, $K_\phi^{(3)}$, ε_{go} , ε_{ev} , and d are the parameters. For the former six parameters, Clementi et al.'s original model uses homogeneous setting, i.e., for each parameter, the same value is used for the entire systems. In CafeMol one can optionally use site-specific parameters, i.e., the parameter values that depend on residues, which allow one to model flexible loops, for example. The summation $\sum_{i < j - 3}^{\text{nat contact}}$

is over the “native contact pairs”, pairs of amino acids that are physically close to each other at the native (or the reference) structure. If one of the nonhydrogen atoms in the i -th amino acid is within a threshold distance (which is 6.5 Å by default) from a nonhydrogen atom in the j -th amino acid, then we define the pair of the i -th and the j -th amino acids as being the native contact. Only nonlocal pairs with $i < j - 3$ are taken into account. We note that, even though CafeMol uses one bead (most often located at C_α atom) per amino acid as the dynamic variable in MD simulations, the native contacts are defined by using all-atom information at the native (reference) structure. Thus, the PDB structure given for the native structure must contain all-atom coordinates. The summation $\sum_{i < j - 3}^{\text{non-native}}$ is over pairs that are not in the native contact pair set. Only nonlocal pairs with $i < j - 3$ are taken into account.

2.2.2. Atomic Interaction Based CG Model. In the Go model, by default, the parameters K_b , K_θ , $K_\phi^{(1)}$, $K_\phi^{(3)}$, and ε_{go} are independent of the residue number i and of secondary structures, which implies that all the information coded by amino acid sequence of a protein is represented by the protein native structure. For better chemical specificity, one may want to use sequence-dependent parameters. Indeed, it was found that in some cases, such a chemical specificity can be crucial for the protein folding and other functional dynamics. Undoubtedly, appropriately implementing the interaction specificity into the Go model can improve the description of the protein dynamics. In CafeMol, based on the work of Li, Wolynes, and Takada,³⁶ we provide a way to implement such a kind of Go model with sequence-specific interactions, of which parameters were determined based on all-atom AMBER energy with an implicit solvent model by using a multiscale approach. This new model is called atomic interaction based CG (AICG) model.

In the AICG model, the interaction strength between the natively interacting residues i and j depends on i and j and is written as $\varepsilon_{go,ij} = \varepsilon_{go} w_{ij}$, with w_{ij} and ε_{go} being the relative weight of the interactions for each pair of contacting residues and the average of the nonlocal native interactions, respectively. The w_{ij} controls the heterogeneity of the nonlocal native interactions. In a simple version of AICG models, CafeMol provides an automatic estimate of w_{ij} via a linear regression to AMBER energy in terms of some atomic details of the residue contacts (see the Supporting Information of ref 36). Alternatively, CafeMol allows the user to provide AMBER-based residue-contact energy precalculated by the user.

In the AICG model, the residues with different secondary structures have different interaction parameters for bond and dihedral angles. Residues are assigned to one of the four major secondary structures, i.e., α -helix, β -strand, turn, and random coil, by the define secondary structure of protein (DSSP) method.⁴⁵ We also assign an independent parameter for the bond and dihedral angles which contain glycine. The nonlocal parameter ε_{go} and all the local parameters are generic and optimized by using the fluctuation matching method for a set of proteins.

2.2.3. Multiple-Basin Model. CafeMol employs the multiple-basin potential of Okazaki et al.,²³ which, based on the Clementi et al.'s off-lattice Go model, represents the energy landscape that has more than one energy basin. We here write the equation of the potential for the two-basin case since this is most often used in real application, although CafeMol can treat the cases with more than two basins.

The multiple-basin potential energy function V_{MB} is defined by the (smaller) eigenvalue of the characteristic equation:

$$\begin{pmatrix} V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_1) & \Delta \\ \Delta & V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_2) + \Delta V \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = V_{\text{MB}} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad (2)$$

where $V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_\nu)$ is essentially the Clementi's off-lattice Go potential $V_{\text{Go}}(\mathbf{R}|\mathbf{R}_\nu)$ but is modified in two respects, as described later. The condition that a nontrivial solution exists leads to a secular equation of which the explicit solution is given as

$$V_{\text{MB}}(\mathbf{R}|\mathbf{R}_1\mathbf{R}_2) = \frac{1}{2}[V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_1) + V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_2) + \Delta V] - \sqrt{\left(\frac{V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_1) - V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_2) - \Delta V}{2}\right)^2 + \Delta^2} \quad (3)$$

Here, Δ is a coupling constant, which smoothed the connection between two Go models (the larger Δ leads to smaller barrier height between two basins), and ΔV is to modulate the relative energies of the two basins. A convenient reaction coordinate that monitors the conformational change is $\chi = \ln(c_2/c_1)$, where (c_1, c_2) are the eigenvector corresponding to V_{MB} .

$V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_\nu)$ is, as was noted, conceptually the same as $V_{\text{Go}}(\mathbf{R}|\mathbf{R}_\nu)$ of Clementi et al. Purely for technical reasons, we need to introduce two modifications. We write the $V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_\nu)$ in the sum of three terms:

$$V_{\text{Go}+}(\mathbf{R}|\mathbf{R}_\nu) = V_{\text{local}}(\mathbf{R}|\mathbf{R}_\nu) + V_{\text{native-attr}}(\mathbf{R}|\mathbf{R}_\nu) + V_{\text{repul}}(\mathbf{R}|\mathbf{R}_\nu) \quad (4)$$

where the first term is

$$V_{\text{local}}(\mathbf{R}|\mathbf{R}_\nu) = \sum_i K_{b,i}(b_i - b_{i,\nu})^2 + \sum_i K_{\theta,i}(\theta_i - \theta_{i,\nu})^2 + \sum_i \{K_{\phi,i}^{(1)}[1 - \cos(\phi_i - \phi_{i,\nu})] + K_{\phi,i}^{(3)}[1 - \cos 3(\phi_i - \phi_{i,\nu})]\} \quad (5)$$

This local potential is slightly different from that of $V_{\text{Go}}(\mathbf{R}|\mathbf{R}_\nu)$: Namely, all of the K 's are now dependent on i in the following ways:

$$K_{b,i}/K_b = \min \left[1, \frac{\varepsilon_{b,\max}}{K_b(b_{i1} - b_{i2})^2} \right] \quad (6)$$

$$K_{\theta,i}/K_\theta = \min \left[1, \frac{\varepsilon_{\theta,\max}}{K_\theta(\theta_{i1} - \theta_{i2})^2} \right] \quad (7)$$

and

$$K_{\phi,i}^{(1)}/K_\phi^{(1)} = K_{\phi,i}^{(3)}/K_\phi^{(3)} \quad (8)$$

$$= \min \left[1, \frac{\varepsilon_{\phi,\max}}{K_\phi^{(1)}[1 - \cos(\phi_{i,1} - \phi_{i,2})] + K_\phi^{(3)}[1 - \cos 3(\phi_{i,1} - \phi_{i,2})]} \right] \quad (9)$$

This is introduced so that the spring constant is weakened where too large local changes are observed between the two reference structures. The thresholds $\varepsilon_{b,\max}$, $\varepsilon_{\theta,\max}$, and $\varepsilon_{\phi,\max}$ define the "large local change". We note that because of this, $V_{\text{local}}(\mathbf{R}|\{\mathbf{R}_\nu\})$

is not just a function of the reference structure \mathbf{R}_ν , but it also depends on other reference structures.

