Supporting Information

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

Jae-In Kim, Sungsoo Na*, and Kilho Eom*

Department of Mechanical Engineering, Korea University, Seoul 136-701, Republic of Korea

* Correspondence should be addressed to S. Na (e-mail: nass@korea.ac.kr) or K. Eom (e-mail: kilhoeom@korea.ac.kr)

Supplementary Method: How to Construct the Constraint Matrix

Here, we provide the detailed procedure to construct the constraint matrix, because it is essential process in component mode synthesis that is employed in our coarse-graining method. For straightforward illustration, instead of considering complex structure such as protein structure, we restrict ourselves to one-dimensional spring system, since protein structure is regarded as elastic networks in our model. Let us consider a chain consisting of 5 nodes, where adjacent nodes are connected by elastic spring with a force constant of k (see Fig. S.1).

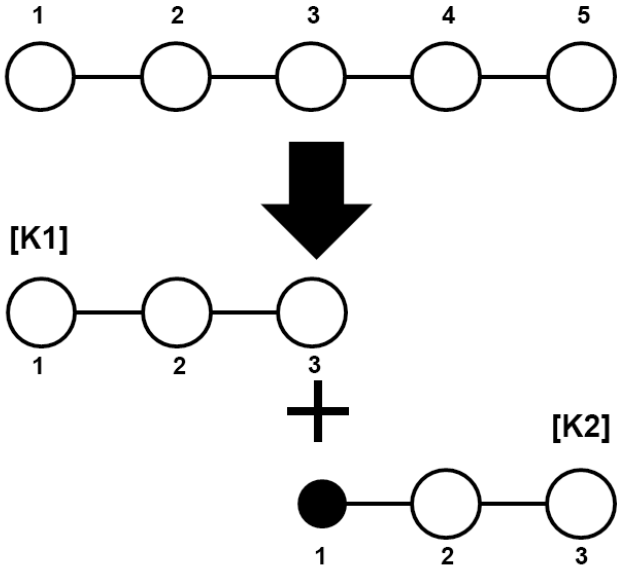


Fig. S.1. Schematic illustration of decomposition of a chain into 2 sub-chains

We denote a displacement for each node as u_j, where a subscript j indicates the index of a node. As shown in Fig.

S.1, a chain is composed into 2 sub-chains such that each sub-chain consists of 3 nodes. We introduce the notation of displacement for each sub-chain such as u^i_j, where superscript i and subscript j indicate the index of sub-chain (i.e. i = 1 or 2) and the index of node in a sub-chain (i.e. j = 1, 2, or 3). It is clear that we have a constraint of $u^1{}_3 = u^2{}_1$.

Now, let us transform the displacement vector into the normal mode space for each sub-chain. With the stiffness matrix of each sub-chain such as $\mathbf{k} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$, the eigen-value problem provides the eigen-values (represented by diagonal matrix Λ) and normal mode space given by

$$\Lambda = \begin{bmatrix} 0 & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & 3k \end{bmatrix} \text{ and } \Phi = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -1 & 1 \end{bmatrix} \tag{S.1}$$

where column vectors of a matrix Φ show the normal modes. The displacement field for nodes in the i-th sub-chain can be represented in the normal mode space such as

$$\begin{bmatrix} v^i_1 \\ v^i_2 \\ v^i_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} u^i_1 \\ u^i_2 \\ u^i_3 \end{bmatrix} \tag{S.2}$$

where v^i_j represents the transformed displacement for j-th node in the i-th sub-chain in the normal mode space. With transformation given by Eq. (S.2), the constraint equation such as $u^1{}_3 = u^2{}_1$ can be transformed into the following equation.

$$v^1_1 - v^1_2 + v^1_3 - v^2_1 + v^2_2 - v^2_3 = 0 \tag{S.3.a}$$

The constraint equation given by Eq. (S.3) can be expressed in the matrix form such as

$$\begin{bmatrix} 1 & -1 & 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} v^1_1 \\ v^1_2 \\ v^1_3 \\ v^2_2 \\ v^2_3 \\ v^2_1 \end{bmatrix} \equiv \begin{bmatrix} \mathbf{P}_1 & \mathbf{P}_2 \end{bmatrix} \begin{bmatrix} \mathbf{s} \\ \mathbf{y} \end{bmatrix} = \mathbf{0} \tag{S.3.b}$$

where $\mathbf{P}_1 = [1 \quad -1 \quad 1 \quad -1 \quad -1]$, $\mathbf{P}_2 = [1]$, $\mathbf{s}^\dagger = [\, v^1{}_1 \quad v^1{}_2 \quad v^1{}_3 \quad v^2{}_2 \quad v^2{}_3]$, and $\mathbf{y} = [v^2{}_1]$.

