Using Harmonic Analysis and Optimization to Study Macromolecular Dynamics

Moon K. Kim, Yunho Jang, and Jay I. Jeong

Abstract: Mechanical system dynamics plays an important role in the area of computational structural biology. Elastic network models (ENMs) for macromolecules (e.g., polymers, proteins, and nucleic acids such as DNA and RNA) have been developed to understand the relationship between their structure and biological function. For example, a protein, which is basically a folded polypeptide chain, can be simply modeled as a mass-spring system from the mechanical viewpoint. Since the conformational flexibility of a protein is dominantly subject to its chemical bond interactions (e.g., covalent bonds, salt bridges, and hydrogen bonds), these constraints can be modeled as linear spring connections between spatially proximal representatives in a variety of coarse-grained ENMs. Coarse-graining approaches enable one to simulate harmonic and anharmonic motions of large macromolecules in a PC, while all-atom based molecular dynamics (MD) simulation has been conventionally performed with an aid of supercomputer. A harmonic analysis of a macroscopic mechanical system, called normal mode analysis, has been adopted to analyze thermal fluctuations of a microscopic biological system around its equilibrium state. Furthermore, a structure-based system optimization, called elastic network interpolation, has been developed to predict nonlinear transition (or folding) pathways between two different functional states of a same macromolecule. The good agreement of simulation and experiment allows the employment of coarse-grained ENMs as a versatile tool for the study of macromolecular dynamics.

Keywords: Conformational pathways, elastic network models, molecular dynamics, normal mode analysis.

1. INTRODUCTION

One of the most interesting research items in modern structural biology is the protein folding problem. It has been believed that a protein's knowledge of "how to fold" is encoded in its amino acid sequence since Anfinsen et al. [1] experimentally showed a denatured protein can reproduce its native folded conformation. As many protein structures have been revealed with high resolution by experiments such as X-ray crystallography and nuclear magnetic resonance (NMR), various computational approaches have been proposed to elucidate the relationship between structure and the folding (or unfolding) dynamics of macromolecules.

Molecular dynamics (MD) simulation is one of the most common tools to predict or explain macromolecular motions at atomic detail as computer power increases rapidly and the empirical potential model which governs the system dynamics becomes accurate. However, such an all-atom simulation is currently limited to produce only early stage (on the timescale of nanoseconds) of timely long-range (microseconds or milliseconds) and topologically large motions. Furthermore, the structural information of large macromolecular assemblies is often obtained by a low-resolution experiment such as cryo-EM (i.e., all-atom coordinates required for MD simulation are not available in this case) and these large systems sometimes suffer from memory problems during computation, even using supercomputers. To overcome these drawbacks, various coarse-graining approaches, most of which are designed for proteins, have been developed and recently reviewed by Tozzini [2].

In this paper, we briefly review a variety of coarse-grained elastic network models (ENMs) based on our previous works. In an ENM, a macromolecule is modeled as a spring network among coarse-grained

Manuscript received November 16, 2005; accepted February 8, 2006. Recommended by Editor Keum-Shik Hong. This work was supported by research program 2006 Kookmin University.

Moon K. Kim and Yunho Jang are with the Department of Mechanical and Industrial Engineering, University of Massachusetts, 160 Governors Dr. Amherst, MA 01003, USA (e-mails: {mkkim, yujang}@ecs.umass.edu).

Jay I. Jeong is with the School of Mechanical and Automotive Engineering, Kookmin University, 861-1 Jeongneung-dong, Songbuk-gu, Seoul 136-702, Korea (e-mail: jayjeong@kookmin.ac.kr)

representatives from single atoms to rigid clusters under a certain connection rule to represent chemical interactions. As the first step toward protein folding research, we have studied vibrational behaviors of a macromolecule using normal mode analysis (NMA), traditional in engineering but newly addressed in structural biology, from its native structure reported on the Protein Data Bank [3]. NMA results reflect the equilibrium dynamics of the given macromolecule. On the other hand, large conformational changes between two functionally different forms of a same macromolecule (over an energy barrier in the chemical viewpoint) have also been deduced by solving an optimization problem called Elastic Network Interpolation (ENI) in which only structural information is considered as constraints for the cost function. We also discuss the robustness of the proposed ENM against the change of stiffness values and present a practical way how to handle large macromolecular systems using sparse functions in Matlab. Both harmonic and anharmonic analyses based on ENM play an important role in understanding the relationship between structure and function of macromolecules. Furthermore, they might serve as a useful computational tool for protein structure prediction and design.

2. METHODS

2.1. Coarse-grained ENM

The mathematical model of a multiple degree-of-freedom (MDOF) linear mass-spring system has been recently utilized to the study of the macromolecular dynamics. In ENM, the system is represented by an elastic network of representatives connected by linear springs. Although elastic network models have been used with all-atom descriptions [4,5], various coarse-graining (sampling) approaches have been proposed to reduce the computational burden of all-atom based simulation for large macromolecules. For instance, Fig. 1 illustrates a conventional C_α coarse-grained ENM of a protein [6-8]. A protein is a long chain of amino acids linked by polypeptide bonds. Only C_α atoms are sampled from each amino acid along the main chain. Here the C_α atom is the first carbon atom in the center of an amino acid to which other functional groups are attached. Likewise, P atoms of nucleotides are usually adopted as representatives to model the DNA or RNA structures with a similar resolution to C_α coarse-graining of the proteins (see Fig. 2). Additionally, heavy atoms in sugar and base ring structures of nucleotides could be sampled to reflect much more detailed base-pair interactions on the ENM [9]. We can also build up a coarse-grained ENM for a polymer by choosing only C atoms in each amide group, in which the primary (covalent) bonds through a polymer backbone and the secondary (hydrogen) bonds between polymer chains can be represented by spring connections.

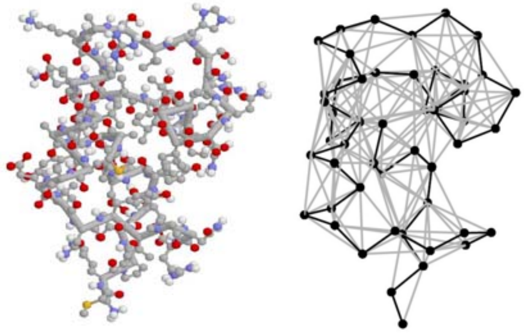


Fig. 1. Representation of protein structure as a coarse-grained elastic network. The lac repressor (PDB code : 1LCC) is illustrated with a ball and stick representation (left). Only C_α atoms are selected as representatives and the spring connections between atoms within a cutoff distance of 8Å are represented by the grey lines (right).

One of the efficient ways to produce large macromolecular machines and assemblies in nature is to assemble repeated units. Most of viruses thereby use some sort of genetic material to encode a small repeated protein unit for their shell structures called capsid. In light of this fact, a symmetry-constrained ENM was developed and applied to reveal the HK97 virus maturation process [10]. The symmetric feature of a system can reduce its DOF significantly because only a repeated unit with symmetry constraints applied is needed to model the whole structure. To implement this symmetric feature on the ENM, group theory has been utilized. The detailed mathematical descriptions are available in our previous paper [10].