Nonlocal potentials are divided into attractive terms $V_{\text{native-attr}}$ and repulsive terms V_{repul} , and the former is given as

$$V_{\text{native-attr}}(\mathbf{R}|\mathbf{R}_\nu) = \varepsilon_{\text{go}} \sum_{i < j - 3}^{\text{nat contact}} \min \left[1, 5 \left(\frac{r_{ij,\nu}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij,\nu}}{r_{ij}} \right)^{10} + 1 \right] \quad (10)$$

where the summation is over the native contact pairs in the same way as that of the single Go model. The repulsive part V_{repul} is further divided into two terms $V_{\text{repul-1}}(\mathbf{R}|\mathbf{R}_\nu)$ and $V_{\text{repul-2}}(\mathbf{R})$:

$$V_{\text{repul}}(\mathbf{R}|\mathbf{R}_\nu) = V_{\text{repul-1}}(\mathbf{R}|\mathbf{R}_\nu) + V_{\text{repul-2}}(\mathbf{R}) \quad (11)$$

where

$$V_{\text{repul-1}}(\mathbf{R}|\mathbf{R}_\nu) = \varepsilon_{\text{go}} \sum_{i < j - 3}^{\text{nat-related}} \max \left[0, 5 \left(\frac{r_{ij,\nu}^{\min}}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij,\nu}^{\min}}{r_{ij}} \right)^{10} \right] \quad (12)$$

and

$$V_{\text{repul-2}}(\mathbf{R}) = \varepsilon_{\text{ev}} \sum_{i < j - 3}^{\text{purely non-nat}} \left(\frac{d}{r_{ij}} \right)^{12} \quad (13)$$

where

$$r_{ij,\nu}^{\min} = \min_{\nu}^{\text{nat contact}} [r_{ij,\nu}] \quad (14)$$

Here, the repulsive-1 term is used for the pairs for which the native contact is formed at least in one of the reference structures ("type 1" and "type 2" pairs in the original paper of Okazaki et al). We termed it as "native-related pairs". For a particular pair ij , if the pair is in native contact in the state ν (state corresponds to basin), then this is a (true) native contact for this state. If the pair is not in the native contact in the state ν , but it makes contact in another state, then we call this pair in the "dummy contact" in the state ν . For the state ν , the "native-related pairs" include both the true native and the dummy contact sets. The repulsive-2 term is for the pairs for which native contact is never formed in any of the reference structures ("type 3" pairs in the original paper). $V_{\text{repul-2}}(\mathbf{R})$ is the same form as that of Clementi et al's Go model.

We note that the attractive part has the same shape as that of Clementi et al but the repulsive part of the Lennard-Jones-like potential is modified for the native-related pairs. Between the two (or all the) states, the repulsive part is identical. The $r_{ij,\nu}^{\min}$ is the smallest distance between i and j among all the states ν for which this pair ij is in native contact. We note that because of this, $V_{\text{repul-1}}(\mathbf{R}|\mathbf{R}_\nu)$ is not just a function of the reference structure \mathbf{R}_ν , but also depends on other reference structures.

2.2.4. Elastic Network Model. The ENM was first proposed by Tirion in all-atom representation.³² CafeMol uses it in one-bead-per-amino-acid CG level. Simply, all the native contact pairs are connected by elastic bonds with their natural lengths $r_{ij,0}$ being equal to those at the native structure. The spring constant K is the same for all the elastic bonds. The same definition of the native contact as that of the off-lattice Go model

is used here:

$$V_{\text{ENM}} = K \sum_{i < j}^{\text{nat contact}} (r_{ij} - r_{ij,0})^2 \quad (15)$$

2.2.5. Electrostatic and Hydrophobic Interactions. In addition to the above four models, we can optionally turn on other interactions that include the electrostatic interaction, a hydrophobic interaction, a pulling force for steering, and others. The electrostatic interaction is expressed as the Debye–Hückel form. The hydrophobic interaction defined below is an empirical form and is qualitatively similar to the accessible surface area model.

The hydrophobic (HP) interaction is modeled by a many-body energy function,²⁵ which has a similar functional form to that was used in Fujitsuka et al.:^{46,47}

$$V_{\text{HP}} = -c_{\text{HP}} \sum_{i \in \text{HP}} \varepsilon_{\text{HP},A(i)} S_{\text{HP}}(\rho_i) \quad (16)$$

Here, the coefficient c_{HP} is to scale the overall strength of the hydrophobic interactions. $A(i)$ stands for the type of the particle i ; 21 types are considered: $A(i) = 1, 2, \dots, 20$ means the 20 types of amino acids and $A(i) = 21$ simply represents all nonamino acid particles. The particle-type-specific parameter that reflects the hydrophobicity of particles is $\varepsilon_{\text{HP},A}$. The “buriedness” $S_{\text{HP}}(\rho_i)$ of the particle i is defined by

$$S_{\text{HP}}(\rho) = \begin{cases} c_{\text{linear}} \rho & \rho \leq \rho_{\text{min}} \\ c_{\text{linear}} \rho + 0.5(1 - c_{\text{linear}}) \left[1 + \cos \frac{\pi(1 - \rho)}{1 - \rho_{\text{min}}} \right] & \rho_{\text{min}} < \rho < 1 \\ 1 & \rho \geq 1 \end{cases} \quad (17)$$

Here, c_{linear} defines the contribution of the linear term, and ρ_{min} is the threshold to introduce nonlinear form. The local density ρ_i for particle i is calculated by

$$\rho_i = \frac{\sum_{j \in \text{HP}, j \neq i} n_{A(j)} u_{\text{HP}}(r_{ij}, r_{\text{min}}, A(i), A(j), r_{\text{max}}, A(i), A(j))}{n_{\text{max}, A(i)}} \quad (18)$$

where n_A is the number of atoms that the particle type A represents, and $n_{\text{max},A}$ is the maximum coordination number for particle type A . The function u_{HP} represents the degree of the contact between particles i and j and is defined as

$$u_{\text{HP}}(r, r_{\text{min}}, r_{\text{max}}) = \begin{cases} 1 & r \leq r_{\text{min}} \\ 0.5 \left(1 + \cos \frac{\pi(r - r_{\text{min}})}{r_{\text{max}} - r_{\text{min}}} \right) & r_{\text{min}} < r < r_{\text{max}} \\ 0 & r \geq r_{\text{max}} \end{cases} \quad (19)$$

where r_{min} (r_{max}) is to define the cutoff for the minimal (maximal) distance, which depends on the particle types $A(i)$ and $A(j)$.

