Chemical Physics 422 (2013) 165–174

Contents lists available at SciVerse ScienceDirect

Chemical Physics

journal homepage: www.elsevier.com/locate/chemphys

Coarse-graining of proteins based on elastic network models

Anton V. Sinitskiy, Gregory A. Voth *

Department of Chemistry, Institute for Biophysical Dynamics, James Franck Institute, and Computation Institute, University of Chicago, Chicago, IL 60637, USA

ARTICLE INFO

Article history:
Available online 29 January 2013

Keywords:
Coarse-grained models
Molecular dynamics
Multiscale modeling
Free energy
Energy landscape
Force-matching
Elastic network models
Potential of mean force (PMF)
Lysozyme

ABSTRACT

To simulate molecular processes on biologically relevant length- and timescales, coarse-grained (CG) models of biomolecular systems with tens to even hundreds of residues per CG site are required. One possible way to build such models is explored in this article: an elastic network model (ENM) is employed to define the CG variables. Free energy surfaces are approximated by Taylor series, with the coefficients found by force-matching. CG potentials are shown to undergo renormalization due to roughness of the energy landscape and smoothing of it under coarse-graining. In the case study of hen egg-white lysozyme, the entropy factor is shown to be of critical importance for maintaining the native structure, and a relationship between the proposed ENM-mode-based CG models and traditional CG-bead-based models is discussed. The proposed approach uncovers the renormalizable character of CG models and offers new opportunities for automated and computationally efficient studies of complex free energy surfaces.

1. Introduction

Computational modeling of large-scale slow motions in biomolecules and biomolecular complexes is an important component of understanding processes occurring on the subcellular level [1,2]. The existing approaches to this problem may be classified into the following three very broad categories. Brute force MD simulations, which rely on stepwise integration of the dynamic equations, have been performed for polypeptides and small proteins on the timescales of up to milliseconds [3,4]. However, for typical size proteins, let alone large biomolecular complexes, the affordable lengths of MD trajectories are much shorter and do not reach the timescale of many biologically relevant processes. Elastic network models (ENMs) [5–7] have proven to be qualitatively or even semi-quantitatively correct [8–10]. The most important advantage of ENMs is very low computational cost; however, systematic improvement of the quality of ENMs requires running expensive atomistic simulations and may be difficult to achieve [11–13]. Enhanced sampling techniques (including umbrella sampling, metadynamics, temperature-accelerated molecular dynamics and many others) [14–19] are widely used to bias large-scale fluctuations in the system under study and accelerate sampling of rare events, while at the same time keeping the resolution of the model at the atomistic level. One of the key measures of success in carrying out enhanced sampling of large molecules (for example, proteins) is a reasonable choice of collective variables, which requires expert knowledge of the studied system and complicates automation of computations.

Coarse-grained (CG) models play versatile roles in studying large-scale motions in biomolecules [2,20–22]. First, by reducing system resolution, CG-ing facilitates understanding of atomistic models or analysis of MD trajectories generated with the use of the above-mentioned methods. Second, CG models can help define collective variables to be used for enhanced sampling. Third, numerical simulations at the CG level can be run using the derived CG models, when expensive all-atom simulations are not possible.

From the pragmatic viewpoint, it is desirable to use CG models as coarse as permissible by the nature of the phenomenon under investigation. In most CG models of biomolecular systems in the literature, one CG site encompasses no more than one amino acid or nucleotide residue. CG models with this resolution level are rather well studied, and general systematic methodologies for building them have been well developed [23–31]. In contrast, CG models with dozens or hundreds of residues per CG variable, which are discussed in this work, are much less studied [32–38].

