Domain Decomposition-Based Structural Condensation of Large Protein Structures for Understanding Their Conformational Dynamics

JAE IN KIM, SUNGSOO NA, KILHO EOM
Department of Mechanical Engineering, Korea University, Seoul 136-701, Republic of Korea

Received 28 October 2009; Revised 9 April 2010; Accepted 28 May 2010
DOI 10.1002/jcc.21613
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: Normal mode analysis (NMA) with coarse-grained model, such as elastic network model (ENM), has allowed the quantitative understanding of protein dynamics. As the protein size is increased, there emerges the expensive computational process to find the dynamically important low-frequency normal modes due to diagonalization of massive Hessian matrix. In this study, we have provided the domain decomposition-based structural condensation method that enables the efficient computations on low-frequency motions. Specifically, our coarse-graining method is established by coupling between model condensation (MC; Eom et al., J Comput Chem 2007, **28**, 1400) and component mode synthesis (Kim et al., J Chem Theor Comput 2009, **5**, 1931). A protein structure is first decomposed into substructural units, and then each substructural unit is coarse-grained by MC. Once the NMA is implemented to coarse-grained substructural units, normal modes and natural frequencies for each coarse-grained substructural unit are assembled by using geometric constraints to provide the normal modes and natural frequencies for whole protein structure. It is shown that our coarse-graining method enhances the computational efficiency for analysis of large protein complexes. It is clearly suggested that our coarse-graining method provides the B-factors of 100 large proteins, quantitatively comparable with those obtained from original NMA, with computational efficiency. Moreover, the collective behaviors and/or the correlated motions for model proteins are well delineated by our suggested coarse-grained models, quantitatively comparable with those computed from original NMA. It is implied that our coarse-grained method enables the computationally efficient studies on conformational dynamics of large protein complex.

© 2010 Wiley Periodicals, Inc. J Comput Chem 00: 000–000, 2010

Key words: normal mode analysis (NMA); coarse graining; large protein dynamics; large protein complex; conformational dynamics; low-frequency normal modes

Introduction

Normal mode analysis (NMA) has been one of important simulation tool kit, which allows the quantitative understanding of protein dynamics such as conformational transitions.[1–5] The key feature of NMA in analysis of protein dynamics is to compute the low-frequency normal modes and the natural frequencies, which are theoretically related to conformational fluctuation described by Debye–Waller factor renowned as B-factors through equilibrium statistical mechanics theory.[6,7] This indicates that, for computationally efficient analysis of protein dynamics, it is a key to find fast the low-frequency normal modes and/or the natural frequencies. The most time-consuming process in NMA is the diagonalization of Hessian (stiffness) matrix based on eigenvalue problem.

In classical NMA, there is other time-consuming process to find the global equilibrium conformation for a given protein structure.[1] Energy minimization process to find the equilibrium conformation becomes computationally expensive process as the protein size is increased. Recently, due to the pioneering work by Tirion,[8] Elastic network model (ENM)[8–12] has allowed one to avoid the energy minimization process, as ENM assumes the harmonic potential near the native conformation. Specifically,

Additional supporting information may be viewed in the online version of this article.

Correspondence to: S. Na; e-mail: nass@korea.ac.kr or K. Eom; e-mail: kilhoeom@korea.ac.kr

Contract/grant sponsor: National Research Foundation of Korea (NRF); contract/grant number: 2009-0071246

Contract/grant sponsor: Korea Ministry of Education, Science, and Technology (Basic Science Research Program); contract/grant number: 2010-0001642 and KRF-2008-314-D00012

© 2010 Wiley Periodicals, Inc.

ENM regards the protein structure as a harmonic spring network in such a way that residues within the neighborhood (cut-off distance) are connected by harmonic springs with a single force constant parameter. In the similar spirit, Bahar and coworkers have introduced the Gaussian network model (GNM),[9,10] which is one-dimensional version of ENM, and/or Anisotropic network model (ANM)[11] spiritually identical to ENM. It is remarkably shown that conformational fluctuation behavior of proteins is well depicted by ENM, quantitatively comparable with experimental data.[8–11] This is attributed to the role of native topology on the protein dynamics, that is, the conformational dynamics is governed by native topology but not the shape of potential energy.[3,5] This conjecture was validated by Teeter and Case,[13] who showed that low-frequency normal modes of a protein are insensitive to details of potential fields. Further, Ma and coworkers[14] have showed that perturbation of Hessian matrix for a protein structure does not induce the significant variations in low-frequency normal modes, as long as native topology of a protein structure is conserved during the perturbation. Moreover, it was importantly reported that low-frequency normal modes obtained from ENM are highly correlated with conformational changes.[15–17] Onuchic and coworkers[18,19] have used the low-frequency normal modes computed from ENM to find the conformational transition pathways in such a way that protein structure is perturbed along the low-frequency normal modes with a constraint. In the similar manner, Bahar and coworkers[20–23] have reported that conformational changes of proteins can be acquired from the perturbation of protein structure along the dominant low-frequency normal modes. Karplus and coworkers[24] have studied the conformational changes of myosin, in the similar spirit, by using low-frequency normal modes along which protein structure is deformed. Kidera and coworkers[25] have described the conformational changes of proteins using perturbation theory using low-frequency normal mode computed from ENM.

Despite theoretical simplicity and robustness in the prediction of conformational dynamics, ENM may be computationally unfavorable for large protein complex because of computationally expensive process to diagonalize the Hessian matrix. One of the pioneering works to attempt the computational improvement in analysis of conformational dynamics is suggested by Doruker and coworkers.[26–28] In their studies,[26–28] ENM structure for protein is empirically reduced in such a way that protein structure is described by nodal points (fewer than total number of residues), within the neighborhood, that are connected by harmonic springs. It is remarkably shown that coarse-grained structure (i.e., more coarsened than ENM) enables the computationally efficient description on conformational fluctuation of protein. In the similar spirit, Eom et al.[29–31] have suggested the model condensation (MC) that allows the coarse graining of ENM by using low-rank matrix approximation inspired from the skeletonization provided by Rokhlin and coworkers.[32] Further, they have also developed the multiscale ENM[33] such that a region near functional sites is described by refined elastic network, whereas the remaining region is depicted by coarse-grained elastic network. Recently, Doruker and coworkers[27,34] have introduced the multiresolution elastic network such that a region near functional sites is described by high-resolution (e.g., atomic resolution) elastic network, whereas the remaining region is delineated by classical ENM (or even coarsened ENM). It is suggested that dynamic motion of functional domains (e.g., loop) can be well described with computational efficiency. It is also shown that conformational changes of protein are computationally efficiently depicted by multiresolution ENM. Further, Bahar and Chennubhotla[35] have introduced the coarse-graining method applicable to protein structure by using Markov statistical method. They have also studied the allosteric signal transduction related to conformational transitions by using the Markov method.[36] In addition, Lu and Ma[37] have recently reported the minimalist network model (MNM) that is able to construct elastic network structure from details of atomic structure. MNM enables one to establish not only the residue-level ENM from atomistic structural model but also the coarsened ENM.

