Available online at www.sciencedirect.com
SCIENCE DIRECT®

Polymer 45 (2004) 649–657

polymer

www.elsevier.com/locate/polymer

Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions

Ozge Kurkcuoglu[a], Robert L. Jernigan[b], Pemra Doruker[a,*]

[a]*Department of Chemical Engineering and Polymer Research Center, Bogazici University, Bebek 34342, Istanbul, Turkey*
[b]*Baker Center for Bioinformatics and Biological Sciences, Iowa State University, Ames, IA 50011-3020, USA*

Received 22 April 2003; received in revised form 16 July 2003; accepted 17 July 2003

Abstract

We perform a mixed coarse-graining approach in a normal mode analysis of protein motions, which enables the modeling of a protein's native conformation with different regions having low and high resolution. As a result, the dynamics of the interesting functional parts within a supramolecular assemblage can be analyzed at high resolution, while the remainder of the structure is represented at poorer resolution, thus keeping the total number of nodes in the system sufficiently low for computational tractability. Our results indicate that the vibrational dynamics of specific components in a large multi-subunit protein are best described by retaining all the components of the structure, whether at higher or lower resolution. It is also shown that similar frequency distributions are obtained for different proteins and at different levels of coarse-graining, at the lower end of the spectrum, where the most significant slowest motions occur.

Keywords: Low resolution models; Collective dynamics; Domain motion

1. Introduction

Biological functions of proteins are controlled by their cooperative motions, typically involving large domains of the structure. The structural dynamics of proteins are most commonly studied by molecular dynamics (MD) simulations and normal mode analyses (NMA) using fully-atomistic, empirical potentials [1–3]. The basic approach in these methods is to focus on the low frequency/ large amplitude modes that are expected to be relevant to function. However, these atomistic techniques become computationally inefficient, and even inadequate, for the largest systems. On the other hand, computational studies on large biological 'supramolecular assemblages' [4], composed of multiple macromolecular units such as the ribosome, are recently becoming of great interest, due to the growing numbers of large structures available from X-ray crystallography and cryoelectron microscopy (EM). For these reasons, coarse-grained protein models and simplified force fields have emerged as powerful, efficient tools to describe the molecular motions of large proteins [5–10].

The fluctuation dynamics of proteins can be effectively described by a coarse-grained normal mode analysis, using the elastic network model. In this model, the nodes of the network are usually taken as the α-carbon sites of the residues and the linkers are harmonic springs between sufficiently close residue pairs. This model is named the Gaussian network model (GNM) in the scalar version, i.e., if the fluctuations are assumed to be isotropic with no directional preferences [5]. The studies have shown that GNM gives results in excellent agreement with X-ray crystallographic Debye–Waller factors [5,6,11,12], H/D exchange free data [13] and the order parameters from NMR-relaxation measurements [14]. However, in reality, it is known that the residue fluctuations are not in general isotropic [15]. This anisotropy can have a great importance in the biological function of the protein. An extension of the GNM, called the anisotropic network model (ANM),

Abbreviations: 3D, Three-dimensional; ANM, Anisotropic network model; EM, Cryoelectron microscopy; GAL, *Escherichia coli* β-galactosidase; GNM, Gaussian network model; HA, Influenza virus hemagglutinin; MD, Molecular dynamics simulations; Ms, Mean square; NMA, Normal mode analysis; XDH, Xanthine dehydrogenease.

* Corresponding author. Tel.: +90-212-3581540x2365; fax: +90-212-2575032.

E-mail address: doruker@boun.edu.tr (P. Doruker).

0032-3861/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.10.071

incorporates the anisotropic effects on fluctuation dynamics [10], yielding the 3-dimensional directions of the motion. The large-scale collective motions obtained from elastic network models are closely related to those extracted from atomistic molecular dynamics studies [16], which lends support to the effectiveness of these coarse-grained models in the analysis of the structure–function relationship of proteins and their complexes.

So far single-node-per-residue representation of the folded protein structure has been commonly used in elastic network models, providing satisfactory results in terms of collective dynamics. Recent ANM studies have indicated that the low-frequency, domain motions can still be obtained with reasonable accuracy for more coarsely grained systems, wherein a single node can represent 2 up to 40 residues [17,18]. Extreme levels of coarse-graining actually reduce the computational time by more than two orders of magnitude thereby making the analysis of large systems feasible. Moreover, by using elastic network models, the vibrational dynamics of proteins can be obtained by reproducing the density distribution of the low-resolution protein structures obtained from EM [19,20] or the overall shape of the molecule on a regular lattice [21].

In this study, we introduce a mixed coarse-graining approach to elastic network model, where the 'interesting' or functional parts of the structure are modeled at a higher resolution than the remainder of the structure, which is represented at lower resolution, in less detail. It will be shown that the collective dynamics of proteins can still be extracted with reasonable accuracy using the mixed-resolution model. As a result, the number of interaction sites can be kept at a reasonable level, so that the normal mode analysis can still be performed with high computational efficiency for large systems. By using such a model, it is possible to focus on the details of interacting interfaces between subunits with ligands.

2. Materials and method

2.1. Proteins

Two relatively large proteins are chosen in this study, as an extension of an earlier work [18], namely hemagglutinin (HA) and β-galactosidase (GAL) with the respective Protein Data Bank [22] file names 2HMG and 1DP0. Influenza virus hemagglutinin (HA) is an integral membrane glycoprotein. The X-ray structure of HA has been determined by Wiley and co-workers at a resolution of 3 Å [23,24]. HA is a cylindrically shaped homo-trimer, comprising 1509 residues in total. Each monomer is composed of two separate chains, HA1 (residues 1–328) and HA2 (329–503) that are linked by two disulfide bridges. The three monomers are assembled into a central coiled coil that forms the stem-like domain, while the three globular heads contain the receptor binding sites. The X-ray structure of *Escherichia coli* β-galactocidase (GAL) has been determined by Matthews and co-workers [25] at 1.7 Å resolution. GAL is a tetramer having four identical subunits, with each monomer having 1023 residues. Its approximate dimensions are 175 Å × 135 Å × 90 Å. The biological function of this enzyme is to hydrolyze lactose and other β-galactosides into monosaccharides.