In real applications, the target system may contain more than one protein molecule. In such a case, the total energy function is the sum of intrachain and interchain interactions. For the former, CafeMol allows us to mix the models; it is possible, for example, that one chain is represented by the Go model, and another chain by the multiple-basin model. Interactions between chains may be modeled by a combination of the few interactions, such as the

contact energy (the same way as the Go model), the electrostatic interaction, the hydrophobic interaction, and the simple repulsion.

We also mention that we can realize a frustrated system even in the framework of the Go model. For example, in modeling a flexible docking of two proteins, one can represent the intrachain interactions of two individual proteins by the Go model using their unbound structures as references, while the interactions between two proteins may be modeled by the native contact taken from the bound complex structure. Even though all the interactions are of Go-type, their reference structures are not the same, thus introducing frustrations. CafeMol has flexible interfaces which allow one to realize such a mixed-reference simulation very easily.

2.2.6. Explicit and Implicit Ligand Models. Protein conformational change is coupled with a ligand binding, and one may want to take into account the ligand binding dynamics into CG simulations. CafeMol has two ways to include ligands: the explicit and implicit ligand models. Explicitly including CG ligands is a straightforward method, whereas the implicit ligand model is a computationally faster alternative proposed in Okazaki and Takada, 2008.¹⁸

In the explicit ligand model, small ligand molecules are modeled as a rigid linear chain, in which the sequence of ligand beads is defined in the input PDB file of the ligand. The energy function for explicit ligands is essentially the same as the local potential term in the Go model, in which the spring constants are set as sufficiently large values to make the ligand rigid. Between the ligand and proteins, we can use the same interactions as those for interprotein interactions.

On the other hand, in the implicit ligand model, we describe that ligand binding in the “two-state” manner. Namely, the ligand is either bound (B) or unbound (U) to a protein. In the unbound state, the protein has just its intraenergy, represented typically by the multiple-basin model, whereas the protein in the bound state has the energy which is the sum of its intraenergy and the ligand binding energy, $V_{\text{imp-lig}}$. This ligand binding energy $V_{\text{imp-lig}}$ does not contain the explicit coordinates of the ligand atoms but is a function of the Cartesian coordinates of ligand-mediated sites of the protein. It takes the negative and large absolute value when the local environment around the binding pocket is close to that of the reference conformation. Concretely, the $V_{\text{imp-lig}}$ is normally defined as

$$V_{\text{imp-lig}} = - \sum_{\substack{\text{ligand-mediated} \\ \text{contact pairs}}} c_{\text{lig}} \varepsilon_{\text{go}} \exp \left[- \frac{(r_{ij}/r_{0ij} - 1)^2}{2(\sigma/r_{0ij})^2} \right]$$

where the “ligand-mediated contact pairs” means the amino acid pairs that satisfy the following three conditions: (a) Both of amino acids in the pair are involved in the binding-sites, which are provided in the input file. (b) The pair is not involved in the native contact. (c) At least one non-hydrogen atom in one amino acid is within 10 Å from at least one non-hydrogen atom in the other amino acid. In time propagation, the protein conformation is moved by the standard MD simulation (described below), whereas the ligand binding state (B or U) is stochastically changed by the rates k_b (binding) and k_u (unbinding) implemented as the metropolis Monte Carlo (MC) scheme in the following way. While in the unbound (U) state, a ligand molecule

reaches the binding pocket at every Δt_b time with probability $p = k_b \Delta t_b$. With this probability, the state changes to the bound (B) one. Here k_b is the apparent first-order rate for binding and thus is proportional to the ligand concentration. While in the bound state, at every Δt_u time, the bound ligand has chance to dissociate at a probability $\exp(-|V_{\text{imp-lig}}|/k_B T)$. The mixed MD–MC scheme thus described here is a convenient way of simulating protein conformational dynamics coupled with ligand binding.

2.3. Simulation Methods. **2.3.1. Units.** CafeMol uses a specific unit. The length unit is Å. The energy unit is kcal/mol. For temperature, we use Kelvin, $1 K = 1.987 \times 10^{-3} \text{ kcal/mol}$. The mass unit is our own one. We set that each amino acid has the mass of 10, which we call 10 cafe-mu (cafe-mol mass unit). Since average mass of 20 amino acids is 137 amu (atomic mass unit), we thus define 1 cafe-mu = 13.7 amu. From these, we can directly obtain the unit time of CafeMol to be 1 cafe-time = $1.809 \times 10^{-13} \text{ s} \sim 200 \text{ fs}$. We need to be cautious in interpreting this time scale. Since intrinsic dynamics is accelerated by coarse graining the energy landscape, this apparent mapping in time scale does not necessarily give a good estimate of the real time.

2.3.2. Constant Temperature MD. The constant temperature MD is the most basic simulation protocol in CafeMol. Currently, one can use either constant-temperature “Newtonian” or Langevin dynamics. The underdamped Langevin dynamics is integrated by the scheme developed by Honeycutt, Guo, and Thirumalai.^{48,49} For the constant-temperature Newtonian dynamics, we employed the velocity-type Verlet algorithm for updating protein structure with the Berendsen thermostat for controlling temperature.³⁸

2.3.3. Searching T_F . CafeMol can automatically estimate T_F , at which the protein is folded (or denatured) with 50% probability, by using the bisection method. Namely, we first specify, in the input file, the lower and the upper bounds of T_F . CafeMol first simulates the protein at the midpoint temperature of the two bounds and sees if the protein is near native conformation for more than half of the simulation time. If yes, then this temperature is set as the new lower bound of T_F . Otherwise, the simulated temperature was set as the new upper bound of T_F . With the new set of the lower and the upper bounds, CafeMol repeats the simulation at their midpoint. This iteration lasts for the required times to narrow the temperature range. For an accurate estimate of T_F , the MD step number needs to be much larger than the folding and unfolding time scale near T_F . Typically, the condition is easily satisfied for proteins with less than 100 amino acids but for proteins with more than 200 residues

2.3.4. Simulated Annealing. The simulated annealing is to search lower (possibly the lowest) energy structure in the simulated system. It uses the constant-temperature MD routine, and the temperature in the routine is decreased at a certain rate.