In recent years, Ma and coworkers[38] have used the substructure synthesis method (SSM) for computationally efficient analysis of conformational dynamics. In a similar spirit, Eom and coworkers[39,40] have suggested the hierarchical component mode synthesis (hCMS) that allows the computationally effective studies on conformational dynamics. The key idea of SSM and/or hCMS is to decompose the protein structure into several substructural units for which NMA is implemented.[41] Once the normal modes and the frequencies for each substructural unit are obtained, such modes and frequencies are assembled based on geometric constraint to find the low-frequency normal modes and their corresponding frequencies for whole protein structure. It is shown that the domain decomposition methods such as SSM and/or hCMS improve the computational efficiency to estimate the B-factors of proteins, quantitatively comparable with those computed from original NMA.

In this study, we have suggested the novel coarse-graining method that combines the MC method[29–31] and domain decomposition method such as hCMS.[39,40] Specifically, a protein structure is decomposed into several substructural units, and then each substructural unit is coarse grained by MC method. NMA is implemented in each coarse-grained substructural unit, and then normal modes and their corresponding frequencies for each substructural unit are assembled based on geometric constraints to find the low-frequency normal modes and their corresponding frequencies for a whole protein structure. For robustness of our method, we have considered $\sim$100 protein structures composed of $\sim 10^4$ residues as model systems. It is remarkably shown that our coarse-graining method enhances the computational efficiency to estimate the conformational fluctuation (e.g., B-factors), quantitatively comparable with those computed from original NMA. Specifically, the computing time to evaluate B-factors of model proteins is enormously reduced (when compared with original NMA or even coarse-grained ENM by MC), whereas the correlation between B-factors obtained from experiment and our coarse-grained method is almost identical to that between experiments and original NMA (with ENM). This validates the robustness of our coarse-graining method that is capable of computationally efficient analysis of large protein dynamics. It is implied that our coarse-graining method may be applicable to biological macromolecules and/or large chemical system for quantitative description of dynamic behavior of such structures.

Theory and Model

Elastic Network Model

ENM assumes that protein structure can be represented by harmonic spring network such that the residues within the neighborhood are connected by harmonic springs with identical force constant.[5,8–11] Despite its simplicity, ENM is very robust in predicting conformational fluctuation (quantitatively comparable with that obtained from experiment), because ENM is sufficient to describe the protein native topology that plays a vital role on protein dynamics.[5,13,14] The potential field, V, for ENM is given by[5,8–11]

$$V = \frac{\gamma}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left[\left|\mathbf{r}_i - \mathbf{r}_j\right| - \left|\mathbf{r}_i^0 - \mathbf{r}_j^0\right|\right]^2 \cdot H\left(r_c - \left|\mathbf{r}_i^0 - \mathbf{r}_j^0\right|\right) \quad (1)$$

where γ is a force constant for a harmonic spring, N is the total number of residues, $\mathbf{r}_i$ is the coordinates of ith residue (α carbon atom), r_c is the cut-off distance that defines the protein topology, $H(x)$ is the Heaviside unit step function defined as $H(x) = 0$ if $x < 0$; otherwise $H(x) = 1$, and superscript 0 indicates the equilibrium conformational state. Here, the cut-off distance is usually in the range of 7–12 Å, which is suitable to describe the protein topology.[11]

To find the conformational fluctuation, NMA has to be implemented. For implementation of NMA, the stiffness (Hessian) matrix has to be computed from a potential field given by Eq. (1). The stiffness matrix $\mathbf{K}$ for ENM is composed of 3 × 3 block matrices $\mathbf{K}_{ij}$ given as[11,40,42]

$$\mathbf{K}_{ij} = -\left[\gamma H\left(r_\mathrm{c} - \left|\mathbf{r}_i^0 - \mathbf{r}_j^0\right|\right)\frac{\left(\mathbf{r}_i^0 - \mathbf{r}_j^0\right)^{\dagger}\left(\mathbf{r}_i^0 - \mathbf{r}_j^0\right)}{\left|\mathbf{r}_i^0 - \mathbf{r}_j^0\right|^2}\right] \times \left(1 - \delta_{ij}\right) - \delta_{ij}\sum_{r\neq i}^{N}\mathbf{K}_{ir} \quad (2)$$

Here, δ_{ij} is the Kronecker delta defined as $\delta_{ij} = 1$ if $i = j$; otherwise $\delta_{ij} = 0$, and a symbol † represents the transpose of a column vector.

Harmonic approximation[43,44] provides the equation of motion for protein dynamics such as $\mathbf{M}(d^2\mathbf{R}/dt^2) + \mathbf{KR} = 0$, where $\mathbf{M}$ is the mass matrix for alpha carbon atoms (i.e., diagonal matrix), $\mathbf{R}$ is a column vector representing the atomic coordinates of alpha carbon atoms, i.e., $\mathbf{R}^{\dagger} = [\mathbf{r}_1^{\dagger}, \mathbf{r}_2^{\dagger}, \ldots \mathbf{r}_N^{\dagger}]$, and $\mathbf{K}$ is the stiffness matrix composed of block matrices $\mathbf{K}_{ij}$ given by Eq. (2). For vibration of protein structure, the atomic coordinates $\mathbf{R}$ can be assumed as $\mathbf{R} = \mathbf{u}\exp[i\omega t]$,[45] where ω is the natural frequency and $\mathbf{u}$ is its corresponding normal mode. With assumption of $\mathbf{M} = m\mathbf{I}$, where m is the molecular weight of alpha carbon atom and $\mathbf{I}$ is the $3N \times 3N$ diagonal matrix, the equation of motion becomes the eigenvalue problem: $\mathbf{Ku} = \omega^2\, m\mathbf{u} \equiv \lambda\mathbf{u}$, where λ is the eigenvalue.

Equilibrium statistical mechanics theory[6,7] enables the computation of fluctuation matrix $\mathbf{Q}$ based on the eigenvalues and their corresponding normal modes for stiffness matrix $\mathbf{K}$ such as

$$\mathbf{Q} = \left\langle(\mathbf{R} - \mathbf{R}_0)^{\dagger}(\mathbf{R} - \mathbf{R}_0)\right\rangle = \sum_{k=7}^{3N}\frac{k_B T}{\lambda_k}\mathbf{u}_k^{\dagger}\mathbf{u}_k \quad (3)$$

where $\langle\mathbf{R}\rangle$ indicates the ensemble average (time average) of quantity $\mathbf{R}$, and $\mathbf{R}_0$ is the atomic coordinates of alpha carbon atoms at equilibrium conformation, i.e., $\mathbf{R}_0 = <\mathbf{R}>$. Herein, it should be noticed that six zero-normal modes corresponding to rigid body motions are excluded for computing the fluctuation matrix. The mean square fluctuation for ith alpha carbon atom is then given as

$$\langle|\mathbf{r}_i - \mathbf{r}_i^0|^2\rangle = \mathbf{Q}_{3(i-1)+1,\,3(i-1)+1} + \mathbf{Q}_{3(i-1)+2,\,3(i-1)+2} + \mathbf{Q}_{3(i-1)+3,\,3(i-1)+3}.$$

The B-factor for ith alpha carbon atom can be easily computed from a relation of $B_i = (8\pi^2/3)\,\langle|\mathbf{r}_i - \mathbf{r}_i^0|^2\rangle$.