2.2. Elastic network model

Elastic network models are constructed based on the folded structure of proteins that presumably approximates the minimum energy conformation, i.e. the native state. In the original model, each residue is represented by a coarse-grained node located at its α-carbon position [5,10]. Then the close neighboring residues in the three-dimensional (3D) structure are connected by harmonic springs.

Explicitly, the total potential energy for a system of N residues is a summation over all harmonic interactions of (i,j) pairs that fall within the cutoff distance of r_c.

$$V = (\gamma/2)\sum_i \sum_j h(r_c - R_{ij})(\Delta\mathbf{R}_j - \Delta\mathbf{R}_i)^2 \tag{1}$$

Here $\mathbf{R}_i$ and $\Delta\mathbf{R}_i$ are, respectively, the position and fluctuation vectors of node i $(1 \le i \le N)$. R_{ij} is the distance between nodes i and j, and $h(x)$ is the Heaviside step function [$h(x) = 1$ if $x \ge 0$, and zero otherwise]. The only adjustable parameter in this model is the force constant, γ, which is taken to be identical for all bonded and non-bonded interactions, as was originally done in Tirion's work [26].

In the following, the residue masses will be taken into account, which were considered identical (equal to one) in earlier elastic network calculations. Our formulation is based on the classical normal mode analysis as applied to proteins [27–29].

For the elastic network model, the mass-weighted fluctuations can be defined as $q_k = \sqrt{m_i}\Delta x_i$, $q_{k+1} = \sqrt{m_i}$ Δy_i, and $q_{k+2} = \sqrt{m_i}\Delta z_i$, where Δx_i, Δy_i, and Δz_i are the Cartesian coordinates of $\Delta\mathbf{R}_i$. Therefore, $\mathbf{q}$ is a $3N$ dimensional vector of fluctuations. The potential energy can be approximated in quadratic form around the minimum energy conformation of the protein indicated by $\mathbf{q} = 0$.

$$V = \frac{1}{2}\sum_{i=1}^{3N}\sum_{j=1}^{3N} \left.\frac{\partial^2 V}{\partial q_i\,\partial q_j}\right|_{q=0} q_i q_j \tag{2}$$

The Lagrangian (L), which is the kinetic energy minus the potential energy, can be written in compact form.

$$L = \frac{1}{2}\dot{\mathbf{q}}^{\mathrm{T}}\dot{\mathbf{q}} - \frac{1}{2}\mathbf{q}^{\mathrm{T}}\mathbf{F}\mathbf{q} \tag{3}$$

Here, $\mathbf{F}$ $(3N \times 3N)$ is the Hessian, or force constant matrix, whose elements are the second derivatives of the potential energy with respect to the mass-weighted coordinates in Eq. (2). T stands for transpose and dot indicates derivative with respect to time.

The real-symmetric matrix **F** can be diagonalized to obtain the canonical form,

$$\mathbf{S}^{\mathrm{T}}\mathbf{F}\mathbf{S} = \boldsymbol{\lambda} \tag{4}$$

where $\boldsymbol{\lambda}$ $(3N \times 3N)$ is a diagonal matrix with the diagonal elements corresponding to the eigenvalues λ_1 to λ_{3N}. **S** $(3N \times 3N)$ is an orthogonal matrix with $\mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{I}$ (**I** being the identity matrix). The columns of **S** are the normalized eigenvectors.

This orthogonal transformation defines the normal coordinates for the elastic network, given by the $3N$ dimensional vector **Q**.

$$\mathbf{Q} = \mathbf{S}^{\mathrm{T}}\mathbf{q} \tag{5}$$

Lagrange's equation of motion in canonical form is:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\mathrm{Q}}_i}\right) = \left(\frac{\partial L}{\partial Q_i}\right) \tag{6}$$

As a result, the dynamics of the $3N$ normal coordinates are obtained as

$$Q_i = A_i \cos(\omega_i t + \varepsilon_i) \tag{7}$$

Here each normal mode has a frequency of $\omega_i = (\lambda_i)^{1/2}$ with phase ε_i and amplitude A_i determined by the initial conditions. It should be noted that six normal modes describing rotational and translational motion of the molecule have zero frequencies, with the remaining $(3N - 6)$ being the internal degrees of freedom.

According to equipartition law, each normal mode has a time-average potential energy of $(1/2)k_{\mathrm{B}}T$ relative to the minimum value.

$$\langle Q_i^2 \rangle = \frac{k_B T}{\omega_i^2} \tag{8}$$

Here k_{B} is the Boltzmann constant and T is the absolute temperature. In terms of the mass-weighted coordinates the mean square (ms) fluctuations are

$$\langle {q_i}^2 \rangle = k_{\mathrm{B}} T \sum_{k=1}^{3N-6} \left(\frac{S_{ik}}{\omega_k}\right)^2 \tag{9}$$

As a result, the ms fluctuations of each residue, $\langle \Delta\mathbf{R}_i^2 \rangle$, can also be calculated, which are usually comparable to the experimental temperature factors (B_i) according to the relationship:

$$B_i = (8\pi^2/3)\langle \Delta\mathbf{R}_i^2 \rangle \tag{10}$$

The only unknown parameter γ in the elastic network model is implicit in ω_k, i.e. $\omega_k^2 = \gamma \omega'^2_k$. The value of γ is determined so as to match the experimental and theoretical ms fluctuations averaged over all residues.

