 CMC, vol.6, no.1, pp.35-42, 2007

Robust Reduction Method for Biomolecules Modeling

Kilho Eom[1], Jeong-Hee Ahn[2], Seung-Chul Baek[2], Jae-In Kim[2], Sungsoo Na[2,3]

Abstract: This paper concerns the application and demonstration of robust reduction methodology for biomolecular structure modeling, which is able to estimate dynamics of large proteins. The understanding of large protein dynamics is germane to gain insight into biological functions related to conformation change that is well described by normal modes. In general, proteins exhibit the complicated potential field and the large degrees of freedom, resulting in the computational prohibition for large protein dynamics. In this article, large protein dynamics is investigated with modeling reduction schemes. The performance of hierarchical condensation methods implemented in the paper is compared with that obtained from full original model, successfully demonstrating robustness of reduction method. The examples presented in these results also show that the computational accuracy of reduction method is maintained, while computational cost is reduced.

Keyword: Biomolecules modeling, model condensation, eigenvalue

1 Introduction

Recent advent of high efficient computational techniques enables us to estimate protein dynamics more accurately in a more realistic environment. Molecular Dynamics simulation is one of the most common methodologies to estimate molecular behaviors at atomic level as computational capability increases enormously [McCammon and Harvey. (1987); Shen and Atluri (2004)]. As it is evident during observations of change of molecular structures, the size and length scales of protein structures of interest increase. For a particular purpose, it may not necessary to account for every microscopic detail in the modeling formulation. Furthermore, Molecular Dynamics simulation is not applicable to large biological structures requiring the large spatial and temporal scales [Cui and Bahar (2005); Kim, Jang and Jeong (2006)]. As a consequence, the normal mode analysis, which has enabled the dynamic analysis in engineering [Xie and Long (2006)], has recently contributed significantly as a standard technique in the analysis of the dynamics of biological macromolecules [Cui, Li, Ma and Karplus (2004); Tama and Brooks (2006); Hayward and Go (1995)]. Normal mode analysis is performed with standard semi-empirical potentials, as far as low frequency normal modes are concerned. Its primary objective is to identify and characterize the conformational fluctuations, dominated by low-frequency normal modes, of large biological macromolecules. Nevertheless, the large structural biological system often suffers from memory problem during computation using normal mode analysis. In this regard, Gaussian network model and/or elastic network model which may be one of the alternatives to the normal mode analysis has become prominent to overcome their limitations of conventional schemes. The related publications have shown that the fluctuation dynamics of proteins in heat bath (e.g. water) can be successfully predicted by Gaussian network model [Tirion (1996); Haliloglu, Bahar and Erman (1997)]. The Gaussian network model assumes the protein structures can be simplified with mass and spring model such that dominant alpha carbon atoms, which are nodes, are connected by elastic harmonic springs. In spite of its

[1] Nano-Bio Research Center, Korea Institute of Science & Technology, Hawolgok-dong, Seongbuk-gu, Seoul, 136-791, Korea (e-mail: eomkh@kist.re.kr).

[2] Department of Mechanical Enginrrering , Korea University, Anam-dong, Sungbuk-gu, Seoul, 136-713, Korea (e-mails: {shibuya, takeoff2, jay414, nass}@korea.ac.kr).

[3] Corresponding author.

simplicity, the Gaussian network model has successfully given promising insights for estimation of dynamics of proteins. However, the Gaussian network model may be still computationally inhibitive for the large proteins with a large number of degrees of freedom. In general, reduction method gains computational efficiency by avoiding the full eigenvalue problem for the entire full model and only dealing with a much more coarse-grained model. The applied reduction method presented in this paper consists of reducing the order of the structure by eliminating insignificant variables (nodes).

In this sense, this paper investigates the results of applying hierarchical condensation method to the simulation of protein dynamics and demonstrates the robustness of reduction methodologies for dynamic behaviors of proteins.

2 Modeling Methods

Gaussian network model (GNM)

The GNM, one dimensional configuration of elastic network model, has been used to construct a bridge between physical reality and mathematical formulation [Cui and Bahar (2006)]. The position of the dominant atoms (nodes) based on GNM are defined by the alpha carbon atom coordinates, and the springs connecting the nodes are representative of the bonded and nonbonded interactions between the pairs of residues located within an interaction range, called cutoff distance, r_c. The cutoff distance is usually taken as $7 \sim 12$ based on the radius of the first coordination shell around residues observed in PDB structures.

