Important fluctuation dynamics of large protein structures are preserved upon renormalization

ARTICLE *in* INTERNATIONAL JOURNAL OF QUANTUM CHEMISTRY · MARCH 2002

Impact Factor: 1.43 · DOI: 10.1002/qua.955

Important Fluctuation Dynamics of Large Protein Structures Are Preserved upon Coarse-Grained Renormalization*

PEMRA DORUKER,[1,2] ROBERT L. JERNIGAN,[2] ISABELLE NAVIZET,[2,3] RIGOBERTO HERNANDEZ[4]

[1]*Chemical Engineering Department and Polymer Research Center, Bogazici University, Bebek 80815, Istanbul, Turkey*
[2]*Molecular Structure Section, Laboratory of Experimental and Computational Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-5677*
[3]*Institut de Biologie Physico-Chimique, 75005 Paris, France*
[4]*Center for Computational Molecular Science and Technology, School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332-0400*

Received 2 October 2001; revised 14 January 2002; accepted 25 January 2002

DOI 10.1002/qua.955

ABSTRACT: The fluctuations and important motions of three large proteins—hemaglutinin, xanthine dehydrogenase, and β-galactosidase—have been considered with a range of models having various levels of detail to represent the structures. Because the slowest modes of motion are the largest contributors to the total motions, and because these motions depend mainly on the shapes of the structures rather than their details, it is possible to replace the real structures with significantly fewer points and still retain the essential features of the structure for these modes of motion. We obtain excellent results, both for the magnitudes of the individual motions as well as for the molecular changes occurring during these motions. Similar results are obtained with another completely different approach where the coarse graining is based on invariant regions of structure found by comparing two structures of the same protein, given as an example here for myosin. Results confirm the important coupling of local functional motions with the large-scale motions, implying important functional roles for the entire protein structure. Int J Quantum Chem 90: 822–837, 2002

Key words: Gaussian network model; anisotropic fluctuations; vibration dynamics; collective motions; hemagglutinin; xanthine dehyrogenase; β-galactosidase; myosin

*Dedicated to the memory of Per-Olov Löwdin.

Correspondence to: R. L. Jernigan; e-mail: jernigan@lmmb.nci.nih.gov.

Contract grant sponsor: National Science Foundation.
Contract grant number: NSF 97-03372.

Introduction

Recently we and others have developed a mechanics approach for studying the motions of proteins [1–14] to obtain the equilibrium fluctuations near an initial structure. The initial structure has usually been determined by crystallography, but other experimental methods, or even modeled structures, could be utilized instead. The underlying assumption in the method is that the starting structure is the minimum energy structure in a local—if not global—minimum. All fluctuations about this form are presumed to be higher in energy, proportional to their mean-square displacements, i.e., the energy form is Gaussian. Within the structure, all close-lying residues (as defined by a cutoff radius) are restrained by an effective spring with a universal force constant and are said to be in contact. Residues nearest in sequence are not distinguished because they necessarily fall within the cutoff radius. The close-lying residue pairs are utilized to form a contact matrix that makes explicit reference to these restraining springs. Because of the simple Gaussian form of the energy, the dynamics can be integrated directly to obtain the mean-square fluctuations of positions, as well as the correlations of the displacements of residue pairs. The required computation is simply the inversion of the contact matrix. This method was initially developed to obtain scalar displacements, but it was readily apparent that the directions of displacement are also important. Recently a three-dimensional version [11] of this approach was developed, and it yields the correlations in the directions of the displacements, with the attendant computational cost from tripling each dimension of the contact matrix.

When structures are coarse-grained at the level of one point per residue, excellent agreement of this approach with experiments has been demonstrated for several proteins with respect to the crystallographic temperature factors [3, 4, 6, 8, 10, 13], as well as with nuclear magnetic resonance (NMR) order parameters [5] and hydrogen exchange data [1]. The computed results reveal that the most important motions are those typically involving large domains such as hinge motions. In addition many other large-scale motions are typically observed, e.g., rotation, stretching, shear, disintegration, and flap motions. Individual residue displacements are observed primarily as components of the motions of these subdomains. Moreover, the relative contributions of the modes involving the largest-scale motions to the observables are significantly larger than that of those modes at the other end of the spectrum, which involve only extremely local motions.

Interestingly, relatively few short-range contacts give rise to the large displacements of other residues by acting as the foci of the motions, such as the hinge foci. These largest-scale motions primarily reflect the shape of the protein rather than details of its internal structure. Some examples we have observed are: thin regions of structure that act as hinge sites, large interior cavities that undergo compression, and small numbers of contacts at subunit interfaces that support interfacial motions such as wobble and counterrotation of two subunits. Since these small numbers of residues involved in the most important motions do not involve the internal structure of the peptide chain, it suggests that coarse graining of the protein structures may readily be performed. We have recently applied this coarse graining, by retaining only 1 of every 40 residues, to haemagglutinin [12], where we have shown that it is possible to reproduce about 73% of the total protein motions. This initial coarse-grained application has raised many issues regarding this procedure. What is the optimal way to perform the coarse graining? In the model, there are only two adjustable parameters, a spring constant and a cutoff distance. How should these be modified or scaled for the coarse-graining renormalization? It is also important to understand how the distance cutoff, determining the spring contacts, scales with the coarse graining, as well as how the spring constant itself ought to be scaled. This work represents a first attempt at answering these questions.

PROTEINS

We have chosen three large proteins to consider in this study, namely β-galactosidase [15] (GAL), xanthine dehyrogenase [16] (XDH), and hemagglutinin [17, 18] (HA), with corresponding pdb file names 1DPO, 1FO4, and 2HMG. The number of residues and number of atoms in the crystal structures in each monomer are, respectively, 1011, 8125; 1299, 10077; and 503, 3957. See Figure 1 for views of these structures. The structural and functional details of these proteins are summarized below, although in this study we will not discuss

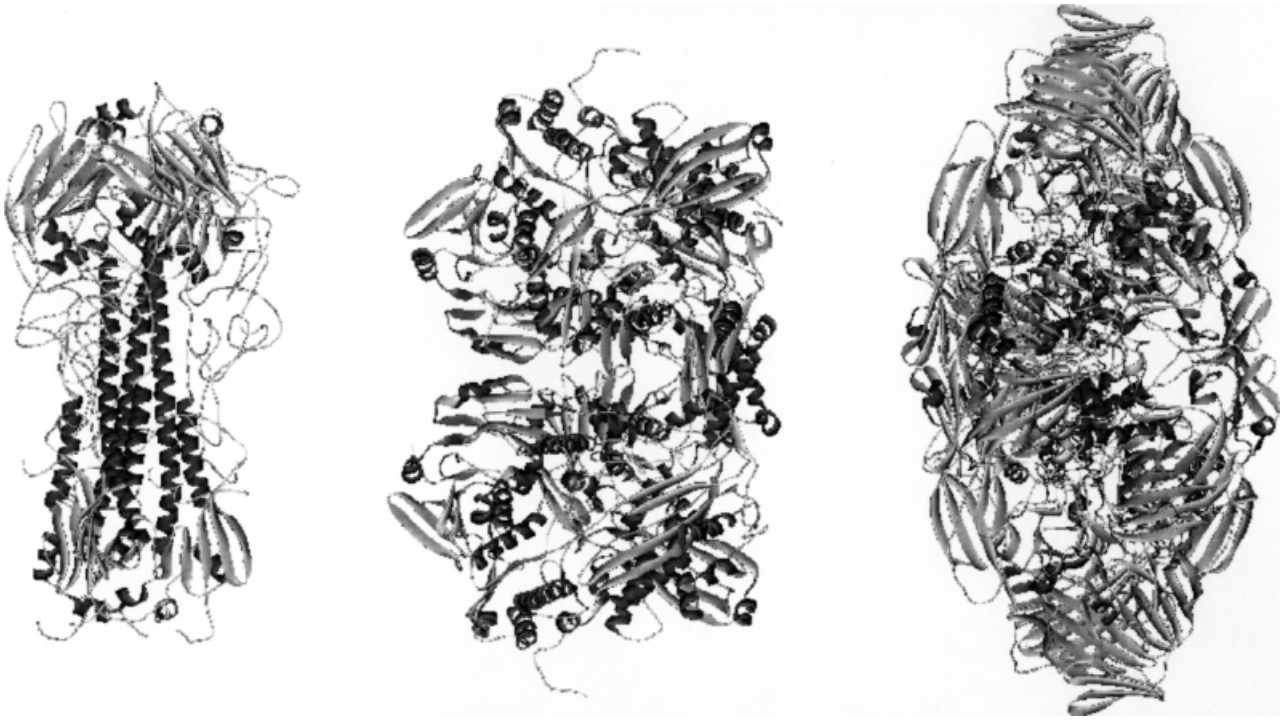


FIGURE 1. Ribbon diagrams of β-galactosidase (right), xanthine dehydrogenase (middle), and influenza virus hemagglutinin (left).

the structure–function relationships of these proteins.