Aggressive (highly) CG models of large biomolecules encounter a number of specific difficulties. Physically meaningful transformation of atomistic to CG variables and, in the case of enhanced sampling all-atom simulations, dynamic coupling between these two sets of variables have to be ensured, in spite of a significant difference between the number of the atomistic and CG variables. Smoothing of the energy landscape has to be manifest, in addition to a reduction of the number of essential variables, so that the number of fitted parameters in the CG model remains small. The lengths of the MD trajectories used to parameterize CG models of proteins and biomolecular complexes will need to be much greater

* Corresponding author. Tel.: +1 773 702 9092; fax: +1 773 795 9106.
E-mail address: gavoth@uchicago.edu (G.A. Voth).

0301-0104/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemphys.2013.01.024

than, for example, in the cases of CG models of liquids or lipids, in order to sufficiently sample large-scale slow motions.

The problem of estimating CG potentials from finite-length MD simulations deserves separate consideration. Boltzmann inversion [23,39,40], a popular method of building CG potentials, has a systematic bias towards overestimation of CG interactions. Consider the simplest case of a particle undergoing Brownian motion in zero external potential. The mean square deviation of the particle from its initial position is finite for finite sampling times and grows linearly in time. The apparent potential obtained by Boltzmann inversion in this case is nonzero, though vanishing in the limit of infinitely long MD trajectories. However, one would prefer to have a method of finding CG potentials free of this bias even for finite simulation lengths. This problem, among others, motivates development and application of methods alternative to Boltzmann inversion, such as force-matching.

The rest of the article is organized as follows. First, an outline of the CG methodology is given, together with the discussion of specific problems arising in building highly coarse-grained, low resolution CG models of large biomolecular systems and possible solutions to these problems. Then, the numerical results on HEW lysozyme are presented. It is shown that the CG interaction potential undergoes renormalization, and an explanation for this effect based on smoothing the rough energy landscape in CG models is offered. The fluctuation modes of the protein turn out to be scale dependent. The entropic contribution to the CG free energy is shown to be of primary importance for maintaining the native folded state of lysozyme. The relationship between proposed models based on ENM modes and traditional CG models with spatially localized CG beads is then discussed. The paper ends with concluding remarks.

2. Theory and methods

2.1. Outline of the general CG methodology

A systematic approach to coarse-graining is presented by the multiscale coarse-graining method (MS-CG) [24,25,41–46]. This method is based on the idea that a CG model should yield the equilibrium distribution of CG coordinates identical to the equilibrium distribution obtained from the atomistic model:

$$p_R(\boldsymbol{R}^N) = \int d\boldsymbol{r}^n p_r(\boldsymbol{r}^n)\delta(M_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N) \propto \exp(-U_{CG}(\boldsymbol{R}^N)/k_BT), \qquad (1)$$

where $\boldsymbol{r}^n$ are the atomistic coordinates, $\boldsymbol{R}^N$ are the CG coordinates, p_r and p_R are the equilibrium distribution probability densities of dynamical states in the all-atom and CG configuration spaces, respectively, M_R^N is the mapping operator from the $3n$-dimensional space of atomistic coordinates into the $3N$-dimensional space of CG coordinates, and $U_{CG}(\boldsymbol{R}^N)$ is the CG potential [for details, see Ref. [25]; $U_{CG}(\boldsymbol{R}^N)$ can also be considered as the multi-dimensional potential of mean force (PMF)].

The CG potential that satisfies the Boltzmann relation in Eq. (1) is not the average of the atomistic potential $u(\boldsymbol{r}^n)$ for those atomistic configurations that map to that CG configuration. This can be shown by considering the average potential energy defined by:

$$\langle U(\boldsymbol{R}^N)\rangle = \frac{\int d\boldsymbol{r}^n u(\boldsymbol{r}^n)p_r(\boldsymbol{r}^n)\delta(M_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N)}{\int d\boldsymbol{r}^n p_r(\boldsymbol{r}^n)\delta(M_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N)} \neq U_{CG}(\boldsymbol{R}^N) \qquad (2)$$