The Gaussian Network Model (GNM) assumes that the protein is fluctuating about the equilibrium state. The fluctuation of the end-to-end distance between residues i and j obeys the Gaussian probability distributions as follows [Eom, Li, Makarov, and Rodin (2003); Eom, Makarov, and Rodin (2005)];

$$p\left(\left|\mathbf{r}_i-\mathbf{r}_j\right|\right)= \left[\frac{\gamma}{2\pi kT}\right]^{\frac{3}{2}}\exp\left[-\frac{\gamma}{2kT}\left(\left|\mathbf{r}_i-\mathbf{r}_j\right|-\left|\mathbf{r}_i^0-\mathbf{r}_j^0\right|\right)^2\right] \tag{1}$$

where $p(r)$ is the probability distribution for r, γ is a force constant for a Gaussian chain connecting residues i and j, k is the Boltzmann constant, T is the temperature, and superscript 0 indicates the equilibrium state. The Gaussian chain following the probability distribution given as Eq. 1 can be modeled as a harmonic spring whose potential is

$$E=\frac{\gamma}{2}\left(\left|\mathbf{r}_i-\mathbf{r}_j\right|-\left|\mathbf{r}_i^0-\mathbf{r}_j^0\right|\right)^2 \tag{2}$$

The GNM of folded proteins presumes that a residue is connected to the residues within a cut-off distance by Gaussian chains, and consequently, the potential field for GNM of folded proteins is

$$V= \frac{\gamma}{2}\sum_i\sum_j\left(\left|\mathbf{r}_i-\mathbf{r}_j\right|-\left|\mathbf{r}_i^0-\mathbf{r}_j^0\right|\right)^2 H\left(r_c-\left|\mathbf{r}_i^0-\mathbf{r}_j^0\right|\right) \tag{3}$$

where $H(r)$ is the Heaviside unit step function, i.e. $H(r) = 1$ if $r \geq 0$; otherwise $H(r) = 0$, and r_c is the cut-off distance.

The fluctuation behavior of a protein is represented as eigenvalue problem for the oscillating GNM such as [Cui and Bahar (2006)]

$$\mathbf{\Gamma} q=\rho\omega^2\mathbf{q} \tag{4}$$

Here, ρ is the mass of an alpha carbon, ω is the natural frequency, and $\mathbf{q}$ is its corresponding normal mode, and Γ is the $N \times N$ stiffness matrix for GNM given as

$$\Gamma_{ij}=-\gamma H\left(r_c-\left|\mathbf{r}_i^0-\mathbf{r}_j^0\right|\right)(1-\delta_{ij})-\delta_{ij}\sum_{k\neq i}^{N}\Gamma_{ik} \tag{5}$$

where δ_{ij} is the Kronecker delta, i.e. $\delta_{ij} = 1$ if $i = j$; otherwise $\delta_{ij} = 0$, and N is the total number of alpha carbons. The equilibrium statistical mechanics theory [Chandler (1987)] allows one to construct the fluctuation matrix $\mathbf{Q}$ such as

$$\begin{aligned} Q_{ij} &= \left\langle\left(\mathbf{r}_i-\left\langle\mathbf{r}_i\right\rangle\right)\cdot\left(\mathbf{r}_j-\left\langle\mathbf{r}_j\right\rangle\right)\right\rangle \\ &= \sum_{n=2}^{N}\frac{kT}{\rho\omega_n^2}q_i^{(n)}q_j^{(n)} \end{aligned} \tag{6}$$

where the angle bracket $\langle \cdots \rangle$ represents the ensemble average (or time average), ω_n is the natural frequency for n-th mode, and $q_i^{(n)}$ is the i-th component of the eigenvector for n-th mode. The exclusion of the first mode for summation in Eq. 6 is to eliminate the rigid body motion for one-dimensional spring network.