The X-ray structure of *Escherichia coli* β-galactosidase determined by Juers and co-workers [15] at 1.7 Å resolution is shown in the left part of Figure 1. This enzyme hydrolyzes lactose and other β-galactosides into monosaccharides. The functional form is a tetramer having 4 identical subunits, with each monomer comprising 1023 residues. The subunits are assembled into a prolate ellipsoidal structure with approximate dimensions of 175 Å $\times$ 135 Å $\times$ 90 Å.

The crystal structure of the dimeric bovine milk xanthine dehyrogenase, displayed in the middle part of Figure 1, has been determined to 2.1 Å resolution [16]. The enzyme catalyzes the hydroxyl addition of hypoxanthine and xanthine, which are the two last steps in the formation of urate. Each monomer has 1332 residues conformed into a butterfly-shaped dimeric enzyme with overall dimensions of approximately 155 Å $\times$ 90 Å $\times$ 70 Å.

The influenza virus hemagglutinin is an integral membrane glycoprotein, which is involved in the binding of virus to target cells and in the fusion of viral and endosomal membranes at low pH. The X-ray structure of the neutral pH form of HA has been determined [17] and refined [18] by Wiley and co-workers to a resolution of 3 Å and is shown in the right part of Figure 1. HA, comprising 1509 residues, is a cylindrically shaped homo-trimer about 135 Å long, varying between 35 and 70 Å in the radial directions. Each monomer consists of 2 polypeptides chains: HA1 (328 residues) and HA2 (175 residues) that are linked by 2 disulfide bonds. The 3 monomers are assembled into a central coiled coil that forms the stemlike domain, and the 3 globular heads containing the receptor binding sites. Each globular head folds into a jelly-roll motif of 8 antiparallel β-strands.

Methods

The coarse graining of structure involves replacing groups of individual points with single points to yield a less detailed structure. This operation resembles the development of an equivalent chain model for polymers, where multiple repeat units of a polymer are coarse-grained into a single unit so as to imitate the behavior of one link of a model chain.

For example, several real bonds of polyethylene, because of their additive flexibility, are equivalent to the enhanced flexibility of a single link in the freely jointed chain model [19]. Such equivalent representations have often been utilized in polymer studies [19]. Applying this concept to the single fixed configurations of segments of a protein is not quite the same physical situation as in a polymeric random coil, since the conformations of the individual segments vary from one to another and cannot uniformly benefit from averaging over conformations, as is the case with polymer models. This is why it is important to see how variable these segments' conformations actually are. In what follows, we first outline the anisotropic network model developed earlier to capture the essential dynamics about the initial (equilibrium) structure and subsequently analyze the degree to which it is invariant to various coarse-graining strategies.

ANISOTROPIC NETWORK MODEL (ANM)

This is a model for protein motions developed as a three-dimensional extension of the Gaussian network model (GNM). It incorporates the anisotropy of fluctuations and yields the directions of each mode of motion; whereas the GNM assumes all fluctuations to be isotropic and gives only the magnitudes of the modes of motion. The potential energy of a structure having N interaction sites is expressed with ANM as a Gaussian form:

$$V = \frac{\gamma}{2}\Delta\mathbf{R}^{\mathrm{T}} H\,\Delta\mathbf{R}, \tag{1}$$

where $\Delta\mathbf{R}$ is a $3N$-dimensional vector of the fluctuations $\Delta\mathbf{R}_i$ in the position vectors $\mathbf{R}_i$ of all sites ($1 \le i \le N$), $\Delta\mathbf{R}^{\mathrm{T}}$ being its transpose, and H the Hessian matrix composed based upon the second derivatives of the potential:

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

The summations will be performed over all interaction sites, $h(x)$ is the Heaviside step function [$h(x) = 1$ if $x \ge 0$, and zero otherwise], R_{ij} is the distance between sites i and j, and r_c is the cutoff distance defining the interactions; H is expressed as a function of N^2 submatrices $\mathbf{H}_{ij}$ in the form

$$\mathbf{H}_{ij} = \begin{bmatrix} \partial^2V/\partial X_i\partial X_j & \partial^2V/\partial X_i\partial Y_j & \partial^2V/\partial X_i\partial Z_j \\ \partial^2V/\partial Y_i\partial X_j & \partial^2V/\partial Y_i\partial Y_j & \partial^2V/\partial Y_i\partial Z_j \\ \partial^2V/\partial Z_i\partial X_j & \partial^2V/\partial Z_i\partial Y_j & \partial^2V/\partial Z_i\partial Z_j \end{bmatrix}, \tag{3}$$

with X_i, Y_i, and Z_i being the components of $\mathbf{R}_i$. Note that $\partial^2V/\partial X_i\partial Y_j = -\partial^2V/\partial X_j\partial Y_i = -\gamma(X_j - X_i)(Y_j - Y_i)/R_{ij}^2$ for $i \ne j$, and $\partial^2V/\partial X_i\partial Y_i = \gamma\sum_j(X_j - X_i)(Y_j - Y_i)/R_{ij}^2$.

In general the correlations between the fluctuations at sites i and j are given by

$$\begin{aligned}\langle\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j\rangle &= \frac{1}{Z}\int(\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j)\exp\{-V/kT\}\,d\{\Delta\mathbf{R}\} \\ &= \frac{3k_BT}{\gamma}\,\mathrm{tr}\left[H^{-1}\right]_{ij},\end{aligned} \tag{4}$$

where k is the Boltzmann constant, Z is the configurational partition function, and $\mathrm{tr}\,[H^{-1}]_{ij}$ is the trace of the ijth submatrix $[H^{-1}]_{ij}$ of $\mathbf{H}^{-1}$; $\langle\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j\rangle$ can be expressed as a sum over the contributions $[\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j]_k$ of the $3N-6$ individual internal fluctuation modes, as $\langle\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j\rangle = \sum_k[\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j]_k$. The contribution of the kth mode is explicitly given by

$$[\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j]_k = \frac{3kT}{\gamma}\,\mathrm{tr}\left[\lambda_k^{-1}\mathbf{u}_k\mathbf{u}_k^{\mathrm{T}}\right]_{ij}, \tag{5}$$

where λ_k is the kth nonzero eigenvalue of H and $\mathbf{u}_k$ is the corresponding eigenvector. The eigenvalues are related to the frequencies of individual modes, and the eigenvectors describe its effect on the positions of the N points of the structure. The eigenvalues are usually organized in ascending order (after removing the six zero eigenvalues), so that λ_1 denotes the lowest frequency, also called the global, mode of motion, and $[\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j]_1$ is the correlation for this mode of motion separately. Actually here we use only the individual residue mean-square (ms) fluctuations for the position at site i for mode k, $[(\Delta\mathbf{R}_i)^2]_k$. Note that zero values can arise either from being uncorrelated or being perpendicular. The slowest modes usually dominate the collective dynamics of the structure and would be the only surviving modes at long times, thus they are particularly relevant to biological function, unless other effects such as anharmonicity interfere.