Domain Decomposition-Based Structural Condensation

For computationally efficient analysis of protein dynamics, we have previously suggested the coarse-graining methods such as MC[29–31] and/or hCMS.[39,40] The key feature of MC is to reduce the degrees of freedom, and consequently, decrease the size of stiffness matrix. On the other hand, the basic idea of hCMS is to decompose the protein structure into several substructural units, and then NMA is implemented in each substructural unit rather than whole structure. In this study, we have developed the novel coarse-graining scheme, which enhances the computational efficiency when compared with previous coarse-graining scheme such as MC, by coupling two features of MC and hCMS.

Figure 1 shows the schematic illustration of our coarse-graining scheme, that is, domain decomposition-based structural condensation. First, a protein structure is decomposed into several substructural units (e.g., two substructural units in Fig. 1). Subsequently, we have used the MC that allows the reduction of degrees of freedom for each substructural unit. NMA is then implemented to coarse-grained substructural units. Finally, natural frequencies and normal modes for each substructural unit are assembled by using geometric constraint. Each process for our suggested coarse-graining scheme is summarized as below.

For straightforward demonstration, we decompose the protein structure into two substructural units. Here, it should be kept in mind that substructural unit should have the sufficient degrees of freedom such that degrees of freedom for substructural unit are much larger than the degrees of freedom for interface between two substructural units. If the degrees of freedom for substructural unit are comparable with the degrees of freedom for interface (i.e., under constraints), the dynamic motion of such a substructural unit would be constrained, which leads to inability to describe the motion of substructural unit.[40] Herein, constraints for residues (colored as green in Fig. 1) belonging to interface between two substructural units are that displacement field for such residues is continuous.

Now, for convenience, the constraint for interface between substructural units is not considered at this moment. Then, the

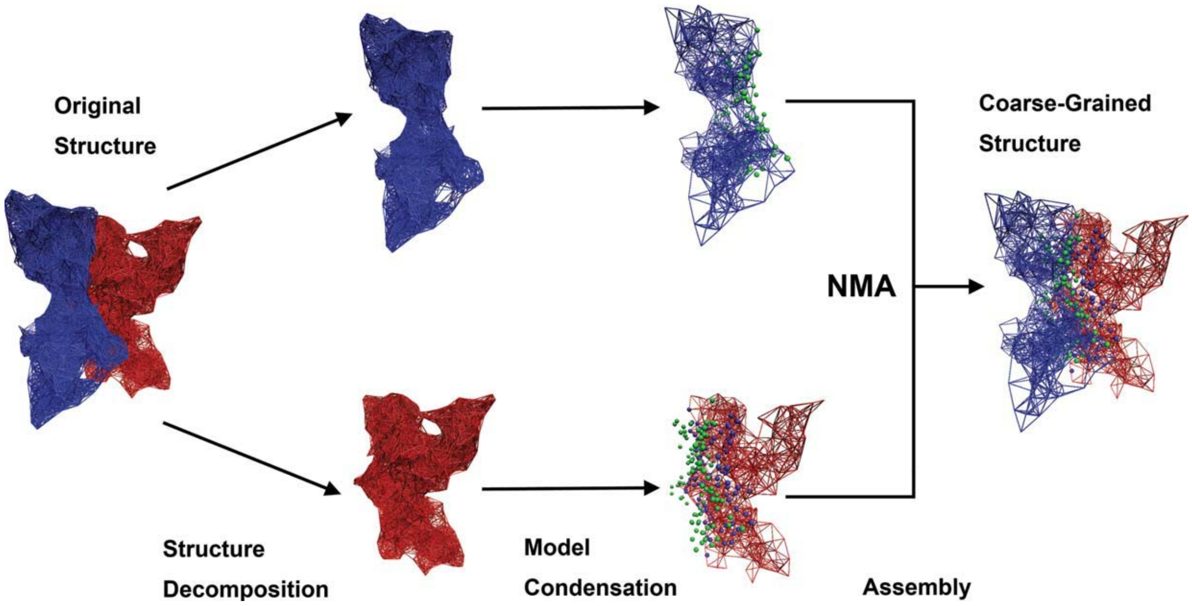

Figure 1. Schematic illustration of domain decomposition-based structural condensation method is shown. First, a protein structure is decomposed into two substructural units, and then each substructural unit is coarse grained by using MC method. Subsequently, NMA is implemented to each coarse-grained substructural unit. Then, by using geometric constraints for interface between two substructural units, the normal modes and their corresponding natural frequencies for coarse-grained substructural units are assembled. Here, green-colored residues belong to the interface between two substructural units. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

potential energy for a system composed of two substructural units without application of constraints is given by

$$V' = \frac{1}{2}\left(\mathbf{v}_{\mathrm{A}}^{\dagger}\mathbf{K}_{\mathrm{A}}\mathbf{v}_{\mathrm{A}} + \mathbf{v}_{\mathrm{B}}^{\dagger}\mathbf{K}_{\mathrm{B}}\mathbf{v}_{\mathrm{B}}\right) \tag{4}$$

where $\mathbf{K}_i$ and $\mathbf{v}_i$ represent the stiffness matrix and the displacement field for *i*th substructural unit (i.e., i = A or B), and prime indicates that constraints imposed to interface are not applied at this moment. In the similar manner, the kinetic energy for a system is represented in the form of

$$T' = \frac{1}{2}\left(\dot{\mathbf{v}}_{\mathrm{A}}^{\dagger}\mathbf{M}_{\mathrm{A}}\dot{\mathbf{v}}_{\mathrm{A}} + \dot{\mathbf{v}}_{\mathrm{B}}^{\dagger}\mathbf{M}_{\mathrm{B}}\dot{\mathbf{v}}_{\mathrm{B}}\right) \tag{5}$$

Here, a symbol dot indicates the time derivative, and $\mathbf{M}_i$ is the mass matrix for *i*th substructural unit.