When we express the displacement field for all nodes in a chain using the displacements for nodes in each sub-chain, we have the redundant degree of freedom before the constraint is imposed. Specifically, the displacement field $\mathbf{z}$ in the normal mode space, which is described using displacements of each sub-chain, is given as

$$\mathbf{z} = \begin{bmatrix} \mathbf{s} \\ \mathbf{y} \end{bmatrix} \tag{S.4}$$

The redundant degree of freedom in the displacement vector $\mathbf{z}$ represented in normal mode space can be eliminated using constraint equation given by Eq. (S.3.b) such as

$$\mathbf{z} \equiv \begin{bmatrix} v_1^1 \\ v_2^1 \\ v_3^1 \\ v_1^2 \\ v_2^2 \\ v_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & -1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} v_1^1 \\ v_2^1 \\ v_3^1 \\ v_2^2 \\ v_3^2 \end{bmatrix} \equiv \begin{bmatrix} \mathbf{I}_5 \\ -\mathbf{P}_2^{-1}\mathbf{P}_1 \end{bmatrix} \mathbf{s} = \mathbf{B}\mathbf{s} \tag{S.4}$$

where **B** is the constraint matrix.

The procedure to construct the constraints in the assembly process for one-dimensional spring system can be extended to our case. As described above, the main key process is to express the constraint equation in the normal mode spaces of each sub-structural unit, and then to remove the redundant degrees of freedom. This will lead to the construction of constraint matrix **B**.

Model Proteins

In this study, we have considered 100 model proteins, most of which exhibits the large degree of freedom ($>10^3$). The model proteins taken into account here are as follows – Deoxy Human Hemoglobin (pdb: 1a3n), A1C12 Subcomplex of F_1F_O-ATP synthase (pdb: 1c17), Bovine Mitochondrial F_1-ATPase (pdb: 1bmf), Scallop Myosin S1 in the pre-power stroke state (pdb: 1qvi), Integrin αvβ3 (pdb: 1jv2), Xanthine Dehydrogenase (pdb: 1fo4), Ground-state Rhodopsin (pdb: 2i36), Methylcitrate Synthase (pdb: 3hwk), Human Phosphofluconate Dehydrogenase (pdb: 2jkv), Mycobacterium Tuberculosis Glutamine (pdb: 2whi), Aplysia ACHBP (pdb: 2wnl), 2,3-dileto-5 methylthiopentyl-1 phosphate enolase (pdb: 2zvi), Human eIF2B (pdb: 3ecs), Tubulin RB3 Stathmin-like domain complex (pdb: 3hkb), E.coli phosphatase YrbI with Mg Tetragonal Form (pdb: 3hyc), bifunctional carbon monoxide dehydrogenase (pdb: 3i01), Proteotive antigen component of Antharax toxin (pdb: 3hvd), Karyopherin nuclear state (pdb: 3icq), Human Thioredoxin reductase I (pdb: 2zz0), Glycogen phosphorylase b R astate AMP (pdb: 3e3n), Nup120 (pdb: 3f7f), Saccharomyces cerevisiae FAS type I (pdb: 3hmj), Apo GroEL (pdb: 1gr5), Betaine Aldehyde Dehydrogenase from Psendomonas (pdb: 2wme), Circadian Clock Protein KaiC (pdb: 3dvl), GAD1 from Arabidopsis Thaliana (pdb: 3hbx), S.aureus Pyruvate Carboxylase in complex with Coenzyme A (pdb: 3ho8), Ectodomain of Human Transferrin Receptor (pdb: 1cx8), Hemochromatosis Protein HFE Complexed with Transferrin Receptor (pdb: 1de4), Beta-Galactosidase (pdb: 1bgm), Glycogen Phosphorylase (pdb: 1noi), Copper Amine Oxidase from Hansenula Polymorpha (pdb: 1a2v), Carbamoyl Phosphate Synthetase from Escherrichia Coli (pdb: 1jdb), R-State Glycogen (pdb: 1abb), Smooth Muscle Myosin Motor Domain Complexed with MGADP ALF4 (pdb: 1br2), Mycobacterium Tuberculosis C171Q KASA vasat (pdb: 2wgf), Arsenite Oxidase (pdb: 1g8k), Unliganded NMRA-Area Zinc Finger Complex (pdb: 2vus), Succinyglutamatedesuccinylase (pdb: 3cdx), Hydrophilic domain of respiratory complex I from Thermus thermophilus (pdb: 3i9v), Delta413-417 (pdb: 3fg1), Pru du amandin, an allergenic protein from prunus dulcis (pdb: 3ehk), E.Coli(lacZ) beta-galactosidse in complex with galactose (pdb: 3e1f), RNA polymerase from Archaea (pdb: 3hkz), Nuclear Export Complex CRM1-Snurportin 1-RanGTP (pdb: 3gjx), Archaeal 13 subunit DNA Directed RNA Pokymerase (pdb: 2waq), Human IDE-inhibitor complex (pdb: 3e4a), hsDDB1-hsDDB2 complex (pdb: 3ei4), T.Thermophilus RNA polymerase holoenzyme in complex with antibiotic myxopyronin (pdb: 3eql), Staphylococcus Aureus Pyruvate Carboxylase (pdb: 3bg5), C3b in complex