A rigid-cluster ENM has also been developed for further simplification of the ENM. In this model, a macromolecule is modeled as a set of rigid bodies interconnected with linear springs. Since the large-scale movements in macromolecules are highly engaged in relative motions among those rigid domains, the rigid-cluster descriptions enable to catch the global dynamics of the given system inexpensively. The computational cost no longer scales with the size of macromolecules. Instead, it depends strongly on the number of rigid domains into which the system can be structurally decomposed. Both rigid-cluster NMA and ENI have been developed and the resulting modes and pathways have been compared with those of C_α coarse graining, respectively [11,12].

It is not necessary that a rigid domain is modeled in the manner of all-atom modeling, even C_α coarse graining. That may cause an unnecessary compu-

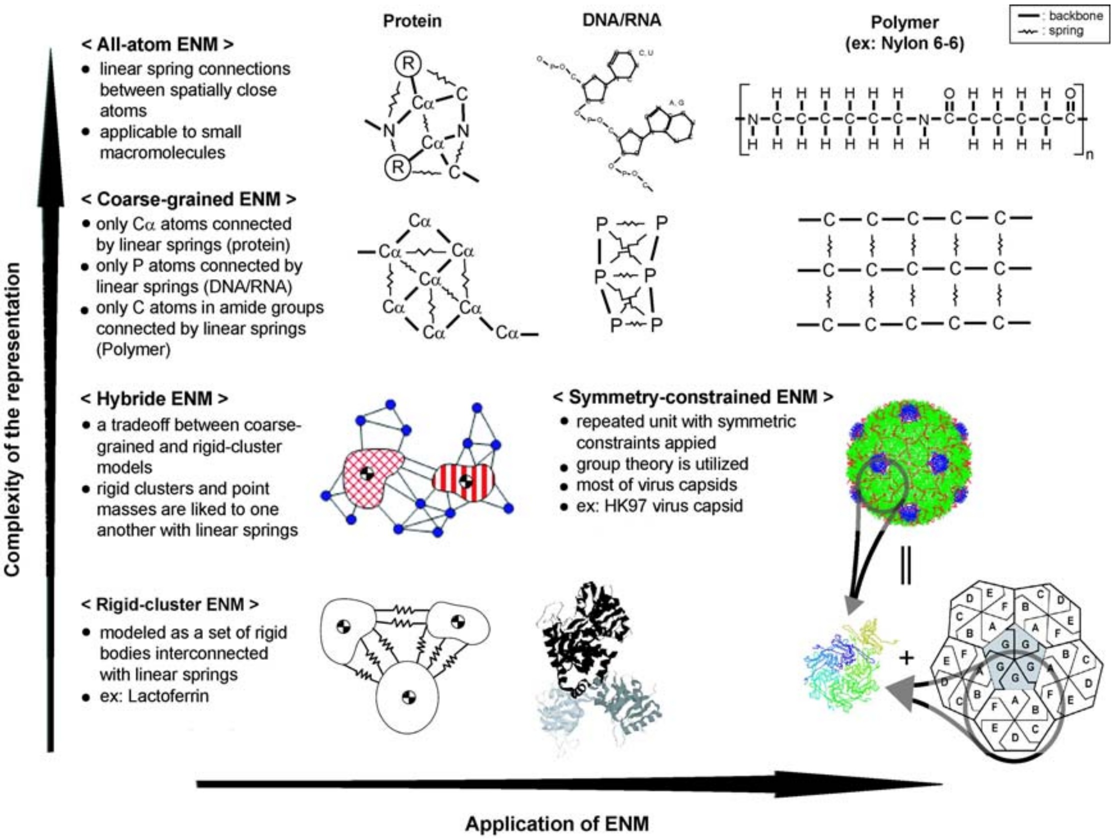

Protein
DNA/RNA
Polymer
(ex: Nylon 6-6)
—: backbone
~: spring
< All-atom ENM >
• linear spring connections between spatially close atoms
• applicable to small macromolecules
< Coarse-grained ENM >
• only Cα atoms connected by linear springs (protein)
• only P atoms connected by linear springs (DNA/RNA)
• only C atoms in amide groups connected by linear springs (Polymer)
< Hybride ENM >
• a tradeoff between coarse-grained and rigid-cluster models
• rigid clusters and point masses are liked to one another with linear springs
< Symmetry-constrained ENM >
• repeated unit with symmetric constraints appied
• group theory is utilized
• most of virus capsids
• ex: HK97 virus capsid
< Rigid-cluster ENM >
• modeled as a set of rigid bodies interconnected with linear springs
• ex: Lactoferrin
Complexity of the representation
Application of ENM

Fig. 2. Illustration of the various coarse-grained ENMs of macromolecules. Main characteristics, schematic representations, and example structures are reported for each model. Symmetry-constrained model can be applied together with any other coarse-grained model.

tational burden. In contrast, oversimplification when using rigid-cluster ENM may destroy the generality of flexible regions of a macromolecule too much. For this reason, hybrid ENM has been proposed as a tradeoff between C_α coarse graining and rigid-cluster modeling in which rigid clusters and point masses are linked to one another with linear springs. The conformational change of GroEL-GroES Complex, a large protein machinery to refold misfolded proteins, has been investigated using the hybrid ENM within reasonable computational time (less than a week in a PC, not like an MD simulation in a supercomputer for a few months or a year) and acceptable resolution of the resulting motions [12].

Fig. 2 summarizes the features of various applications of coarse-grained ENMs. One can choose a proper coarse-graining level (from all-atom description to rigid-cluster representation) as the tradeoff between computational efficiency and physical reality of the model.

2.2. Modeling parameters in ENM

The dynamic behaviors of ENMs of macromolecules might vary with modeling parameters such as linking matrix k and its spring coefficient k_{ij} Three different methods have been developed to create linking matrices as follows: the distance-cutoff method [6], the number-cutoff method [7,8], and the bond-cutoff method [13].

The simplest one is to use a cutoff distance with a constant spring constant as described in the previous section. Large cutoff values increase the number of interacting pairs and the linking matrices become denser (i.e., the elastic networks tend to be stiffer). Consequently, the computation time tremendously increases and the amplitude of fluctuations decreases. On the other hand, short cutoff values strongly force the residues to be in contact with local neighbors only. We may obtain more than six zero eigenvalues in NMA. Six zeros obviously correspond to the rigid-body motions in space, whereas the others result from the lack of spring connections which represent the constraints of the system. For a C_α coarse-grained ENM of a protein, it is empirically observed that a cutoff value should be longer than 11Å to guarantee the system stability. Here Å (=$10^{-10}m$) is the default unit of length in proteins.