2.3.5. Replica Exchange Method. CafeMol can run one- and two-dimensional replica exchange method (REM).^{50–52} The standard replica-specifying parameters are the temperature and the ionic strength, although one can easily adapt it for other replica-specifying parameters. In addition, CafeMol equips the feedback-optimized REM,⁵³ by which the replica temperatures can automatically be optimized.

2.3.6. Switching Go Model. Many proteins work by changing their conformations depending on their interactions with their partner molecules. For example, proteins change their conformation upon binding to ligands. How binding and conformational change are coupled in proteins is in itself a subject to be studied. However, for studying more complex biological phenomena, we

may want to enforce such a conformational change by design. The switching Go model is proposed for this type of simulations.

In the switching Go model, we first simulate a protein by the Go model $V_{\text{Go}}(\mathbf{R}|\mathbf{R}_A)$ with a structure A being the reference (i.e., native) structure. At room temperature, the protein usually resides nearby the structure A, $\mathbf{R} \sim \mathbf{R}_A$. At a certain time, we suddenly change the reference structure of the Go model to another structure B, resulting in a new Go potential $V_{\text{Go}}(\mathbf{R}|\mathbf{R}_B)$. The protein jumps from the bottom of the Go model $V_{\text{Go}}(\mathbf{R}|\mathbf{R}_A)$ to the downhill slope of the new Go potential $V_{\text{Go}}(\mathbf{R}|\mathbf{R}_B)$. Note that the simulated structure \mathbf{R} does not change suddenly but changes continuously. Right after the switch, the protein starts to relax its conformation from $\mathbf{R} \sim \mathbf{R}_A$ to $\mathbf{R} \sim \mathbf{R}_B$. This mimics the conformational change from the structure A to the structure B, in a simple and crude way. CafeMol is ready to switch Go models in any fixed times.

2.3.7. Switching Bias in Multiple-Basin Potential. The switching Go model is perhaps the simplest way to realize some large-scale conformational change in biomolecules. However, in switching Go models, the protein is “excited” to the new potential, and the resulting procedure is nothing but the relaxation on the new Go potential surface. This resembles to the photoactivated process. Biologically, however, many events are thermally activated, and thus overcoming the energy barrier by thermal fluctuation may be of essential importance, in some cases. To realize this thermally activated conformational transition, CafeMol uses the multiple-basin model. We first simulate a protein with the multiple-basin model setting the structures A and B as reference structures of the two basins, i.e., $V_{\text{MB}}(\mathbf{R}|\mathbf{R}_A, \mathbf{R}_B, \Delta V)$, where ΔV is positive and sufficiently large so that the basin A is more stable. The protein mostly resides in the basin A, i.e., $\mathbf{R} \sim \mathbf{R}_A$. At a certain time, we suddenly change the bias ΔV to a negative and sufficiently large absolute value so that the basin B is now more stable. Soon after the switch, the protein would still fluctuate in the basin A, but after a while, it overcomes the barrier to reach the more stable basin B, through thermal fluctuations. We can also couple the switch bias in multiple-basin potential and switch the reference structure as in the case of switching Go model.

3. RESULTS

Here, after investigating parallel performance we illustrate prototypical simulations by CafeMol. Six basic examples are followed by simulations of two protein machines at work. Technical details on how to write input files are described in the Supporting Information and in Tables S1–S4.

3.1. Parallel Performance. The CafeMol code is parallelized by MPI and OpenMP commands. The parallel efficiency strongly depends on the simulated system. As in all other MD simulations, the bottleneck in computation is the force calculation and thus parallelization in the force calculations determines the overall parallel performance. For simple Go model simulations without additional interactions, forces from the native contact and excluded volume interactions are dominant. When the electrostatic interaction is turned on, the computational bottleneck shifts to the force calculation of the electrostatic interaction, as in normal MD simulations. This is because the native contact is relatively short ranged, while the electrostatic interactions are long ranged.

Here, we show benchmark tests of parallel efficiency up to 64 cores for 2 viral capsid protein complexes using the Go model

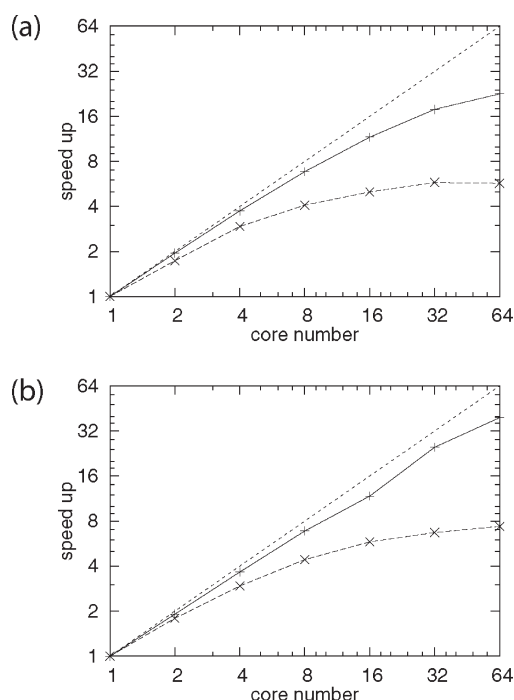


Figure 1. Performance of parallel computation of Go model simulations with (solid curves) and without (dashed curves) electrostatic interactions. (A) STMV capsid proteins containing 8820 residues. (B) Norwalk viral capsid proteins containing 89700 residues. The acceleration rate is plotted against core numbers in the log–log scale. The dotted line indicates the ideal linear scaling.

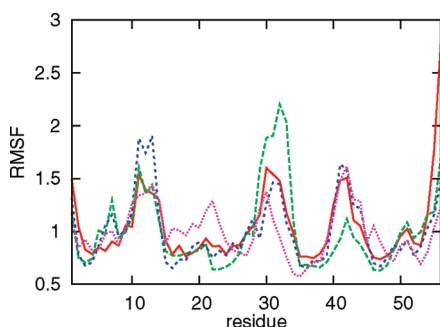


Figure 2. The RMSF of src SH3 domain in its native state at 300 K. Results by Go model (the red solid line), by all-atom MD with an implicit solvent model (the green long-dashed line), and by all atom MD with explicit water solvent (the blue short-dashed line) together with the estimate from the X-ray B-factors (the pink dotted line) are compared. All the curves are scaled so that their averages become unity.

with and without electrostatic interactions. We used a standard PC cluster with infiniband networks. OpenMP is used up to four cores, whereas the MPI/OpenMP hybrid is employed beyond four cores, in which OpenMP always deals with four cores. A smaller system, satellite tobacco mosaic virus (STMV) capsid protein complex (the pdb code 1A34), is made of 60 identical proteins of 147 amino acids (excluding unstructured terminus), containing 8820 CG particles. A larger system, capsid protein complex of Norwalk virus (the pdb code 1ihm), contains 180 subunits of about 500 residue proteins which results in 89700 CG particles (about 10 times as large as STMV). For each of the two protein complexes, we measured the parallel performance of