Then, we use the MC that enables the coarse-graining of each substructural unit. We partition the residues for a substructure into two sets of residues, one of which is maintained during MC while the rest is eliminated during MC. The potential energy for a substructural unit A can be represented in the form of

$$V'_{\mathrm{A}} = \frac{1}{2}\mathbf{v}_{\mathrm{A}}^{\dagger}\mathbf{K}_{\mathrm{A}}\mathbf{v}_{\mathrm{A}} = \frac{1}{2}\left[\left(\mathbf{v}_{\mathrm{A}}^{\alpha}\right)^{\dagger}\ \left(\mathbf{v}_{\mathrm{A}}^{\beta}\right)^{\dagger}\right]\begin{bmatrix}\mathbf{K}_{\mathrm{A}}^{\alpha\alpha} & \mathbf{K}_{\mathrm{A}}^{\alpha\beta}\\ \mathbf{K}_{\mathrm{A}}^{\beta\alpha} & \mathbf{K}_{\mathrm{A}}^{\beta\beta}\end{bmatrix}\begin{bmatrix}\mathbf{v}_{\mathrm{A}}^{\alpha}\\ \mathbf{v}_{\mathrm{A}}^{\beta}\end{bmatrix} \tag{6}$$

where Greek symbols α and β indicate the set of master residues (that are maintained during MC) and the set of slave residues (that are supposed to be eliminated during MC), respectively. Based on the MC suggested in ref. 29, the effective stiffness matrix for coarse-grained substructural unit A is given by

$$\bar{\mathbf{K}}_{\mathrm{A}} = \mathbf{K}_{\mathrm{A}}^{\alpha\alpha} - \mathbf{K}_{\mathrm{A}}^{\alpha\beta}\left(\mathbf{K}_{\mathrm{A}}^{\beta\beta}\right)^{-1}\mathbf{K}_{\mathrm{A}}^{\beta\alpha} \tag{7}$$

In the similar manner, it is straightforward to find the effective stiffness matrix for coarse-grained substructural unit B. In addition, the effective mass matrix for coarse-grained substructural unit A can be represented in the form of $\overline{\mathbf{M}}_{\mathrm{A}} = m\mathbf{I}_{\mathrm{A}}^{\alpha}$, where m is the molecular weight of alpha carbon atom, and $\mathbf{I}_{\mathrm{A}}^{\alpha}$ indicates the $(N_{\mathrm{A}}/n) \times (N_{\mathrm{A}}/n)$ identity matrix with N_{A} being the total number of alpha carbon atoms for substructural unit A (before coarse graining) and n being the degrees of MC (e.g., n = 2, 3, ….). As a consequence, the potential energy for a system composed of coarse-grained substructural units is given as

$$V' = \frac{1}{2}\left[\left(\mathbf{v}_{\mathrm{A}}^{\alpha}\right)^{\dagger}\bar{\mathbf{K}}_{\mathrm{A}}\mathbf{v}_{\mathrm{A}}^{\alpha} + \left(\mathbf{v}_{\mathrm{B}}^{\alpha}\right)^{\dagger}\bar{\mathbf{K}}_{\mathrm{B}}\mathbf{v}_{\mathrm{B}}^{\alpha}\right] \tag{8}$$

Here, $\mathbf{v}_{\mathrm{A}}^{\alpha}$ represents the displacement field for master residues in substructural unit A. Then, the displacement field $\mathbf{v}_{\mathrm{A}}^{\alpha}(\mathbf{x}_{\mathrm{A}}^{\alpha},\ t)$ where $\mathbf{x}_{\mathrm{A}}^{\alpha}$ is the coordinates of master residues of substructural unit A, is assumed to be in the form of $\mathbf{v}_{\mathrm{A}}^{\alpha}(\mathbf{x}_{\mathrm{A}}^{\alpha},\ t) = \Phi_{\mathrm{A}}(\mathbf{x}_{\mathrm{A}}^{\alpha}) \cdot \mathbf{z}_{\mathrm{A}}^{\alpha}(t)$, where Φ_{A} is the $(N_{\mathrm{A}}/n) \times (N_{\mathrm{A}}/n)$ matrix whose column vector indicates the normal mode of effective stiffness matrix $\overline{\mathbf{K}}_{\mathrm{A}}$, i.e., $\overline{\mathbf{K}}_{\mathrm{A}}\Phi_{\mathrm{A}} = \Phi_{\mathrm{A}}\ \Lambda_{\mathrm{A}}$. Here, Λ_{A} is the diagonal matrix whose component is the eigenvalue of effective stiffness matrix $\overline{\mathbf{K}}_{\mathrm{A}}$. With linear transformation, the potential energy and the kinetic energy without geometric constraints are given by

$$V' = \frac{1}{2}\left[\left(\mathbf{z}_{\mathrm{A}}^{\alpha}\right)^{\dagger} \left(\mathbf{z}_{\mathrm{B}}^{\alpha}\right)^{\dagger} \right] \begin{bmatrix} \Lambda_{\mathrm{A}} & \mathbf{0} \\ \mathbf{0} & \Lambda_{\mathrm{B}} \end{bmatrix} \begin{bmatrix} \mathbf{z}_{\mathrm{A}}^{\alpha} \\ \mathbf{z}_{\mathrm{B}}^{\alpha} \end{bmatrix} \equiv \frac{1}{2}\mathbf{z}^{\dagger}\Lambda\mathbf{z} \tag{9a}$$

$$T' = \frac{1}{2}\left[\left(\dot{\mathbf{z}}_{\mathrm{A}}^{\alpha}\right)^{\dagger} \left(\dot{\mathbf{z}}_{\mathrm{B}}^{\alpha}\right)^{\dagger} \right] \begin{bmatrix} \Phi_{\mathrm{A}}^{\dagger}\bar{\mathbf{M}}_{\mathrm{A}}\Phi_{\mathrm{A}} & \mathbf{0} \\ \mathbf{0} & \Phi_{\mathrm{B}}^{\dagger}\bar{\mathbf{M}}_{\mathrm{B}}\Phi_{\mathrm{B}} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{z}}_{\mathrm{A}}^{\alpha} \\ \dot{\mathbf{z}}_{\mathrm{B}}^{\alpha} \end{bmatrix} \equiv \frac{1}{2}\dot{\mathbf{z}}^{\dagger}\mathrm{D}\dot{\mathbf{z}} \tag{9b}$$

Here, it should be noticed that the transformed effective mass matrix for a coarse-grained system, **D**, is not a diagonal matrix.

