Phys. Biol. **2** (2005) S173–S180 doi:10.1088/1478-3975/2/4/S12

Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies

Chakra Chennubhotla, A J Rader, Lee-Wei Yang and Ivet Bahar

Department of Computational Biology, School of Medicine, University of Pittsburgh,
W1040 BST 200 Lothrop Street, Pittsburgh, PA 15261, USA

E-mail: bahar@pitt.edu

Received 3 June 2005
Accepted for publication 5 August 2005
Published 9 November 2005
Online at stacks.iop.org/PhysBio/2/S173

Abstract

With advances in structure genomics, it is now recognized that knowledge of structure alone is insufficient to understand and control the mechanisms of biomolecular function. Additional information in the form of dynamics is needed. As demonstrated in a large number of studies, the machinery of proteins and their complexes can be understood to a good approximation by adopting Gaussian (or elastic) network models (GNM) for simplified normal mode analyses. While this approximation lacks chemical details, it provides us with a means for assessing the collective motions of large structures/assemblies and perform a comparative analysis of a series of proteins, thus providing insights into the mechanical aspects of biomolecular dynamics. In this paper, we discuss recent applications of GNM to a series of enzymes as well as large structures such as the HK97 bacteriophage viral capsids. Understanding the dynamics of large protein structures can be computationally challenging. To this end, we introduce a new approach for building a hierarchical, reduced rank representation of the protein topology and consequently the fluctuation dynamics.

1. Introduction

With recent advances in structural genomics, a considerable fraction of the complete set of folds assumed by proteins seems to be within reach. While the newly elucidated structures provide significant insights that could not be acquired from the examination of sequences alone, an emerging view is that knowledge of structure alone may not be sufficient in many cases for assessing the mechanism, or origin, of biomolecular function. Proteins do not function as static entities or in isolation, but they are engaged in functional motions and interactions both within and between molecules, which permit them to achieve their function. An enzyme would not bind a ligand, for example, unless it possessed the conformational flexibility to accommodate the binding molecule and/or its interaction energy by suitable changes in conformation. These changes can range from single amino acid side chain reorientations (local) to concerted domain–domain motions (global) that would bury, for example, the ligand in a cleft between two domains.

While motions on a local scale can be explored to a good approximation by conventional molecular simulations with full atomic potentials and explicit solvent, those at the global scale are usually beyond the range of such simulations and are more efficiently, and in some cases more accurately, assessed by simplified models used in analytical methods. A prime example of the utility of such coarse-grained approaches is the normal mode analysis (NMA) of ribosome using an elastic network (EN) model [1, 2]. These studies provided valuable insights on the complex machinery of the ribosomal complex, consistent with experimental data. They furthermore set the framework for similarly combining experimental and theoretical studies to explore the dynamics of other supramolecular structures, such as the viral capsid [3–5].

Many studies now demonstrate the utility of methods based on EN models [6]. EN-based NMA are used not only for exploring the structural dynamics of biomolecules and their complexes, or extracting the dominant modes of motions and key sites that control their collective machinery, but they are also being used in the reconstruction or refinement

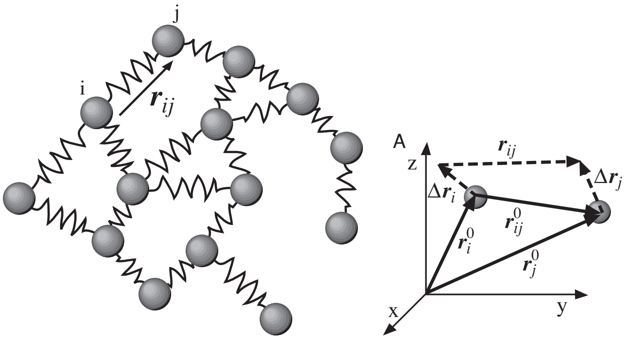
1478-3975/05/040173+08$30.00 © 2005 IOP Publishing Ltd Printed in the UK

of low-resolution cryo-EM structures [7–9] and improving the efficiency of molecular dynamic simulations [10–12]. Yet, as the application and extension to more complex processes or larger systems is undertaken, it becomes increasingly expensive, from both computational memory and time viewpoints, to adopt the residue-based EN models that have been originally introduced [13–18]. Instead, coarse-grained models have been adopted. Examples are the hierarchical coarse-graining (HCG) approach in which sequences of m consecutive amino acids are condensed into unified nodes, thus reducing the computing time and memory by a factor of m^3 and m^2 [19]; adopting a mixed coarse graining in which the substructures of interest are modeled at single-residue-per-node level and the surrounding structural units at a lower resolution of m residues per node; representation of the structure by rigidly translating and rotating blocks (RTB) [20] or the so-called block normal mode analysis (BNM) [21].

In this paper, we present two recent applications of the Gaussian network model (GNM) [13, 14] and also introduce an extension aimed at improving the accuracy and efficiency of the method when exploring supramolecular structures. An overview of the GNM is presented first, with emphasis on the significance of the fluctuations and correlations that are usually computed with the GNM. Section 3 illustrates the use of the GNM for elucidating the close correspondence between chemically active (catalytic) sites known from experiments and key mechanical sites predicted by the GNM. These findings emphasize the functional role of a coupling between chemistry and mechanics for enzymatic activity suggested by experiments [22, 23] and provide insights into the mechano-chemical nature of protein's function. In section 4, we analyze the HK97 bacteriophage procapsid, a supramolecular structure composed of $\approx 10^5$ residues. The large size of the structure precluded a residue-level assessment of the global motions [4, 5]. To improve our understanding of the mechanisms of maturation of the procapsid, we present a hierarchical coarse-grained anisotropic network model (ANM) based analysis [16, 17]. The NMA of supramolecular structures of this size is a challenging task, even with the use of EN models. In the last section, we introduce a novel hierarchical clustering algorithm that provides an efficient means of exploring the collective dynamics, with minimal loss in accuracy. The approach is based on a Markov process description of the communication (or affinity) between interacting residues and has the additional advantage of including residue specificity via consideration of atom–atom interactions. We illustrate the validity and utility of the method by an application to influenza virus protein.