with a C3b specific Fab (pdb: 3g6j), Complete ectodomain of integrin aIIBb3 (pdb: 3fcs), E.Coli(lacZ) beta-galactosidase (H418N) (pdb: 3dyp), Uba1-Ubiquitin Complex (pdb: 3cmm), Archaeal RNA polymerase from Sulfolobus solfataricus (pdb: 2pmz), Sodium-Potassium Pump (pdb: 3b8e), Karyopherin beta2/transportin (pdb: 2qmr), Yeast Fatty Aacid Synthase (pdb: 2pff), Transcription regulation by alarmone ppGpp (pdb: 1smy), RNA Polymerase II from Schizosaccharomyces pombe (pdb: 3h0g), Myosin-V inhibited state (pdb: 2dfs), Tricorn Protease (pdb: 1k32), Carbamoyl phosphate synthetase : small subunit mutant c269s with bound glutamine (pdb: 1c3o), Human milk xanthine oxidoreductase (pdb: 2ckj), Glutamate synthase (pdb: 2vdc), Reovirus core (pdb: 1ej6), Human Complement Component 5 (pdb: 3cu7), Glutamine Synthetase (pdb: 2gls), Ribulose1,5Bisphosphate Carboxylase Oxygenase from Nicotiana Tabacum in the Activated state (pdb: 4rub), ribulose 1, 5 bisphosphate carboxylase from synechococcus pcc6301 (pdb: 1rbl), Feedback Inhibition of fully unadenylated Glutamine Synthetase from Salmonella Typhimurium by Glycine, Alanine, And Serine (pdb: 2lgs), A bacterial non-haem iron hydroxylase that catalyses the biological oxidation of methane (pdb: 1mmo), an effector induced inactivated state of Ribulose Bisphosphate Carboxylase Oxygenase: The binary complex between Enzyme (pdb: 1rsc), Bacterial Chaperonin GroEL (pdb: 1grl), Activated Unliganded Spinach Rubisco (pdb: 1aus), Murine Polyomavirus complexed with 3'Sialyl lactose (pdb: 1sid), D-Amino Acid Oxidase From Pig Kidney (pdb: 1kif), Bovine Heart Cytochrome C Oxidase At the Fully Oxidized State (pdb: 1occ), Nitrogenase MOFE Protein from Azotobacter Vinelandii, Oxidized State (pdb: 2min), Methane Monooxygenase hydroxylase From Methylococcus Capsulatus(BATH) (pdb: 1mty), Nitrogenase Complex from Azotobacter Vinelandii Stabilized by ADP-Tetrafluoroaluminate (pdb: 1n2c), EF-TU EF-TS Complex from Thermus Thermophilus (pdb: 1aip), Betaine Aldehyde Dehydrogenase from Cod Liver (pdb: 1a4s), Escherichia Coli Glutamine Phosphoribosylpyrophosphate (PRPP) Amidotransferase Complexed with 2 GMP, 1 MG per Subunit (pdb: 1ecb), Cytochrome BC1 complex from chicken (pdb: 1bcc), Bovine F1-ATPase Covalently Inhibited with 4-Shloro-7-Nitrobenzofurazan (pdb: 1nbm), Orthorhombic Crystals of Beef liver Catalase (pdb: 4blc), Carbamoyl phosphate synthetase : Caught in the act of Glutamine Hydrolysis (pdb: 1a9x), E.Coli Glycerol Kinase and the Mutant A65T in an Inactive Tetramer (pdb: 1glf), Methyl-Coenzyme M Reductase (pdb: 1mro), Bovine Heart Cytochrome C Oxidase at the Fully Oxidized State (pdb: 2occ), Cystathionine Gamma-Synthase from Nicotiana Tabacum (pdb: 1qgn), Glutamate Dehydrogenase from Thermococcus Litoralis (pdb: 1bvu), Rubisco Alcaligenes Eutrophus (pdb: 1bxn), Activated Ribulose 1,5-Bisphosphate Carboxylase/Oxygenase(RUBISCO) Complexed with the Reaction Intermediate Analogue 2-Carboxyarabinitol 1,5-Bisphosphate (pdb: 1bwv), L-Amino Acid Oxidase From Calloselasma Rhodostoma, Complexed with Three Molecules of O-Aminobenzoate (pdb: 1f8s), Glutamate Dehydrogenase (pdb: 1b26), Human Ubiquitous Mitochondrial Creatine Kinase (pdb: 1qk1), Orthorhombic Crystal form of Heat Shock Locus U (HSLU) from Escherichia Coli (pdb: 1do0), and Acetohydroxyacid Isomeroreductase Complexed with its reaction product Dihydroxy-Methylvalerate, Managanese and ADP-Ribose (pdb: 1qmg).

Elastic Network Model

As described in the article, the protein structure is regarded as harmonic spring network in such a way that the residues within the neighborhood are connected by harmonic springs with identical force constant. Here, the force constant is determined by fitting of B-factors computed from ENM to those obtained from experiments.

Fig. S2 shows the force constants for 20 representative model proteins.