COARSE GRAINING OF THE ANM

Here we take N to be the number of residues in the total structure (protein), s the number of coarse-grained segments, and n the number of residues in one coarse-grained segment, so that

$$N = sn. \tag{6}$$

The cutoff distance r_c defining interactions (springs) needs to be sufficiently large to include the s

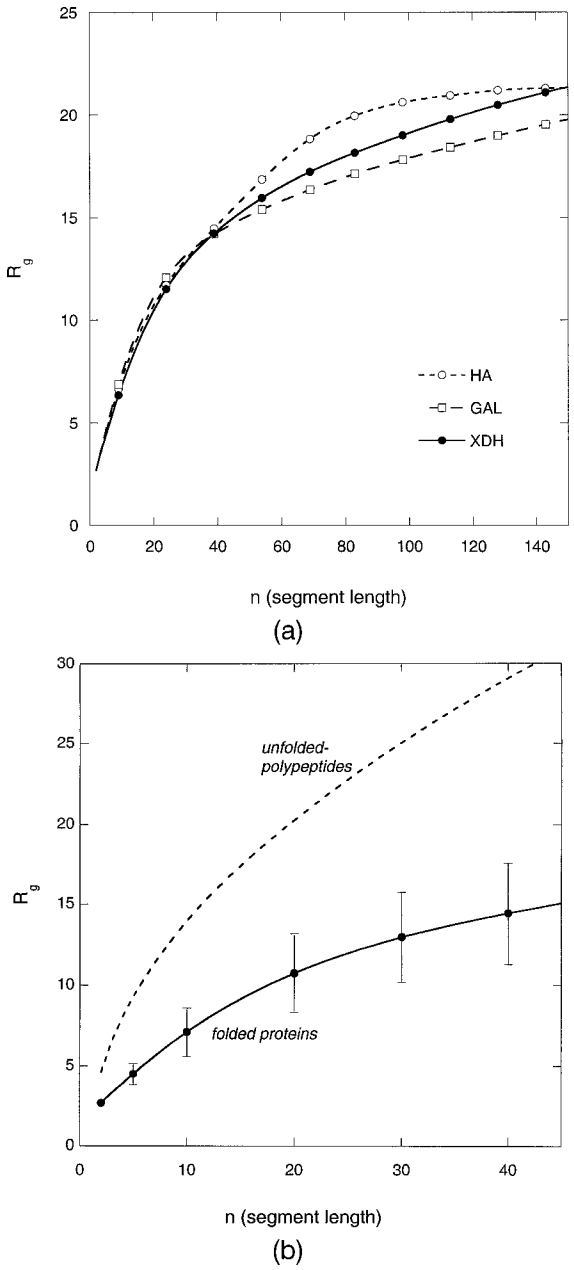


FIGURE 2. (a) Radius of gyration of chain segments in the folded proteins GAL, XDH, and HA. (b) comparison of the radius of gyration of chain segments in random coil polypeptides and folded proteins, where values given on the lower curve are average values for the three proteins, with the bars showing the standard deviations.

residues in each of the n segments. For this purpose we compute R_G the radius of gyration for each of the segments in the three proteins. See Figure 2(a) for segments up to 140 residues in length. Because of the finite size of the proteins, the values converge to a clear limit. This behavior is reminiscent of the behavior of flexible polymer chains of different lengths. Despite the heterogeneity in each of the segments (or links), the three proteins behave similarly up to the coarse-graining level of 40 residues.

RADIUS OF GYRATION OF FOLDED CHAIN SEGMENTS

A point of comparison for the R_G values of the protein segment size is found in the R_G values of the random coil model for homopolymers consisting of N peptide units [20, 21]. The average dimension, expressed as the characteristic ratio, from an average of several experiments, for several different polypeptides having β carbons, is

$$\langle r^2 \rangle / NL^2 = 9, \tag{7}$$

where r is the end-to-end distance, and L is the virtual bond length. For a long Gaussian chain, the radius of gyration is related to the mean square of the end-to-end distance by

$$\langle R_G^2 \rangle = \tfrac{1}{6} \langle r^2 \rangle. \tag{8}$$

Thus

$$R_G / \sqrt{NL^2} = 1.225, \tag{9}$$

where, as before, N is the number of residues and L is the virtual bond length.

In Figure 2(b), the random coil limit for R_G appears as the smooth upper curve. As might be expected, all of the protein segments are more compact than the random coil peptide. The bars show the range of individual values for segments of different sizes, all of which are significantly more compact than the random polypeptide case.

It would be interesting to learn the origin of the variations in the R_G values for a fixed size segment. Are the locally compact segments determined by their own sequences or by more global considerations? Do the segments with the lowest R_G values include glycines, which could facilitate turns, or do they have more hydrophobic residues on average, which could contribute to collapsed forms? Or are there other composition effects?

In order to further coarse-grain folded proteins, it is helpful to know how the overall dimensions of the chain segments in folded proteins change as a function of segment length. This will indicate how the cutoff radius in the ANM calculations should be

adjusted for further coarse graining along the backbone of the protein.

For the three proteins that are considered in this study, we calculate the mean-square radius of gyration, $\langle R_G^2 \rangle$, for segments of various lengths. This calculation is carried out separately for the 6, 2, and 4 chains that make up HA, XDH, and GAL, respectively. And the average is calculated by moving the starting point of each segment along the chain backbone one by one toward the end of the chain. Therefore, for a single chain composed of N_c residues, the radius of gyration is averaged over $(N_c - n + 1)$ frames for a segment of length n.

In Figure 2(a), the radius of gyration, R_G, is plotted as a function of segment length for the three proteins. The behavior is similar up to $n = 40$, presumably reflecting the average behavior of peptides. For $n > 40$, differences begin to be manifested which occur because of the differences in the overall sizes and shapes of proteins.

For $n < 40$, the data can be fit with the form

$$R_G = an^b. \tag{10}$$

These parameter values are found to be $a = 1.778$ and $b = 0.595$ from a fit to the average over the three log–log plots of R_G vs. n for HA, XDH, and GAL. The $n = 1$ limit of Eq. (10) corresponds to a single monomer whose radius of gyration must be a, suggesting that the average bond length is approximately equal to $2a$ (=3.556 Å), which is in close consistency with the virtual bond length between sequential α-carbon atoms of 3.8 Å.

In Figure 2(b), the lower curve gives the radius of gyration averaged over all segments of a given size in the three folded proteins (HA, XDH, and GAL), and the error bars are shown for some representative values of n. Here, the standard deviation for a specific value of n has been calculated over the frames of all possible segments in the three proteins. The dashed curve in the same figure gives the R_G of unfolded segments of length n, as predicted by the model for polyalanine developed by Flory [21].

In earlier work, a cutoff radius of 13 Å was found to be suitable for ANM calculations, in which all α-carbon atoms in the protein structure were retained [11]. In the current study, as we further coarse-grain the structures, we recognize that the renormalized sites are interacting at longer ranges because their effective sizes have grown. The cutoff radius should thus equal the sum of the renormalized radii of each site plus the invariant contact distance R_0 between the sites, i.e.,

$$r_c = 2R_G + R_0, \tag{11}$$

where R_G is obtained according to Eq. (10) with the parameters found above. To be consistent with our earlier work, R_0 should be set to a value of (13 Å $- 2a$), but for simplicity, in what follows we have used the value of 13 Å instead. This choice leads to little change in the results since they are only modestly dependent on R_0, while being strongly dependent on the growth of R_G with N. Results for the three illustrative proteins of this study are shown in Table I.

TABLE I
Details of coarse graining.

Segment length n	Cutoff radius[a] r_c (Å)	s, Number of segments		
		GAL	XDH	HA
1	13.0	4044	2587	1509
2	18.4	2024	1294	756
5	22.3	812	518	303
10	27.0	408	260	153
20	34.1	204	130	78
30	39.9	136	88	51
40	44.9	104	66	42
80	61.2	52	34	24

[a] Cutoff radius is calculated according to $r_c = 2R_G + 13$ Å, where R_G is found from Eq. (10).

Results and Discussion

X-RAY CRYSTALLOGRAPHIC TEMPERATURE FACTORS

The relationship between an individual residue's fluctuations and its temperature factor is

$$B_i = (8\pi^2/3)\langle \Delta R_i^2 \rangle. \tag{12}$$

In Figure 3, these experimental temperature factors measured by X-ray crystallography (solid curves) are compared to those predicted by the ANM (dashed curves). For each of the three proteins, each monomer exhibits practically the same behavior both in experiment and calculation. Therefore, the fluctuations of residues are presented as averages over all monomers. The overall agreement is excellent as has often been observed with this model.