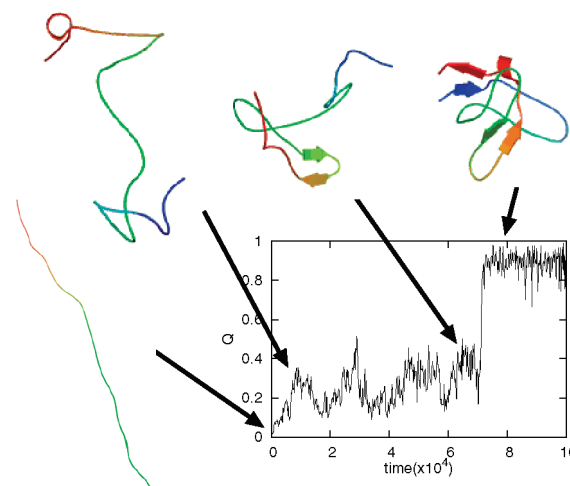


Figure 3. Time course of src SH3 domain folding simulated by the Go model. The nativeness is quantified by the Q score, the fraction of the formed native contacts. Some representative snapshot structures are illustrated. The molecular images in this article were created with PyMOL⁵⁸ and VMD.⁴³

constant temperature MD simulations of the Go model with and without the electrostatic interactions (Figure 1). Without electrostatic interactions, the parallel performance saturated at around 8 cores and the asymptotic speed-up was about 6- and 7- fold for STMV and for the Norwalk virus, respectively (dashed curves). When the electrostatic interactions were turned on, parallel efficiency was markedly improved (solid curves); 24- and 40- fold speeds-up for STMV and for the Norwalk virus, respectively, with 64 cores. Thus, as expected, for larger molecular systems with electrostatic interactions, the parallel performance is more pronounced.

In addition, the replica exchange MD is parallelized naturally by MPI. Namely, different replicas are dealt with different nodes. One can combine the force parallelization and the parallel treatment of replicas, resulting in two-dimensional parallelization.

3.2. Near-Native Dynamics. We start with the constant temperature MD simulation of src SH3 domain protein by the Go model to calculate fluctuations in the native state. We perform a 1×10^6 step MD simulation with the time step 0.2 cafe-time. See Supporting Information, Table S1 for technical details.

Figure 2 represents the root-mean-square fluctuation (RMSF) of every amino acids in src SH3 domain protein calculated by the Go model together with the results obtained by some standard all-atom MD simulations. They contained the 10 ns MD result by all-atom protein model of the AMBER ff99SB⁵⁴ with the modified generalized Born implicit solvent model⁵⁵ as well as those by the same all-atom model with explicit water molecules of TIP3P.⁵⁶ Apparently in Figure 2, the agreement among the three methods is overall quite good. Not surprisingly, both N- and C-termini residues are intrinsically flexible with quite a large RMSF, and they can differ, to more extent, among three results. Ignoring 3 residues in each terminus, we obtained the correlation coefficient between the RMSFs by the Go model and those by the all-atom MD (explicit water) to be 0.88. In contrast, the correlation coefficient between the RMSFs by all-atom MD with explicit waters and those with the implicit water model was 0.65 for this protein.

3.3. Protein Folding. The second example is a folding simulation of src SH3 domain protein by the Go model at 300 K.

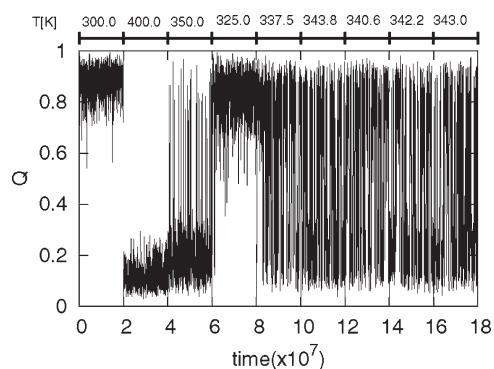


Figure 4. The automatic T_F search by the bisection method for src SH3 domain. The constant temperature MD is performed at each temperature (specified at the top of the figure) for 2×10^7 cafe-time, and the time series of the Q score is plotted.

Figure 3 shows time series of an overall folding reaction coordinate, i.e., Q -score, the fraction of the formed native contacts. We see that the folding occurs quite cooperatively, and the folding nucleus is found in the so-called distal β -hairpin (light-green to orange segments in the cartoon of Figure 4), which agrees quite well with the results of experimental ϕ -value analysis.⁵⁷

3.4. Automatic T_F Search. As described above, CafeMol has a function to automatically estimate T_F for modest size proteins. Here, we exemplify it for src SH3 domain protein. Figure 4 shows the result. Since the upper and the lower temperature bounds are 500 and 100 K, the initial midpoint temperature is 300 K where the first MD simulation is conducted for 2×10^7 cafe-time. This results in the native state for almost the entire time. Then, 300 K is set as the lower bound. In the second round, the new midpoint temperature is 400 K. The same procedure continues nine times. The employed temperatures in every step are shown at the top of the figure. See Supporting Information, Table S2, for technical details.

3.5. Conformational Change by Multiple-Basin Potential. We then illustrate a conformational change simulation of the glutamine binding protein by the multiple-basin potential. The glutamine binding protein is a typical molecule that exhibits close/open conformational changes upon ligand binding/dissociation, and both the open and the closed structures are available. The multiple-basin potential has two empirical and important parameters, ΔV and Δ , which we tuned by a trial-and-error approach so that the resulting MD produces reversible transitions between the open and the closed structures and that the residential time in each state is nearly 0.5. Here, we chose $\Delta V = -8$ and $\Delta = 55$. The result of a MD simulation is shown in Figure 5 where the reaction coordinate χ (defined above) is plotted as a function of time. The χ is negative (positive) when the protein is in the basin of open (closed) state. Here, we observed quite abrupt and reversible conformational transitions between the two states, thus showing that the two states separated by a low free energy barrier are in near equilibrium. See Supporting Information, Table S3, for technical details.

