JCTC Journal of Chemical Theory and Computation

Models To Approximate the Motions of Protein Loops

Aris Skliros, Robert L. Jernigan, and Andrzej Kloczkowski*

L. H. Baker Center for Bioinformatics and Biological Statistics, Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011

Received March 17, 2010

Abstract: We approximate the loop motions of various proteins by using a coarse-grained model and the theory of rubberlike elasticity of polymer chains. The loops are considered as chains where only the first and the last residues thereof are tethered by their connections to the main structure, while, within the loop, the loop residues are connected only to their sequence neighbors. We applied these approximate models to five proteins. Our approximation shows that the loop motions can usually be computed locally which shows these motions are robust and not random. But most interestingly, the new method presented here can be used to compute the likely motions of loops that are missing in the structures.

Introduction

Coarse-grained elastic network models (ENM) have been extremely successful in predicting the large-scale motions of proteins, RNA, and other biological structures, even for the largest complexes such as the ribosome. The predicted fluctuations of the positions of amino acids in the coarse-grained representations usually give excellent agreement with the experimental *B*-factors reported by crystallographers,[1−3] the ensembles reports by NMR scientists,[4] and the variability in structures manifested in the known multiple structures of the same protein.[4,5] The only information required in the ENMs is the structure of the protein, to furnish the coarse-grained coordinates of essential atoms, usually those of the of C^{α} atoms (but they could be other points representing the amino acids, such as centers of mass of side chains) for residue-level coarse graining. It has been shown that fluctuations of residues in proteins depend mostly on the protein's shape.[6] Because of this, the ENMs give excellent results even for relatively low-resolution structural data, such as electron micrographs.

The problem of modeling the conformations of external protein loops is a really important problem. These are generally the most mobile parts of the protein localized on their surface and are the functional sites for many protein activities, particularly for encounter complexes and binding. Because of their relatively high mobilities they are often unresolved in the crystal structures, particularly for larger loops. Because of this, frequently the PDB coordinates for residues in loops are either missing or have alternative positions. When obtaining an experimental structure, often the loops are the most uncertain parts of the structure.

Functions of biomolecules depend on their structures and are often exerted through functional motions. This makes understanding loop motions in proteins a particularly critical problem. Often interactions with other proteins or ligands can lead to apparent rearrangements of loops to accommodate functional ligands. For drug binding, the reconfiguration of target protein loops is relevant. Thus, the frequent involvement of external loops in function makes the prediction of their conformations essential for many structural applications in molecular biology, medicine, pharmacy, and drug design. The motions of loops computed by using ENMs have been intriguing, since often these move together with motions of the large-scale domains, either as rigid parts of domains or as separate parts moving in an anticorrelated way, but controlled by the domain motions. We have observed that the functionally meaningful loops most often appear to move under the control of the entire structure and its domain motions. Protein loops have been the focus of many previous studies. Modeling of loops has a significant role in making comparisons between protein structures,[7] since these may be found in one structure and missing in the other. The field of structural genomics often requires building loop structures.[8] Methods for the automated classification of the structures of protein loops have been developed.[9] In principle, the study of loops can aid the understanding of protein evolution.[10] Panchenko and Madej[10] noted that protein loops are far from being random coils, regardless of their size.

10.1021/ct1001413
Published on Web 09/16/2010

Changes in the conformations of protein loops have also been a subject of some specific study.[11] The importance of protein loops for protein function has been widely acknowledged.[12] Conformations of loops play a large role in protein docking as has been pointed out in refs 13−16. The motion of protein loops, especially where they are flexible, is an important factor for understanding the various roles that proteins play. The Web site https://simtk.org/home/looptk provides a toolkit to model the kinematics of protein loops. In ref 17 a novel approach for loop prediction was presented and analyzed. Kolodny et al.[18] describe an algorithm for generating conformations of candidate loops for a gap of a given size. A similar work also appeared in ref 19. Protein loops are also essential for protein folding.[20] Conformational evaluation of loops and their major role in protein design was discussed in ref 21. The importance of loop prediction was emphasized also in ref 22. In this paper, we devise a method for specifying how a protein loop can adopt different configurations. Our simple approximate model accounts only for the sequential connections within the loop and the loop's connections at the two ends, and this is a surprisingly successful model for generating loop forms. The issues of other interactions of the loop with the body of the protein are not explicitly taken into account in this work.