2.3. Mixed coarse-graining procedure of the elastic network model

The mixed-coarse graining procedure is composed of three main parts: (i) uniform coarse-graining of the protein structure at a series of hierarchical levels by retaining N, $(N/2)$, $(N/5)$, $(N/10)$, $(N/20)$ and $(N/40)$ residues of the original X-ray structure; (ii) establishing the relation between the force constant and the cutoff radius for these different levels of coarse-graining; and (iii) coarse-graining the 'interesting' parts at a higher resolution and the rest of the structure at a lower resolution.

In this study, each chain or monomer of a multi-subunit protein is coarse-grained separately along its chain backbone, starting with the first residue. For example, HA is composed of three HA1 and three HA2 chains, which are separately transformed into linear chains of coarse-grained nodes, each containing the reduced number of residues, taken in a sequentially linear way.

Cutoff radius as a function of segment length. The uniformly coarse-grained structure can be viewed as a collection of s coarse-grained segments, each containing n residues. Thus the total number of residues in the original structure is $N = sn$. This closely resembles the way in which polymer chains have been coarse-grained to construct equivalent chain models for simpler conformational models of random coils. One example of this is the Kuhn statistical link similarly made up of n units of the chain [30].

In earlier work [18], the relationship between the radius of gyration (R_{G}) and the segment length was shown to be very similar for three large proteins up to $n = 40$, with the functional form given by

$$R_{\mathrm{G}} = an^b \tag{11}$$

And the parameters for $1 < n \leq 40$ were found to be $a = 1.778$ and $b = 0.595$ from a fit to the average behavior for three proteins, namely HA, GAL and xanthine dehydrogenase (XDH) [18]. Due to the necessarily longer interaction ranges of the renormalized sites, the cutoff radius was adjusted as the sum of the renormalized radius of each site plus the invariant contact distance R_0 between the sites,

$$r_{\mathrm{c}} = 2R_{\mathrm{G}} + R_0 \tag{12}$$

where R_{G} is calculated according to Eq. (11). R_0 was determined as 13 Å in earlier work for $n = 1$, which is also adopted in this work for consistency [10]. We will also adopt the functional form and parameters given by Eqs. (11) and (12) for the determination of r_{c} for $n \geq 2$.

Fluctuations of coarse-grained nodes. We have to determine the effective temperature factor B_i for a group of residues in order to calculate the effective force constant for different segment lengths. The fluctuation of the center of mass of a group of n residues that have different molecular weights and fluctuations in Cartesian coordinates (not in mass-weighted coordinates) can be defined as

$$\Delta\mathbf{R}_{cm} = \frac{m_1\Delta\mathbf{R}_1 + \cdots + m_n\Delta\mathbf{R}_n}{\sum_{i=1}^{n} m_i} \quad (13)$$

Neglecting the cross-correlations terms between residue fluctuations, it can be shown that

$$\langle\Delta\mathbf{R}_{cm}^2\rangle_n = \frac{\sum_{i=1}^{n} m_i^2\langle\Delta\mathbf{R}_i^2\rangle}{\left(\sum_{i=1}^{n} m_i\right)^2} \quad (14)$$

where summations are performed over the n residues in the collective unit. In a similar fashion, based on $B_{cm,n} = (8\pi^2/3)\langle\Delta\mathbf{R}_{cm}^2\rangle_n$, the temperature factor for a node composed of n residues can be approximated as

$$B_{cm,n} = \frac{\sum_{i=1}^{n} m_i^2 B_i}{\left(\sum_{i=1}^{n} m_i\right)^2} \quad (15)$$

If all residue masses were identical, we would similarly obtain

$$B_{cm,n} = (1/n^2)\sum_{i=1}^{n} B_i$$

Force constant as a function of segment length. After fixing the cutoff radius for the segment length $n \geq 2$, the effective force constant for the renormalized nodes can be adjusted by maximizing the match between the average values of the ms fluctuations predicted by theory and derived for coarse-grained nodes using Eq. (15) and experimental B_i. As the level of coarse-graining increases, the distance between the nodes and the mass of the nodes increases, and as a result the renormalized force constants become stronger, which will be discussed in the next section.

3. Results and discussion

3.1. Uniform coarse-graining

Force constants. In this study, each coarse-grained node is placed at the center of mass of its constituent residues/atoms. In addition, the residue center of mass is calculated considering all atoms except hydrogens. From a comparison of experimental and theoretical fluctuations, the force constant is determined for different segment lengths (n). Fig. 1 exhibits the relation between the force constant and the cutoff distance, i.e. the size of coarse-grained nodes. HA and GAL exhibit almost the same behavior up to $n = 20$.

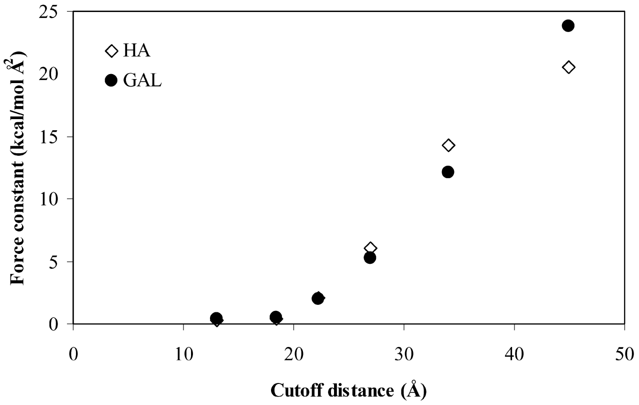

Fig. 1. Relationship between the force constant and the cutoff distance for the two proteins hemagglutinin (HA) and β-galactosidase (GAL) for the elastic network models.