2. The Gaussian network model

The GNM [13] was proposed to explore the role and contribution of purely topological constraints, defined by the 3D structure, on the collective dynamics of proteins around their *equilibrium* configurations. Each protein is modeled by an EN (figure 1(*a*)), the dynamics of which is entirely defined by network topology. The position of the nodes of the EN are defined by the C^α-atom coordinates and the springs connecting the nodes are representative of the bonded and non-bonded interactions between the pairs of residues located within an interaction range, or cutoff distance, of r_c. The cutoff distance is usually taken as 7 Å, based on the radius of the first coordination shell around residues observed in PDB structures [24, 25], and confirmed with an extensive comparison of GNM-predicted B factors with those observed by x-ray crystallography [26].



(*a*) (*b*)

Figure 1. Description of the Gaussian network model (GNM). (*a*) Every residue is represented by a node and connected to spatial neighbors by uniform springs. These springs determine the $N-1$ degrees of freedom in the network and the structure's modes of vibration. (*b*) Schematic representation of the equilibrium positions of the ith and jth nodes, $\boldsymbol{r}_i^0$ and $\boldsymbol{r}_j^0$, with respect to a laboratory-fixed coordinate system ($\boldsymbol{xyz}$). The instantaneous fluctuation vectors $\boldsymbol{\Delta r}_i$ and $\boldsymbol{\Delta r}_j$ are shown by the dashed arrows, as well as the instantaneous separation vector $\boldsymbol{r}_{ij}$ between the positions of the two residues. $\boldsymbol{r}_{ij}^0$ is the equilibrium distance vector from node i to j.

To study the dynamics of such a network, we define the equilibrium position of a node i by vector $\boldsymbol{r}_i^0$ and its instantaneous position by $\boldsymbol{r}_i$ (figure 1(*b*)). The fluctuation, or deformation, from this mean position is defined by the vector $\boldsymbol{\Delta r}_i = \boldsymbol{r}_i - \boldsymbol{r}_i^0$. The deformation in the distance vector $\boldsymbol{r}_{ij}$ that extends from residue i to j is defined as $\boldsymbol{\Delta r}_{ij} = \boldsymbol{r}_{ij} - \boldsymbol{r}_{ij}^0 = \boldsymbol{\Delta r}_j - \boldsymbol{\Delta r}_i$. We are interested in the statistical (ensemble) properties of these pairwise fluctuations, given in a matrix form: $\boldsymbol{\Delta r}_{ij}\boldsymbol{\Delta r}_{ij}^{\mathrm{T}}$. Using the components of the deformation vector

$$\begin{aligned}\boldsymbol{\Delta r}_{ij} &= [(\Delta x_j - \Delta x_i) \quad (\Delta y_j - \Delta y_i) \quad (\Delta z_j - \Delta z_i)] \\ &= [\Delta x_{ij} \quad \Delta y_{ij} \quad \Delta z_{ij}],\end{aligned}$$

the fluctuation matrix can be written as

$$\boldsymbol{\Delta r}_{ij}\boldsymbol{\Delta r}_{ij}^{\mathrm{T}} = \begin{pmatrix} \Delta x_{ij}\Delta x_{ij} & \Delta x_{ij}\Delta y_{ij} & \Delta x_{ij}\Delta z_{ij} \\ \Delta y_{ij}\Delta x_{ij} & \Delta y_{ij}\Delta y_{ij} & \Delta y_{ij}\Delta z_{ij} \\ \Delta z_{ij}\Delta x_{ij} & \Delta z_{ij}\Delta y_{ij} & \Delta z_{ij}\Delta z_{ij} \end{pmatrix}. \qquad (1)$$

GNM makes two assumptions on the distribution of these pairwise fluctuations: first, they are *isotropic*, and second, the distribution is Gaussian. The isotropic assumption implies that the variations in Δx_{ij}, Δy_{ij} and Δz_{ij} are independent of each other and hence the fluctuation matrix is diagonal

$$\boldsymbol{\Delta r}_{ij}\boldsymbol{\Delta r}_{ij}^{\mathrm{T}} = \begin{pmatrix} \Delta x_{ij}\Delta x_{ij} & 0 & 0 \\ 0 & \Delta y_{ij}\Delta y_{ij} & 0 \\ 0 & 0 & \Delta z_{ij}\Delta z_{ij} \end{pmatrix}. \qquad (2)$$

Next, the potential of the network V_{GNM} is defined as a function of the fluctuation covariance arising from interacting residue pairs,

$$\begin{aligned} V_{\mathrm{GNM}}(i,j) &= \frac{\gamma}{2}\,\mathrm{Tr}\left(\boldsymbol{\Delta r}_{ij}\boldsymbol{\Delta r}_{ij}^{\mathrm{T}}\right)\mathcal{H}\left(r_{\mathrm{c}}-\left\|\boldsymbol{r}_{ij}^{0}\right\|\right)\\ &= \frac{\gamma}{2}\boldsymbol{\Delta r}_{ij}^{\mathrm{T}}\boldsymbol{\Delta r}_{ij}\mathcal{H}\left(r_{\mathrm{c}}-\left\|\boldsymbol{r}_{ij}^{0}\right\|\right)\\ &= \frac{\gamma}{2}\{(\Delta x_{ij})^2+(\Delta y_{ij})^2+(\Delta z_{ij})^2\}\mathcal{H}\left(r_{\mathrm{c}}-\left\|\boldsymbol{r}_{ij}^{0}\right\|\right). \end{aligned} \tag{3}$$

Here $V_{\mathrm{GNM}}(i,j)$ is the contribution to the overall potential from a residue pair (i,j), γ is the force constant taken to be uniform for all network springs, $\mathcal{H}(\cdot)$ is the Heavyside step function that is 1 only if the argument is positive and zero otherwise, $\|\cdot\|$ denotes the norm of a vector and $\mathrm{Tr}(\cdot)$ designates the trace of the enclosed matrix. The Heavyside function allows bonded and non-bonded interactions between residues that are within a cutoff distance r_{c} from each other.