To overcome the difficulties with loop predictions, here we have applied the analytical theory of fluctuations in Gaussian Phantom Networks originally developed in the rubberlike elasticity theory of polymers by James and Guth[23−26] and others.[27−33] The theory assumes that polymer chains are phantomlike, i.e., they can pass freely through one another, so that excluded volume effects are to be completely neglected. It is also assumed that the distributions of the end-to-end vectors for polymer chains are Gaussian. This means that mechanically the network behaves as a collection of nodes (junctions) connected by simple Hookean springs and both chains and junctions fluctuate harmonically around their mean positions. Kloczkowski et al.[28] obtained analytical solutions for this model by assuming that all junctions in the network have the same connectivity ϕ (i.e., each junction is connected to ϕ other chains) and that a polymer network has the topology of an infinite tree. The theory provides analytical expressions for fluctuations of chains and junctions in such networks and for correlations of instantaneous fluctuations of two different points within the network.

The theory of phantom elastomeric networks was successfully adapted to treat protein motions originally as the Gaussian network model (GNM) by Bahar and Erman[1] and others.[3,6,34−38] Their coarse-grained model was based on an earlier work of Tirion[39] who proposed that both nonbonded and covalently bonded atomic contacts in proteins could be modeled using a universal single spring constant in a harmonic analysis of protein dynamics. She assumed that two atoms are connected by a spring if they are separated by a distance smaller than a specified cutoff value. This defines a connectivity matrix for a system of nodes connected by springs. The coarse-grained GNM model enables computation of fluctuations of residues around their mean positions in protein structures directly from this connectivity matrix. Fluctuations of residues are simply expressed by the diagonal elements of the inverse of the connectivity matrix. Theoretical predictions are usually in quite good agreement with crystallographic temperature factors (*B*-factors) that measure the extent of disorder in crystallographically determined positions of atoms resulting from thermal motions. Several variations of the elastic network approach to treat protein dynamics have been proposed recently[4,35,36,40,41] that additionally improve agreement of theoretical results with *B*-factors.

In the present paper, we will apply analytical results from the theory of polymer networks obtained originally for treelike networks to the external loops in proteins, and then we will compare these results with results from GNM computations (based on the known packing details within a protein structure). It is worthwhile mentioning that both the original theory of phantom polymer networks and the elastic network models of proteins are based on the assumption that excluded volume effects are completely negligible. The theory of phantom Gaussian networks, although developed for polymer networks with the topology of an ideal infinite tree, works well for real polymer networks that contain many loops. This means that the detailed topology of the network is really not so essential for studies of individual chains.

The theory of phantom Gaussian networks provides analytical expressions for fluctuations of chains and junctions in the polymer network having connectivity ϕ (where ϕ is the number of polymer chains connected at each junction), which is constant. However, since each end of the chain in an exterior loop of a protein can be connected to a different number of springs, the original theory[28] had to be modified to reflect having junctions at the two opposite ends of the loop with different functionalities ϕ_1 and ϕ_2. We analytically compute the mean-square fluctuations and correlations of the instantaneous fluctuations for junctions and points along the polymer chains in such a treelike network,[42] and here apply these analytical results to treat protein loops. The comparison of analytical predictions with the results of GNM computations for proteins with known crystallographic coordinates of loops overall show an excellent agreement. Our results demonstrate that it is possible to model theoretically the motions of protein loops using the Gaussian model from the polymer network without knowing the structural details of the loop itself.

The structure of the present paper is as follows. First, we describe briefly the Gaussian theory of random polymer networks and show how fluctuations of chains and junctions (cross-links) and covariances among them can be computed analytically for a network with an ideal treelike topology. In the next section we discuss the Gaussian network model (GNM) of proteins and its relationship to the earlier discussed theory of random polymer networks. Later, we compare the results from the GNM for several proteins with large external loops having known structures with the analytical results based on the theory of random polymer networks and the experimental *B*-factors. Other possible applications of our method for computing the structures of loops could be made to treat loops in nucleic acids, and other loops in large