Frequency distribution. The density of the vibrational frequencies, $g(\omega)$, is defined as the number of modes per frequency range divided by the total number of modes. In Fig. 2(a), the frequency distributions of HA and GAL are shown for $n = 1$, which are very similar particularly in the low frequency range. This is consistent with previous findings on the behavior of smaller proteins (with at most 375 residues) that fall under a universal curve [31]. Fig. 2(b) and (c) show $g(\omega)$ distributions at different levels of coarse-graining for HA and GAL, respectively. The distributions look quite similar, although there is some scatter at the high coarse-graining levels of HA, which is a smaller protein compared to GAL. In Fig. 2(d), $G(\omega)$, which is the cumulative density of modes up to frequency ω, is plotted as a function of ω for the low-frequency region (comprising about 3–4% of the total number of modes). Again the two curves for HA and GAL are very similar at $n = 1$. On the logarithmic curve, an exponent $k \cong 2.2$ is found for $G(\omega) \sim \omega^k$, which is close to the value of 2 found from atomistic normal mode analysis on smaller proteins [31]. Previous elastic network model results reported on smaller proteins (containing up to 164 residues) indicate $k \sim 1.63$ [11]. In this work, the lowest frequency is found as 1 cm^{-1} for HA (0.7 cm^{-1} for GAL) at $n = 1$, which depends on the value of γ extracted using Eqs (9) and (10).

3.2. Mixed coarse-graining

In the mixed coarse-graining procedure, the interesting parts of the system are analyzed at higher resolution, whereas the remaining parts are modeled at lower resolution. Higher resolution corresponds to smaller values of segment length (n), cutoff radius and force constant. Since segments with at least two different lengths (n_1 and n_2) exist in the mixed system, the cutoff distance determining the range of interaction between node types 1

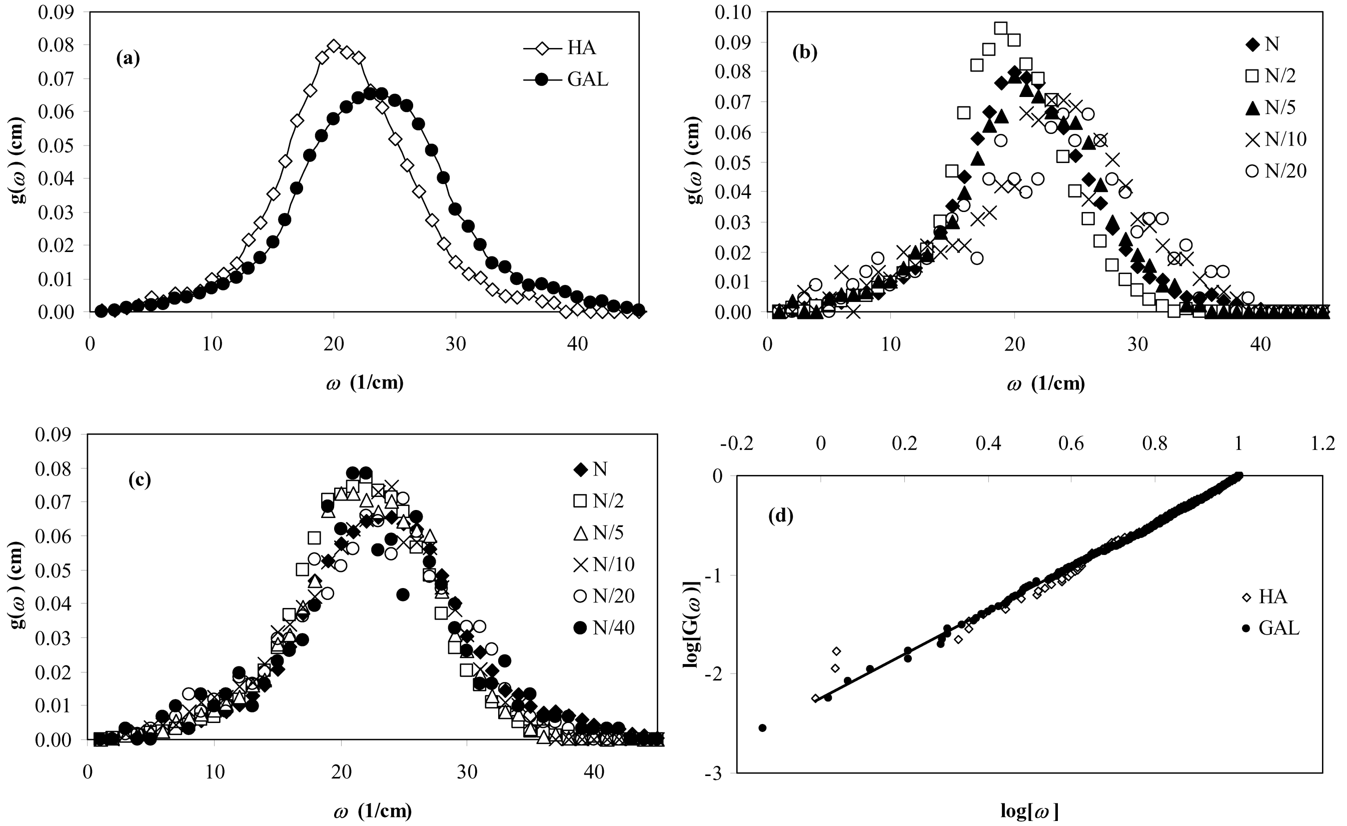

Fig. 2. Density of vibrational frequencies, $g(\omega)$, based on uniform coarse-graining for: (a) HA and GAL with $n = 1$, (b) HA with different segment lengths, (c) GAL with different segment lengths, and (d) HA and GAL at $n = 1$, showing a log–log plot of cumulative distributions, $G(\omega)$, at the low frequency end of the spectrum.

and 2 is adjusted as:

$$r_{c,12} = \left(\frac{r_{c,1}^3 + r_{c,2}^3}{2} \right)^{1/3} \qquad (16)$$

Here the different size nodes are considered to have the same density, with the masses of nodes being proportional to the volume of spherical residues. As a result, the force constant acting between node types 1 and 2 can be determined from Fig. 1, which corresponds to the new cutoff value, $r_{c,12}$.