A convenient way to capture the interactions between *all* residue pairs in the network is to define a Kirchhoff matrix $\boldsymbol{\Gamma}$ as

$$\boldsymbol{\Gamma}_{ij} = \begin{cases} -\mathcal{H}\left(r_{\mathrm{c}}-\left\|\boldsymbol{r}_{ij}^{0}\right\|\right), & i \neq j,\\ \sum_{j(\neq i)}\boldsymbol{\Gamma}_{ij}, & i = j. \end{cases} \tag{4}$$

In terms of individual fluctuations, the Kirchhoff matrix simplifies the expression for the potential of the entire network of n residues to

$$\begin{aligned} V_{\mathrm{GNM}} &= \frac{\gamma}{4}\left[\sum_{i,j}^{n}\boldsymbol{\Gamma}_{ij}\{(\Delta x_i-\Delta x_j)^2\right.\\ &\qquad \left.+(\Delta y_i-\Delta y_j)^2+(\Delta z_i-\Delta z_j)^2\}\right]\\ &= \frac{\gamma}{2}[\boldsymbol{\Delta x}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Delta x}+\boldsymbol{\Delta y}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Delta y}+\boldsymbol{\Delta z}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Delta z}]. \end{aligned} \tag{5}$$

The fluctuation vectors $\boldsymbol{\Delta x}^{\mathrm{T}}$, $\boldsymbol{\Delta y}^{\mathrm{T}}$ and $\boldsymbol{\Delta z}^{\mathrm{T}}$ have components $[\Delta x_1\ \ \Delta x_2\ \cdots\ \Delta x_n]$, $[\Delta y_1\ \ \Delta y_2\ \cdots\ \Delta y_n]$ and $[\Delta z_1\ \ \Delta z_2\ \cdots\ \Delta z_n]$, respectively. As pointed out above, the fluctuations predicted by the GNM are isotropic and hence there is no information on the 'directions' of motions in different modes, but just their sizes. In order to assess the directions of motions, an extension of GNM has been introduced in [16, 17] called ANM. It is identical to a NMA with an elastic network model at the C^{α} level [18]. The corresponding potential, V_{ANM}, differs from V_{GNM}, in that the term $\boldsymbol{\Delta r}_{ij}^{\mathrm{T}}\boldsymbol{\Delta r}_{ij} = \left(\boldsymbol{r}_{ij}-\boldsymbol{r}_{ij}^{0}\right)^{\mathrm{T}}\left(\boldsymbol{r}_{ij}-\boldsymbol{r}_{ij}^{0}\right)$ is replaced by $\left(\|\boldsymbol{\Delta r}_{ij}\| - \left\|\boldsymbol{\Delta r}_{ij}^{0}\right\|\right)^2$. While the second derivatives of V_{GNM} with respect to the $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ components of the fluctuation vectors lead to a diagonal matrix with identical elements (i.e. uniform force constants along the three directions), the same operation applied to V_{ANM} yields *anisotropic* fluctuations. Thus, GNM cannot be used for predicting the directions of fluctuations, as protein motions are not isotropic, but it can provide an accurate estimate of the *size* of fluctuations. ANM, on the other hand, provides information on the directions of fluctuations. The size of the matrix (Hessian) decomposed in ANM is $3N \times 3N$, which leads to an increase in computing time by about 3^3.

The overall potential V_{GNM} gives rise to a probability distribution over fluctuations $\boldsymbol{\Delta r}$ that is Gaussian,

$$p(\boldsymbol{\Delta r}) = p(\boldsymbol{\Delta x})p(\boldsymbol{\Delta y})p(\boldsymbol{\Delta z}),$$

with

$$\begin{aligned} p(\boldsymbol{\Delta x}) &= \frac{1}{Z_{\boldsymbol{\Delta x}}}\exp\left\{-\frac{\gamma}{2k_{\mathrm{B}}T}\boldsymbol{\Delta x}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Delta x}\right\}\\ &= \frac{1}{Z_{\boldsymbol{\Delta x}}}\exp\left\{-\frac{1}{2}\boldsymbol{\Delta x}^{\mathrm{T}}\left[\frac{k_{\mathrm{B}}T}{\gamma}\boldsymbol{\Gamma}^{-1}\right]^{-1}\boldsymbol{\Delta x}\right\} \end{aligned} \tag{6}$$

and the partition function associated with the fluctuations along the x direction, $Z_{\boldsymbol{x}}$, is given by the determinant

$$Z_{\boldsymbol{\Delta x}} = (2\pi)^{N/2}\left[\det\left(\frac{k_{\mathrm{B}}T}{\gamma}\boldsymbol{\Gamma}^{-1}\right)\right]^{1/2}.$$

Because of the isotropic assumption, similar expressions are valid for $p(\boldsymbol{\Delta y})$ and $p(\boldsymbol{\Delta z})$. So, the partition function for GNM can be written as

$$Z_{\boldsymbol{\Delta r}} = (2\pi)^{3N/2}\left[\det\left(\frac{k_{\mathrm{B}}T}{\gamma}\boldsymbol{\Gamma}^{-1}\right)\right]^{3/2}. \tag{7}$$

From the Gaussian distribution, we can infer the correlations in the residue fluctuations as

$$\left\langle\boldsymbol{\Delta r}_i^{\mathrm{T}}\boldsymbol{\Delta r}_i\right\rangle = \frac{3k_{\mathrm{B}}T}{\gamma}\boldsymbol{\Gamma}_{ii}^{-1}, \tag{8}$$

and the cross-correlations between residues as

$$\left\langle\boldsymbol{\Delta r}_i^{\mathrm{T}}\boldsymbol{\Delta r}_j\right\rangle = \frac{3k_{\mathrm{B}}T}{\gamma}\boldsymbol{\Gamma}_{ij}^{-1}. \tag{9}$$

The determinant of the Kirchhoff matrix $\boldsymbol{\Gamma}$ is 0 and hence care must be taken in computing the inverse $\boldsymbol{\Gamma}^{-1}$. One approach is to diagonalize the symmetric $\boldsymbol{\Gamma}$ matrix and obtain an orthogonal eigenbasis $\boldsymbol{U}$ and a diagonal eigenvalue matrix $\boldsymbol{\Lambda}$,

$$\boldsymbol{\Gamma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}. \tag{10}$$