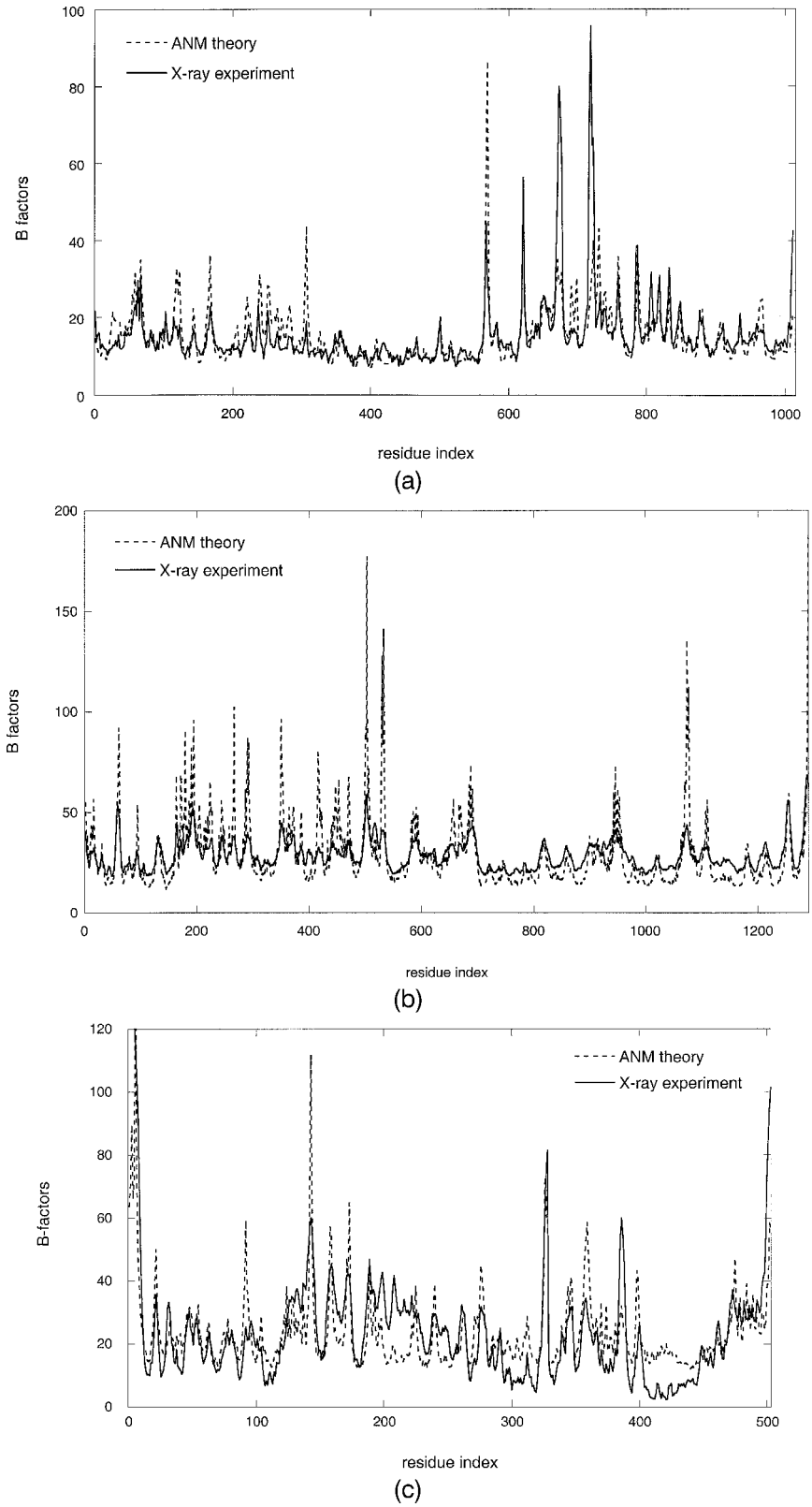


FIGURE 3. Comparison of temperature factors from X-ray crystallography and those calculated with ANM calculations for (a) β-galactosidase, (b) xanthine dehydrogenase, and (c) hemagglutinin.

TABLE II
Force constants γ for coarse-grained ANM calculations.

n	GAL	XDH	HA
1	—	0.688	0.890
2	0.874	0.496	0.644
5	1.442	0.758	1.176
10	1.768	0.953	1.641
20	2.009	1.048	1.876
30	2.571	1.390	1.909
40	2.333	0.901	1.654
80	1.971	1.139	1.457

Once the cutoff radius for the interactions is fixed, the force constant γ is the only remaining parameter in the calculations. In turn its value is fixed by requiring a match between the average values of the mean-square fluctuations predicted by ANM and the experimental B factors. In Figure 3, such adjustments were made in order to compare the experimental and theoretical results. The experimental B factor, B^n of a coarse-grained segment composed of n residues is calculated as the average of the B factors of its n constituent residues. And the force constant is extracted by a comparison of the coarse-grained B factors with the mean-square fluctuations calculated with ANM. Table II gives the force constant values. As our previous experience with a large number of proteins has indicated, γ varies among proteins by no more than a factor of 2. However, as the coarse graining is applied, the force constants become stronger monotonically, upon passing from the scaling at $n = 2$ to $n = 30$.

Parenthetically, it should be noted that in the case of β-galactosidase [Fig. 3(a)], only an $N/2$ calculation was carried out instead of an all-residue calculation because of the large size of this protein (4044 residues in total). Although an $n = 1$ calculation is feasible, this has not been executed here. And the experimental B factors, for comparison, were averaged over neighboring pairs of residues.

COMPARISON OF ANM RESULTS AT DIFFERENT LEVELS OF COARSE GRAINING

B Factors

Figure 4(a) compares the temperature factors from coarse-grained calculations $N/2$ and $N/10$ for GAL. Higher levels of coarse graining lead to smoother curves, but the basic structure of the peaks is readily apparent at the level of $N/10$ calculations. Figure 4(b) shows the calculated B factors at the same $N/10$ level for xanthine dehydrogenase. From these results it is clear that the essential structure of fluctuations is retained after the coarse graining.

First Mode

The slowest mode shapes obtained with $N/2$ and $N/10$ calculations are displayed in Figure 5(a) for GAL. There is a remarkable match between the curves, which have been normalized to match the scales. Figure 5(b) shows a comparison of the N and $N/10$ calculations for hemagglutinin. Clearly, the general features of the first mode shape are obtained. As a result of these comparisons, it is evident that the functionally important collective mode shapes can still be reproduced quite well at higher levels of coarse-graining.

Eigenvalues

Figure 6 compares the weighted contribution of each mode to the mean-square fluctuations at the different levels of coarse graining employed for GAL, XDH, and HA. The modes are sorted and indexed starting from the slowest mode having the largest contribution and running up to higher frequencies. In order to capture the same collective modes at higher levels of coarse graining, the fractional contributions at the low-frequency end of the spectrum need to be similar. And thisis exactly what we observe in these logarithmic plots. In Table III, the cumulative contributions of the first three modes are listed. As the level of coarse graining increases, the cumulative contribution of slowest modes increases because there are fewer modes at the high-frequency end of the distribution. Yet the fractional contributions of the collective modes appear to be comparable after renormalization.