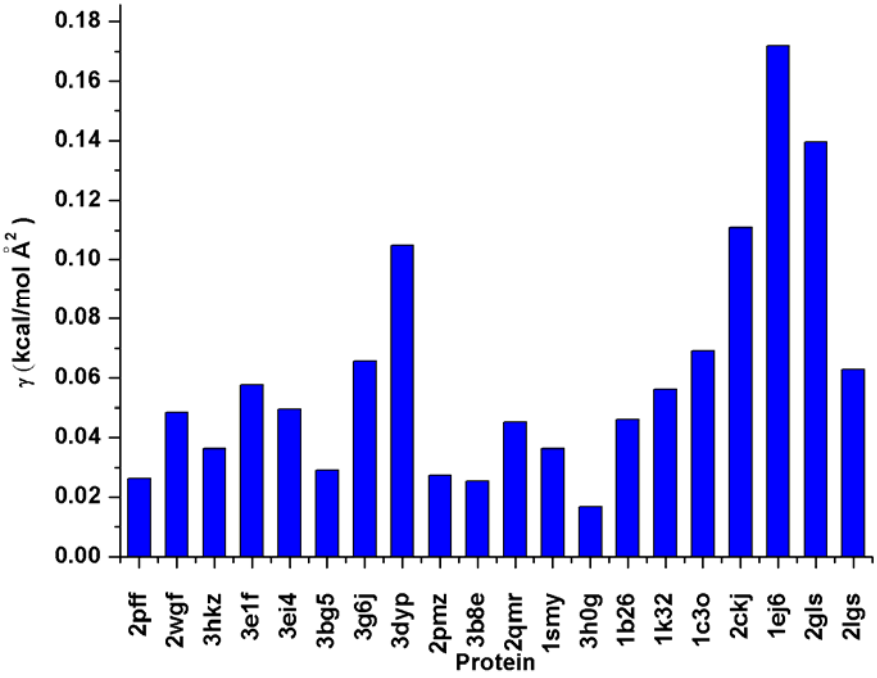


Fig. S.2. Force constants for ENM obtained from fitting of B-factors to those obtained from experiments for 20 representative model proteins.

B-Factors

For robustness of our coarse-graining method for analysis of conformational fluctuation, we have taken into account the Debye-Waller factors (B-factors) computed from our coarse-grained method in comparison with original NMA as well as experimental data. As a model system, we have first considered the hemoglobin. Our coarse-graining method is implemented to hemoglobin such that a structure of hemoglobin is decomposed into 2 sub-structural units, and then each sub-structural unit is coarse-grained such that coarse-grained sub-structural unit is described by $N/2$ nodal points, where N is the total number of residues for a given sub-structural unit. The B-factors obtained from our coarse-grained method is presented in Fig. 2(a). It is shown that B-factors obtained from our coarse-grained method is qualitatively comparable to those computed from original NMA as well as those estimated from the model condensation (MC) method in our previous work[1]. This indicates that our coarse-grained method reliably allows the computationally efficient acquisition of B-factors of proteins. Fig. S.3 shows the B-factors of model proteins such as F_0-ATPase motor protein (b), F_1-ATPase motor protein (c), and scallop myosin (d). It is provided that B-factors for such model proteins (composed of ~10^3 residues) are well depicted by our coarse-grained method.