3 Reduction Schemes

3.1 Hierarchical condensation methods

When the dominant atoms in protein structure may be modeled by lumped masses, lumped mass matrices are generated by point masses and eliminate unnecessary nodes if there is a good reason to believe that those particular atoms are not dominant. Hence, the diagonal entries associated with insignificant atoms are generally zero. The fact that certain diagonal entries in the lumped mass matrix are zero is an indication that corresponding displacements are not vital to the fluctuation and can be discarded from the eigenvalue problem. The elimination process is based on hierarchically condensed method, which exhibits the reasonable predictions of fluctuations comparable to original full model. It may be related to the structural characteristics of proteins, that is, the protein structure consists of several domains that are relatively rigid when compared with flexible region such as hinge region. One may consider the protein structures of which the relatively rigid domains are connected with the soft springs. In this respect, the dynamic behavior of protein can be decomposed into two parts; one is fluctuation of the soft region such as hinge region, and the other is internal fluctuation of each relatively rigid domain. That is, the model condensation of protein samples was to be implemented mostly for the domains while the inter-domain interactions are maintained, so that the model condensation allows one to obtain the fast computation of thermal fluctuation with accuracy.

3.2 Model Condensation I

We assumed the inertia forces associated with some of the displacements are known to be much smaller than those associated with others, so that the importance of the former class of displacements in an overall solution can be regarded as being relatively insignificant.

In this sense, we provide Model Condensation I (MC I) that enables us to reconstruct the low-resolution structure consisting of much less number of alpha carbons from the GNM of proteins. We identified the residues that are retained in the low-resolution structure, referred to as master residues, whereas the other residues to be removed in the model condensation are referred to as slave residues. The equation of motion given as Eq. 4 can be described by the kinetic energy K and potential energy V given as [Meirovitch (1980)].

$$K = \frac{1}{2\rho}\mathbf{p}\cdot\mathbf{p} = \frac{1}{2\rho}\begin{bmatrix}\mathbf{p}_m & \mathbf{p}_s\end{bmatrix}\begin{bmatrix}\mathbf{p}_m \\ \mathbf{p}_s\end{bmatrix} \tag{7a}$$

$$V = \frac{1}{2}\mathbf{q}^{\dagger}\mathbf{\Gamma}\mathbf{q} = \frac{1}{2}\begin{bmatrix}\mathbf{q}_m & \mathbf{q}_s\end{bmatrix}\begin{bmatrix}\mathbf{\Gamma}_{mm} & \mathbf{\Gamma}_{ms} \\ \mathbf{\Gamma}_{sm} & \mathbf{\Gamma}_{ss}\end{bmatrix}\begin{bmatrix}\mathbf{q}_m \\ \mathbf{q}_s\end{bmatrix} \tag{7b}$$

where $\mathbf{p}$ is the momenta vector for alpha carbons, $\mathbf{q}$ is the displacement vector, subscript m represents the master residues, and subscript s indicates the slave residues. It is assumed that the slave residues are in the equilibrium.

$$\frac{\partial V}{\partial \mathbf{q}_s} = \mathbf{\Gamma}_{sm}\mathbf{q}_m + \mathbf{\Gamma}_{ss}\mathbf{q}_s = 0 \tag{8}$$

The equation of motion represented by Eq. 4 with kinetic energy and potential energy given as Eq. 7 can be re-written in the form of

$$\begin{bmatrix}\mathbf{\Gamma}_{mm} & \mathbf{\Gamma}_{ms} \\ \mathbf{\Gamma}_{sm} & \mathbf{\Gamma}_{ss}\end{bmatrix}\begin{bmatrix}\mathbf{q}_m \\ \mathbf{q}_s\end{bmatrix} = \rho\omega^2\begin{bmatrix}\mathbf{q}_m \\ \mathbf{q}_s\end{bmatrix} \tag{9}$$

From Eq. 8 and Eq. 9, one can obtain the equation of motion for the low-resolution structure consisting of master residues.

$$\tilde{\mathbf{\Gamma}}\mathbf{q}_m = \rho\omega^2\mathbf{q}_m \tag{10}$$

Here $\tilde{\mathbf{\Gamma}}$ is the effective stiffness matrix for the low-resolution structure, given as

$$\begin{aligned}\tilde{\mathbf{\Gamma}} &\equiv \mathbf{\Psi}\mathbf{\Gamma} \\ &= \begin{bmatrix}\mathbf{I}_m & -\mathbf{\Gamma}_{ms}\mathbf{\Gamma}_{ss}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{\Gamma}_{mm} & \mathbf{\Gamma}_{ms} \\ \mathbf{\Gamma}_{sm} & \mathbf{\Gamma}_{ss}\end{bmatrix}\begin{bmatrix}\mathbf{I}_\mu \\ \mathbf{0}\end{bmatrix}\end{aligned} \tag{11}$$

where $\mathbf{I}_\mu$ is the $\mu \times \mu$ identity matrix (μ = number of master residues), and Ψ is the transition operator that maps the stiffness matrix Γ for the original structure to the effective stiffness matrix $\tilde{\mathbf{\Gamma}}$ for the low-resolution structure.