The eigenvectors describe the modes of free vibration accessible to structures in their native conformations. The eigenvalues determine the mode frequencies. The first eigenvector is constant, indicative of an absence of internal motions and the corresponding eigenvalue is 0. Ignoring the first eigenvector, the pseudo-inverse of the Kirchhoff matrix can be written as

$$\boldsymbol{\Gamma}^{-1} = \sum_{k=2}^{N}\lambda_k^{-1}\boldsymbol{u}_k\boldsymbol{u}_k^{\mathrm{T}}, \tag{11}$$

from which the residue correlations can be extracted,

$$\left\langle\boldsymbol{\Delta r}_i^{\mathrm{T}}\boldsymbol{\Delta r}_i\right\rangle = \frac{3k_{\mathrm{B}}T}{\gamma}\sum_{k=2}^{N}\lambda_k^{-1}\boldsymbol{u}_{ik}^{2}, \tag{12}$$

$$\left\langle\boldsymbol{\Delta r}_i^{\mathrm{T}}\boldsymbol{\Delta r}_j\right\rangle = \frac{3k_{\mathrm{B}}T}{\gamma}\sum_{k=2}^{N}\lambda_k^{-1}\boldsymbol{u}_{ik}\boldsymbol{u}_{jk}. \tag{13}$$

Equation (11) expresses the inverse Kirchhoff matrix as a sum over the $n-1$ non-zero mode contributions. In most applications, it is of interest to extract the contribution of the most cooperative, usually the slowest, modes. The (normalized) distribution of residue squared mobilities in mode k (also called the shape of mode k) is simply given

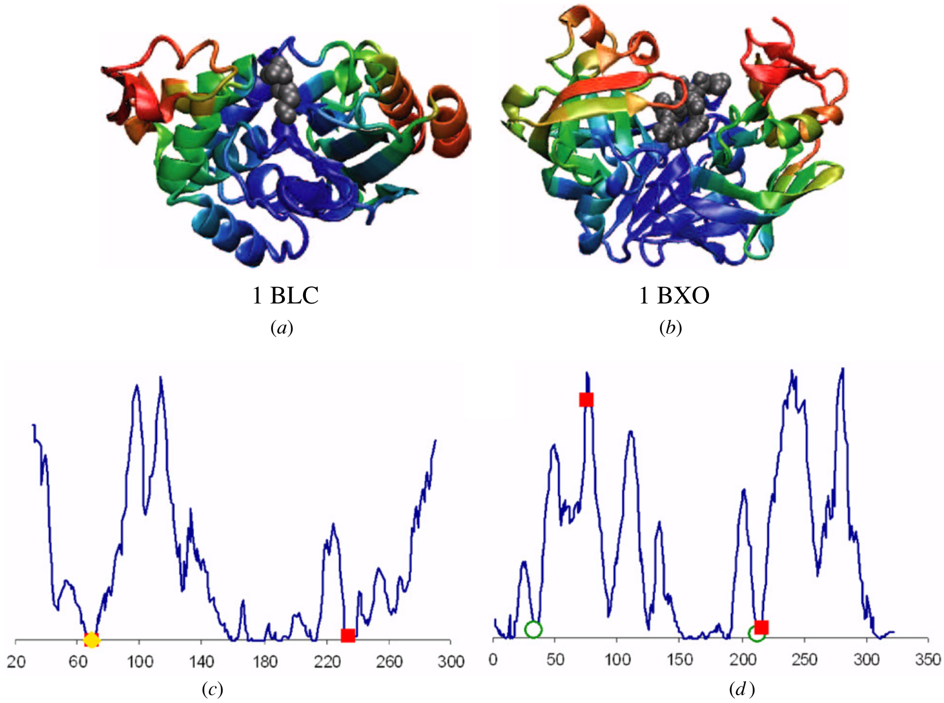


Figure 2. Color-coded ribbon diagrams for two enzymes, β-lactamase (1BLC) (*a*) and penicillopepsin (1BXO) (*b*), illustrating the mobility of residues in the first (lowest frequency) GNM mode. The color code is blue–red–yellow–green in the order of increasing mobility. Both enzymes contain an inhibitor (shown in space filling, gray) bound near the most constrained (lowest mobility) region. (*c*) and (*d*) Corresponding square fluctuation profiles and positions of catalytic and inhibitor-binding residues. See table 1 for the list of chemical (from experiments) and mechanical (from computations) key residues. Residues directly involved in catalytic function at active sites are shown by the green open circles, inhibitor-binding residues are shown by the red squares and residues serving both catalytic and inhibitor-binding functions are marked by the orange diamond.

by the diagonal elements $\boldsymbol{u}_{ik}^2$, for $i = \{1, \ldots, n\}$, of the matrix $\boldsymbol{u}_k\boldsymbol{u}_k^{\mathrm{T}}$. Below, we illustrate the slowest mode shapes of two enzymes (section 3) and the effect of a subset of dominant modes on HK97 procapsid dynamics (section 4), and discuss their relevance of these modes of motion to biological function. Finally, it should be noted that the GNM methodology cannot describe the transition between two states separated by an energy barrier, but applies to fluctuations in the neighborhood of a single energy minimum, the latter being approximated by a quadratic functional form.

3. Catalytic site recognition by GNM

Our recent study demonstrates the existence of a coupling between enzyme catalytic residues and protein mechanical hinge sites [27]. The dynamics of a set of 98 enzymes (93 monomers and 5 multimers) representative of different EC classes was analyzed with the Gaussian network model (GNM) and the mobilities of catalytic residues were investigated. The experimentally identified catalytic residues are defined, following [28, 29]. A given residue is catalytic if (i) it is directly involved in a catalytic function, (ii) it affects the residues or water molecules directly involved in catalysis, (iii) it can stabilize a transient intermediate, or (iv) it interacts with a substrate or cofactor that facilitates the local chemical reaction. The inhibitor-binding sites, on the other hand, are those reported in previous experimental studies to bind inhibitors (ligands). They may, or may not, overlap with an active site. The result showed that more than 70% of the catalytic residues in examined monomeric enzymes are found to be co-localized with the *global hinge centers* predicted by the GNM. Global hinge centers are taken to be the zero-crossings in the low-frequency mode shapes [27]. Moreover, 94% (87/93) of the examined enzymes have at least one global hinge center in their active site.