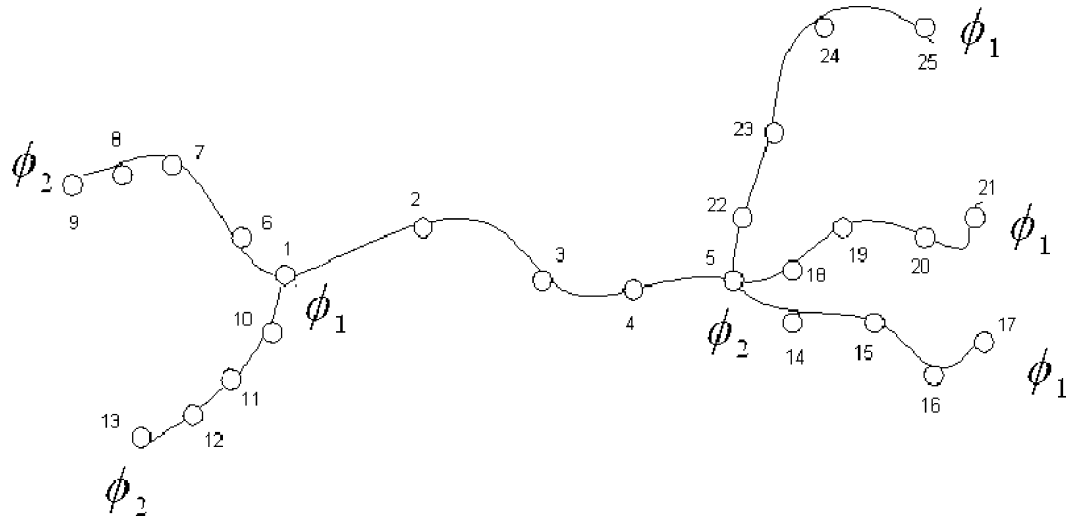

Figure 1. Treelike network with alternating functionalities separating each chain into four segments of equal length by three additional 2-functional junctions.

biological structures, such as the ribosome, for the prediction of protein function and for drug design.

Methods

Theory of Random Polymer Networks. The first theory of rubber elasticity was proposed by Kuhn in late 1930s.[43] The theory was further developed by Treloar.[30,31] It was based on the assumptions that the rubber network consists on ν freely jointed Gaussian chains, which are cross-linked. It was also assumed that positions of the junctions (points of the chemical cross-links) deform affinely upon mechanical deformation of the rubber. The theory of phantom networks was developed in the 1940s by James and Guth.[23−26,44] They also considered the network to be composed of cross-linked Gaussian chains. Additionally they assumed that there are two types of network junctions. Junctions which are at the surface of the rubber are fixed and deform affinely with the macroscopic strain, while the junctions inside the network are free to fluctuate around their mean positions. They assumed that the behavior of the network is determined only by the connectivities of network chains and neglected the effect of the excluded volume of the chains. The chains in their model are phantom-like; i.e., they may pass freely through one another.

Chain dimensions and fluctuations in random elastomeric networks were studied by Flory,[27] and by Kloczkowski, Mark, and Erman[28] who examined in detail the behavior of phantom Gaussian networks in the undeformed state. It can be shown that the mean-square fluctuations in position of a junction i $\langle(\Delta\mathbf{R}_i)^2\rangle$ are related to the element ${\Gamma_{II}}^{-1}$ of the inverse of the connectivity matrix $\mathbf{\Gamma}$, and more generally the covariances in positions of points i and j $\langle(\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j)\rangle$ are related to ${\Gamma_{II}}^{-1}$

$$\langle(\Delta\mathbf{R}_i\cdot\Delta\mathbf{R}_j)\rangle = \frac{3\langle r^2\rangle_0}{2}{\Gamma_{ij}}^{-1} \qquad (1)$$

The elements of the inverse matrix have been calculated analytically for the network with the topology of an infinite tree, composed of chains of equal length (unimodal network), with equal mean square end-to-end distances $\langle r^2\rangle_0$ in the undeformed state. It is assumed that the network has functionality ϕ, i.e., that each free junction connects exactly ϕ chains.

Examples of unifunctional networks, recurrence relations between fluctuations of junctions in the neighboring tiers of the tree, recurrence relations between fluctuations of two junctions m and n separated by d other junctions along the path joining m and n, recurrence relations between fluctuations of points along the chains in the network and covariances of fluctuations among such points, recurrence relations between the elements of the inverse connectivity matrix $\mathbf{\Gamma}^{-1}$ are given by Kloczkowski et al. in ref 28 and presented briefly in Appendix A in the Supporting Information.

Because most models of real polymer networks use phantom network as a reference state for the construction of the real network models, these results are significant for rubber elasticity. The Gaussian network model has been also extended to proteins, as will be described later.