For MC I of large proteins, the large number of slave residues is not appropriate because the computing expense to estimate the transition operator Ψ is proportional to $O(l^3)$, where l is the number of slave residues. As a consequence, we implemented MC I in the iterative manner as follows: (*i*) Identify the slave residues whose number is much less than that of specified slave residues, (*ii*) Calculate the effective stiffness matrix $\tilde{\mathbf{\Gamma}}$ from Eq. 11, and (*iii*) Repeat the steps (*i*)-(*ii*) until one obtains the effective stiffness matrix $\tilde{\mathbf{\Gamma}}$ for the low-resolution structure consisting of specified master residues. Moreover, MC I was implemented in the hierarchical manner by retaining $N/2$, $N/4$, and $N/16$ alpha carbons, where N is the total number of residues for the original protein structure.

3.3 Model Condensation II

In the eigenvalue problem associated with model condensation II (MC II), the mass matrix generally neglects point masses associated with slave residues. The fact that certain diagonal entries in the mass matrix are zero is an indication that the corresponding displacements are not significant to the solution and can be eliminated from the eigenvalue problem formulation. Its net result is to reduce the order of the eigenvalue problem. According to the present MC II, this enables one to partition the eigenvalue problem as follows:

$$\begin{bmatrix} K_{mm} & K_{ms} \\ K_{sm} & K_{ss} \end{bmatrix} \begin{bmatrix} \mathbf{q}_m \\ \mathbf{q}_s \end{bmatrix} = \omega^2 \begin{bmatrix} M_{mm} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{q}_m \\ \mathbf{q}_s \end{bmatrix} \tag{12}$$

Herein K_{mm} represents the stiffness of substructure including master residues, K_{ss} denotes stiffness of substructure associated with slave residues, while K_{ms} indicates the interaction between master substructure and slave substructure. Furthermore, $q_m$and q_s represent master displacement and slave displacement, respectively. Eq. (12) can be separated into the two equations [Eom, Baek, Ahn, and Na (2007)].

$$K_{mm}q_m + K_{ms}q_s = \omega^2 M_{mm}q_m \tag{13a}$$

$$K_{sm}q_m + K_{ss}q_s = 0 \tag{13b}$$

Solving Eq. (13 b) for q_s, one obtains

$$q_s = -K_{ss}^{-1}K_{sm}q_m \tag{14}$$

so that, substituting Eq. (14) into Eq. (13 a), one obtains the reduced eigenvalue problem via MC II as follows

$$K_1 q_m = \omega^2 M_1 q_m \tag{15}$$

where

$$K_1 = K_{mm} - K_{ms}K_{ss}^{-1}K_{sm}, \quad M_1 = M_{mm} \tag{16}$$

3.4 Mean square fluctuations

The model condensation allows us to obtain the mean-square fluctuation driven by thermal energy for the low-resolution structure. The mean square fluctuation induced by thermal energy for a specific residue i about the equilibrium state is defined as

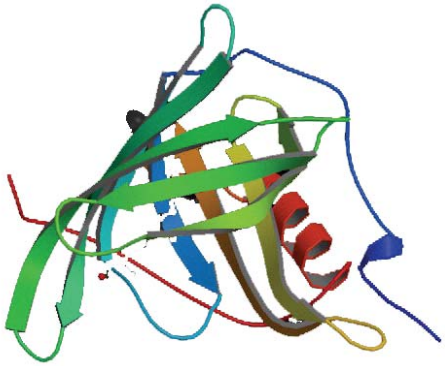
$$\left\langle (\Delta r_i)^2 \right\rangle = Q_{ii} \tag{17}$$

In order to compare the mean-square fluctuation of the low-resolution structure with that of the original structure, the reconstruction procedure for mean-square fluctuation from the low-resolution structure is implemented as follows: (*i*) Calculate the normal modes and mean-square fluctuations for the low-resolution structure consisting of $N/8$ alpha carbons, and then (*ii*) the step (*i*) is repeated for 8 times in order to obtain the mean-square fluctuation for all residues of proteins.