If one focuses on the lowest frequency modes and normalizes the associated square fluctuations of residues (equation (12)) in each protein and rate the most mobile residue as 100% and the least as 0%, a low translational mobility ($<$7%) is observed for the catalytic residues consistent with the fine-tuned design of enzymes to achieve precise mechano-chemical activities. Two illustrative examples are presented in figure 2. The corresponding chemical and mechanical properties are listed in table 1. The ribbon diagrams are color coded to reflect the relative mobilities of the residues in the slowest mode of motion ($k = 2$) in each case, and the curves display the normalized distributions of square mobilities induced in these modes. The odds ratio in the last column shows the enhancement p/p_0 in the probability p of finding a catalytic residue among the key mechanical sites, as opposed to that p_0 for randomly finding one among all residues in a given enzyme. On the other hand, ligand-binding residues, while generally closely neighboring catalytic sites, enjoy a moderate flexibility to accommodate

Table 1. Correlation between functional sites from experiments and computations.

			Experimental data[c]		Theoretical data[d,e]	
PDB[a]	Protein	Size[b]	Catalytic residues	Ligand-binding residues	Key mechanical sites	Odds ratio
1BLC	β-Lactamase	31–290	**70**	**69**, **70**, **234**	65–72, 206–215	6.2
1BXO	Penicillopepsin	323	**33**, **213**	75, **216**	146–180	5.3

[a] 1BLC [30, 31].

[b] 1BLC PDB coordinates refer to the indicated range.

[c] Bold face residues have mobility scores <0.10 in GNM mode analysis.

[d] Hinge residues with mobility scores <0.05, from regions that are at the intersection of positive and negative displacements in slow mode 2.

[e] Odds ratio = $\frac{p}{p_0}$ is the probability of finding a catalytic residue among key mechanical sites in comparison to that among all residues.

the incoming substrates. Nevertheless, highly mobile ligand-binding residues are occasionally observed in the cases for binding a wide range of ligands or serving as part of the proton-shuttling machinery. These findings could establish new criteria for assessing drug-binding residues and lessen the computational burden of substrate docking searches.

4. Elastic networks of supramolecular structures: viral capsid dynamics

There are several reasons for the increased computational efficiency of residue-level elastic network models over atomic-level NMA. First, the number of nodes in the network is typically reduced by an order of magnitude producing a decrease of 10^3 in computing time. Second, the input structure is assumed to be at a minimum, eliminating the need for costly energy minimization as a function of all atom coordinates. However, the analysis of very large structures such as viral capsids containing over 10^5 residues still requires specialized computational techniques due to memory and calculation limitations.

A typical solution to this problem involves further coarse graining the structure of interest by assuming that groups of residues function as a single node [19] or rigid block [20]. Such a reduced model, where each protein that forms the virus capsid is represented by a single block, has been adopted for a series of virus to compare their dynamics [4]. An alternative approach exploits the symmetry of viral capsids to elucidate local dynamics of a subunit of the capsid structure [32] but disregards the most cooperative, potentially functional, symmetry-breaking motions.

Recently, the dynamics of the HK97 bacteriophage viral capsid including all residues has been analyzed using the GNM [5]. This capsid contains 420 copies of a single protein monomer arranged to form 12 pentamers and 60 hexamers. During maturation, the spherical procapsid (Prohead II) [33] expands into an icosahedral, mature form (Head II) [34]. Our results provide information on the relative sizes of the fluctuations of all residues and the contributions to their mobilities due to different modes of motions. However, the GNM analysis does not provide information concerning the directions of these fluctuations. Here we report the ANM results that have been calculated using a coarse graining of $n/6$ for the Prohead II structure. Using the method described by [19], groups of six residues are represented by a single node, the coordinate of which is identified by every sixth residue along the sequence.

Panels (*a*) and (*b*) in figure 3 illustrate the deformations accessible by the slowest non-zero mode for the capsid using this coarse-grained ANM comprised of 17 920 nodes. The capsid nodes are colored from red (most mobile) to blue (most constrained) following the size of motions predicted by this mode, and the two diagrams represent two alternative conformations between which the capsid fluctuates, by the action of this mode. This motion is essentially an elongation/contraction along the vertical axis. The mobilities are consistent with the slowest GNM mode presented in our earlier study, a three-fold degenerate mode driving the same distribution of mobilities along the x, y and z axes. We note that this is an icosahedrally asymmetric mode: it identifies a pentamer-centered region at each pole as the most mobile (red).

As the frequency increases, the associated modes become more localized and less collective. This is reflected by the deformations shown for the succeeding low-frequency modes in figures 3(*c*)–(*f*). The number of mobile regions observed increase as their sizes decrease. Additionally, because of the high degree of symmetry in the capsid, many calculated modes are degenerate (i.e. they have the same frequency). For example, the modes indicated in figures 3(*c*)–(*f*) are four-fold or three-fold degenerate, each oriented along a different axis of the structure. We note that the successive modes activate different regions of the procapsid. In particular, hexameric regions are observed to be set in motion in panels (*b*) and (*c*), whereas panels (*d*) and (*e*) show a more complex motion that jointly activates subsets of pentamers and hexamers.

Of interest is to see the result from a subset of dominant modes. A superposition of the slowest modes indeed provides an insightful description of the range of deformations accessible by such global modes. Figure 3(*g*) depicts the fluctuations obtained by the superposition of the 11 slowest modes from the GNM. The combination of these dominant modes identifies the 12 pentamers as the most mobile (red) regions in the procapsid, in accord with the icosahedrally symmetric structure of the capsid. The high mobilities of the pentamers are in agreement with experimental cryo-EM data of intermediates between the prohead and head conformations which indicate a large degree of motion for the pentamers during expansion [35].