To find the natural frequencies and their corresponding normal modes for a coarse-grained system, we have to impose the constraints to the interface, which is the assembly process. Specifically, the displacement field at interface between two coarse-grained substructural units should be continuous. Such constraints for interface can be represented in the form of $\mathbf{Pz} = \mathbf{0}$. Here, a vector **z** has the redundant degrees of freedom, since the alpha carbon atoms belonging to interface were redundantly enumerated. The constraint equation is then given as $\mathbf{Pz} \equiv \mathbf{P}_1\mathbf{s} + \mathbf{P}_2\mathbf{y} = 0$, where **s** is the independent variable and **y** is the dependent variable. The detailed procedure to obtain the matrices **P**, $\mathbf{P}_1$, and $\mathbf{P}_2$, which represent the geometric constraints, is described in Supporting Information. After imposition of constraints, the potential energy and the kinetic energy become

$$V = \frac{1}{2}\mathbf{s}^{\dagger}\left(\mathbf{B}^{\dagger}\Lambda\mathbf{B}\right)\mathbf{s} \equiv \frac{1}{2}\mathbf{s}^{\dagger}\mathbf{L}\mathbf{s} \tag{10a}$$

$$T = \frac{1}{2}\dot{\mathbf{s}}^{\dagger}\left(\mathbf{B}^{\dagger}\mathbf{D}\mathbf{B}\right)\dot{\mathbf{s}} \equiv \frac{1}{2}\dot{\mathbf{s}}^{\dagger}\mathrm{G}\dot{\mathbf{s}} \tag{10b}$$

where the matrix **B** is the constraint matrix defined as

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} \\ -\mathbf{P}_2^{-1}\,\mathbf{P}_1 \end{bmatrix} \tag{10c}$$

Here, matrices **L** and **G** indicate the stiffness matrix and the mass matrix, represented in the space spanned by normal modes of coarse-grained substructural units, respectively. The potential energy and the kinetic energy given by eqs. [10(a)] and [10(b)], respectively, are the exact form for a coarse-grained system composed of coarse-grained substructural units, as the geometric constraints are prescribed.

NMA for a coarse-grained protein structure is represented as the eigenvalue problem as follows: $\mathbf{LU} = \mathbf{GU}\Omega$, where **U** is the modal matrix and Ω is the diagonal matrix whose components are the eigenvalues of a coarse-grained protein structure. To describe the protein dynamics based on low-frequency normal modes, the modal matrix **U** has to be transformed to the matrix **W**, whose column vectors represent the normal modes, given as $\mathbf{W} = \Phi\mathbf{BU}$, where $\Phi^{\dagger} = [\Phi_{\mathrm{A}}^{\dagger}\ \Phi_{\mathrm{B}}^{\dagger}]$. Then, the conformational dynamics of coarse-grained protein structure can be easily understood from equilibrium statistical mechanics theory that provides the fluctuation matrix given by Eq. (3).

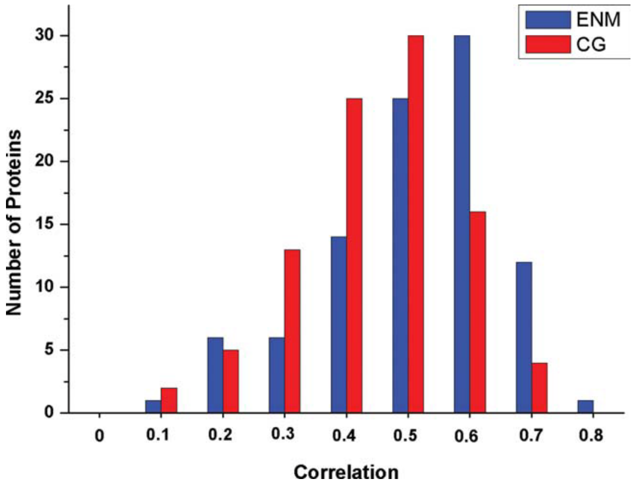

Figure 2. Histogram of correlations between B-factors (for ~100 model proteins) obtained from experiments and ENM is presented. In addition, the histogram of correlations between B-factors obtained from experiments and our proposed coarse-graining (CG) method is also shown. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Results and Discussion

Model Proteins

For validation of our coarse-graining method as well as its robustness, we have considered ~100 protein complexes that consist of $\sim 10^4$ residues. The smallest protein complex which we took into account, is the hemoglobin composed of 572 residues. The largest protein complex considered in this study is the glutamate synthase comprised of 11,568 residues. Most (>90%) of protein complexes taken into account have $>10^3$ residues. The details of model proteins are suggested in Supporting Information. Model proteins are treated with ENM that straightforwardly provides the stiffness matrix. The parameters of ENM are cut-off distance that is prescribed as ~10 Å and force constant that is determined by fitting of B-factors computed from ENM and those obtained from experiments. The force constants of model proteins are summarized in (Supporting Information) Figure S.2.

B-factors

For reliability and robustness of our coarse-graining method, we have compared the B-factors of ~100 model proteins computed from our coarse-graining method with those obtained from experiments (see also Supporting Information, Fig. S.3). For quantitative comparison, we have introduced the correlation parameter *r* defined as[46]

$$r = \frac{\sum_{i=1}^{N}\left(B_i^{\mathrm{exp}} - \langle B_i^{\mathrm{exp}}\rangle\right)\left(B_i^{\mathrm{sim}} - \langle B_i^{\mathrm{sim}}\rangle\right)}{\sqrt{\sum_{i=1}^{N}\left(B_i^{\mathrm{exp}} - \langle B_i^{\mathrm{exp}}\rangle\right)^2 \sum_{j=1}^{N}\left(B_i^{\mathrm{sim}} - \langle B_i^{\mathrm{sim}}\rangle\right)^2}} \tag{11}$$

where B_i^{exp} and B_i^{sim} represent the B-factors for *i*th residue obtained from experiments and simulations (e.g., original

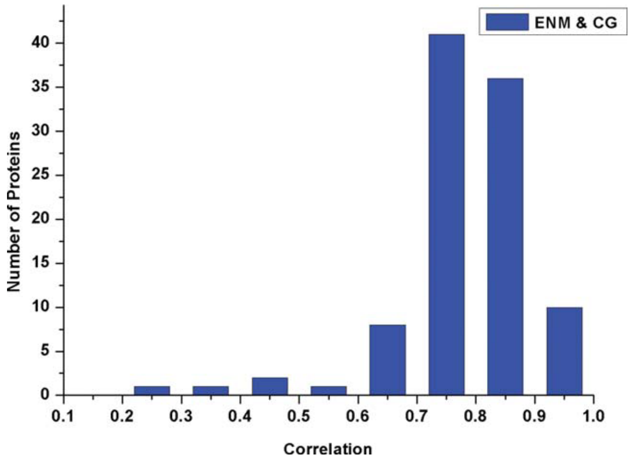

Figure 3. Histogram of correlations between B-factors computed from ENM and our proposed coarse-graining (CG) method for 100 model proteins is presented. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

NMA or our coarse-grained method), respectively. Figure 2 depicts the histogram of correlations between B-factors obtained from experiment and our coarse-graining method for ~100 large protein structures. A histogram of the correlation between B-factors obtained from experiment and original NMA for ~100 model proteins is also presented in Figure 2. As stated earlier, in this study, we have only considered the large protein complexes composed of $\sim 10^4$ residues. It is remarkably shown that the histograms of two correlations (i.e., correlation between experiment and original NMA and correlation between experiment and our coarse-graining method) are similar to each other. This indicates that the performance in prediction of B-factors by our coarse-graining method is close to that by original NMA. The number of proteins having the correlation of >0.5 is 50 (out of 95) by using our coarse-graining method, whereas the number of proteins have a correlation of >0.5 is 68 (out of 95) based on original NMA with ENM.