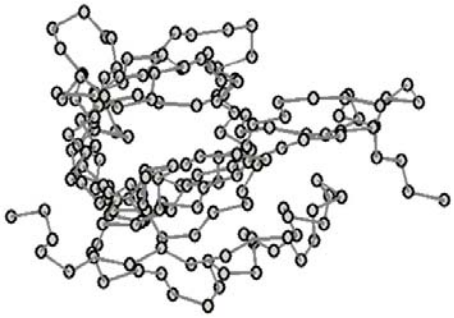
4 Simulation Results

The model condensation that transforms the protein molecular structures into the low-resolution (coarse-grained) protein structures is presented. Here, the model condensation implements the coarse model of ENM to build the low-resolution molecular structure consisting of much less number of alpha carbons. To validate the model condensation scheme for biomolecules in this study,

we considered three proteins such as retinol binding protein (pdb code: 1aqb), hemoglobin (pdb code: 1gzx), and Hiv-1 reverse transcriptase (pdb code: 1tkz). These proteins have the degrees of freedom in the order of $10^2 \sim 10^3$ such that application of reduction method is computationally challenging for simulating the dynamic characteristics of proteins. Fig. 1(a) shows three dimensional image of retinol binding protein and the number of alpha carbons represented by point mass is 175. Fig. 1(b) displays the corresponding mass-spring model composed only by dominant alpha carbons. The two different condensation methods were performed in a hierarchical manner such that we reduce the degree of freedom of biomolecules from N to $N/8$.



(a) molecular structure



(b) mass-spring model

Figure 1: Retinol binding protein **(RBP)** modeling

Fig. 2 represents the quantitative comparison of mean square fluctuation predicted by full model implemented by GNM and experimental data obtained by X-ray crystallography. The result indicates that the fluctuation behavior of model protein can be predicted well by GNM.

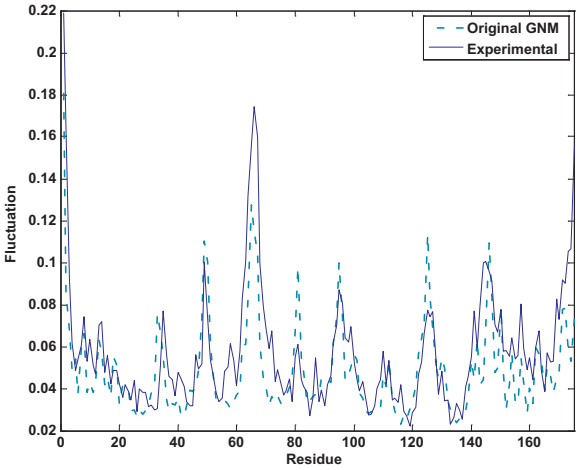


Figure 2: Comparison of experimental data and full GNM model of retinol binding protein

Figs. 3-5 denote the comparisons of mean-square fluctuations between the original full model of structure and the low-resolution structure of $N/8$. It is of great interest in that the fluctuation pattern of the low-resolution structure is qualitatively similar to that of the original structure except the amplitude of the mean-square fluctuations. The larger amplitude for the low-resolution protein structure is obvious because our model condensation scheme is based on the elimination of elastic springs in the original structure so as to soften the protein structure. Furthermore, the amplitude of mean-square fluctuations is not of significance as long as the pattern of mean-square fluctuations is comparable to that of the original structure, because the force constant γ is empirically determined by the curve-fitting to the experimental data. Accordingly, one may obtain the mean-square fluctuations of the low-resolution structure qualitatively and quantitatively comparable to the original structure by adjusting the force constant γ. In addition, from Figs. 3-5, it may suggest that the low-resolution structure consisting of less number of entropic springs for small number of residues is very sufficient to represent the native topology of proteins related to the thermal fluctuations.