Theory of Random Polymer Networks with Alternating Functionality. The theory was developed in Skliros et al.[42,45] and is presented briefly in Appendix B in the Supporting Information. To study fluctuations of points along the chain, we follow the method proposed in refs 28 and 46 and assume that all chains consist of n equal length segments and of $n - 1$ junctions of functionality 2, which connect these segments. Figure 1 illustrates this approach and the method of numbering all junctions for a treelike network with alternating multifunctional functionalities composed of two tiers.

Although we have obtained the most general solution of the problem when two points i and j can be separated by several multifunctional junctions,[42] we show here only the results when these two points belong to the same chain; i.e., there are no multifunctional junctions between them. Additionally, since the network is assumed to be infinite, we will concentrate on the case for the central first tier shown in the center of Figure 1.

The positions of 2-functional junctions i and j can be expressed as the fraction of the chain between ϕ_1-functional and ϕ_2-functional junctions, counted from the closest ϕ_1-functional junction on the left of points i or j in Figure 1; $\zeta = (i - 1)/n$; $\theta = (j - 1)/n$.

The final result is

$$\begin{bmatrix} \langle(\Delta R_i)^2\rangle & \langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle \\ \langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle & \langle(\Delta R_j)^2\rangle \end{bmatrix} = \frac{3n}{2\gamma_0} \times$$

$$\begin{bmatrix} \dfrac{\phi_2(\phi_1-1)}{\phi_1(\phi_1\phi_2-\phi_1-\phi_2)} + \dfrac{\zeta(1-\zeta)(\phi_1\phi_2-\phi_1-\phi_2)+\zeta(\phi_1-\phi_2)}{\phi_1\phi_2} & \dfrac{\phi_2(\phi_1-1)}{\phi_1(\phi_1\phi_2-\phi_1-\phi_2)} + \dfrac{(\phi_1\phi_2-\phi_1-\phi_2)}{\phi_1\phi_2}[\min(\zeta,\theta)-\zeta\theta] + \dfrac{\min(\zeta,\theta)}{\phi_2} - \dfrac{\max(\zeta,\theta)}{\phi_1} \\ \dfrac{\phi_2(\phi_1-1)}{\phi_1(\phi_1\phi_2-\phi_1-\phi_2)} + \dfrac{(\phi_1\phi_2-\phi_1-\phi_2)}{\phi_1\phi_2}[\min(\zeta,\theta)-\zeta\theta] + \dfrac{\min(\zeta,\theta)}{\phi_2} - \dfrac{\max(\zeta,\theta)}{\phi_1} & \dfrac{\phi_2(\phi_1-1)}{\phi_1(\phi_1\phi_2-\phi_1-\phi_2)} + \dfrac{\theta(1-\theta)(\phi_1\phi_2-\phi_1-\phi_2)+\theta(\phi_1-\phi_2)}{\phi_1\phi_2} \end{bmatrix} \quad (2)$$

Gaussian Network Model of Proteins. The Gaussian network model (GNM) was originally developed for the theory of rubberlike elasticity of random polymer networks[27,28] to calculate fluctuations of junctions and chains inside the network. That physical situation is quite different from that prevailing in a protein because the polymer chains have random forms and the protein may be a more fixed form. The model has been adapted to coarse-grained proteins in 1997 by Bahar and Erman[1,47] based on the earlier result of Tirion[39] with a single harmonic force parameter, which successfully described the large-scale motions in proteins.

The GNM is based on coarse-grained modeling of protein structure, with a single site per residue representing proteins. Positions of these sites are usually identified with the coordinates of C^α atoms in proteins, and it is assumed that both bonded and nonbonded contacts in protein structure are connected by uniform massless harmonic springs. Significantly, the atomic version gives only slightly better results than the coarse-grained model,[3] indicating that the motions are mostly representative of the overall structure, and not so much of it details.

To define which sites are in contact, a uniform cutoff distance R_c is used.[1,38,40,47] Residues separated by this distance or closer than R_c (including neighbors along the sequence) are assumed to be in contact and are connected with identical springs. This leads to the elastic network representation of a protein in the folded state that bears a resemblance to a random polymer network. While this model of a protein is closely similar to that of a rubbery network, the main difference is that in the rubber the coordinations are defined by covalent links whereas in the GNMs and ENMs the connections are primarily nonbonded contacts arising from close packing within the structure. While the GNM formally neglects the excluded volume, regions with a higher density of atoms are represented by higher density of springs, while less dense regions are represented by few springs.