Mean-square fluctuations. In Fig. 3(a), the temperature factors predicted by theory and experiments are plotted for HA. The overall agreement seems reasonable, as in previous work with HA, where the residue masses were taken to be uniform [32]. The incorporation of the different residue masses in the elastic network model does not seem to affect the ms fluctuations, resulting in almost identical behavior as seen in the previous study [32].

Fig. 3(b)–(e) compares the temperature factors from mixed coarse-grained calculations with those of uniform coarse-graining with $n = 1$. HA is a homo-trimer with each monomer consisting of two chains: HA1 and HA2. In our first trial, one of the HA1 chains is modeled at relatively higher resolution, i.e. it contains all 328 residues as the nodes, whereas the remaining five chains are modeled at lower resolution, either $n = 5$ or 20. Fig. 3(b) shows the results for the whole protein at $n = 1$–5 level. There is a slight level difference between the average ms fluctuations of the high and low-resolution parts of the model, specifically the $n = 1$ portion exhibits lower ms fluctuations than the experimental values, whereas the reverse behavior is observed for the rest. This may result from the

Table 1
Correlation coefficients for mixed and uniform coarse-graining results

Protein/monomer	Segment lengths n	B-factors	Slowest mode
HA1	1, 5	0.92	0.96
HA1	1, 20	0.40 (0.60[a])	0.77
HA1 only[b]	1	0.26	0.47
HA2	1, 5	0.80	0.94
HA2	1, 20	0.61	0.95
HA2 only[b]	1	0.03	0.04
GAL	2, 10	0.92	0.97
GAL	2, 20	0.77	0.91
GAL	2, 40	0.74	0.88
GAL only[b]	2	0.16 (0.77[a])	0.60

The uniform coarse-graining results are for $n = 1$ (HA) and $n = 2$ (GAL).

[a] A few high peaks as in Fig. 3(d) are removed to improve the correlations.

[b] Only a single chain (HA1 or HA2) or a single monomer from GAL is modeled by neglecting the remaining chains.

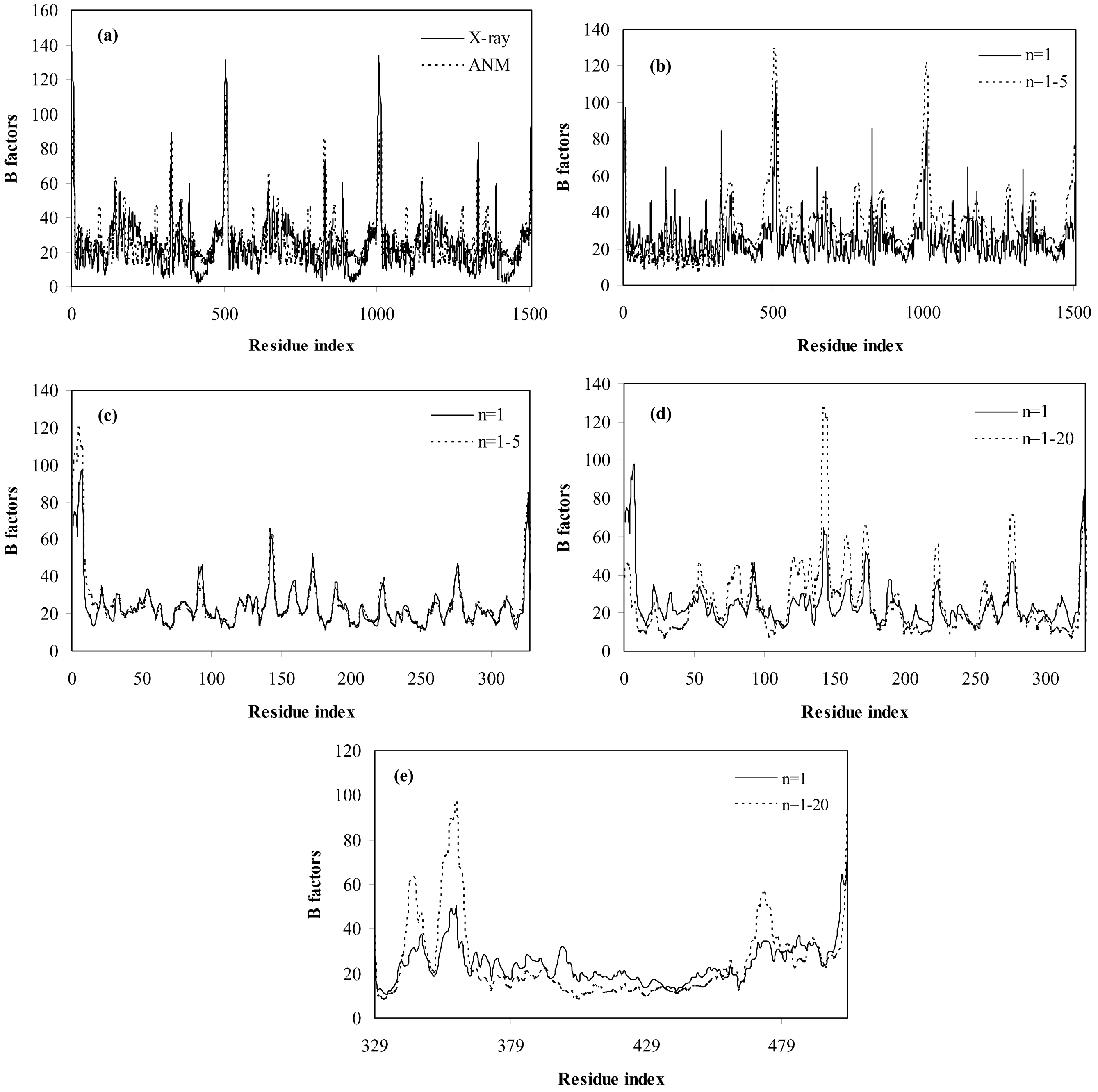

Fig. 3. Dependence of B-factors on residue index (a) original HA results with $n = 1$ in comparison with experimental B-factors, (b) HA at $n = 1, 5$ level (HA1 at higher resolution), (c) HA at $n = 1, 5$ level showing only HA1 residues, (d) HA at $n = 1, 20$ level showing only HA1 residues, (e) HA at $n = 1, 20$ level showing only the HA2 residues that are modeled at higher resolution.