Mechanisms of Motion

In Figure 7the two extreme positions for the first two slowest modes of β-galactosidase are shown at two different levels of coarse graining, $N/2$ and $N/10$. It is amply clear from these figures that the same motions occur, despite the coarse graining. The first mode is for bending at the "waist" of the protein, and the second is a stretching–compression type of motion that we

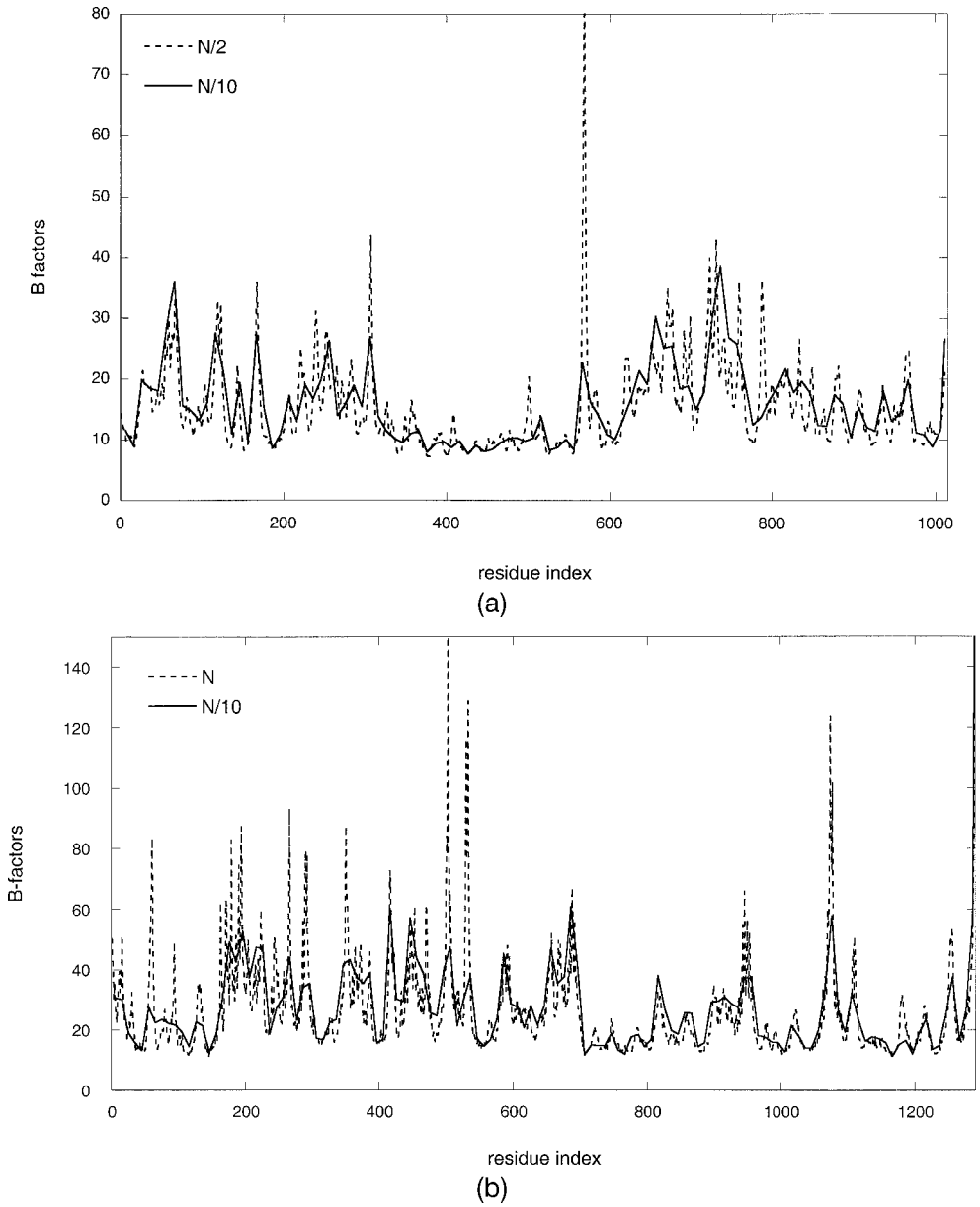


FIGURE 4. Comparison of temperature factors predicted by ANM at different levels of coarse graining for (a) β-galactosidase and (b) xanthine dehydrogenase.

have often observed in asymmetric elongated protein structures.

The correlations computed between the motions with the coarser-grained models and with the single residue–single point results are high. For hemagglutinin (see Table IV) it can be seen that, whereas the total motions are not so well represented (at the 49% level for the 1 out of every 40 models), the representations of the first, slowest mode remain above 90% for even the 1 out of every 40-residue model. Thus the coarse-grained results are most viable for motions having the largest displacements.

Structure-Based Coarse Graining

Finally we consider a completely structure-based approach, which requires multiple structures to specify which parts of the structure are to be coarse-grained. The parts of the two structures having the smallest differences are identified directly to deter-

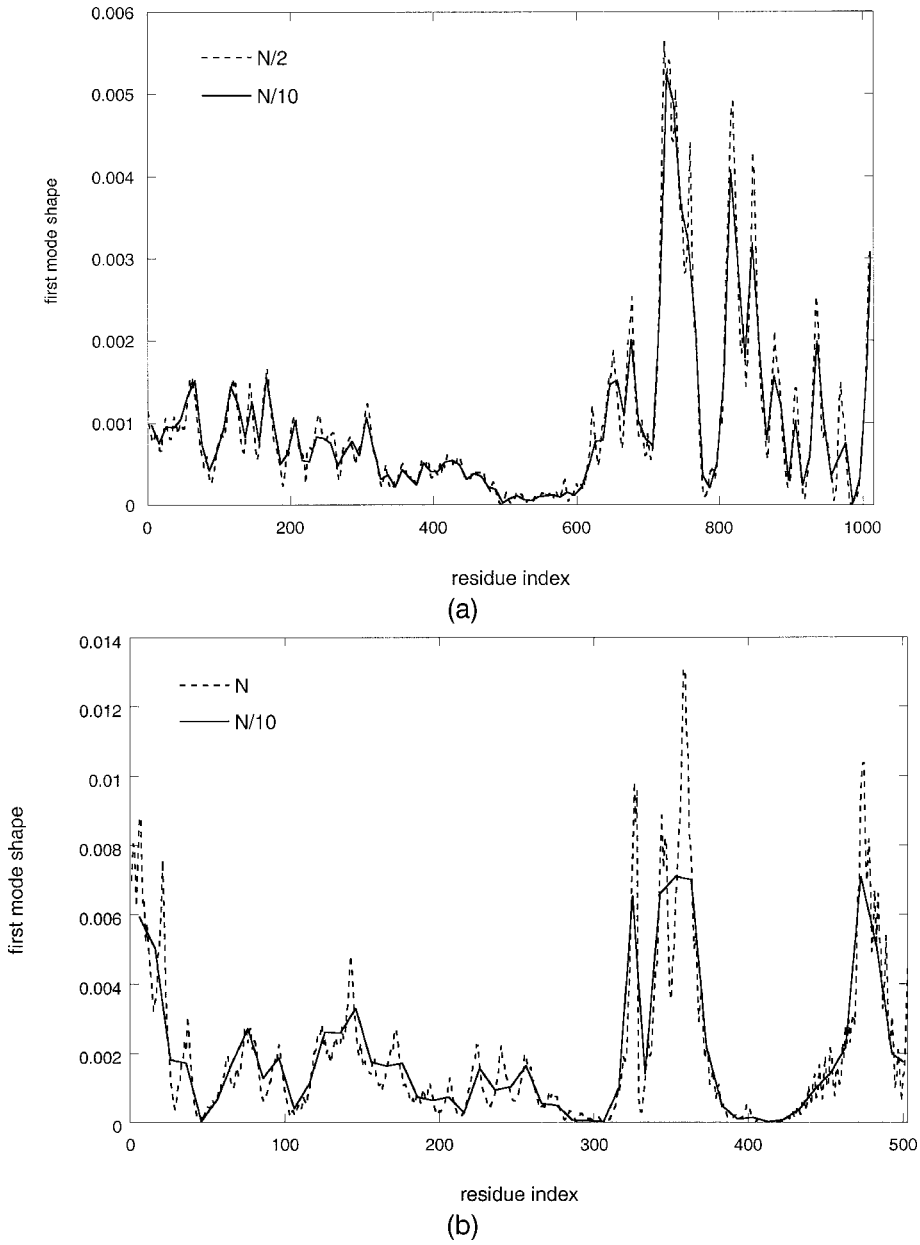


FIGURE 5. Slowest mode shapes predicted by ANM at different levels of coarse graining for (a) β-galactosidase and (b) hemagglutinin.