The distance vector between the ith and jth sites is $\mathbf{R}_{ij}$, with $\Delta\mathbf{R}_{ij}$ being the instantaneous displacement of $\mathbf{R}_{ij}$ from the mean value $\boldsymbol{R}_{ij}^0$, and $(\Delta\boldsymbol{R}_{ij})^2$ is given by the scalar product $(\Delta\boldsymbol{R}_{ij}{}^T \cdot \Delta\boldsymbol{R}_{ij})$. The reference structure is usually the crystal structure taken from the Protein Data Bank (PDB), but could be a modeled structure or even the shape of a structure from an electron micrograph, which was filled with lattice points.[48] It can be shown[27,28] that

$$\langle \Delta\boldsymbol{R}_i \cdot \Delta\boldsymbol{R}_j \rangle = \frac{3k_B T}{2\gamma}(\mathbf{\Gamma}^{-1})_{ij} \quad (3)$$

where $(\mathbf{\Gamma}^{-1})_{ij}$ is the ijth element of the inverse of the connectivity matrix $\mathbf{\Gamma}$.

It should be note that the connectivity matrix $\mathbf{\Gamma}$ has been defined so that all elements in every row (or column) sum to zero. Because of this $\det \mathbf{\Gamma} = 0$, the matrix is singular, and only the pseudoinverse of $\mathbf{\Gamma}$ can be computed by the use of the singular value decomposition method. The pseudoinverse of $\mathbf{\Gamma}$ may be written as $\mathbf{\Gamma}^{-1} = \mathbf{U}(\mathbf{\Lambda}^{-1})\mathbf{U}^T$ where $\mathbf{U}$ is the matrix composed of eigenvectors $\mathbf{u}_i$ ($1 \le i \le N$) of $\mathbf{\Gamma}$, and $\mathbf{\Lambda}$ is the matrix having eigenvalues of $\mathbf{\Gamma}$ on the diagonal, and zeros off-diagonal. Additionally, it can be proven that all eigenvalues λ_i of $\mathbf{\Gamma}$ are nonnegative.

Mean-square fluctuations of each C^α computed from eq 3 can be compared with the Debye−Waller factors for the C^α atoms. These temperature factors are frequently measured by X-ray crystallography for all heavy atoms in the protein structure and are deposited in the Protein Data Bank as temperature B-factors. The computed B-factors for the ith residue are given by:

$$B_i = 8\pi^2\langle(\Delta R_i)^2\rangle/3 \quad (4)$$

The B-factors computed by the GNM usually are in excellent agreement with experimental data,[2] although even better agreement is found when compared with the averages of internal distances from NMR ensembles.[4,49]

The matrix $\mathbf{\Gamma}^{-1}$ can be written as the sum of contributions from individual modes[50]

$$\mathbf{\Gamma}^{-1} = \sum_k \lambda_k{}^{-1}\boldsymbol{u}_k\boldsymbol{u}_k^T \quad (5)$$

where the zero eigenvalues (physically corresponding to motions of the center of mass of the system) are excluded from the sum. The ith component of the eigenvector $\boldsymbol{u}_k$ (corresponding to the kth normal mode) specifies the magnitude of fluctuational motions of the Ith residue in the protein exerted by the kth mode. If the eigenvalues are ordered according in an ascending order starting from zero, then the most meaningful contributions in eq 5 are given by the smallest nonzero eigenvalues λ_k, which correspond to the large-scale slow modes. The slowest modes play a dominant role in the fluctuational dynamics of structures, because their contributions to the mean-square fluctuations scale with $\lambda_k{}^{-1}$. It has been shown that the most essential motions of proteins[51−53] or large biological structures such as the ribosome,[54−57] which are associated with their biological function, are clearly identifiable within a few of the slowest modes of the GNM or ENM. The large-scale

changes of protein conformations between "open" and "closed" forms, or domain swapping in proteins, can be also explained well by these ENMs.[58,59] The Gaussian network model is the simplest version of several different ENMs. It has been extended to treat anisotropic fluctuations with vector directions for the motions,[40] and hierarchical[35] or mixed[36] levels of coarse graining.