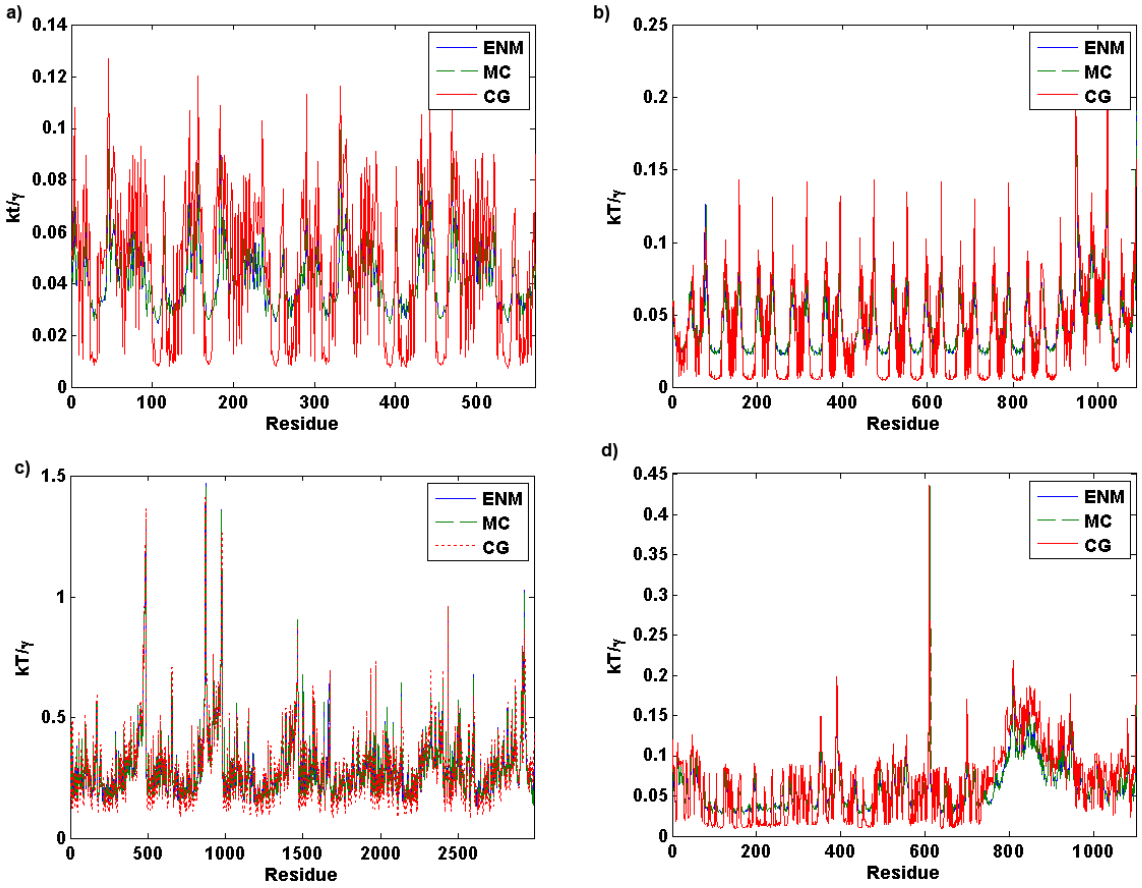


Fig. S.3. Debye-Waller factors (B-factors) computed from ENM, Model Condensation (MC) reported in Ref. [1], and our proposed coarse-graining (CG) method are presented for **(a)** hemoglobin, **(b)** F0-ATPase, **(c)** F1-ATPase, and **(d)** scallop myosin

Lowest-Frequency Normal Mode

Since the low-frequency normal modes obtained from NMA with ENM allow the description of conformational dynamics such as conformational changes,[2-4] we have considered the lowest-frequency normal mode (excluding zero normal modes) obtained from our coarse-grained method and ENM. Fig. S4 shows the lowest-frequency normal modes, computed from NMA with ENM, our previous MC method, and our current coarse-grained method, for 4 representative model proteins. Fig. S4(a) depicts the first low-frequency normal mode, for hemoglobin, that is estimated from ENM, MC, and our coarse-grained method. It is remarkably shown that lowest-frequency normal mode obtained from ENM is almost similar to that computed from our coarse-grained method. The collective dynamic motion of domains *A* and *B*, for hemoglobin,can be clearly seen in the lowest-frequency normal mode obtained from our coarse-grained method as similar to the case of ENM. In the similar manner, as shown in Fig. S4(b) – (d), the lowest-frequency normal modes for other model proteins such as F_0-ATPase, F_1-ATPase, and scallop myosin are well delineated by our coarse-grained method, quantitatively comparable to those computed from NMA with ENM.

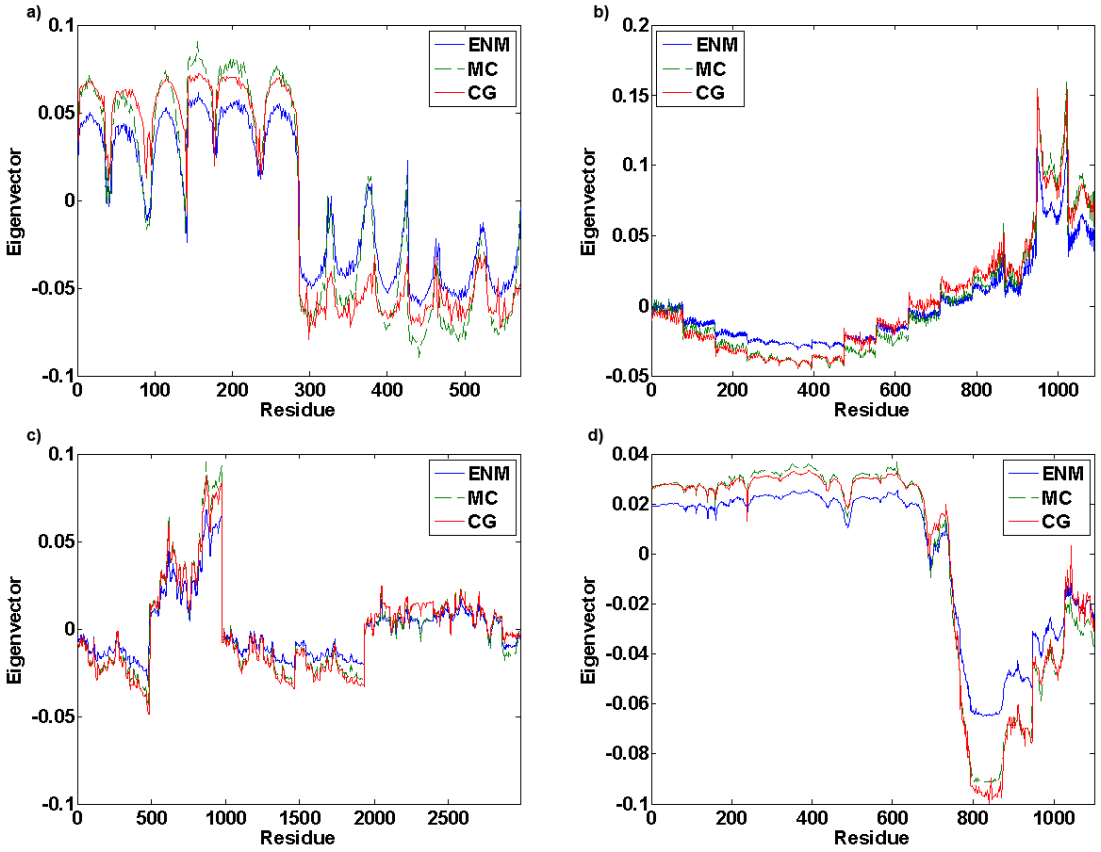


Fig. S.4. Lowest-frequency normal modes obtained from ENM, MC, and our proposed coarse-grained method for **(a)** hemoglobin, **(b)** F0-ATPase, **(c)** F1-ATPase, and **(d)** scallop myosin