mine the blocks to be coarse-grained. Then, within these most constant blocks, the spring constants are increased to prevent intrablock motions. Another way of implementing this approach would be to treat these fixed blocks as "fat" rigid elements including many more than usual contacts with the other individual residues. This approach is applied here for demonstration purposes to two structures of myosin (pdb names 1B7T [22] and 1DFL [23]). The blocks defined by this approach are shown in Figure 8 within which the changes in distances have been limited to a maximum of 0.1 Å. The invariant regions are identified in different colors in Figure 8, with the few remaining residues not included within the rigid blocks are shown in gray. Importantly this approach yields nearly identical computed temperature factors, to those computed with the individual one point per residue model (see

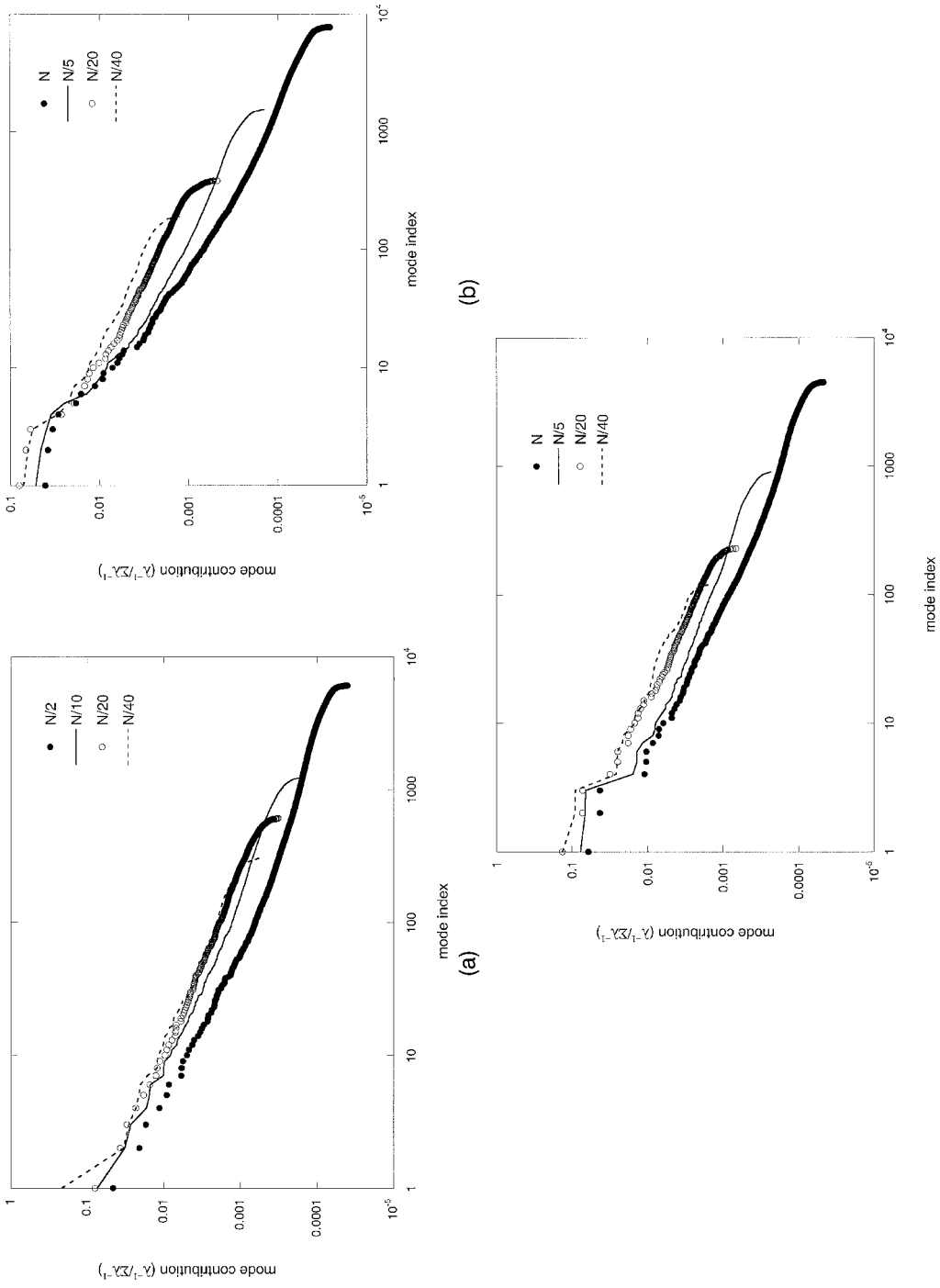


FIGURE 6. Contributions of the modes at different levels of coarse graining for (a) β-galactosidase, (b) xanthine dehydrogenase, and (c) hemagglutinin. All plots log–log plots to emphasize that only the lowest indexed modes are significant contributors to the overall motions. Also notable is the extent agreement in the dominant mode contributions between the models, regardless of the level of coarse graining.

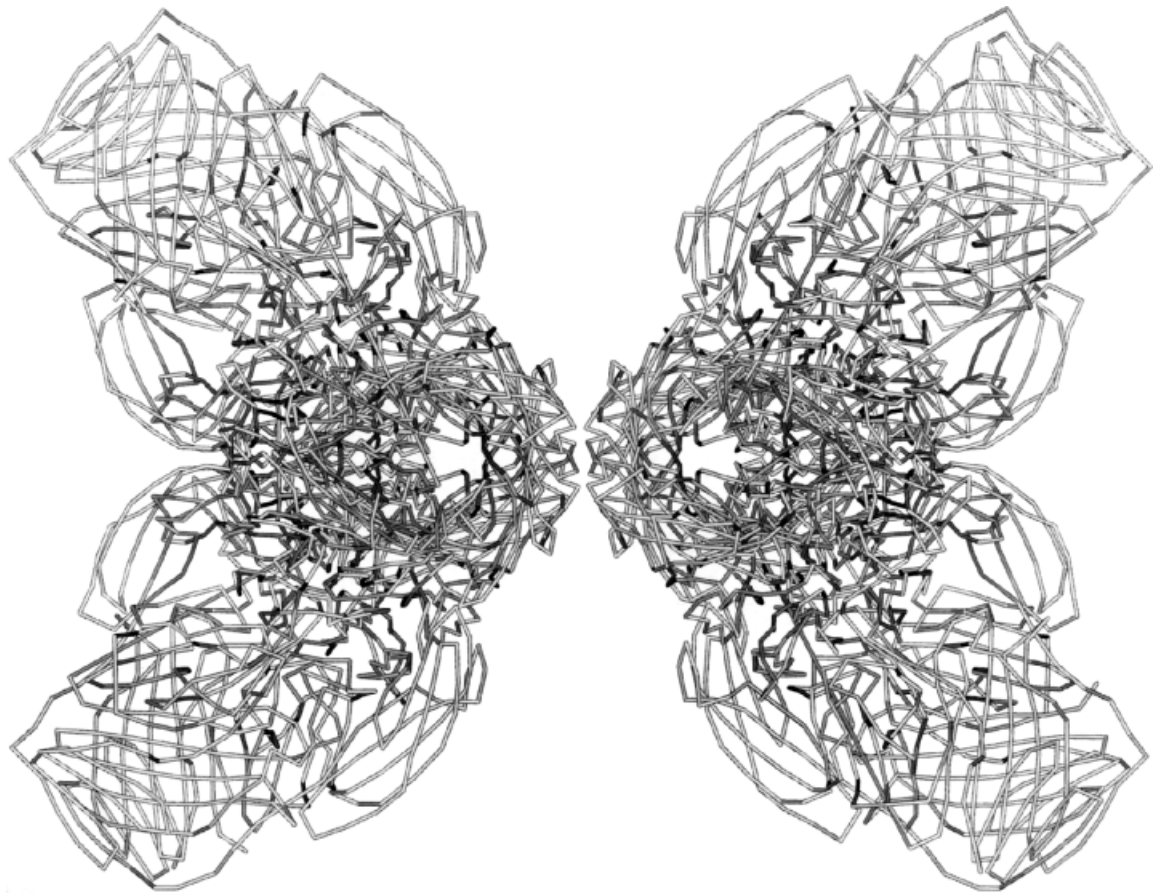
TABLE III
Total fractional contribution of the slowest three modes to the mean-square fluctuations.

n	GAL	XDH	HA
1		0.112	0.145
2	0.084	0.123	0.142
5	0.113	0.137	0.209
10	0.138	0.169	0.262
20	0.146	0.206	0.279
40	0.279	0.191	0.313

Fig. 9). Consequently, this model represents an alternative coarse-grained model that has its basis in two different structures. It is noteworthy that the most rigid regions of the structure are clearly clustered within these local domains.