3.6. Conformational Change by Switching Go Model. Although we can realize reversible conformational transitions between two structures for small proteins by the multiple-basin model, as was described in the previous subsection, this may not easily be realized for a very large-amplitude conformational change in very large proteins. For such cases, we can induce

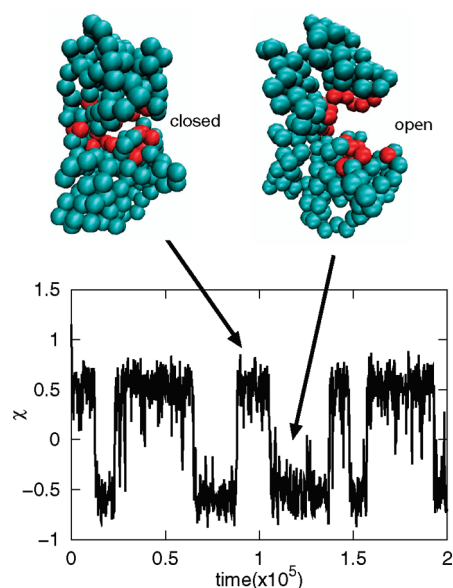


Figure 5. Reversible conformational transitions of glutamine binding protein by the multiple-basin model. The parameters are tuned so that the open and closed states are almost equally stable and the conformational changes are sufficiently frequent. The χ monitors the conformational change and takes the positive (negative) value when it is in the closed (open) state. The representative snapshots are drawn where the binding site residues are in red.

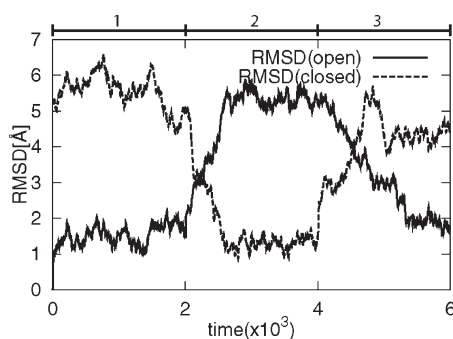


Figure 6. Conformational changes of glutamine binding protein by the switching Go model. The y-axis is the rmsd from the open (solid) and the closed (dashed) states. The potential is switched every 2×10^3 time.

such large-scale conformational changes by switching the Go model, although this is less sophisticated than the multiple-basin model. Here, we exemplify it for the glutamine binding protein. First, the protein is simulated with the Go model with the reference structure being the open structure. At $t = 2 \times 10^3$, we switch the reference structure of the Go model to the closed structure, which immediately initiates a closing motion. Later, we switch the native structure back to the open structure inducing the opening conformational change. See Supporting Information, Table S4, for technical details.

Figure 6 depicts the result, where the y-axis monitors the root-mean-square deviation (rmsd) from two reference structures. We clearly observe here that, after switching, the protein gradually makes conformational change to the new native structures.

3.7. Replica Exchange MD. The next example is a simple folding and unfolding simulation of protein G (61 residue small

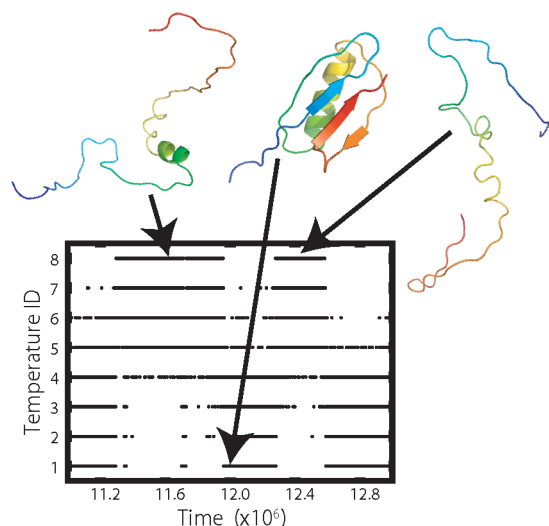


Figure 7. Itinerary of replica 1 in the replica MD simulations of protein G with 8 replicas. The temperature ID of the replica 1 is plotted as a function of time. Some representative snapshot structures are illustrated.

protein, pdb id 2igd) by the REM. Although the automatic T_F search work well for this small protein, the REM is the only one available method in CafeMol one can use to determine the transition temperature T_F for larger proteins. Here, we use 8 replicas of which temperatures are distributed exponentially with the lower and the upper bound temperatures being 300 and 380 K, respectively.

Figure 7 shows temperature id itinerary of the replica 1 showing that the replica travels all the temperature id's in quite high frequency. Some representative structures are shown in the figure.

3.8. Molecular Motor at Work. The last two examples are simulations of protein machines at work. Here, we illustrate a simulation of a rotary motor F_1 -ATPase. F_1 -ATPase, a half portion of ATP synthase, hydrolyzes ATP to rotate its central stalk. The X-ray structure of the minimal catalytic complex, $\alpha_3\beta_3\gamma$ was first solved by Walker's group.⁵⁹ In the complex, the central stalk, γ -subunit, is surrounded by the hexameric ring $\alpha_3\beta_3$ where the α - and β -subunits are arranged alternatively. The ATP hydrolysis reaction is catalyzed at each interface between the α - and β -subunits, with the catalytic residues primarily in β -subunits. In the Walker's structure, three $\alpha\beta$'s differ in their nucleotide-bound states as well as their structures: one $\alpha\beta$ binds ATP (called $\alpha\beta_{TP}$), another $\alpha\beta$ binds ADP (called $\alpha\beta_{DP}$), and the last $\alpha\beta$ does not bind any nucleotide (called $\alpha\beta_E$, where E stands for "empty"). Sequential and coordinated chemical and structural changes in three $\alpha\beta$'s suggested a 120° rotation of the central γ -subunit. This rotary mechanism, first proposed by Boyer,⁶⁰ was finally proved by the direct observation of the γ rotation by a single-molecule experiment.⁶¹ Based on the structural and single-molecule experiments, we have previously conducted CG MD simulations of the rotary motion using an ancestor program of CafeMol.¹⁴ Our trial-and-error computer experiments ended up with a prediction of a particular mechanochemical coupling scheme, the always bisite model, which is consistent with several experimental results published beforehand as well as afterward.^{62–64}

Here, we present a simple simulation of the rotation of γ in $\alpha_3\beta_3\gamma$ complex by the switching Go model (Figure 8) (We note

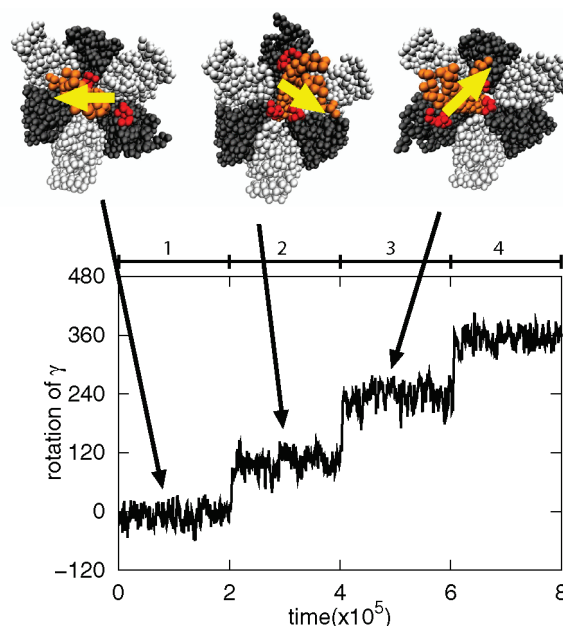


Figure 8. Rotary motion of F_1 -ATPase studied by the switching Go model. The rotation angle of γ -subunit is plotted as a function of time. The potential is switched every 2×10^5 cafe-time when the reference structures of all three $\alpha\beta$'s are cyclically changed. The representative snapshots are shown where the three β 's are drawn in dark gray, the three α 's are in light gray, the central γ is in orange. The yellow arrow is the vector that monitors the rotary angle of γ . In β 's, segments that make large conformational changes are marked in red.