assumptions considered in the formulation of Eq. (15), and the corresponding force constants extracted. However, if we concentrate on the high-resolution region separately and scale its fluctuations so as to match the average value of the uniform coarse-graining, excellent agreement is observed in Fig. 3(c) for $n = 1, 5$ level. Fig. 3(d) shows similar results for the case $n = 1, 20$. Finally, one of the HA2 chains (175 residues) is modeled at high resolution and Fig. 3(e) shows the case $n = 1, 20$.

Table 1 (third column) gives the linear correlation coefficients between several mixed coarse-graining calculations and uniform coarse-graining results for HA ($n = 1$), and GAL ($n = 2$). In this table, the correlation coefficients are also given for the case where only the high-resolution chain is modeled without incorporating rest of the protein at lower resolution. The correlation coefficients are significantly lower if the remaining chains are not considered, the worst case being HA2. This indicates the utility of the mixed-coarse graining procedure, which can be applied to supra-molecular structures, and demonstrated that in order

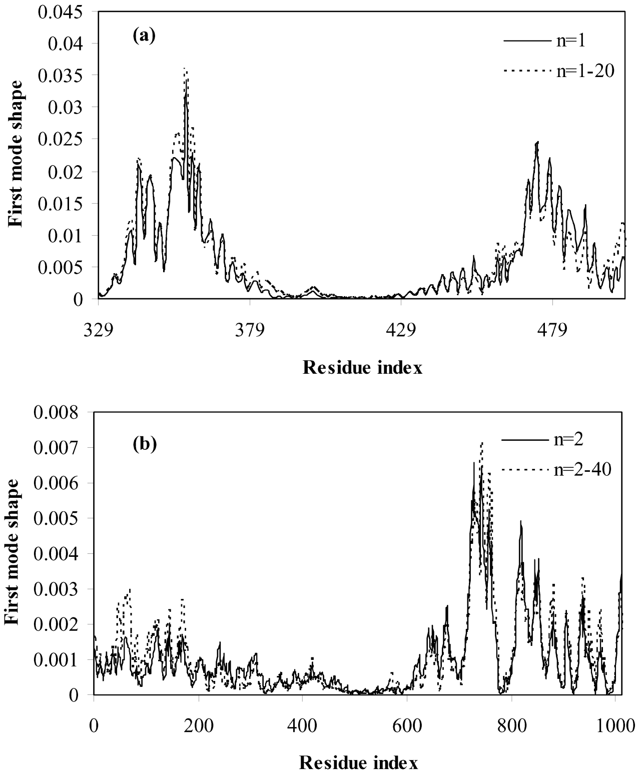

Fig. 4. Comparison of the mean square fluctuations due to slowest mode for (a) HA2 at $n = 1$ and 1,20 level, (b) first monomer of GAL at $n = 2, 40$ level of mixed-coarse graining.

to be effective, the entire structure needs to be included in the calculation.

Slow modes. The low-frequency, collective modes of motion are especially significant, since they are the ones most often related to protein function. Therefore, in this section we will focus on the determination of the slowest modes by the mixed coarse-graining approach. Fig. 4(a) and (b) show the residue fluctuations resulting from only the first, lowest frequency mode for HA2 ($n = 1, 20$) and GAL ($n = 2, 40$), respectively. Correlation coefficients for the slowest mode fluctuations are also given in Table 1 (last column), which are in general higher than the correlation of all modes (B-factors).

Alternative conformations of the molecule originating in the action of slowest modes need to be investigated in order to assure that the collective deformations are similar to the original motions. Fig. 5(a) shows the structure of HA ($n = 1$) with HA1 (left panel) and HA2 (right panel) chains colored in black. The first mode of motion, which is a twisting of the whole molecule along its cylindrical axis, is shown in Fig. 5(b). The left panel shows original calculations ($n = 1$), whereas the right panel is for the mixed coarse-grained model ($n = 1, 5$) with HA1 modeled at high resolution. In both panels, HA1 is colored black and its deformations are similar. Similarly, Fig. 5(c) indicates the bending deformation in the second mode with the HA2 chain modeled at higher resolution in a mixed system of $n = 1, 5$ (right panel), in comparison to uniform coarse-graining. The implications of these slow modes on the structure–function relationships of HA related to binding and membrane fusion have been discussed in detail by Isin et al. [32].