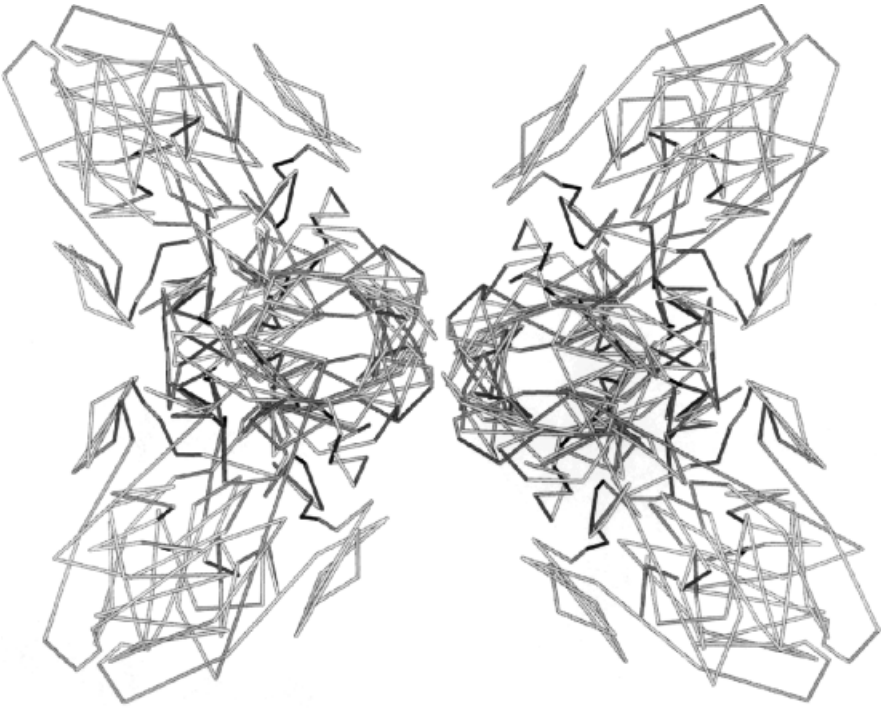
Discussion

One of the most important findings from these types of computations is the occurrence of functional "local motions" not independently but within one of the slowest most important motions. Examples that we have previously observed include flaps opening and closing over small molecule bind-

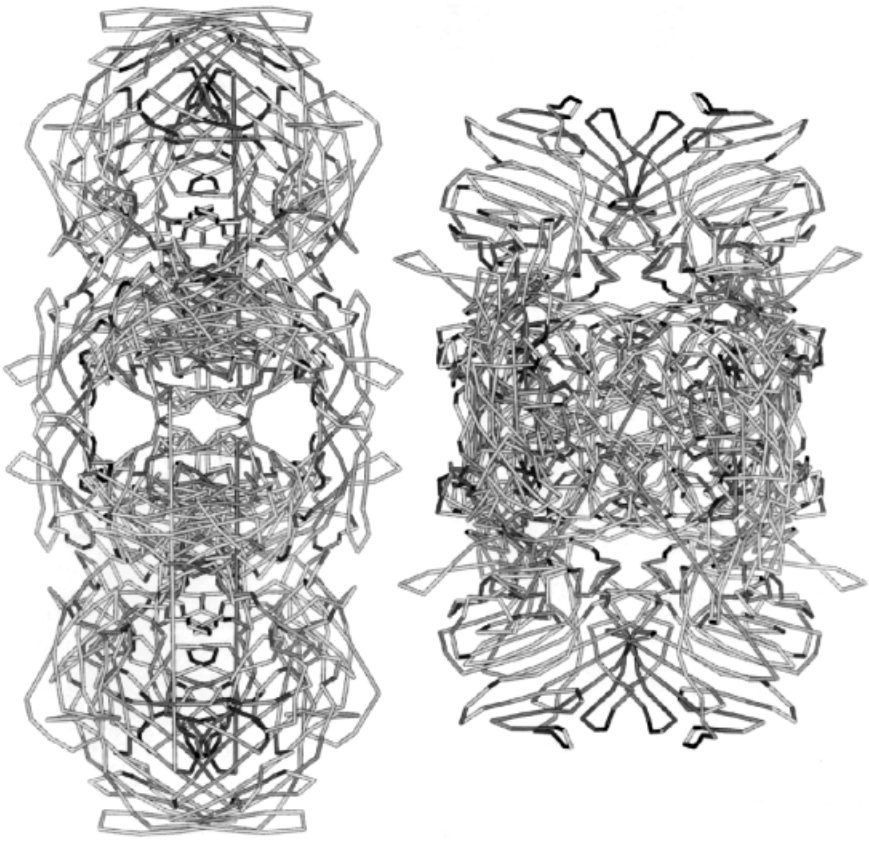


(a)

FIGURE 7. First (a), (b) and second (c), (d) modes of motion for β-galactosidase at $N/2$ (*a*), (c) and $N/10$ (b), (d) levels of coarse graining. Note that in parts (a) and (c) only half of the α-carbon positions are shown (and used) and in parts (b) and (d) only 1 out of every 10 residue is shown (and used in the computations). The first mode is a bending of the molecule along its activating interface, and the second mode is a stretching–compression type of motion. Loops often are opened and closed during these large-scale motions. This can be seen most clearly at the top and bottom of the structure in the stretching–compression mode of motion.

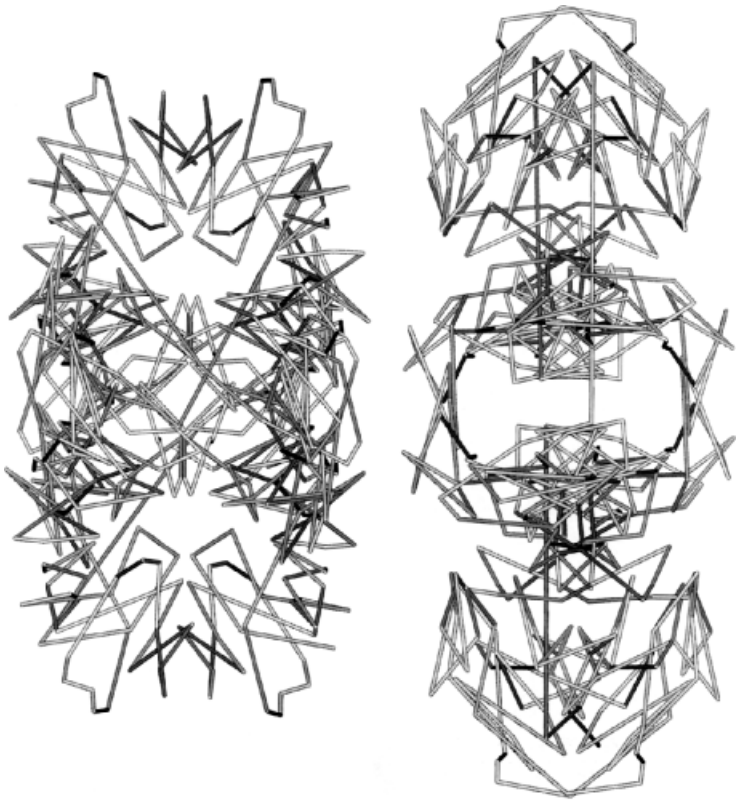


(b)



(c)

FIGURE 7. (Continued.)