that the simulation here is much simpler one than in ref 14, and it is only for illustration of the utility of CafeMol). We start the CG MD with the Walker's complex structure, where both intrasubunit and intersubunit interactions in $\alpha_3\beta_3$ rings are modeled by the Go model, the central stalk γ is also represented by the Go model, and the interactions between $\alpha_3\beta_3$ and γ are modeled by the simple repulsion. At $t = 2 \times 10^5$, we switch the reference structures of the $\alpha_3\beta_3$ ring so that each $\alpha\beta$ changes its structure one step ahead, i.e., in a site $\alpha\beta_{TP}$ is induced to change to $\alpha\beta_{DP}$, in another site $\alpha\beta_{DP}$ is changed to $\alpha\beta_E$, and in the third site, $\alpha\beta_E$ is changed to $\alpha\beta_{TP}$. After the switch, the three $\alpha\beta$'s indeed made the corresponding conformational changes, which further induced rotary motion of γ by 120° (Figure 8). At $t = 4 \times 10^5$ and 6×10^5 , we further switched the reference structures in the same way, observing totally the 360° rotation of γ -subunit.

3.9. Transporter at Work. As the final example, we show a simulation of a transporter at work. Here, we use a multidrug transporter that exports a drug molecule. Specifically, the simulated system is the bacterial multidrug transporter AcrB together with an antibiotic, minocycline. We note that the simulation here is based on, but is not identical to, our earlier work.²⁵

AcrB is a bacterial multidrug transporter that exports a broad range of hydrophobic drug molecules from cytoplasmic side to the cell outside driven by the proton motive force across the inner membrane. It was experimentally clarified that AcrB forms a homotrimer, and the trimeric complex takes asymmetric structures, where each subunit has different conformation:⁶⁵ one subunit takes the state B as the drug-bound state, another subunit takes the state E as the drug extrusion state, and the third subunit takes the state A as the drug access state. It has been postulated that the cyclic conformational change from state A, to

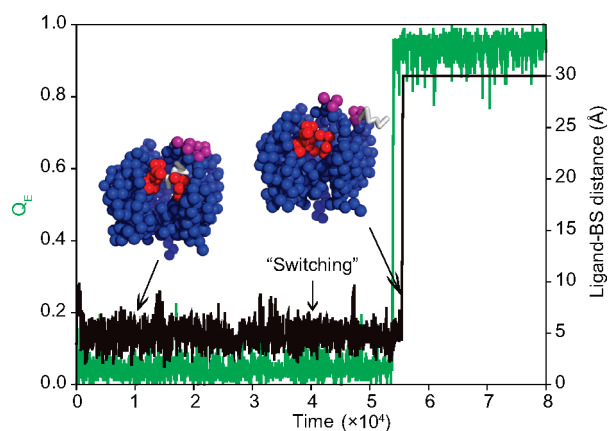


Figure 9. Simulation of multidrug transporter AcrB monomer with a drug molecule. The conformational change of AcrB is quantified as Q_E score (green curve with the scale on the left axis), which is defined as the fraction of native contacts that are observed in the E state but do not exist in the B or A state. Simultaneously, motion of the drug molecule is monitored by the distance between the geometric centers of drug and binding sites (BS) (black curve with the scale on the right axis). The simulation started in the B state and then at $t = 4 \times 10^4$ the potential function is switched to induce conformational change to the E state. The two snapshot structures are illustrated where AcrB is viewed from the side (some residues are not drawn to make the protein interior visible). The red spots in the middle are the drug binding site, whereas the purple spots at the top shows the drug exit point. The drug is drawn by white stick representation, which is near the binding pocket in the left cartoon and is near the exit point in the right cartoon.

state B, and to state E in each subunit can transport one drug molecule from the cytoplasmic side of the membrane to the cell outside per subunit and that, as the homotrimer, it proceeds from the asymmetric complex ABE to BEA, to EAB, and to ABE, which is called the functional rotation mechanism. This is analogous to the rotary motion of the $\alpha_3\beta_3$ ring in F_1 -ATPase.

Here, only for an illustration of a simulation with a small explicit ligand, we use one subunit of the drug binding domain, called the porter domain, of AcrB (see cartoon in Figure 9). We use the multiple-basin model with three states for AcrB and the hydrophobic interaction and the excluded volume terms for the interactions between AcrB and the drug. The simulation starts with the B state, and the drug (shown by the stick representation in Figure 9) is bound near the binding pocket (the left cartoon in Figure 9). Initially, the parameters are set so that the three states are equally probable. The energy function is switched at $t = 4 \times 10^4$ to stabilize the E state. Somewhat after the switching, AcrB makes the transition from the B to E state at around $t = 5.4 \times 10^4$ (monitored by the Q score shown in the green curve), which soon induces the drug export toward the outside of the cell at $t = 5.5 - 5.6 \times 10^4$ (the black curve).

4. DISCUSSIONS AND CONCLUSION

We have described the CafeMol method and software, which is characteristic, among many existing MD packages, in several aspects. First, it provides the MD code specific to the structure-based CG MD. In particular, the multiple-basin model is able to simulate conformational change of proteins when the two end structures are given, which is not easily realized by standard MD packages. Second, CafeMol provides various flexible and convenient means to mimic protein motions driven by external

energy source. Third, CafeMol equips the AICG model that takes into account atomic interactions in the framework of CG models, thus making it much more accurate without much of extra computational cost.

Yet, in the current version, some parameters in the energy functions were determined in a heuristic way. In particular, the parameters in the multiple-basin models are quite heuristic. It is highly desired that these parameters are derived systematically, for example, based on all-atom simulations.

Many more functions are currently under development in CafeMol. First, we are implementing CG models of DNA, RNA, and lipid membrane. Once realized, we can simulate all the macromolecular components of the cell in an integrated way. For better description of rotation and translation of macromolecules, the hydrodynamic interaction is crucial. We are working on implementing methods that take into account it.

■ ASSOCIATED CONTENT

S Supporting Information. Summary on the structure of the code and used file formats. Technical details on how to write input files. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: takada@biophys.kyoto-u.ac.jp.