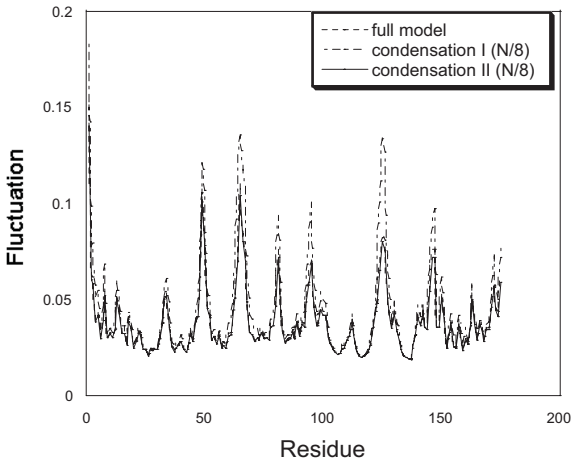
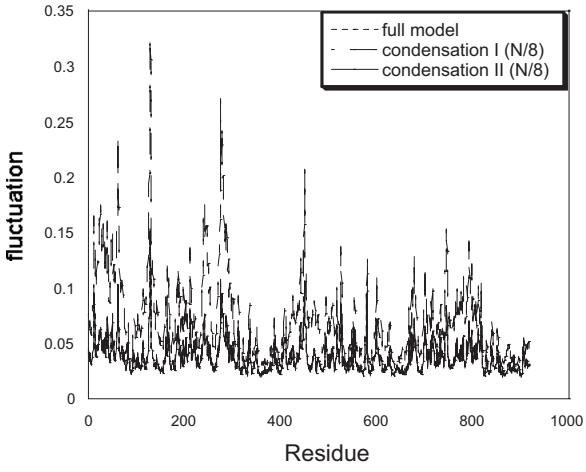


Figure 3: Mean square fluctuation of retinol binding protein **(RBP)**

(a) molecular structure



(b) mean square fluctuation

Figure 4: HIV-1 reverse transcriptas

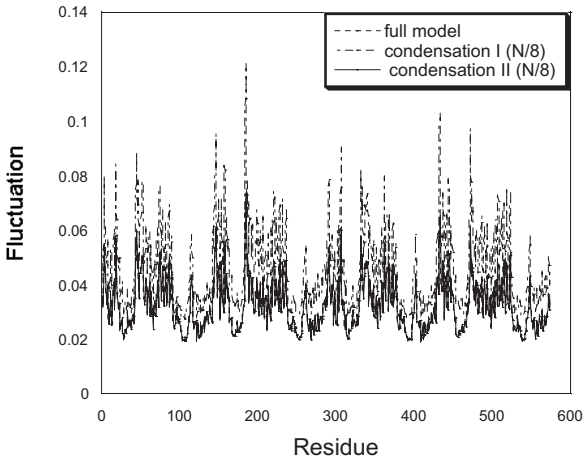


Figure 5: Mean square fluctuation of T state hemoglobin ($N/8$)

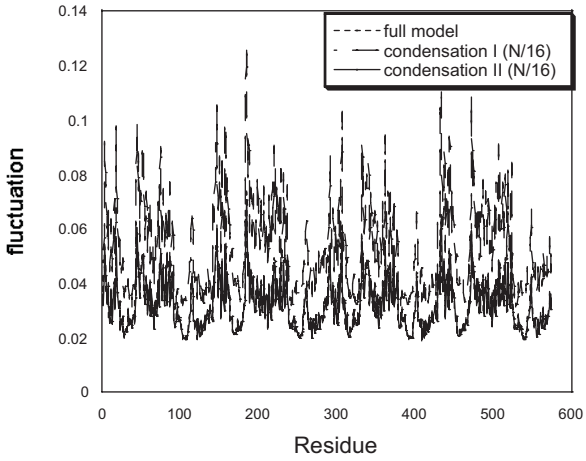


Figure 6: Mean square fluctuation of T state hemoglobin ($N/16$)

In Fig. 6, it is shown that the coarser structures exhibit the qualitatively consistent mean-square fluctuation regardless of resolution of structures as long as N ranges from 8 to 16. It is quite interesting in that the low-resolution structure through model condensation is sufficient to the normal mode study of large proteins.