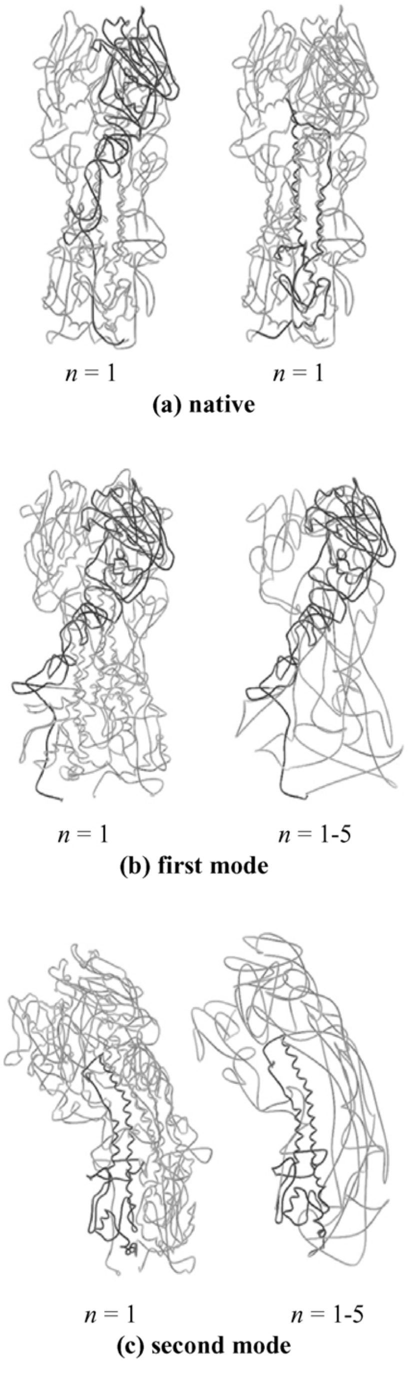

Fig. 5. Mixed coarse grained representations of HA: (a) The undeformed, native structure with HA1 (left) and HA2 (right) colored in black, (b) First mode: global twisting around the longitudinal axis. Uniform structure ($n = 1$, left), and mixed coarse-grained structure with high resolution HA1 ($n = 1, 5$, right), (c) second mode: bending motion of the whole molecule. Uniform structure ($n = 1$, left), and mixed coarse-grained structure with high resolution HA2 ($n = 1, 5$, right).

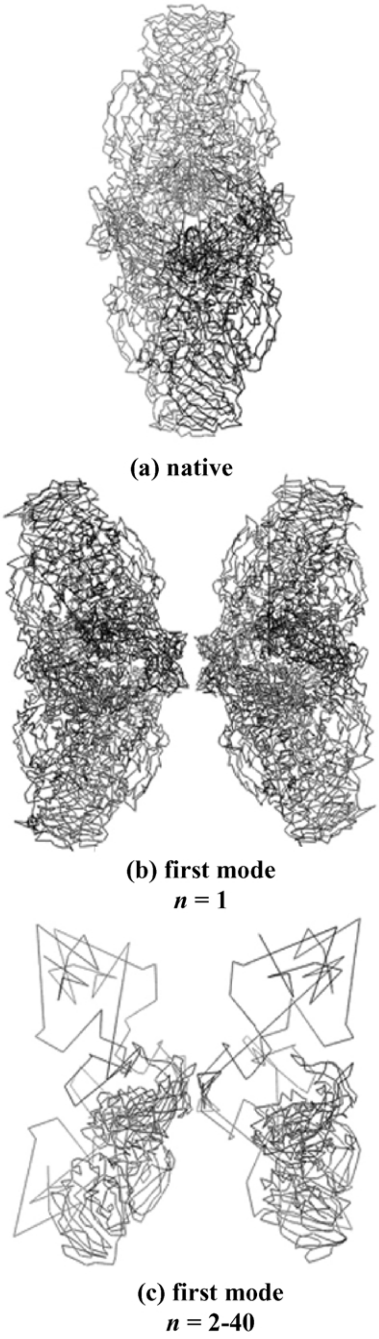

Fig. 6. Mixed coarse grained representations of the structure of GAL: (a) undeformed, native structure with one monomer colored in black, (b) first mode: bending along the activating interface of the protein for all residue system. Uniform coarse-graining with $n = 1$, (c) first mode for mixed coarse-graining at $n = 2, 40$ level.

The dominant correlated deformations in the first mode of GAL are compared in Fig. 6. Fig. 6(a) exhibits the side view of the protein in the native state (one monomer colored black). The alternative deformations from uniform coarse-graining ($n = 1$), which are shown in Fig. 6(b), indicate a bending of the whole molecule along its activating interface, as also indicated previously [18]. Similar deformations are obtained for mixed coarse-graining ($n = 2, 40$) in Fig. 6(c).

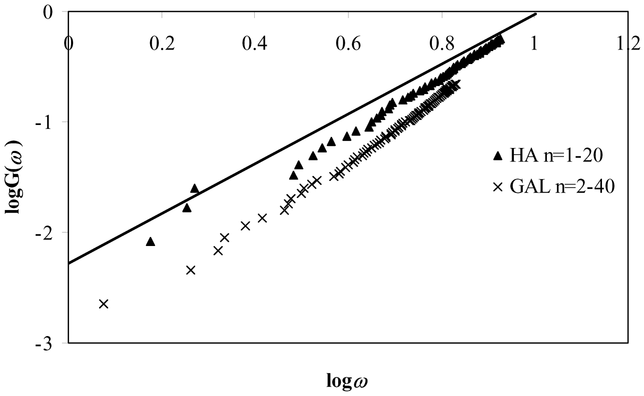

Fig. 7. Cumulative fraction of modes, $G(\omega)$, up to frequency ω for mixed coarse-grained systems at the low frequency end of the spectrum (logarithmic plot). The solid line, of slope 2.2, is based on the uniform coarse-graining results with $n = 1$. The slopes for the coarse-grained models exhibit relatively small deviations from the line shown.

Frequency distribution. Fig. 7 compares the cumulative frequency distributions, $G(\omega)$, from mixed coarse-graining calculations for HA ($n = 1, 20$) and GAL ($n = 2, 40$). The behavior including 70 slowest modes is shown since the low frequency range is important for our purposes. The straight line with a slope of 2.2 is drawn through the data points of uniform coarse-graining calculations at $n = 1$, i.e. the same line in Fig. 2(d). The slopes in mixed coarse-graining calculations fall in the range of $k \cong 2.2-2.7$ for $G(\omega) \sim \omega^k$. These distributions seem reasonable considering the extreme levels of coarse-graining employed.