(d)

FIGURE 7. (Continued.)

ing sites. These motions do not occur locally and independently but rather together with a highly coordinated motion of the entire protein. This type of motion can be clearly seen in Figure 7(b) where the flaps at the top and bottom of the structure open upon compression and close upon stretching, whereas opposite behavior can be observed for surface flaps in the center of the structure.

TABLE IV
Correlations at different levels of coarse graining.

	All modes (HA)	First mode (HA)
N/2	0.93	1.00
N/10	0.73	0.99
N/20	0.53	0.96
N/40	0.49	0.91

Two alternative approaches for coarse graining have been presented, one based on scaling the size of the cutoff distance based on the average dimensions of protein segments and the other more empirically based on actual changes between two experimental structures.

In many protein studies there has been a focus on functional sites while the remainder of the protein structure has been substantially ignored. The present work emphasizes that there is a truly important role for the entire protein in controlling these critical functional motions. In our view, the *raison d'être* for protein structure is that a fold pattern leads to its shape, which in turn controls the important functional motions of the protein. It is furthermore important that it be possible to substantially ignore the details of the structure in extracting these largest-scale motions. A secondary implication is that high-resolution structures may not be required in order to infer the important motions of proteins.

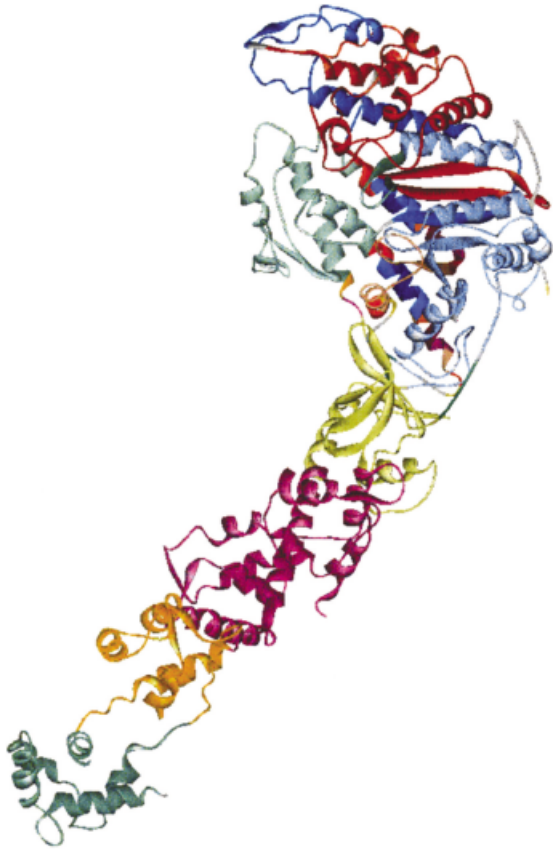


FIGURE 8. Ribbon diagram of the myosin head structure [22] 1B7T. Residues in the same block are shown in the same color. The few residues in gray are those not included in any blocks.

ACKNOWLEDGMENTS

R.H. is supported through the National Science Foundation (Grant No. NSF 97-03372) and is presently an Alfred P. Sloan Fellow and Research Corporation Cottrell Scholar. P.D. is partially supported by the Bogazici Research Fund (project 01HA501), and she thanks O.T. Turget for helpful occasions.

References

1. Bahar, I.; Wallqvist, A.; Covell, D. G.; Jernigan, R. L. Biochemistry 1998, 37, 1067–1075.
2. Demirel, M. C.; Atilgan, A. R.; Jernigan, R. L.; Erman, B.; Bahar, I. Protein Sci 1998, 7, 2522–2532.
3. Bahar, I.; Jernigan, R. L. J Mol Biol 1998, 281, 871–884; Bahar, I.; Atilgan, A. R.; Erman, B. Folding Des 1997, 2, 173–181.
4. Bahar, I.; Erman, B.; Jernigan, R. L.; Covell, D. G. J Mol Biol 1999, 285, 1023–1037.
5. Haliloglu, T.; Bahar, I. Proteins 1999, 37, 654–667.
6. Bahar, I.; Jernigan, R. L. Biochemistry 1999, 38, 3478–3490.
7. Jernigan, R. L.; Demirel, M. C.; Bahar, I. Int J Quantum Chem (B. Pullman Memorial Volume) 1999, 75, 301–312.
8. Keskin, O.; Jernigan, R. L.; Bahar, I. Biophys J 2000, 78, 2093–2106.
9. Jernigan, R. L.; Bahar, I.; Covell, D. G.; Atilgan, A. R.; Erman, B.; Flatow, D. T. J Biomol Struct Dyn, Conversation 11, Issue 1, 2000, 49–55.
10. Keskin, O.; Bahar, I.; Jernigan, R. L. Biochemistry, to appear.
11. Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Biophys J 2001, 80, 505–515.
12. Doruker, P.; Jernigan, R. L.; Bahar, I. J Comput Chem 2002, 23, 119–127.

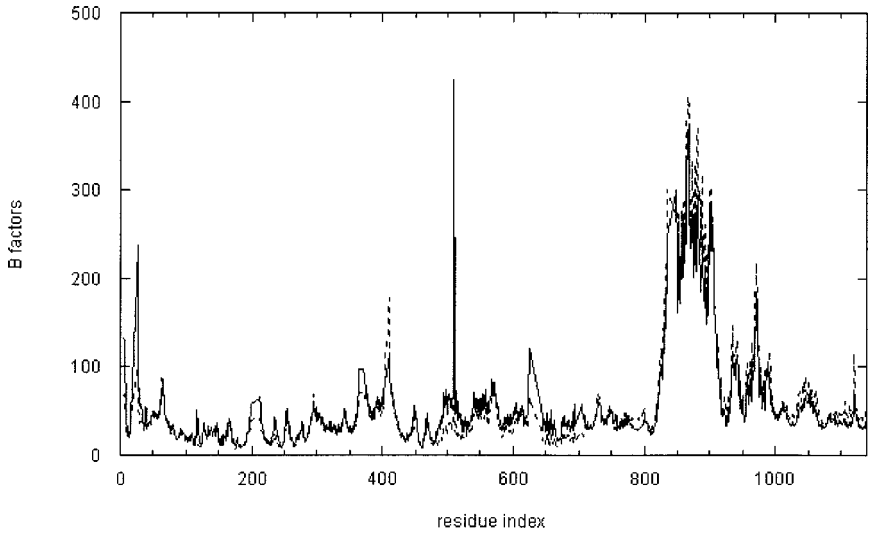


FIGURE 9. Comparison of temperature factors of myosin predicted from calculations taking into account the blocks (solid) and the full non-coarse-grained single-residue calculations (dashed).

13. Doruker, P.; Atilgan, A. R.; Bahar, I. Proteins 2000, 40, 512–524.
14. Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y.-H. Proteins 2000, 41, 1–7.
15. Juers, D. H.; Jacobson, R. J.; Wigley, D.; Zhang, D.-J.; Huber, R. E.; Tronrud, D. E.; Matthews, B. W. Protein Sci 2000, 9, 1685–1699.
16. Enroth, C.; Eger, B. T.; Okamoto, K.; Nishino, T.; Nishino, T.; Pai, E. F. Proc Natl Acad Sci USA 2000, 97, 10723–10728.
17. Wilson, I. A.; Skehel, J. J.; Wiley, D. C. Nature 1981, 289, 366–373.
18. Weis, W. I.; Brünger, A. T.; Skehel, J. J.; Wiley, D. C. J Mol Biol 1990, 212, 737–761.
19. Flory, P. J. Statistical Mechanics of Chain Molecules; Interscience: New York, 1969; Vol. 12, pp. 326–328.
20. Brant, D. A.; Flory, P. J. J Am Chem Soc 1964, 87, 2788–2800.
21. Flory, P. J. Statistical Mechanics of Chain Molecules; Interscience: New York, 1969; p. 277.
22. Houdusse, A.; Kalabokis, V. N.; Himmel, D.; Szent-Gyorgyi, A. G.; Cohen, C. Cell 1999, 97, 459–470.
23. Houdusse, A.; Szent-Gyorgyi, A. G.; Cohen, C. Proc Natl Acad Sci USA 2000, 97, 11238–11243.