The number-cutoff method has been, alternatively, developed to create a uniformly sparse linking matrix. The linking matrix can be generated by imposing a cutoff on the number of connections. That is, a representative atom is connected to neighboring atoms

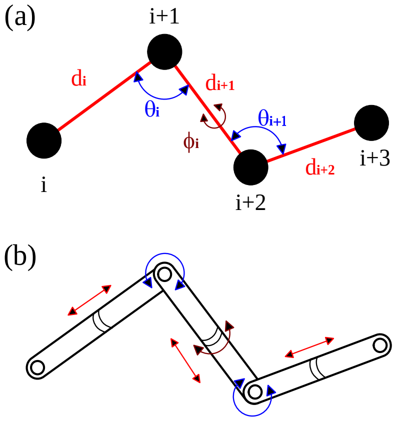


Fig. 3. Labeling internal variables of a polymer chain and the analogous serial robot mechanism. (a) Bond lengths d_i, bond angles θ_i, and torsion angles ϕ_i are, respectively, determined by two, three, and four consecutive atoms. (b) Prismatic and revolute joints of the serial link corresponds to the internal variables in (a).

in order from the closest one until the fixed number of connection is reached. This method cannot reflect local dynamics of the system on the ENM because all atoms have the same number of connections, but reduce computational cost for generating realistic conformational transitions in ENI.

Internal coordinate representation has been widely used to describe the conformational changes of a polymer. It is analogous to a serial robot mechanism connected with different types of joints. Given a system composed of n point masses, the total DOF is $3n$. Of course, 6 are for the rigid-body motions, and $3n-6$ are internal DOF composed of $n-1$ bond lengths (prismatic joints), $n-2$ bond angles (revolute joints at each pivot point), and $n-3$ torsion angles (revolute joints in the middle of each link) as shown in Fig. 3. These internal variables are strongly localized in the sequence of a macromolecule and the distance constraints in ENMs can be described as functions of those variables, even nonlinear. Thus, the spring connections from one residue up to 3 neighbors along the backbone of a protein fully guarantee no loss of DOF. That is,

$$k_{ij} = 1, \quad if \; |i-j| \le 3, \qquad (1)$$

where obviously $i \neq j$ for all i and the total number of spring connections is equal to that of internal variables. Since the distance between two consecutive C_α atoms is approximately 3.8Å [14], the backbone connections suggested here provide an analytic explanation about our earlier observation of the minimum distance-cutoff value needed to stabilize the elastic network.

Using this fact, the bond-cutoff method has been recently proposed to better reflect the chemical interactions within a macromolecule on its ENM. In this model, the system stability is accommodated by spring connections along the backbone according to (1) and then the other connections are added based on the chemical bond information such as sulfide bonds, hydrogen bonds, ionic bonds, and van der Waals interactions. This method provides more accurate and sparser ENMs than the distance-cutoff method [13].

Table 1 shows the density of linking matrices and computation time of NMA. Here we use 12Å as a distance cutoff. As the size of a protein becomes larger, computational time for calculating the full sets of normal modes increases tremendously. To accommo-

Table 1. Linking matrix density and computational efficiency of NMA with the cutoff distance of 12Å.

PDB code	No. of residues	Density[a](%)	Time1[b](sec)	Time2[c](sec)
1LCC	51	42.4	1.0	1.2
1HHP	99	23.5	5.4	2.3
1ATN	372	8.1	148.9	10.7
1LFH	691	4.6	872.2	27.9
1KJU	994	3.4	2542.5	39.4

[a] Percentage of nonzero elements in the linking matrix.
[b] Elapsed time for calculating all normal modes on a 1.5GHz Pentium with 512MB memory.
[c] Elaspsed time for the first 20 nonrigid-body modes selectively calculated by using the "eigs" command in Matlab with the sparse form of a stiffness matrix.

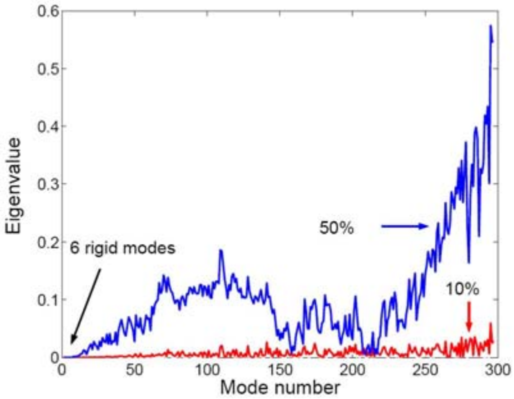


Fig. 4. Sensitivity of eigenvalues to perturbations of stiffness value. The nominal stiffness value is 1 for all springs. This value is randomly perturbed within 50%. That is, stiffness values vary from 0.5 to 1.5. The 10% and 50% plots show the errors between the nominal eigenvalues and the perturbed eigenvalues, respectively (the 5% plot is not displayed here). These changes are negligible compared to the highest nominal eigenvalue of 15.5.

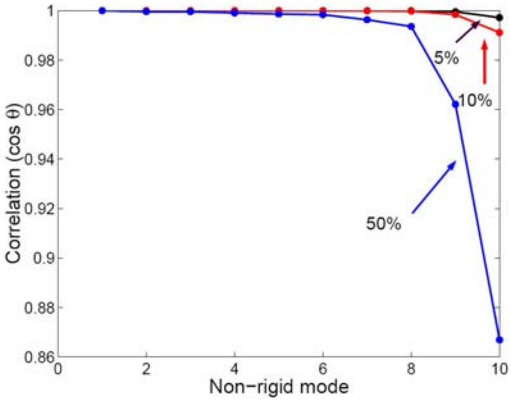


Fig. 5. Sensitivity of non-rigid normal modes to perturbations of stiffness value. The first 10 non-rigid modes, denoted as $\vec{v}_j^{'}$, are taken from each perturbation. They are used to approximate each nominal normal mode denoted as $\vec{v}_i$. The correlation between each nominal normal mode and the approximated one, denoted as $\vec{v}_i^{app} = \sum_{j=1}^{10} c_{ij}\vec{v}_j^{'}$, is presented by the cosine value of the angle difference between two vectors denoted as θ. It is observed that low-frequency modes are not sensitive to changes of stiffness value.

date this problem, we alter the matrix into the sparse form and use the "eigs" command in Matlab to find only a few low-frequency modes with relatively little computational cost.

As another variation of ENM, each stiffness value k_{ij} is randomly perturbed from the unity in this context. Fig. 4 shows an example of the sensitivity of eigenvalues when the stiffness value is perturbed by 5%, 10%, and 50% from the nominal value, respectively. The ENM used here is constructed using the distance-cutoff method for the HIV-1 protease (PDB code: 1HHP). It appears that the change of eigenvalues is negligible. Also, the low-frequency modes are insensitive to perturbations of stiffness value as shown in Fig. 5. Hence, ENM may serve as a robust tool to capture the main dynamics of macromolecules in spite of the lack of physical properties such as spring constant.

2.3. Harmonic fluctuation computed by NMA

The complete mathematical descriptions of NMA and ENI based on ENM were already addressed in our previous papers [7,8]. Nevertheless, the following equations are reproduced here to make a clear explanation.