Collective Dynamics

Low-frequency normal modes have been reported to play a pivotal role on collective dynamics.[5,6] For the robustness of our coarse-grained method, we consider the collective dynamics dictated by our coarse-grained method in comparison with those predicted by ENM. Here, the collectivity parameter κ_i representing the contribution of i-th normal mode to the collective dynamics is defined as[5]

$$\kappa_i = \frac{1}{N^*}\exp\left[-\sum_{j=1}^{N^*}\left|q_i^j\right|^2 \log\left|q_i^j\right|^2\right] \qquad (12)$$

where N^* is the total number of normal modes (i.e. degrees of freedom for Hessian matrix), q_i^j is the j-th component of the i-th normal mode $\mathbf{q}_i$. The collectivity parameter is in the range between $1/N^*$ and 1. A small value of κ_i close to $1/N^*$ represents the localized motion, whereas κ_i close to 1 indicates the collective (global) motion. Fig. S5 describes the collectivity parameters (with respect to normal mode index) for 4 representative model proteins. For example, a shown in Fig. S5(a) for a case of hemoglobin, the collectivity parameter, computed from both ENM and our coarse-grained method, for the first low-frequency normal mode is close to 1, which implies that collective dynamics is attributed to the first low-frequency normal mode. Moreover, the collectivity parameter estimated from our coarse-grained method for 100-th normal mode is quantitatively similar to that computed from ENM. This indicates that localized motion (high-frequency motions) of

hemoglobin is also well described by our coarse-grained method, quantitatively comparable to that predicted by ENM. In the similar manner, the collective dynamic motions of other model proteins are well delineated by our coarse-grained method, quantitatively comparable to those evaluated from ENM.