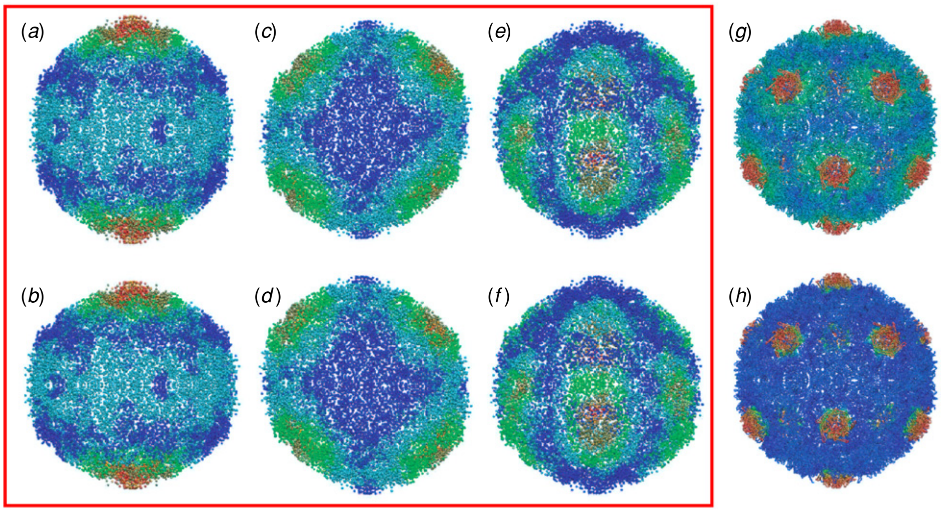


Figure 3. Dynamics of the HK97 bacteriophage viral capsid using ANM (panels (*a*)–(*f*)) and GNM (panels (*g*) and (*h*)). (*a* and *b*) The coarse-grained, $n/6$ ANM of the Prohead II form of this viral capsid represents the 107 520 residues by 17 920 nodes. Deformations along the slowest (global) mode indicate an extension/compression at opposite poles. The structure is colored such that the most mobile residues are red and least mobile are blue. (*c* and *d*) Representative deformations due to next slowest (degenerate) mode indicate four equally spaced most mobile regions, rather than two in the slowest mode. (*e* and *f*) This trend of increasing localization and number of patches of mobile residues continues as one examines higher modes. Representative deformations due to the next slowest (degenerate) mode illustrate a more complicated motion and begin to suggest a mechanism for maturation of the viral capsid. (*g*) GNM has been calculated for the entire HK97 viral capsid containing 107 520 residues. The linear combination of the slowest 11 modes, weighted by their eigenvalues, indicates that the 12 pentamers are the most mobile (red) regions. (*h*) The first non-degenerate (i.e. icosahedrally symmetric) GNM mode (number 31) also distinguishes the pentamers by their high mobility.

We note that the above-described slow/global modes lack spherical and/or icosahedral symmetry, and as a result, they do not correlate, on an individual basis, with the overall structural change (from an approximately spherical to an icosahedrally symmetric shape) undergone by the capsid during its maturation. Comparison of the individual ANM modes with the experimentally observed structural change points to the role of a small subset of icosahedrally symmetric (non-degenerate) modes in effectuating the structural changes [4, 5].

Figure 3(*h*) illustrates the mobilities in the first (slowest) non-degenerate mode computed with the GNM for the full (all residues) capsid. This mode (mode 31) is icosahedrally symmetric and also identifies the 12 pentamers as the most mobile regions. However, because the frequency of mode 31 is at least three times that of the first 11 modes, its fluctuations are more localized and its contribution to the potential structural fluctuations is small relative to that of these slower, asymmetric modes.

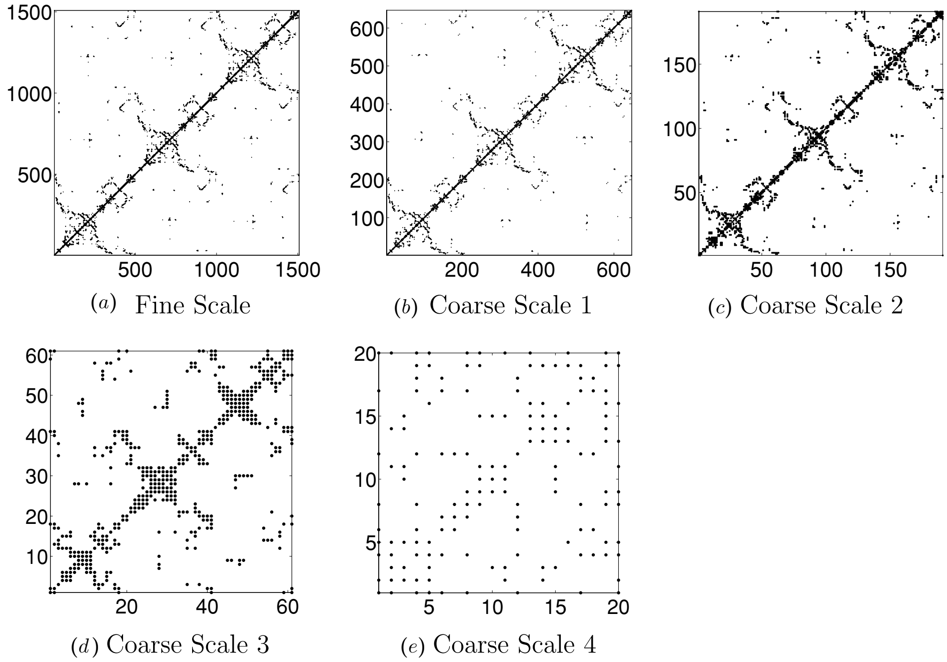
As shown by this study, HCG-based methods can be useful in exploring large systems. However, the parameters associated with HCG are arbitrarily set. For example, the level of HCG is chosen beforehand and the inclusion/exclusion of bonded/non-bonded neighbors that are equally distant in space cannot be justified in general. Similarly, in the case of RTB [20] or BNM [21] approaches, all atomic, or residue, level information is lost and substructures that may contain internal degrees of freedom—some of which functional—are assumed to move as a rigid block. The choice of these rigid blocks potentially affects or biases the predicted motions. To this end, in the next section, we introduce a novel hierarchical clustering algorithm that provides an efficient means of exploring the collective dynamics, with minimal loss in accuracy. The approach is based on a Markov process description of the communication (or affinity) between interacting residues and has the additional advantage of including residue specificity via consideration of atom–atom interactions.