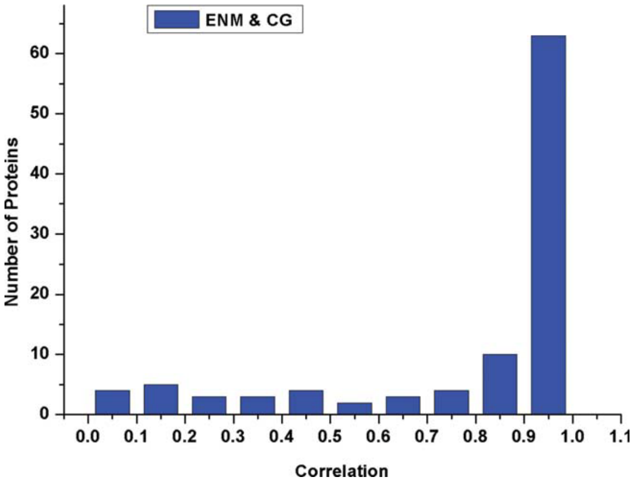

Figure 4. Histogram of correlations between lowest-frequency normal modes estimated from ENM and our proposed coarse-graining (CG) method, for 100 model proteins, is presented. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

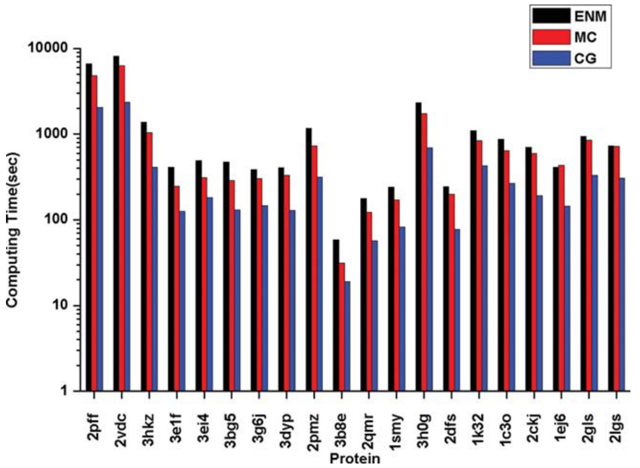

Figure 5. Computation times for estimation of B-factors using original NMA with ENM, MC, and our coarse-graining (CG) method for 20 representative model proteins, are presented. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

For further evaluation of robustness of our coarse-graining method, we have taken into account the correlation between B-factors computed from our coarse-grained method and NMA with ENM. Figure 3 shows the histogram of correlation between B-factors obtained from our coarse-graining method and NMA with ENM for 100 model proteins. It is interestingly shown that the correlation between our coarse-graining method and NMA with ENM is relatively high. Remarkably, the percentage of proteins having the correlation (between our coarse-graining method and NMA with ENM) of >0.7 is 87%. This indicates that the B-factors computed from our coarse-graining method are well correlated with those estimated from NMA with ENM.

Low-Frequency Normal Mode

For robustness of our coarse-graining scheme, based on 100 model proteins, we take into account the correlation between lowest-frequency normal modes obtained from ENM and our coarse-graining method. The lowest frequency normal modes, which are computed from NMA, MC (our previous method[29–31]), and our current coarse-graining method, respectively, for some model proteins are presented in Figure S.4. Figure 4 shows the histogram of correlations (between lowest frequency normal modes computed from ENM and our coarse-graining method) for 100 model proteins. The percentage of proteins having the correlation of >0.7 is 77%. High correlations of >0.9 (between normal modes computed from ENM and our coarse-grained model) are found in 63% of model proteins. This implies that our coarse-graining method is allowable for quantitative description