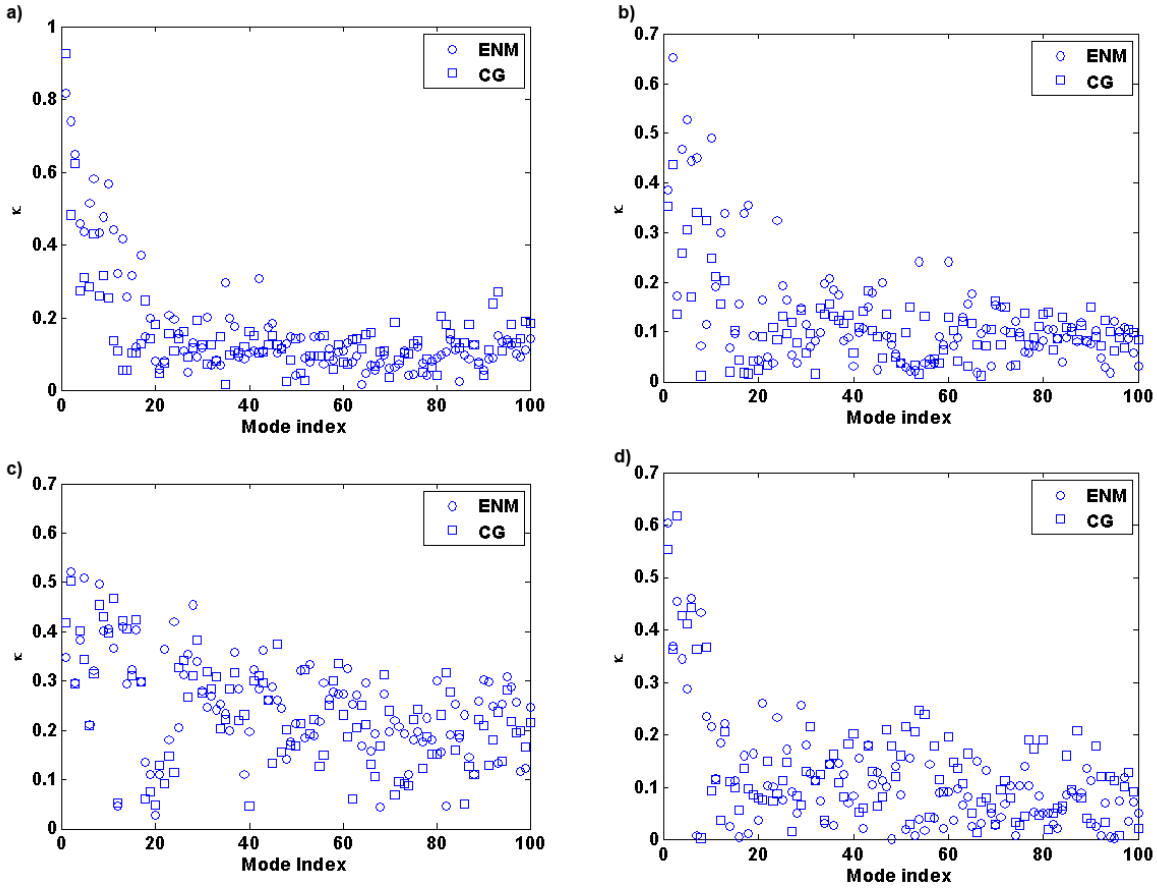


Fig. S.5. Collectivity parameters evaluated from ENM and our proposed coarse-graining (CG) method for model proteins such as **(a)** hemoglobin (pdb: 1a3n), **(b)** F_0-ATPase (pdb: 1c17), **(c)** F_1-ATPase (pdb: 1bmf), and **(d)** scallop myosin (pdb: 1qvi).

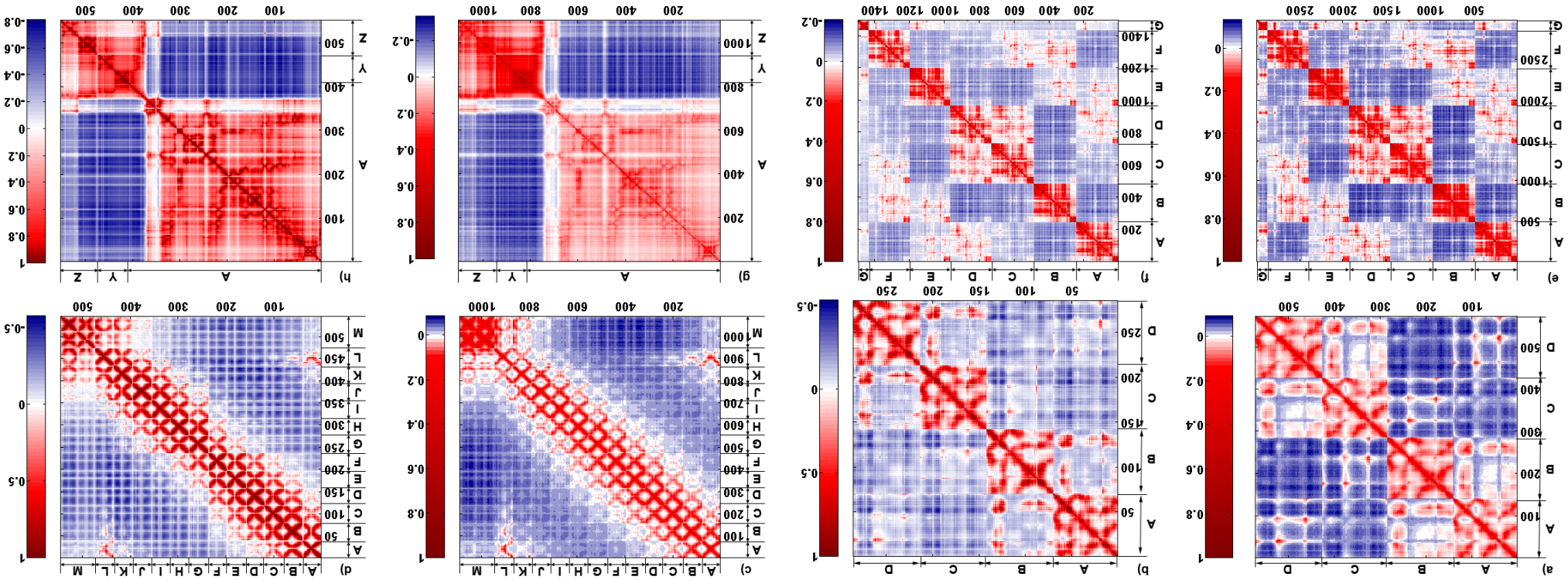
Correlated Motion

We have considered the correlated motions between protein domains. Such correlated motion is well illustrated by cross-correlation defined as[1]

$$c_{ij} = \frac{\left\langle \left(\mathbf{r}_i - \mathbf{r}_i^0\right) \cdot \left(\mathbf{r}_j - \mathbf{r}_j^0\right) \right\rangle}{\sqrt{\left\langle \left|\mathbf{r}_i - \mathbf{r}_i^0\right|^2 \right\rangle \left\langle \left|\mathbf{r}_j - \mathbf{r}_j^0\right|^2 \right\rangle}} \tag{13}$$

Here, c_{ij} indicates correlation between motions of residues i and j. A value of c_{ij} close to 1 represents the correlated motion between residues i and j, while the value of c_{ij} close to 0 shows the uncorrelated motion or orthogonal motion between such two residues, and the value of c_{ij} close to –1 describes the anti-correlated motion between such two residues. Fig. S.6 provides the cross-correlation map computed from both ENM and our coarse-grained method for 4 representative model proteins. It is interestingly shown that our coarse-grained method provides the correlated motions of proteins, quantitatively comparable to those predicted by ENM. For instance, as shown in Fig. S.6(a) for case of hemoglobin, the correlated motion between two domains *A* and *B* is well depicted by our coarse-grained method, quantitatively comparable to that described by ENM. The anti-correlated motion between domains *A* and *C* is well dictated by our coarse-grained method, quantitatively comparable to that depicted by ENM. This indicates that our coarse-grained method enables the robust prediction of correlated motion of proteins.