5. Hierarchical coarse graining of the protein topology

It is natural to consider the elastic net in terms of graph theory and undertake random walks [36, 37]. In particular, each residue in the protein is seen as a node in a weighted, undirected graph and the edges in the graph represent interactions between residues. Using a random walk analogy, we can build a hierarchical, coarse-grained representation of the protein graph. The adjacent levels of the hierarchy are connected by assuming that a random walk undertaken at the coarse scale induces a random walk, simultaneously, at the fine scale. The passage between the two levels is ensured by

$$\boldsymbol{\pi} = \boldsymbol{K}\boldsymbol{\delta}, \tag{14}$$

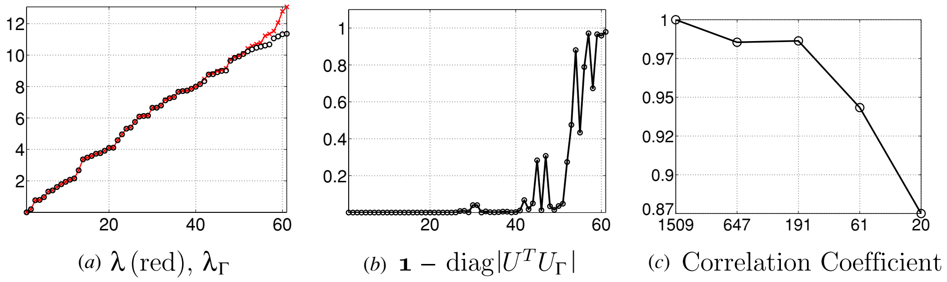
where $\boldsymbol{\delta}$ is an unknown probability distribution of length m, $\boldsymbol{K}$ is an $n \times m$ non-negative *kernel* matrix whose columns are probability distributions that each sum to 1 and m is low dimensional ($m \ll n$). The kernel matrix acts as an *expansion* operator, mapping a low-dimensional distribution $\boldsymbol{\delta}$ to a high-dimensional probability distribution $\boldsymbol{\pi}$. Furthermore, if the random walk reaches an equilibrium distribution $\boldsymbol{\delta}$ at the coarse scale, it will correspond to reaching a stationary distribution $\boldsymbol{\pi}$ at the fine scale. The equilibrium distribution $\boldsymbol{\pi}$ at the fine



(*a*) Fine Scale (*b*) Coarse Scale 1 (*c*) Coarse Scale 2

(*d*) Coarse Scale 3 (*e*) Coarse Scale 4

Figure 4. Contact matrix hierarchy for the influenza virus (PDB: 2HMG). The contact matrices, which are otherwise real valued, are shown here as dot plots to highlight the similarity in the structure of the matrix across the hierarchy. The sizes of the contact matrices are (*a*) 1509 × 1509, (*b*) 647 × 647, (*c*) 191 × 191, (*d*) 61 × 61 and (*e*) 20 × 20.



(*a*) λ (red), λ_Γ (*b*) $\mathbb{1} - \text{diag}|U^T U_\Gamma|$ (*c*) Correlation Coefficient

Figure 5. Hierarchical eigensolver results (*a*) comparing eigenvalues λ_Γ from a direct decomposition of the Γ (circles) with multi-scale eigensolver spectrum λ (red line). For the direct eigen decomposition, we use the Matlab program `svds.m` which invokes the compiled ARPACKC routine [39], with a default convergence tolerance of 1×10^{-10}. (*b*) Eigenvector mismatch: $\mathbb{1} - \text{diag}(|U^{\mathrm{T}} U_\Gamma|)$, between eigenvectors U derived by the multi-scale eigensolver and the direct decomposition of U_Γ. (*c*) Comparing the mean-square fluctuations of the residues at the finest scale with those obtained from successive levels of coarse graining.

scale is analytically determined by the diagonal elements of the Kirchhoff matrix $\mathbf{\Gamma}$. More details will be presented in a forthcoming paper.

Next, knowing $\boldsymbol{\pi}$ we can solve the above equation by an expectation-maximization type algorithm [37, 38]. Knowing $\boldsymbol{K}$ and $\boldsymbol{\delta}$, we can then derive a coarse-scale symmetric, real-valued affinity matrix $\widetilde{\boldsymbol{A}}$

$$\widetilde{\boldsymbol{A}} = (\text{diag}(\boldsymbol{\delta})\boldsymbol{K}^{\mathrm{T}})(\text{diag}(\boldsymbol{K}\boldsymbol{\delta})^{-1})(\boldsymbol{K}\text{diag}(\boldsymbol{\delta})), \qquad (15)$$

which maintains the same contact topology as the affinity matrix $\boldsymbol{A}$ corresponding to the high-resolution (single-residue-per-node) model. Here $\boldsymbol{A}$ is derived from the negative of the Kirchhoff matrix $\mathbf{\Gamma}$, with the diagonal elements set to zero.

For the purpose of demonstration, we build a five-level coarse-grain hierarchy on the protein influenza virus (PDB: 2HMG) having 1509 residues. The structure of the fine-scale affinity matrix $\boldsymbol{A}$ is shown in figure 4(*a*). Recursive application of the hierarchical algorithm gives rise to coarse-scale affinity matrices shown in figures 4(*b*)–(*f*). The affinity matrices, which are otherwise real valued, are shown here as *dot* plots. Note the *similarity* in the structure of the affinity matrix across the hierarchy. This suggests that the dynamics at the fine scale can be related to the dynamics at the coarse scale.