Instead, it can be demonstrated by direct calculation that the following equation holds:

$$\langle U(\boldsymbol{R}^N)\rangle = U_{CG}(\boldsymbol{R}^N) - T\frac{\partial}{\partial T}U_{CG}(\boldsymbol{R}^N) \qquad (3)$$

which is similar to the well-known thermodynamic relationship between the free energy A and the internal energy E, namely $E = A - T\cdot\partial A/\partial T$. For this reason, the multidimensional surfaces defined by $U_{CG}(\boldsymbol{R}^N)$ and $\langle U(\boldsymbol{R}^N)\rangle$ are further interpreted as the free energy landscape and the (potential) energy landscape, respectively. The difference between $U_{CG}(\boldsymbol{R}^N)$ and $\langle U(R^N)\rangle$, or the entropic contribution to the free energy [46], comes from the "fast" degrees of freedom integrated out as a result of coarse-graining.

It has been shown [25] that $U_{CG}(\boldsymbol{R}^N)$ can be found by the force-matching procedure, specifically by variational minimization of the residual

$$\chi^2[\boldsymbol{F}] = \frac{1}{3N}\int d\boldsymbol{r}^n p_r(\boldsymbol{r}^n)\sum_{I=1}^{N}|\mathcal{F}_I(\boldsymbol{r}^n) - \boldsymbol{F}_I(M_{\boldsymbol{R}}^N(\boldsymbol{r}^n))|^2 \qquad (4)$$

where N is the number of CG sites, $\mathcal{F}_I$ is the atomistic force acting on the I-th CG site, and $\boldsymbol{F}_I$ is the CG force field given by $\boldsymbol{F}_I(\boldsymbol{R}^N) = -\partial U_{CG}(\boldsymbol{R}^N)/\partial \boldsymbol{R}_I$. According to the ergodic hypothesis, averaging over an equilibrium ensemble in Eq. (4) can be replaced by averaging over time, which provides a practical way to find the CG potential from atomistic MD trajectories by force-matching [24,41].

The MS-CG method has been successfully applied to water [44,47], liquid methanol [41,46,47], lipid bilayers [24,46], ionic liquids [41], monosaccharides [48], peptides, proteins [27,49] and other systems. However, highly low resolution CG models of biomolecules have not been built in this way thus far.

2.2. Highly coarse-grained models of proteins

CG-ing of large biomolecules faces a number of challenges that do not arise in MS-CG-ing of smaller size molecules at a higher CG resolution. First, in low resolution CG models there is a significant gap between the number of CG and atomistic variables (~1–2 orders of magnitude), which may make a straightforward transition from an atomistic to a CG model complicated. Indeed, the thermal fluctuations in proteins at the level of amino acid residues may demonstrate a scale-free character (in the sense defined in Ref. [37]), which is not captured by rigid-body domain models. If nevertheless some amino acid residues are assigned solely to one CG domain, some other residues solely to another CG domain, and sharp interdomain borders separate some successive amino acid residues [34], the resulting CG model may be nonoptimal for certain purposes. For example, if such a CG model is used to enhance sampling of rare events, then the application of external forces to the centers of mass of different CG beads may create unphysical stress between neighboring amino acids if these amino acids are assigned to different CG domains, but will have no effect if the residues belong to the same CG domain. By contrast, in the case of finer grained MS-CG models, such as earlier developed models of lipids [24,46], no problems due to the relaxation of external forces throughout the molecule arise, since atoms within each CG domain are separated by a small number (often, only one) covalent bond.

Second, the spirit of coarse-graining implies not only reducing the number of essential variables, but also smoothing the energy as a function of the remaining variables. There is no sense in keeping narrow local maxima and minima on CG energy surfaces if they have characteristic heights $\leqslant k_BT$ and if their effect is negligible in comparison to the errors introduced by projecting out the "fast" degrees of freedom. In general, this may refer to a CG model of any molecule, but in the case of biomolecules, the energy landscapes are well-known to be very rugged [50–52] and smoothing them is of particular importance.

Third, the question remains open of whether large-scale slow motions are sufficiently sampled in MD trajectories on a timescale of tens to hundreds of nanoseconds, even if these motions are not associated with transitions between different metastable