The successful reproduction of dynamic motion of proteins through model condensation may suggest that the coarse-grained structure is quite sufficient to provide the protein structure topology for protein dynamics. This conjecture may provide that the lower degrees of freedom is appropriate to understand the dynamic motion of pro-

teins, indicating that the low-resolution structure described by the lower degrees of freedom is very sufficient for the studies of protein dynamics. Regarding how many residues are sufficient to represent the native topology of proteins that is related to the dynamic behavior of proteins, from our work, it is stated that the low-resolution structure consisting of small number of residues is able to reproduce the low-frequency modes, that is, functional modes. This indicates that the degree of model condensation may be related to the number of functional modes. Specifically, the number of residues retained in the low-resolution structure may be associated with the number of functional modes of protein dynamics.

5 Conclusions

The model condensation scheme may suggest the further model condensation of protein structures which might be applicable to the large proteins that are hardly tractable with conventional models such as normal mode analysis. This may provide the hierarchical model reduction of protein structures to build the low-resolution structures consisting of the minimal number of atoms for the studies of protein dynamics and structure predictions. For the future perspective, the model condensation may enable one to study the dynamics of the biological supramolecular complexes. Specifically, the dynamics of large proteins inaccessible with conventional NMA with ENM may be computationally tractable by model condensation to reconstruct the protein structures in the low-resolution.

Acknowledgement: This work was supported in part by Nano-Bio Research Center in KIST (to K.E.), and LG YONAM FOUNDATION (to S.N.). S.N. also acknowledges the support by Basic Research Program of the Korea Science & Engineering Foundation (KOSEF) under grant No. R01-2007-000-10497-0.

References

McCammon, J. A.; Harvey, S. (1987): *Dynamics of proteins and nucleic acids*, Cambridge University Press, Cambridge.

Shen, S.; Atluri, S. N. (2004): Computational nano-mechanics and multi-scale simulation, *CMC: Computers, Materials, and Continua*, 1, pp 59-90.

Cui, Q.; Bahar, I. (2005): *Normal Mode Analysis: Theory and Application to Biological and Chemical Systems*, CRC Press.

Kim, M. K.; Jang, Y.; Jeong, J. I. (2006): Using Harmonic Analysis and Optimization to Study Macromolecular Dynamics, *Int. J. of Control, Automation, and Systems*, Vol. 4, No. 3, pp. 382-393.

Xie, G. Q.; Long, S. Y. (2006): Elastic vibration behaviors of carbon nanotubes based on micropolar mechanics, *CMC: Computers, Materials & Continua*, pp. 11-20.

Cui, Q.; Li, G.; Ma, J.; Karplus, M. (2004): A normal mode analysis of structural plasticity in the biomolecular motor F(1)-ATPase, *J. Mol. Biol.*, 340, pp 345-372.

Tama, F.; Brooks, C. L. (2006): Symmetry, form, and shape: Guiding principles for robustness in macromolecular machines, *Annu. Rev. Biophys. Biomol. Struct.*, 35, pp. 115-133.

Hayward, S.; Go, N. (1995): Collective variable description of native protein dynamics, *Annu. Rev. Phys. Chem.*, 46, pp. 223-250.

Tirion, M. M. (1996): Large amplitude elastic motions in proteins from a single-parameter atomic analysis, *Phys. Rev. Lett.*, 77, pp. 1905-1908.

Haliloglu, T.; Bahar, I.; Erman, B. (1997): Gaussian dynamics of folded protein, *Phys. Rev. Lett.*, 79, pp.3090-3093.

Eom, K.; Li, P.-C.; Makarov, D. E.; Rodin, G. J. (2003): Relationship between the mechanical properties and topology of cross-linked polymer molecules: Parallel strands maximize the strength of model polymers and protein domains, *J. Phys. Chem. B.*, 107, pp. 8730-8733.

Eom, K.; Makarov, D. E.; Rodin, G. J. (2005): Theoretical studies of the kinetics of mechanical unfolding of cross-linked polymer chains and their implications for single-molecule pulling ex-

periments, *Phys. Rev. E.*, 71, 021904.

Eom, K.; Baek, S.; Ahn, J.; Na, S. : On the coarse-graining of protein structures for the normal mode studies, *J. Comput. Chem.*, 28, pp. 1400-1410.

Chandler, D. (1987): *Introduction to modern statistical mechanics*, Oxford University Press.

Meirovitch, L. (1980): *Computational methods in structural dynamics*, Sijthoff & Noordhoff.