Indeed, the fluctuation dynamics are derived from the topology. In particular, diagonalizing the Kirchhoff matrix gives the modes of vibration (equation (10)). But the Kirchhoff matrix at the fine scale can be very large and its eigen decomposition can be expensive even if we are interested in only in a subset of modes. But given the hierarchical representation, using the coarse-scale affinity matrix $\widetilde{\boldsymbol{A}}$ we can build a corresponding Kirchhoff matrix $\widetilde{\mathbf{\Gamma}}$ which is much smaller in size compared to $\mathbf{\Gamma}$ at each level of the hierarchy. Furthermore, by performing eigen decomposition of the Kirchhoff matrix at the coarsest level of hierarchy and propagating information to the fine scale, we achieve a *hierarchical GNM*. More details will be presented in a forthcoming paper.

In figure 5(*a*), we compare the eigen spectrum λ obtained from propagating information from level 4 of the hierarchy (red line) with the spectrum λ_Γ obtained by a direct eigen decomposition of the fine-scale Kirchhoff matrix $\mathbf{\Gamma}$ (circles). There is an excellent agreement in the leading (slow) eigenvalues. For a quantitative comparison between the eigenvectors, we plot in figure 5(*b*) the following measure: $\mathbf{1} - \mathrm{diag}(|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}_\Gamma|)$, where $\boldsymbol{U}$ is the matrix of eigenvectors obtained by the multi-scale approximation, $\boldsymbol{U}_\Gamma$ is the matrix obtained from a direct eigen decomposition of the fine-scale Kirchhoff matrix $\mathbf{\Gamma}$ and $\mathbf{1}$ is a vector of all ones. The relative error plot demonstrates a close match, except for the last few eigenvectors, which suggests that the information propagation over the hierarchy has not clearly separated them from other directions. However, the contribution of these modes to the overall fluctuation dynamics is relatively small, as the contribution of each mode scales with its inverse eigenvalue. To demonstrate the effect of coarse graining on fluctuation dynamics, we compare the mean-square fluctuations obtained from different levels of the hierarchy with those derived *before* coarse graining. As shown in figure 5(*c*), a correlation coefficient value of 0.87 is achieved in the mean-square fluctuation values after mapping the structure of 1509 residues into a representative network of 20 nodes. Thus, the fluctuation behavior of individual residues is accurately maintained despite a significant reduction in the complexity of the examined network.

6. Conclusion

In this paper, we presented recent applications of the GNM to a series of enzymes as well as large structures such as the HK97 bacteriophage viral capsids. Understanding the dynamics of large protein structures can be computationally challenging and to this end, we presented a new approach for building a hierarchical, reduced rank representation of the protein topology and consequently the fluctuation dynamics. The new methodology permits the network representation in terms of models that are lower in complexity, without any distinguishable change in the frequency distribution and shape of the dominant modes. The reduction in complexity with minimal loss in accuracy illustrated here for the influenza virus supports the utility of the new methodology for exploring the dynamics of supramolecular assemblies.

References

[1] Tama F *et al* 2003 *Proc. Natl Acad. Sci. USA* **100** 9319
[2] Wang Y *et al* 2004 *J. Struct. Biol.* **147** 302
[3] Tama F and Brooks C L III 2002 *J. Mol. Biol.* **318** 733
[4] Tama F and Brooks C L III 2005 *J. Mol. Biol.* **345** 299
[5] Rader A J *et al* 2005 *Structure* **13** 413
[6] Ma J 2005 *Structure* **13** 373–80
[7] Delarue M and Dumas P 2004 *Proc. Natl Acad. Sci. USA* **101** 6957–62
[8] Tama F *et al* 2004 *J. Struct. Biol.* **147** 315–26
[9] Hinsen K *et al* 2005 *Biophys. J.* **88** 818–27
[10] Hang Z *et al* 2003 *Biophys. J.* **84** 3583–93
[11] He J *et al* 2003 *J. Chem. Phys.* **119** 4005–17
[12] Tatsumi R *et al* 2004 *J. Comput. Chem.* **25** 1995–2005
[13] Bahar I *et al* 1997 *Folding Des.* **2** 173
[14] Haliloglu T *et al* 1997 *Phys. Rev. Lett.* **79** 3090
[15] Hinsen K 1998 *Proteins* **33** 417
[16] Doruker P *et al* 2000 *Proteins: Struct. Funct. Genet.* **40** 512–24
[17] Atilgan A R *et al* 2001 *Biophys. J.* **80** 505
[18] Tama F and Sanejouand Y H 2001 *Protein Eng.* **14** 1
[19] Doruker P *et al* 2002 *J. Comput. Chem.* **23** 119
[20] Tama F *et al* 2000 *Proteins* **41** 1
[21] Li G H and Cui Q 2002 *Bipohys. J.* **83** 2457
[22] Benkovic S J and Hammes-Schiffer S 2003 *Science* **301** 1196–202
[23] Daniel R M *et al* 2002 *Annu. Rev. Biophys. Biomol. Struct.* **32** 69–92
[24] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 534
[25] Bahar I and Jernigan R L 1997 *J. Mol. Biol.* **266** 195
[26] Kundu S *et al* 2002 *Biophys. J.* **83** 723
[27] Yang L-W and Bahar I 2005 *Structure* **13** 893–904
[28] Bartlett G 2002 *J. Mol. Biol.* **324** 105–21
[29] Porter C *et al* 2004 *Nucl. Acids Res.* **32** D129–33
[30] Chen C C and Herzberg O 1992 *J. Mol. Biol.* **224** 1103–13
[31] Khan A R *et al* 1998 *Biochemistry* **37** 16839–45
[32] Kim M K *et al* 2003 *J. Struct. Biol.* **143** 107–17
[33] Conway J F *et al* 2001 *Science* **292** 744
[34] Wikoff W R *et al* 2000 *Science* **289** 2129
[35] Lata R *et al* 2000 *Cell* **100** 253
[36] Chennubhotla C and Jepson A 2005 *NIPS* **17** 273–80
[37] Chennubhotla C 2004 *PhD Thesis* Department of Computer Science, University of Toronto, Canada
[38] McLachlan G J and Basford K E 1988 *Mixture Models: Inference and Applications to Clustering* (New York: Dekker)
[39] Lehoucq R B *et al* 1996 *ARPACK User Guide* (Department of CAM, Rice University)