Given a set of coordinates of n representative atoms (e.g., C_α atoms in the case of proteins), the global mass matrix and the global stiffness (i.e., Hessian) matrix can be derived. The position of the i^{th} atom at time t is denoted as

$$\vec{x}_i(t) = [x_i(t), y_i(t), z_t(t)]^T \in R^3 . \tag{2}$$

The total kinetic energy has the form

$$T = \frac{1}{2}\sum_{i=1}^{n} m_i \left\|\dot{\vec{x}}(t)\right\|^2 . \tag{3}$$

If we define $\vec{\delta}_i(t)$ as a vector of small displacement from the initial position $\vec{x}_i(0)$ such that

$$\vec{x}_i(t) = \vec{x}_i(0) + \vec{\delta}_i(t) , \tag{4}$$

then (3) can be rearranged such that

$$T = \frac{1}{2}\dot{\vec{\delta}}^T M \dot{\vec{\delta}} , \tag{5}$$

where

$$\vec{\delta} = \left[\vec{\delta}_1^T , \ldots , \vec{\delta}_n^T\right]^T \in R^{3n} . \tag{6}$$

In the present case, the global mass matrix M is diagonal.

The total potential energy has the form

$$T = \frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\left\{\left\|\vec{x}_i(t) - \vec{x}_j(t)\right\| - \left\|\vec{x}_i(0) - \vec{x}_j(0)\right\|\right\}^2 \tag{7}$$

and

$$k_{ij} = \begin{cases} 1 \ \ if \ \left\|\vec{x}_i - \vec{x}_j\right\| \le d \\ 0 \ \ if \ \left\|\vec{x}_i - \vec{x}_j\right\| > d , \end{cases} \tag{8}$$

where d is a cutoff distance between C_a atoms and k_{ij} is the (i^{th}, j^{th}) element of k called the linking matrix being unity for all contacting pairs and zero for pairs not in contact. For the small deformations, it can be approximated in a classical quadratic form

$$V = \frac{1}{2}\vec{\delta}^T K \vec{\delta} , \tag{9}$$

where K is called the stiffness matrix (see [7] for more details). This matrix is nothing more than the second derivatives of the harmonic potential function with respect to the generalized coordinates.

Finally, the equations of motion for a coarse-grained protein model can be obtained as

$$M\ddot{\vec{\delta}} + K\vec{\delta} = \vec{0} . \tag{10}$$

The fluctuation dynamics of the structure can be obtained by solving (10). Eigenvalues of the weighted stiffness matrix are the squared natural frequencies of the harmonic motions and the corresponding eigenvectors reflect the mode shapes.

According to statistical mechanics, we can predict that the contribution to a conformational change due to the motion along a normal mode is inversely proportional to the square root of the corresponding eigenvalue (see Appendix). In other words, the global slow motions of structures are dominantly ruled by a few low-frequency modes. Since only a few lowest modes are of interest, the full sets of eigenvalues and eigenvectors are not necessary in this context. In practice, we have used the “eigs” command in Matlab to numerically find a few eigenvalues and eigenvectors of a large but sparse matrix. NMA enables one to infer the global motions and functions of a macromolecule from its structural information.

2.4. Anharmonic pathway generated by ENI

Conformational transitions between two forms of a same macromolecule are very important to understand the connection between its structure and function. Since NMA is not able to predict large anharmonic motions and pathways of macromolecules, several interpolation techniques have been utilized to generate plausible intermediate conformations. The pros and cons of Cartesian coordinate (linear) interpolation, internal coordinate interpolation, and other energy based approaches have been discussed in [7].

The key idea of ENI is to interpolate two sets of distances between spatially close point masses by solving an optimization problem. Suppose the sets of Cartesian coordinates describing representative C_α atoms in two different conformations are denoted as $\vec{x}_i$ and $\vec{y}_i$, respectively. Then we introduce a penalty (cost) function as follows

$$C(\vec{\delta}) = \frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{ij}\left\{\left\|\vec{x}_i + \vec{\delta}_i - \vec{x}_j - \vec{\delta}_j\right\| - l_{ij}\right\}^2, \quad (11)$$

where the linking matrix k is the “union” of the two linking matrices for $\{\vec{x}_i\}$ and $\{\vec{y}_i\}$ built by the number-cutoff method [8]. This cost function can be slightly modified to be fitted with symmetry-constrained and rigid-cluster ENMs, respectively [10, 12]. The value l_{ij} is the distance constraint between i and j, which can be chosen as

$$l_{ij} = (1-\alpha)\left\|\vec{x}_i - \vec{x}_j\right\| + \alpha\left\|\vec{y}_i - \vec{y}_j\right\|, \quad (12)$$

where $\alpha (0 < \alpha < 1)$ sets the coefficient specifying how far a given state is along the transition from $\{\vec{x}_i\}$ to $\{\vec{y}_i\}$.

An intermediate conformation can be obtained by finding $\vec{\delta}$ which minimizes the cost function. Assuming the movements are pretty small, (11) can be approximated as

$$C(\vec{\delta}) \approx \frac{1}{2}\vec{\delta}^T \Gamma \vec{\delta} + \frac{1}{2}\vec{\gamma}\vec{\delta} + B, \quad (13)$$

where Γ is a $3n \times 3n$ matrix, $\vec{\gamma}$ is a $3n$-dimensional row vector, and B is a constant (for more details, refer to [7]). To minimize $C(\vec{\delta})$ with respect to $\vec{\delta}$, differentiation results in the following constraint equation:

$$\frac{\partial C(\vec{\delta})}{\partial \vec{\delta}} = \Gamma\vec{\delta} + \frac{1}{2}\vec{\gamma}^T = \vec{0}. \quad (14)$$

From (14), the solution $\vec{\delta}$ is calculated when $\alpha = 0.01$ so that we can obtain the first intermediate conformation denoted as $\{\vec{x}_i^1\}$ by adding $\vec{\delta}$ to the initial set of coordinates. That is,

$$\vec{x}_i^1 = \vec{x}_i + \vec{\delta}_i, \quad (15)$$

where $\vec{x}_i^1$ is the position of the i^{th} atom out of the set $\{\vec{x}_i^1\}$. In general, the remaining conformations are iteratively generated by increasing α of (12) with increment of 0.01. However, this incremental step size can be adjusted depending on the magnitude of conformational difference between two end conformations measured by root-mean-square deviation (RMSD).

3. SIMULATION RESULTS

3.1. NMA for G-actin and GroEL-GroES complex

To test coarse-grained ENMs, we particularly choose two proteins, G-actin and GroEL-GroES complex, which represent the small and large structures, respectively. We will briefly discuss topological features of those examples and do NMA.