■ ACKNOWLEDGMENT

This work is supported by Research and Development of the Next-Generation Integrated Simulation of Living Matter, a part of the Development and Use of the Next-Generation Supercomputer Project of the Ministry of Education, Culture, Sports, Science, and Technology. Some of the simulations were performed in RIKEN Integrated Cluster of Clusters.

■ REFERENCES

- (1) Brooks, B.; Brucoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (2) Pearlman, D.; Case, D.; Caldwell, J.; Ross, W.; Cheatham, T.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1–41.
- (3) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (4) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.
- (5) Izvekov, S.; Voth, G. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.
- (6) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.
- (7) Tan, R.; Petrov, A.; Harvey, S. *J. Chem. Theory Comput.* **2006**, *2*, 529–540.
- (8) Ayton, G.; Noid, W.; Voth, G. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.
- (9) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.
- (10) Gorecki, A.; Szymowski, M.; Dlugosz, M.; Trylska, J. *J. Comput. Chem.* **2009**, *30*, 2364–2373.
- (11) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694–698.
- (12) Clementi, C.; Nymeyer, H.; Onuchic, J. *J. Mol. Biol.* **2000**, *298*, 937–953.
- (13) Koga, N.; Takada, S. *J. Mol. Biol.* **2001**, *313*, 171–180.
- (14) Koga, N.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5367–5372.

- (15) Takagi, F.; Koga, N.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11367–11372.
- (16) Hyeon, C.; Lorimer, G.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18939–18944.
- (17) Hyeon, C.; Onuchic, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2175–2180.
- (18) Okazaki, K.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 11182–11187.
- (19) Noguchi, H.; Takasu, M. *Phys. Rev. E* **2001**, *64*, 041913.
- (20) Shelley, J.; Shelley, M.; Reeder, R.; Bandyopadhyay, S.; Moore, P.; Klein, M. *J. Phys. Chem. B* **2001**, *105*, 9785–9792.
- (21) Marrink, S.; de Vries, A.; Mark, A. *J. Phys. Chem. B* **2004**, *108*, 750–760.
- (22) Marrink, S.; Risselada, H.; Yefimov, S.; Tieleman, D.; de Vries, A. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.
- (23) Okazaki, K.; Koga, N.; Takada, S.; Onuchic, J.; Wolynes, P. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11844–11849.
- (24) Koga, N.; Kameda, T.; Okazaki, K.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 18237–18342.
- (25) Yao, X.; Kenzaki, H.; Murakami, S.; Takada, S. *Nat. Commun.* **2010**, *1*, 117.
- (26) Bryngelson, J.; Wolynes, P. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524–7528.
- (27) Bryngelson, J.; Onuchic, J.; Socci, N.; Wolynes, P. *Proteins* **1995**, *21*, 167–195.
- (28) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.
- (29) Leopold, P.; Montal, M.; Onuchic, J. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8721–8725.
- (30) Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11698–11700.
- (31) Taketomi, H.; Ueda, Y.; Go, N. *Int. J. Pept. Res.* **1975**, *7*, 445–459.
- (32) Tirion, M. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.
- (33) Haliloglu, T.; Bahar, I.; Erman, B. *Phys. Rev. Lett.* **1997**, *79*, 3090–3093.
- (34) Best, R.; Chen, Y.; Hummer, G. *Structure* **2005**, *13*, 1755–1763.
- (35) Lu, Q.; Wang, J. *J. Am. Chem. Soc.* **2008**, *130*, 4772–4783.
- (36) Li, W.; Wolynes, P.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 3504–3509.
- (37) Magnasco, M. *Phys. Rev. Lett.* **1993**, *71*, 1477–1481.
- (38) Berendsen, H.; Postma, J.; Van Gunsteren, W.; DiNola, A.; Haak, J. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (39) Noel, J.; Whitford, P.; Sanbonmatsu, K.; Onuchic, J. *Nucleic Acids Res.* **2010**, *38*, W657–W661.
- (40) Karanicolas, J.; Brooks, C. *Protein Sci.* **2002**, *11*, 2351–2361.
- (41) Moritsugu, K.; Smith, J. *Comput. Phys. Commun.* **2009**, *180*, 1188–1195.
- (42) Yang, L.-W.; Rader, A.; Liu, X.; Jursa, C.; Chen, S.; Karimi, H.; Bahar, I. *Nucleic Acids Res.* **2006**, *34*, W24–W31.
- (43) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (44) Plimpton, S. *J. Comput. Phys.* **1995**, *117*, 1–19.
- (45) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.
- (46) Fujitsuka, Y.; Takada, S.; Luthey-Schulten, Z.; Wolynes, P. *Proteins* **2004**, *54*, 88–103.
- (47) Fujitsuka, Y.; Chikenji, G.; Takada, S. *Proteins* **2006**, *62*, 381–398.
- (48) Honeycutt, J.; Thirumalai, D. *Biopolymers* **1992**, *32*, 695–709.
- (49) Guo, Z.; Thirumalai, D. *Biopolymers* **1995**, *36*, 83–102.
- (50) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (51) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.
- (52) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058–9067.
- (53) Katzgraber, H.; Trebst, S.; Huse, D.; Troyer, M. *J. Stat. Mech.: Theory Exp.* **2006**, *2006*, P03018.
- (54) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.
- (55) Onufriev, A.; Bashford, D.; Case, D. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.
- (56) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (57) Riddle, D.; Grantcharova, V.; Santiago, J.; Alm, E.; Ruczinski, I.; Baker, D. *Nat. Struct. Biol.* **1999**, *6*, 1016–1024.
- (58) *The PyMOL Molecular Graphics System*, version 1.2; Schrödinger, LLC: New York, 2008.
- (59) Abrahams, J.; Leslie, A.; Lutter, R.; Walker, J. *Nature* **1994**, *370*, 621–628.
- (60) Boyer, P. *BBA Libr.* **1993**, *1140*, 215–250.
- (61) Noji, H.; Yasuda, R.; Yoshida, M.; Kinosita, K., Jr. *Nature* **1997**, *386*, 299–302.
- (62) Adachi, K.; Oiwa, K.; Nishizaka, T.; Furuike, S.; Noji, H.; Itoh, H.; Yoshida, M.; Kinosita, K. *Cell* **2007**, *130*, 309–321.
- (63) Okuno, D.; Fujisawa, R.; Iino, R.; Hirono-Hara, Y.; Imamura, H.; Noji, H. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20722–20727.
- (64) Masaike, T.; Koyama-Horibe, F.; Oiwa, K.; Yoshida, M.; Nishizaka, T. *Nat. Struct. Mol. Biol.* **2008**, *15*, 1326–1333.
- (65) Murakami, S.; Nakashima, R.; Yamashita, E.; Matsumoto, T.; Yamaguchi, A. *Nature* **2006**, *443*, 173–179.