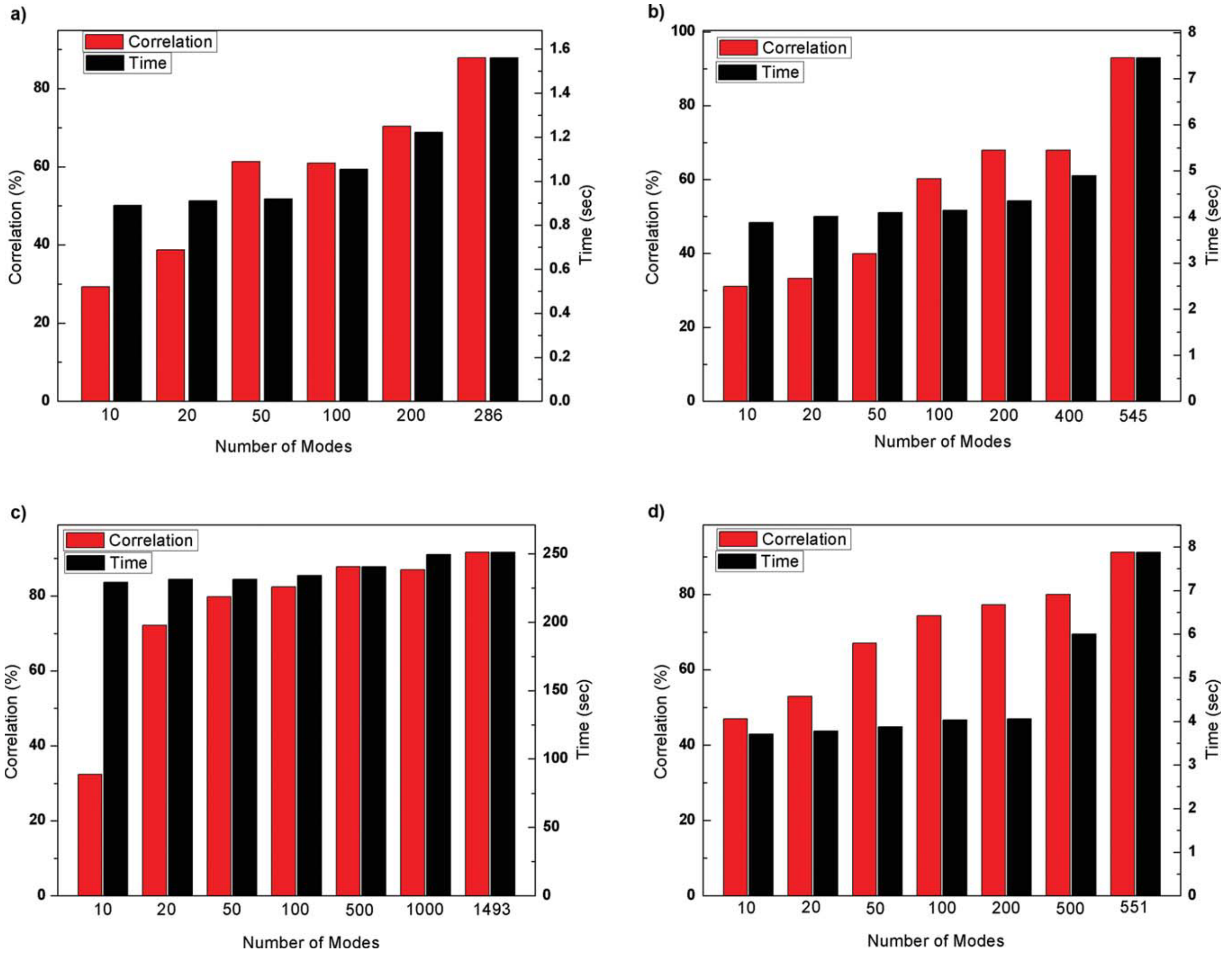

Figure 6. Correlation between B-factors computed from ENM and our coarse-graining (CG) method, using different number of normal modes of coarse-grained substructural units, is presented. In addition, the computing time to estimate B-factor by using our coarse-graining (CG) method that uses the different number of normal modes of coarse-grained substructural units is shown. (a) Hemoglobin, (b) F0-ATPase motor protein, (c) F1-ATPase motor protein, and (d) scallop myosin. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

of low-frequency normal mode relevant to conformational dynamics.

Collective behaviors found in low-frequency motions for model proteins are well described by our coarse-graining method, quantitatively similar to those predicted from NMA with ENM (see Supporting Information, Fig. S.5). For quantitative comparison, we have introduced the collectivity parameters (described in Supporting Information). It is remarkably shown that collective parameters for low-frequency modes, computed from our coarse-graining method and NMA with ENM, are also similar to each other. This indicates that our coarse-graining method is robust in computationally efficient predictions on collective dynamics based on low-frequency motion. Moreover, the correlated motion predicted from both NMA (with ENM) and our coarse-graining method is presented in Supporting Information (see Supporting Information, Fig. S.6). It is shown that correlated motion for protein domains is well depicted by our coarse-graining method, quantitatively comparable with that anticipated from NMA with ENM.

Computational Efficiency

To validate the computational efficiency of our coarse-graining method, we have measured the computing time to calculate the low-frequency normal modes for large protein complex by using our coarse-graining method, our previous MC method, and original NMA. Here, 20 model proteins are considered such that such proteins exhibit $>10^3$ residues. It is remarkably shown that the computing time to evaluate low-frequency normal modes is in the order of [our coarse-graining method] $<$ [MC method[29–31]] $<$ [original NMA] (see Fig. 5). This indicates that our coarse-graining method, which couples the key feature of MC and hCMS, enhan-

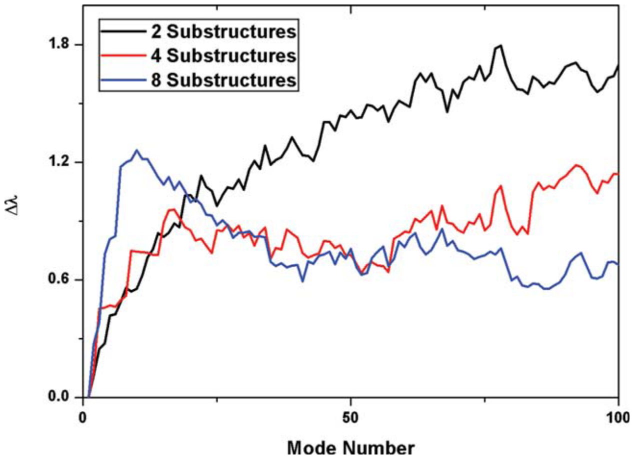

Figure 7. Differences between eigenvalues, for hemoglobin, computed from NMA with ENM and our coarse-graining methods, Herein, our coarse-grained models are established such that hemoglobin structure is decomposed into N_d substructural units, where N_d = 2, 4, or 8, and then each substructure is condensed. It is shown that, for N_d = 8, the differences in eigenvalues corresponding to low-frequency motions become significant. This indicates that coarse-graining using N_d = 8 is inappropriate to describe the low-frequency motion, related to conformational changes, of hemoglobin. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ces the computational efficiency enormously especially for large protein complex. Herein, the computing time is measured based on the workstation system with Intel Xeon 2.0 GHz Quadcore ×2 and RAM 8 GB.

Conformation Dynamics Described by Normal Modes

Our coarse-graining method allows the computationally efficient analysis of large protein dynamics by using the key features of MC and hCMS. The feature of MC is to reduce the degrees of freedom for protein structure, whereas the key feature of hCMS is to decompose the protein structure into several substructural units, and then transform the stiffness matrix in Cartesian coordinates into that represented in the space spanned by normal modes of substructural units. In our previous study,[40] it is shown that the reduced space spanned by certain number of normal modes (rather than all normal modes) of substructural units enhances the computational efficiency. In this study, our coarse-graining method may be computationally improved by using the reduced space spanned by certain number of normal modes. However, it should be also kept in mind that, if one uses the reduced space spanned by few low-frequency normal modes, the conformational dynamics of proteins cannot be well predicted by such reduced space.

We have considered the B-factors for four representative model proteins computed from ENM and our coarse-graining method by using reduced space spanned by certain number of normal modes of substructural units. It is interestingly shown that the reduced space spanned by <50 low-frequency normal modes does not enable the description of B-factors. At least >100 normal modes should be considered for reduced space into which the stiffness matrix is transformed (see Supporting Information, Fig. S.7). We have also taken into account the correlation between B-factors computed from ENM and our coarse-graining method by using reduced space spanned by certain number of normal modes with their computing times (see Fig. 6). Herein, for convenience, the model protein structure is decomposed into two substructural units. Except F_1-ATPase, the reduced space spanned by >100 normal modes (rather than using all normal modes) improves the computing time to calculate the B-factors, quantitatively comparable with those computed from ENM. As F_1-ATPase is decomposed into two substructural units, each substructural unit still exhibits the large degree of freedom which leads to the computationally expensive process to reduce the structure using MC method. The computational efficiency for F_1-ATPase will be enhanced when such a large protein is decomposed into several substructural units rather than two substructural units.

Degree of Domain Decompositions

As presented in our previous study,[40] we can decompose a protein structure into more than two substructural units in our coarse-graining process, which consists of two dominant processes of CMS and MC. However, it should be reminded that, if the structure is decomposed into too many substructural units, the dynamic behavior of a protein domain cannot be well described. This is attributed to the fact that decomposition into many substructural units leads to induce the more constraints on protein dynamics so as to degrade the dynamic characteristics such as low-frequency motion. As discussed in our previous study,[40] the appropriate number of decomposition is comparable with the number of domains in a protein structure.