Fig. S.6. Cross-correlation map obtained from NMA with ENM and our coarse-graining (CG) method (a-b) cross-correlation for hemoglobin estimated from ENM (a) and our CG method (b), (c-d) cross-correlation map for F_0-ATPase computed from ENM (c) and CG method (d), (e-f) cross-correlation map for F_1-ATPase obtained from ENM (e) and our CG method (f), and (g-h) cross-correlation map evaluated from ENM (g) and our CG method (h)



B-Factors Computed from Our Coarse-Graining Method Using Certain Number of Normal Modes

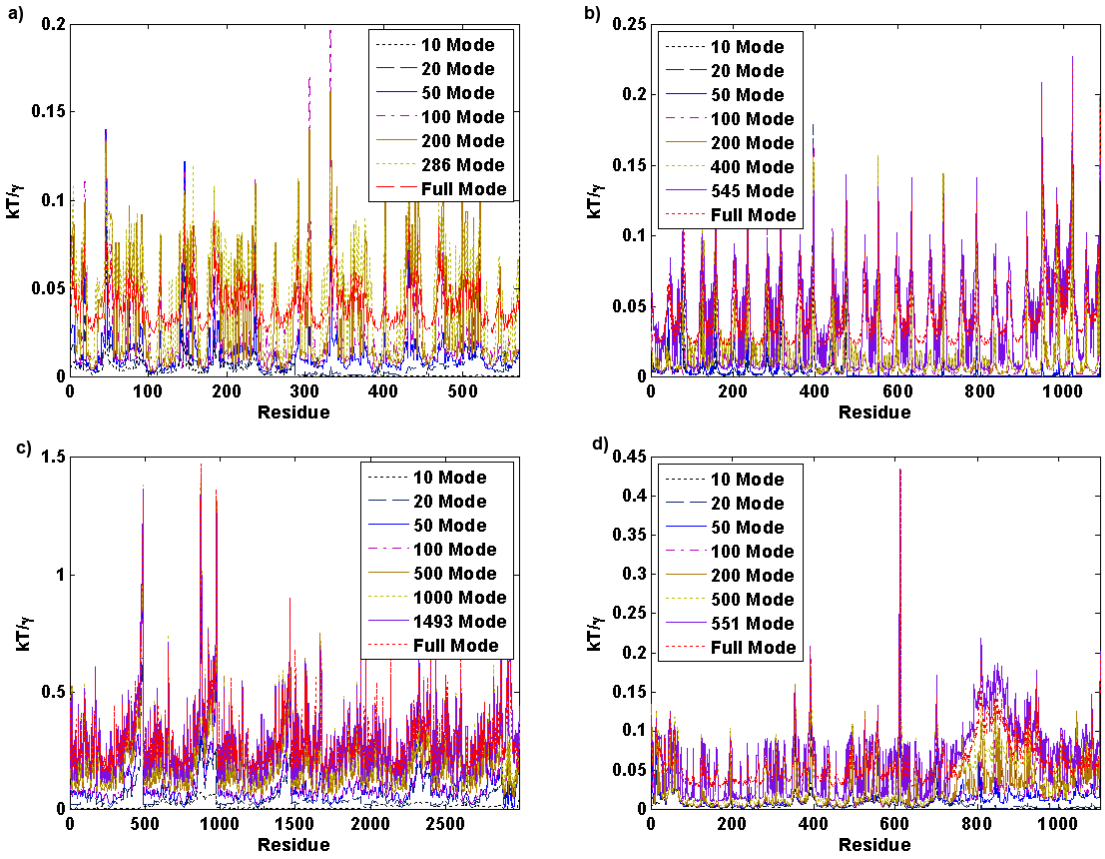


Fig. S.7. B-factors estimated from our coarse-graining (CG) method using different number of normal modes of coarse-grained sub-structural units are shown: **(a)** hemoglobin (pdb: 1a3n), **(b)** F_0-ATPase (pdb: 1c17), **(c)** F_1-ATPase (pdb: 1bmf), and **(d)** scallop myosin (pdb: 1qvi).

References

1. Eom, K.; Baek, S.-C.; Ahn, J.-H.; Na, S. J Comput Chem 2007, 28, 1400.
2. Tama, F.; Sanejouand, Y. H. Protein Eng 2001, 14, 1.
3. Tobi, D.; Bahar, I. Proc Natl Acad Sci USA 2005, 102, 18908.
4. Bahar, I.; Chennubhotla, C.; Tobi, D. Curr Opin Struct Biol 2007, 17, 633.
5. Lienin, S. F.; Bruschweiler, R. Phys Rev Lett 2000, 84, 5439.
6. Navizet, I.; Lavery, R.; Jernigan, R. L. Proteins: Struct Funct Genet 2004, 54, 384.