The X-ray crystallography provides Cartesian coordinates of G-actin at atomic level (PDB code: 1ATN). We use only C_α positions to represent each residue (i.e., amino acid) of this protein. G-actin consists of 2 domains, each of which has 2 subdomains (see Fig. 6). The small domain is divided into subdomains 1 (ASP1 to PRO32, GLU72 to ALA144, and SER338 to ARG372) and 2 (SER33 to TYR69). Here the first three-letter code indicates the abbreviation of each amino acid name and the following number labels its location along the protein backbone chain. The large domain is also divided into subdomains 3 (SER145 to LEU180 and GLU270 to TYR337) and 4 (ALA181 to MET269) [15]. All mass

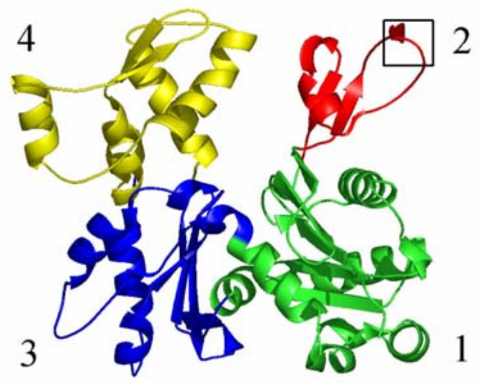


Fig. 6. Schematic of 3D structure of the G-actin with ribbon representation. The large domain is on the left hand side and the small domain is on the right hand side. The two small subdomains are 2 and 4 and the two large subdomains are 1 and 3. The most flexible turn is marked as a box in subdomain 2. The nucleotide binding site is located at the center of the structure (not displayed here).

elements are set to be unity and a linking matrix is made by the cutoff distance of 12Å. Despite this simplification the low-frequency normal modes still provide information on the large domain motions that we are interested in.

We take a closer look at the lowest four modes and identify their corresponding residue motions. Fig. 7 shows which residues dominantly fluctuate in modes 1 and 2. It appears that the residue motions are highly concentrated on a turn structure of subdomain 2 (MET44 through MET47), which is the most flexible part of G-actin. The inset cartoon illustrates corresponding mode shapes. Likewise, Fig. 8 shows the characteristics of modes 3 and 4. We still have the traces of modes 1 and 2, even less concentrated. However, a new active region is found at subdomain 4, as well as many scattered peaks throughout the rest of the plot. They produce a scissor-like bending motion between subdomains of each domain. These global motions were consistently observed by Tirion and Benavraham [4]. They performed an all-atom based NMA with an empirical potential function parameterized not only main-chain torsion angles but also side-chain torsion angles.

Functionally, G-actin is transformed to the thin filament of muscle fiber called F-actin. During this polymerization, the bound Adenosine Triphosphate (ATP), the major energy currency of the cell, is hydrolyzed to the bound Adenosine Diphosphate (ADP) and the cleaved phosphate group. F-actin plays a role in muscle contraction and relaxation in conjunction with myosin. This mechanochemical event may be facilitated with the exposure of the nucleotide binding site induced by the twist and bending motions of G-actin identified by NMA.

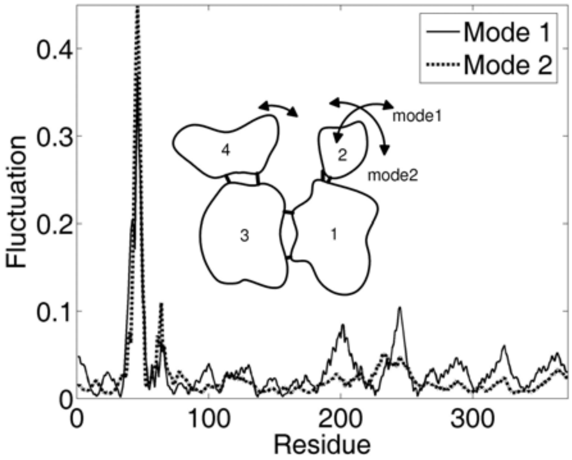


Fig. 7. Characteristics of G-actin in modes 1 and 2. Magnitude of each residue due to modes 1 and 2 is, particularly, presented here. The corresponding mode shapes show the high concentration on subdomain 2 which bends and twists significantly. The mobility of the other subdomains is relatively small.

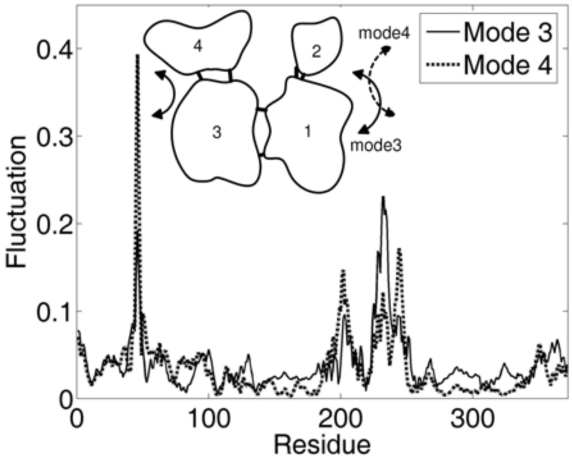


Fig. 8. Characteristics of G-actin in modes 3 and 4. The traces of modes 1 and 2 still remain but residue motions are delocalized in modes 3 and 4. The large and small domains bend to either the same direction (mode 3) or the opposite direction (mode 4). These motions make the ADP binding site open and closed.

The next example protein is the GroEL-GroES complex. It is widely believed that this protein assists the folding of misfolded proteins with the consumption of ATP. Using X-ray crystallography it has been revealed that this complex are formed by GroEl, GroES, and seven bound ADP molecules (PDB code: 1AON) [16]. The overall shape looks like a seven-fold bullet composed of the GroES cap and two GroEL ring structures stacked back to back as shown

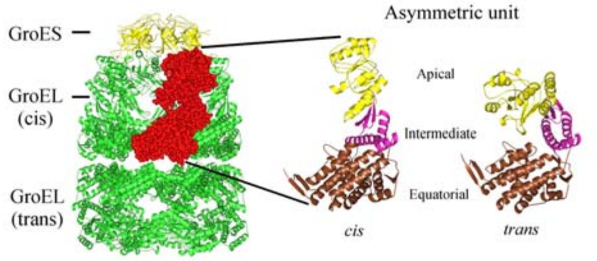


Fig. 9. A cartoon of the GroEL-GroES complex. An asymmetric unit of GroEL is illustrated by space-filling representation. Seven identical subunits comprise a ring structure.

in Fig. 9 [17]. The major deformation of the GroEL ring is associated with the GroES binding. When the GroES binds to the top surface of the GroEL ring, it results in the expansion of the cavity volume. It is called the *cis* ring. In contrast, the release of the GroES from the *cis* ring triggers the conformational change back to the unliganded GroEL called *trans* ring [18,19].

The total number of residues is 8,015 so that the matrix size is 24,045. Since this number is too huge for a matrix in Matlab, it should be divided into submatrices of which the size is small enough to handle within a reasonable time on a PC. We limit the size of a submatrix below 3,000 in this context. To avoid memory limitations, all the matrices are stored as sparse forms. This enables to collect all non-zero elements of submatrices and then put them together to build a whole stiffness matrix in a sparse form. Now we apply the "eigs" function to this large and sparse matrix in order to find the low-frequency normal modes.