For brief elucidation, we have considered the dynamic behavior of hemoglobin, described by our coarse-grained models. Here, our models are constructed such that hemoglobin structure is first decomposed into N_d substructural units (where N_d = 2, 4, 8), and then each substructural unit is condensed. Figure 7 depicts the differences between eigenvalues computed from our coarse-graining methods (using different degree of decomposition) and NMA with ENM. It is interestingly shown that it is acceptable until one decomposes the hemoglobin structure into four substructural units. As shown in Figure 7, if the hemoglobin structure is decomposed into eight substructural units, the difference in eigenvalues (corresponding to low-frequency motion) becomes larger when compared with the cases, where hemoglobin is decomposed into two or four substructural units. This indicates that it is unacceptable that hemoglobin is decomposed into more than four substructural units. In other words, it is acceptable when a protein structure is decomposed into the number of domains. This is ascribed to the fact that, if a protein structure is decomposed into more than number of domains, such decomposition will severely constrain on the low-frequency functional motion of a protein structure.

Conclusion

In this study, we have suggested the novel coarse-graining method, that is, domain decomposition-based structural condensation method by using the key features of MC method[29–31] and hCMS method.[39,40] Specifically, a large protein structure is decomposed into the several substructural units, and then such substructural units are coarse grained by using MC. Subsequently, the stiffness matrix for protein structure is transformed into that represented by the space spanned by normal modes of substructural units. It is clearly shown that domain decomposition-based structural condensation method allows the computationally efficient analysis of large protein dynamics. This proposed an approach that can enable the computationally effective calculations of normal modes relevant to conformational transition. This approach can be, thus, applicable to the prediction of conformational transition pathway based on the perturbation of structure along the low-frequency normal modes. In conclusion, our proposed approach allows for studying dynamic behavior of large chemical structures and/or macromolecular structures based on NMA.

References

1. Cui, Q.; Bahar, I. Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems; CRC Press, Boca Raton, 2005.
2. Bahar, I.; Rader, A. J. Curr Opin Struct Biol 2005, 15, 586.
3. Ma, J. P. Structure 2005, 13, 373.
4. Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. Chem Rev 2010, 110, 1463.
5. Tama, F.; Brooks, C. L. Annu Rev Biophys Biomol Struct 2006, 35, 115.
6. Weiner, J. H. Statistical Mechanics of Elasticity; Dover Publication, 1983.
7. Chandler, D. Introduction to Modern Statistical Mechanics; Oxford University Press, 1987.
8. Tirion, M. M. Phys Rev Lett 1996, 77, 1905.
9. Haliloglu, T.; Bahar, I.; Erman, B. Phys Rev Lett 1997, 79, 3090.
10. Bahar, I.; Atilgan, A. R.; Demirel, M. C.; Erman, B. Phys Rev Lett 1998, 80, 2733.
11. Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Biophys J 2001, 80, 505.
12. Eom, K.; Yoon, G.; Kim, J.-I.; Na, S. J Comput Theor Nanosci 2010, 7, 1210.
13. Teeter, M. M.; Case, D. A. J Phys Chem 1990, 94, 8091.
14. Lu, M. Y.; Ma, J. P. Biophys J 2005, 89, 2395.
15. Tama, F.; Sanejouand, Y. H. Protein Eng 2001, 14, 1.
16. Hall, B. A.; Kaye, S. L.; Pang, A.; Perera, R.; Biggin, P. C. J Am Chem Soc 2007, 129, 11394.
17. Kirillova, S.; Cortes, J.; Stefaniu, A.; Simeon, T. Proteins 2008, 70, 131.
18. Miyashita, O.; Onuchic, J. N.; Wolynes, P. G. Proc Natl Acad Sci USA 2003, 100, 12570.
19. Whitford, P. C.; Miyashita, O.; Levy, Y.; Onuchic, J. N. J Mol Biol 2007, 366, 1661.
20. Xu, C. Y.; Tobi, D.; Bahar, I. J Mol Biol 2003, 333, 153.
21. Tobi, D.; Bahar, I. Proc Natl Acad Sci USA 2005, 102, 18908.
22. Bahar, I.; Chennubhotla, C.; Tobi, D. Curr Opin Struct Biol 2007, 17, 633.
23. Isin, B.; Schulten, K.; Tajkhorshid, E.; Bahar, I. Biophys J 2008, 95, 789.
24. Cecchini, M.; Houdusse, A.; Karplus, M. PLoS Comput Biol 2008, 4, e1000129.
25. Ikeguchi, M.; Ueno, J.; Sato, M.; Kidera, A. Phys Rev Lett 2005, 94, 078102.
26. Doruker, P.; Jernigan, R. L.; Bahar, I. J Comput Chem 2002, 23, 119.
27. Kurkcuoglu, O.; Jernigan, R. L.; Doruker, P. Polymer 2004, 45, 649.
28. Kurkcuoglu, O.; Jernigan, R. L.; Doruker, P. Qsar Comb Sci 2005, 24, 443.
29. Eom, K.; Baek, S.-C.; Ahn, J.-H.; Na, S. J Comput Chem 2007, 28, 1400.
30. Eom, K.; Ahn, J. H.; Baek, S. C.; Kim, J. I.; Na, S. CMC: Comput Mater Continua 2007, 6, 35.
31. Eom, K.; Na, S.In Computational Biology: New Research;Russe, A. S., Eds.; Nova Science Publisher: New York, 2009; p. 193.
32. Cheng, H.; Gimbutas, Z.; Martinsson, P. G.; Rokhlin, V. SIAM J Sci Comput 2005, 26, 1389.
33. Jang, H.; Na, S.; Eom, K. J Chem Phys 2009, 131, 245106.
34. Kurkcuoglu, O.; Turgut, O. T.; Cansu, S.; Jernigan, R. L.; Doruker, P. Biophys J 2009, 97, 1178.
35. Chennubhotla, C.; Bahar, I. In Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2006, p. 379.
36. Chennubhotla, C.; Bahar, I. Mol Syst Biol 2006, 2, Article No 36.
37. Lu, M.; Ma, J. Proc Natl Acad Sci USA 2008, 105, 15358.
38. Ming, D.; Kong, Y.; Wu, Y.; Ma, J. Proc Natl Acad Sci USA 2003, 100, 104.
39. Kim, J.-I.; Eom, K.; Kwak, M.-K.; Na, S. CMC: Comput Mater Continua 2008, 8, 67.
40. Kim, J.-I.; Na, S.; Eom, K. J Chem Theor Comput 2009, 5, 1931.
41. Meirovitch, L. Computational Methods in Structural Dynamics; Sijthoff & Noordhoff: Rockville, Maryland, 1980.
42. Yoon, G.; Park, H.-J.; Na, S.; Eom, K. J Comput Chem 2009, 30, 873.
43. Brooks, B.; Karplus, M. Proc Natl Acad Sci USA 1983, 80, 6571.
44. Janezic, D.; Venable, R. M.; Brooks, B. R. J Comput Chem 1995, 16, 1554.
45. Meirovitch, L. Analytical Methods in Vibrations; Macmillan: New York, 1967.
46. Kondrashov, D. A.; Cui, Q.; Phillips, G. N., Jr. Biophys J 2006, 91, 2760.