4. Conclusions

Normal mode analysis is an established technique to study the fluctuation dynamics of proteins around their native conformations. Collective deformations or domain motions of large proteins can be effectively determined by NMA using coarse-grained potentials [6,7,33]. Lately, it has been shown that extreme levels of coarse-graining and low-resolution protein structures from EM yield satisfactory results in terms of collective motions, which permits the application of the elastic network model to extremely large proteins [17–20]. Based on these accomplishments, we have aimed here at introducing a mixed coarse-graining approach in the elastic network model, where different parts of the protein structure are represented at lower or higher resolutions. This approach enables us to focus on certain regions of the structure that are of interest without neglecting rest of the structure. As a result, a reduced number of nodes can be retained in NMA in more efficient computations and still obtain meaningfully similar results to those obtained from more detailed computations. Here we performed our computations with two quite large proteins, namely hemagglutinin and β-galactosidase. It is also possible to treat significantly larger supramolecular assemblages with this methodology.

The relatively large proteins (HA and GAL) exhibit similar frequency distributions, specifically in the low frequency region. These results are in conformity with previous NMA results on much smaller proteins, which have indicated universal behavior in terms of their frequency distributions [31]. We have shown that the cumulative density of modes up to frequency ω, scales as $G(\omega) = \omega^k$ with $k \sim 2.2$. This exponent is also very similar to the atomistic results on smaller proteins [31].

The relationship between the cutoff radius to be selected and the coarse-grained segment length has been reported in our previous study [18]. Here, we demonstrate how the effective force constant between coarse-grained nodes changes as a function of the cutoff radius, with similar trends found for the two proteins. Based on these relationships, our mixed coarse-graining results indicate that the temperature factors and the slow modes of motion can be successfully reproduced, and that they exhibit high correlations with the original uniform, but less coarsely grained results. With this methodology, it is possible to reduce the computational time by 2–4 orders of magnitude depending on the level of mixed coarse-graining applied. Thus, it becomes feasible to apply this methodology routinely to supramolecular assemblages such as the ribosome or larger. Furthermore, it is also possible to model certain parts such as the interfaces or active sites in atomistic detail by adjusting the parameters and using the same mixed coarse-graining methodology (results not shown here). Thus, it would be possible to observe the structural changes at specific regions of the large proteins resulting from the deformations in the slow modes, which would not be possible by fully atomistic approaches.

Acknowledgements

This work has been supported by the Bogazici University B.A.P. (03A501-D and 03R104), DPT Project (01K120280), and the Turkish Academy of Sciences in the framework of the Young Scientist Award Program (PD-TUBA-GEBIP/2002-1-9). PD specially thanks O.T. Turgut for helpful discussions on the subject.

References

[1] McCammon A, Harvey SC. Dynamics of proteins and nucleic acids. Cambridge: Cambridge University Press; 1987.
[2] Kitao A, Go N. Curr Opin Struct Biol 1999;9:164.
[3] Berendsen HJC, Hayward S. Curr Opin Struct Biol 2000;10:165.
[4] Elcock AH. Curr Opin Struct Biol 2002;12:154.
[5] Bahar I, Atilgan AR, Erman B. Fold Des 1997;2:173.
[6] Bahar I, Jernigan RL. J Mol Biol 1998;281:871.
[7] Hinsen K, Thomas A, Field MJ. Proteins 1999;34:369.
[8] Keskin O, Jernigan RL, Bahar I. Biophys J 2000;78:2093.
[9] Tama F, Gadea FX, Marques O, Sanejouand YH. Proteins 2000; 41:1.
[10] Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Biophys J 2001;80:505.
[11] Haliloğlu T, Bahar I, Erman B. Phys Rev Lett 1997;79:3090.
[12] Jernigan RL, Demirel MC, Bahar I. Int J Quantum Chem 1999;75: 301.
[13] Bahar I, Wallqvist A, Covell D, Jernigan RL. Biochemistry 1998;37: 1067.
[14] Haliloğlu T, Bahar I. Proteins 1999;37:654.
[15] Kuriyan J, Petsko GA, Levy RM, Karplus M. J Mol Biol 1986;190: 227.
[16] Doruker P, Atilgan AR, Bahar I. Proteins 2000;40:512.
[17] Doruker P, Jernigan RL, Bahar I. J Comput Chem 2002;23:119.
[18] Doruker P, Jernigan RL, Navizet I, Hernandez R. Int J Quantum Chem 2002;90:822.
[19] Ming D, Kong Y, Lambert MA, Huang Z, Ma J. Proc Natl Acad Sci USA 2002;99:8620.
[20] Tama F, Wriggers W, Brooks CL. J Mol Biol 2002;321:297.
[21] Doruker P, Jernigan RL. Proteins 2003;53:174.
[22] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000;28:235.
[23] Wilson IA, Skehel JJ, Wiley DC. Nature 1981;289:366.
[24] Weis WI, Brünger AT, Skehel JJ, Wiley DC. J Mol Biol 1990;212: 737.
[25] Juers DH, Jacobson RJ, Wigley D, Zhang D-J, Huber RE, Tronrud DE, Matthews BW. Prot Sci 2000;9:1685.
[26] Tirion MM. Phys Rev Lett 1996;77:1905.
[27] Levitt M, Sander C, Stern PS. J Mol Biol 1985;181:423.
[28] Hayward S. Normal mode analysis of biological molecules. In: Becker OM, MacKerell AD Jr, Roux B, Watanabe M, editors. Comput biochem biophys. New York: Marcel Dekker; 2001. p. 153–68.
[29] Hayward S, Kitao A, Go N. Prot Sci 1994;3:936.
[30] Flory PJ. Configurational statistics of chain molecules. New York: Wiley; 1969.
[31] ben-Avraham D. Phys Rev 1993;47:21.
[32] Isin B, Doruker P, Bahar I. Biophys J 2002;82:569.
[33] Keskin O. Biomol Struct Dyn 2002;20(3):333.