As a result, the first 10 lowest modes are sketched in Fig. 10. In the first mode, two GroEL rings rotate about the symmetry axis but each ring moves to the opposite direction. The second and third modes look like sliding motions between two rings. The only different thing is that the direction of motion in the second mode is perpendicular to that of the third mode. Likewise, the fourth and fifth modes complement each other. In these modes, the bullet shape is deformed so that the upper ring is somewhat skewed. In the sixth mode, the large motion of the upper *cis* ring looks like a squeezing motion. This theoretically calculated mode shape has good agreement with the experimentally observed motions between *cis* and *trans* states during the reaction cycle [17]. In the seventh mode, the swing motion of the GroES is observed. It might be related to its docking and releasing motions. The eighth mode is the elongation along the symmetry axis, whereas the ninth mode is the expansion perpendicular to the axis. These two modes are related to the volume change of the *cis* ring. In the tenth mode, the conformational change of the lower *trans* ring is observed.

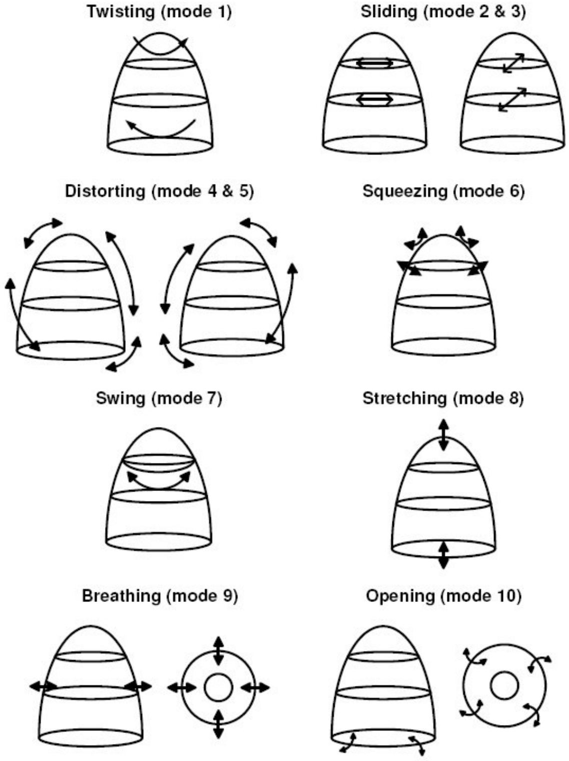


Fig. 10. Pictorial representation of the first 10 lowest mode shapes of the GroEL-GroES chaperonin complex.

From these two examples, we note that only a few low-frequency modes are required to capture well the essential dynamics of a protein and also are strongly related to its biological function. Animations for those mode shapes are presented at a web server (http://biomechanics.ecs.umass.edu/umms.html) called UMass Morph Server (UMMS).

3.2. ENI for a chaperonin called rosettasome

As an example of symmetry-constrained ENI, we choose a chaperonin protein called rosettasome. Chaperonins are multi-subunit protein complexes referred to as heat shock proteins (HSPs). They are synthesized in a cell to increase tolerance for heat and other HSP-inducing stresses. It is widely believed that chaperonins play an important role in folding of newly synthesized proteins and refolding of misfolded proteins damaged by HSP-inducing stresses in the cytoplasm of the cell in vivo [20,21]. However, it has been recently proposed that the rosettasome, a double-ring complex HSP synthesized at lethal temperatures over 90ºC, may function as a membrane skeleton to change the permeability of the membrane in the hyperthermophilic archaeon *Sulfolobus shibatae* [22, 23].

Both "closed" and "open" conformations of rosettasome are isolated from cells as double-ring

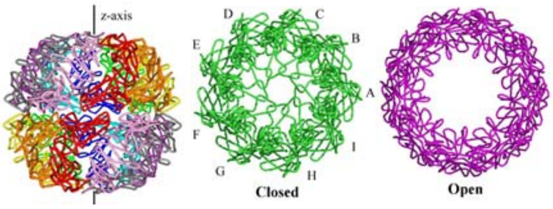


Fig. 11. The double-ring structure of rosettasome. Each ring is 160Å in diameter and 75Å in height. It has been recently discovered that the double-rings associate to form filaments [23]. The closed and open conformations of rosettasome are also displayed (top-down view).

complexes as shown in Fig. 11. Each ring consists of 9 asymmetric units (Chains A through I). An asymmetric unit consists of 503 residues from LYS28 to ALA530. To visually compare the structures of closed and open conformations, VAL529 and ALA530 are discarded from the open conformation because they are not provided in the closed conformation.

The 9-fold ring structure of rosettasome is reconstructed by copying an asymmetric unit, denoted as Chain A, with rigid-body rotation about the z-axis every 40º in this context. Initially the position and orientation of Chain A are fitted well in space not causing steric clash problems with other copying units. If the given two rosettasome structures are perfectly symmetric and symmetry constraints always hold during conformational changes, then the symmetry-constrained ENI can be applied to generate intermediate without having to consider the whole structure, but rather only a single asymmetric unit together with a representation of how it interacts with neighboring units. For example, Chain A is surrounded by Chains B and I.

Symmetry-constrained ENI efficiently generates a feasible pathway between the closed and open conformations in Fig. 12. The RMSD of each intermediate is calculated with respect to the initial conformation. Fig. 13(a) shows that this RMSD increases monotonically. To test the possibility of steric clashes between atoms during the transition, the minimum distance over all possible pairs of C_α atoms within three consecutive Chains A, B, and I is computed. Fig. 13(b) shows the minimum distances of each intermediate conformation. It is observed that those minimum contacts do not occur in inter-connections between two chains, but in intra-connections within a chain. That is, there is no steric conflict between chains during the transition.

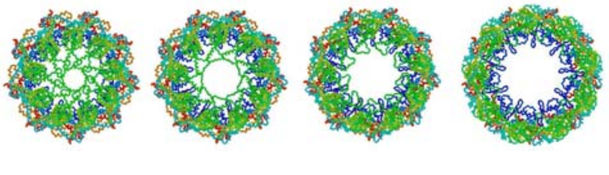
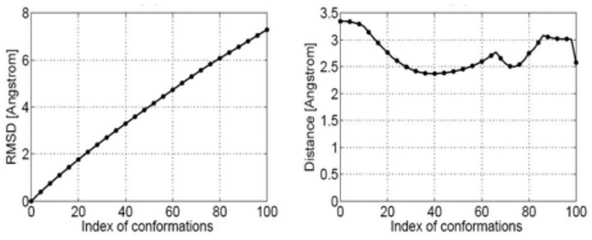


Fig. 12. The conformational transition from closed (left) to open (right) conformations of rosettasome.



(a) (b)

Fig. 13. Characteristics of the simulated pathway of rosettasome. (a) The RMSD value of intermediate conformations with respect to the closed form increases monotonically. (b) The minimum distance between all possible pairs of C_a atoms in intermediate conformations shows that the symmetry-constrained ENI observes steric constraints well.

4. CONCLUSIONS

Coarse-grained ENMs are addressed as a new tool for the study of macromolecular structure and dynamics. For example, only C_a atoms are treated as representatives of each residue of a protein and the interaction between proximal residues is modeled with a linear spring. We also discuss symmetry-constrained, rigid-cluster, and hybrid ENMs for improving both computationally efficiency and physical reality. Using a harmonic potential function based on the proposed ENM, NMA is performed to elucidate an equilibrium dynamics of a macromolecule analytically. Simulation results show that several low-frequency modes dominantly contribute to the global motions of a macromolecule, but they are also insensitive to perturbations of stiffness values. The ENI method is developed to optimally examine the conformational transition between two meta-stable conformations of a same macromolecule by minimizing a cost function derived from its ENM. Unlike dynamics-based methods such as MD and NMA, ENI is purely geometric so that the number of required intermediate frames can be dictated only by the difference in shape between the two conformations. ENI reliably generates anharmonic pathways without steric clashes. Both NMA and ENI results are posted at UMMS (http://biomechanics.ecs.umass.edu/umms.html). One can request harmonic or anharmonic analysis of a macromolecule of interest through this online server. In the near future, this insightful approach will be further explored to i) establish a topology-based folding theory in which ENI might serve as a tool for prediction of the folding pathway from a molten globule (i.e., a stable, partially folded protein state) to the native structure, ii) search new drugs based on ligand binding simulations, iii) understand the role of point mutations and associated large conformational changes in human inherited diseases, and iv) develop a hierarchical system including both discrete and continuum models for the study of mechanical behaviors of polymer composites with multi-scale (nano and micron sized) reinforcements.

APPENDIX A

Magnitude of RMS fluctuations

Given a C_α coarse-grained ENM with *n* residues, a set of eigenvalues and eigenvectors can be obtained from NMA. The k^{th} normalized eigenvector is given by

$$\vec{v}_k = \left[\left(\vec{s}_k^{\,1}\right)^T, \left(\vec{s}_k^{\,2}\right)^T, \ldots, \left(\vec{s}_k^{\,n}\right)^T\right]^T \in R^{3n}, \tag{16}$$

where the motion of residue i due to mode k is given by $\vec{s}_k^{\,i} \in R^3$. In this context, a subscript indicates a mode number, whereas a superscript indicates a residue number. If the amplitude of vibration due to mode k is denoted as α_k, the displacement of residue *i* due to mode *k* is defined as

$$\vec{\delta}_k^{\,i}(t) = \alpha_k \vec{s}_k^{\,i} \sin(\sqrt{\lambda_k}\, t + \phi_k), \tag{17}$$

where λ_k is the k^{th} eigenvalue and ϕ_k is the phase difference at mode k. The RMS fluctuation of residue i due to mode k is defined as

$$\begin{aligned} \sigma_k^i &= \sqrt{\frac{1}{p}\int_0^p \left\| \alpha_k \vec{s}_k^{\,i} \sin(\sqrt{\lambda_k}\, t + \phi_k) \right\|^2 dt} \\ &= \sqrt{\frac{1}{p}\alpha_k^2 \left\| \vec{s}_k^{\,i} \right\|^2 \int_0^p \sin^2(\sqrt{\lambda_k}\, t + \phi_k) dt} \\ &= \sqrt{\frac{1}{2}\alpha_k^2 \left\| \vec{s}_k^{\,i} \right\|^2}, \end{aligned} \tag{18}$$

where $p = 2\pi / \sqrt{\lambda_k}$. The RMS fluctuation of residue *i* due to all modes is defined as

$$\sigma^i = \sqrt{\sum_{k=1}^{3n} (\sigma_k^i)^2}. \tag{19}$$

The cross-correlation between residues fluctuations (i.e., the displacement covariance) is theoretically calculated from

$$< \vec{\delta}_i \cdot \vec{\delta}_j >= \int_{\vec{\delta}} \vec{\delta}_i \cdot \vec{\delta}_j\, f(\vec{\delta}) d\vec{\delta}. \tag{20}$$

Here $\vec{\delta}_i$ is the displacement of residue *i* and $f(\vec{\delta})$ is the probability density

$$f(\vec{\delta}) = \frac{1}{Z_c}\exp[-\frac{1}{k_B T} V(\vec{\delta})], \tag{21}$$

where the conformational partition function Z_c= $\int_{\vec{\delta}} \exp[-\frac{1}{k_B T} V(\vec{\delta})] d\vec{\delta}$ and $V(\vec{\delta})$ is the potential energy function in (9). The fluctuation vector of residue i due to all modes can be written as

$$\vec{\delta}_i = [\vec{s}_1^{\,i} \quad \vec{s}_2^{\,i} \quad \cdots \quad \vec{s}_{3n}^{\,i}]\vec{q} = Q_i \vec{q}, \tag{22}$$

where Q_i is a $3 \times 3n$ matrix and $\vec{q}$ is the generalized coordinates of the system. Substitution into (20) yields

$$\begin{aligned} &< \vec{\delta}_i \cdot \vec{\delta}_j > \\ &= \frac{1}{Z_c}\int_{\vec{q}} \vec{q}^{\,T} Q_i^T Q_j \vec{q} \exp[-\frac{1}{2k_B T}\sum_{k=1}^{3n} \lambda_k q_k^2] d\vec{q}\,. \end{aligned} \tag{23}$$

If we denote $S_{i,j} = Q_i^T Q_j$, then the (a^{th}, b^{th}) element of $S_{i,j}$, denoted as $s_{i,j,a,b}$, is of the form

$$s_{i,j,a,b} = \vec{s}_a^{\,i} \cdot \vec{s}_b^{\,j}. \tag{24}$$

Using this notation, (23) can be changed to the summation of integrals such that

$$\begin{aligned} &< \vec{\delta}_i \cdot \vec{\delta}_j > \\ &= \frac{1}{Z_c}\sum_{a=1}^{3n}\sum_{b=1}^{3n}\left(\int_{\vec{q}} s_{i,j,a,b}\, q_a q_b \exp[-\frac{1}{2k_B T}\sum_{k=1}^{3n}\lambda_k q_k^2] d\vec{q}\right). \end{aligned} \tag{25}$$

Finally, we get

$$< \vec{\delta}_i \cdot \vec{\delta}_j >= k_B T \sum_{k=1}^{3n} \frac{s_{i,j,k,k}}{\lambda_k}. \tag{26}$$

The RMS fluctuation of residue i, denoted as σ^i, is calculated by setting $i = j$ in the above equation such that

$$\begin{aligned} \sigma^i &= \sqrt{k_B T \sum_{k=1}^{3n} \frac{s_{i,i,k,k}}{\lambda_k}} \\ &= \sqrt{k_B T \sum_{k=1}^{3n} \frac{s_{i,i,k,k}}{\lambda_k}} \\ &= \sqrt{\sum_{k=1}^{3n} \frac{k_B T}{\lambda_k} \left\| \vec{s}_k^{\,i} \right\|^2}. \end{aligned} \tag{27}$$

From (19) and (27), σ_k^i can be written in another form such that

$$\sigma_k^i = \sqrt{\left(\frac{k_B T}{\lambda_k} \left\| \bar{s}_k^i \right\|^2 \right)}. \qquad (28)$$

Substitution into (18) yields

$$\alpha_k = \sqrt{\left(\frac{2k_B T}{\lambda_k} \right)}. \qquad (29)$$

Additionally, the RMS fluctuation of all residues due to mode k, denoted as σ_k, is obtained as

$$\sigma_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\sigma_k^i)^2} = \sqrt{\frac{k_B T}{n\lambda_k}} \propto \frac{1}{\sqrt{\lambda_k}}. \qquad (30)$$

Hence the magnitude of RMS fluctuations is inversely proportional to the vibrational frequencies ($\omega_k = \sqrt{\lambda_k}$) of the system.

REFERENCES

[1] M. Sela, F. H. White, and C. B. Anfinsen, “Reductive clevage of disulfide bridges in ribonuclease,” *Science*, vol. 125, pp. 691-692, 1957.

[2] V. Tozzini, “Coarse-grained models for proteins,” *Current Opinion in Structural Biology*, vol. 15, no. 2, pp. 144-150, 2005.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The protein data bank,” *Nucleic Acids Research*, vol. 28, pp. 235-242, 2000.

[4] M. M. Tirion and D. Ben-Avraham, “Normal mode analysis of G-actin,” *Journal of Molecular Biology*, vol. 230, pp. 186-195, 1993.

[5] M. M. Tirion, “Large amplitude elastic motions in proteins from a single-parameter, atomic analysis,” *Physical Review Letters*, vol. 77, pp. 1905-1908, 1996.

[6] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, “Anisotropy of fluctuation dynamics of proteins with an elastic network model,” *Biophysical Journal*, vol. 80, pp. 505-515, 2001.

[7] M. K. Kim, G. S. Chirikjian, and R. L. Jernigan, “Elastic models of conformational transitions in macromolecules,” *Journal of Molecular Graphics and Modelling*, vol. 21, pp. 151-160, 2002.

[8] M. K. Kim, R. L. Jernigan, and G. S. Chirikjian, “Efficient generation of feasible pathways for protein conformational transitions,” *Biophysical Journal*, vol. 83, pp. 1620-1630, 2002.

[9] M. K. Kim, W. Li, B. A. Shapiro, and G. S. Chirikjian, “A comparison between elastic network interpolation and MD simulation of 16S ribosomal RNA,” *Journal of Biomolecular Structure and Dynamics*, vol. 21, pp. 395-405, 2003.

[10] M. K. Kim, R. L. Jernigan, and G. S. Chirikjian, “An elastic network model of HK97 capsid maturation,” *Journal of Structural Biology*, vol. 143, pp. 107-117, 2003.

[11] A. D. Schuyler and G. S. Chirikjian, “Normal mode analysis of proteins: A comparison of rigid cluster modes with C-alpha coarse graining,” *Journal of Molecular Graphics and Modelling*, vol. 22, pp. 183-193, 2004.

[12] M. K. Kim, G. S. Chirikjian, and R. L. Jernigan, “Rigid cluster models of conformational transitions in macromolecular machines and assemblies,” *Biophysical Journal*, vol. 89, pp. 43-55, 2005.

[13] J. I. Jeong, Y. Jang, and M. K. Kim, “A connection rule for alpha-carbon coarse-grained elastic network models using chemical bond information,” *Journal of Molecular Graphics and Modelling*, vol. 24, pp. 296-306, 2006.

[14] C. Breanden and J. Tooze, *Introduction to Protein Structure*, Garland Publishing, New York, 1998.

[15] W. Kabsch, H. G. Mannherz, D. Suck, E. Pai, and K. C. Holmes, “Atomic structure of the actin: I Complex,” *Nature*, vol. 347, pp. 37-44, 1990.

[16] H. S. Rye, S. G. Burston, W. A. Fenton, J. M. Beechem, Z. Xu, P. B. Sigler, and A. L. Horwich, “Distinct actions of *cis* and *trans* ATP within the double ring of the chaperonin GroEL,” *Nature*, vol. 388, pp. 792-798, 1997.

[17] P. B. Sigler, Z. Xu, H. S. Rye, S. G. Burston, W. A. Fenton, and A. L. Horwich, “Structure and function in GroEL-mediated protein folding,” *Annual Review of Biochemistry*, vol. 67, pp. 581-608, 1998.

[18] Z. Xu, A. L. Horwich, and P. B. Sigler, “The crystal structure of the asymmetric GroEL-GroES-(ADP)$_7$ chaperonin complex,” *Nature*, vol. 388, pp. 741-750, 1997.

[19] Z. Xu and P. B. Sigler, “GroEL/GroES: Structure and function of a two-stroke folding machine,” *Journal of Structural Biology*, vol. 124, pp. 129-141, 1998.

[20] F. U. Hartl and M. Hayer-Hartl, “Molecular chaperones in the cytosol: From nascent chain to folded protein,” *Science*, vol. 295, pp. 1852-1858, 2002.

[21] J. D. Trent, H. K. Kagawa, and T. Yaoi, “The role of chaperonins in vivo: The next frontier,” *Annals of the New York Academy of Science*, vol. 851, pp. 36-47, 1998.

[22] H. K. Kagawa, T. Yaoi, L. Brocchieri, R. A. McMillan, T. Alton, and J. D. Trent, “The composition, structure and stability of a group II

chaperonin are temperature regulated in a hyperthermophilic archaeon," *Molecular Mice-obiology*, vol. 48, pp. 143-156, 2003.

[23] J. D. Trent, H. K. Kagawa, C. D. Paavola, R. A. McMillan, J. Howard, L. Jahnke, C. Lavin, T. Embaye, and C. E. Henze, "Intracellular localization of a group II chaperonin indicates a membrane-related function," *Proc. of the National Academy of Science of the United States of America*, vol. 100, pp. 15589-15594, 2003.

Moon K. Kim received the B.S. and M.S. degrees in Mechanical Engineering from Seoul National University in 1997 and 1999, respectively, and the M.S.E. and Ph.D. degrees from The Johns Hopkins University in 2002 and 2004, respectively. He has been an Assistant Professor of Department of Mechanical and Industrial Engineering, University of Massachusetts Amherst since 2004. His current research interests are focused on computational structural biology based on robot kinematics including protein dynamics, folding prediction, and drug design for conformational diseases.

Yunho Jang received the B.S. degree in Mechanical Engineering from Myoungji Univeristy in 1996. He received the M.S degree in Mechanical Engineering from University of Massachusetts Amherst, and he is currently purruing the Ph.D. degree at the same univeristy. His current research interests are focused on computational structural biology and multi-scale polymer composites based on robot kinematics and elastic network representation.

Jay I. Jeong received the B.S., M.S., and Ph.D. degrees in Mechanical Engineering from Seoul National University in 1995, 1997, and 2002, respectively. From 2003 to 2006, he was a Post-Doctoral Researcher in The Johns Hopkins University, where he was researching on protein kinematics. He has been with School of Mechanical and Automotive Engineering in Kookmin University, Seoul, Korea as a Full-Time Lecturer since 2006. His research interests include computational biology using elastic network representation, parallel kinematics machines, and calibration of mechanisms.