Results and Discussion

Prediction of Motions of Loops in Proteins. We have studied in detail the motions of external loops in five different proteins: tubulin (PDB code 1tub, tubulin α/β dimer), reverse transcriptase (1n5y), triose phosphate isomerase (1tph, human triose phosphate isomerase), protease (1j71, extracellular aspartic proteinase from *Candida tropicalis* yeast), and myoglobin (2v1k, ferrous deoxymyoglobin at pH 6.8). Tubulin, reverse transcriptase, and triose phosphate isomerase are composed of two monomers and have 867, 910, and 496 residues, respectively. Protease and myoglobin each contain single chains with 338 and 163 residues, respectively. We first locate loops of these proteins that are on the surface and compute the mean-square fluctuations of all the residues in these loops, and their cross correlations (covariances) between fluctuations of two different residues by using both the GNM (eq 4) and the analytical formula (eq 2) derived for a polymer network with alternating functionality. More specifically, we consider the loops as chains where the first and the last residue are junctions with functionalities equal to the actual connectivities for these residues with the remainder of the protein as given for a cut off distance (7 Å); the other residues of the loop are considered as junctions having functionality two. Now since the functionality of the two terminal loop residues may be different, in order to find the auto and cross covariances of the loop residues we use eq 2. For the case of the GNM model, we find the connectivity matrix (for the whole protein including PDB data for residues forming loops) and then we find the pseudoinverse thereof by using singular value decomposition. The fluctuations and the covariances of the residues of each loop residue are found based on eq 4.

We identify protein loops by first excluding helices and β-strands in the protein structure, leaving only coils. The criterion for a loop is the requirement that four or more consecutive coil residues are located on the protein surface. We illustrate these loops in protein structures for three of the studied proteins by coloring them blue in Figure 2.

We calculated covariances of instantaneous fluctuations of residues in loops both from eq 2 (polymer theory of rubberlike elasticity) and from eq 4 (GNM computations based on the complete protein structure). The results obtained are shown in Figures 3−7. Curves with squares show covariances calculated from eq 2 using only information on the connectivity of the terminal residues of protein loops (functionality of their junctions) and the length of a loop, while curves with dots display covariances calculated from eq 4 for GNM applied to the whole protein. The pattern for the residue−residue indexing is as follows: initially the index shows the covariance of the first residue in the loop with itself and with all others, then of the second residue with itself and with all others (except the first one), etc. For a loop composed of n residues the residue−residues index changes from 1 to $n(n + 1)/2$.

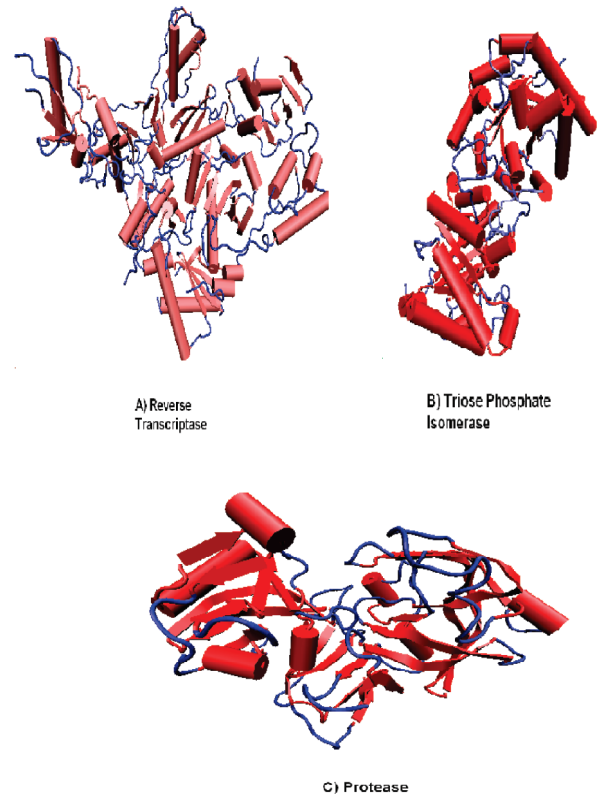

Figure 2. (A) Loops (colored in blue) of reverse transcriptase (A), triose phosphate isomerase (B), and protease (C).

Figure 3 shows the values of covariances calculated from polymer rubberlike elasticity theory (squares) and from GNM (dots) for the following loops (for indexing of all loops see Appendix C in the Supporting Information): (a) reverse transcriptase loop number 4 ($n = 7$), (b) tubulin loop number 4 ($n = 5$), (c) triose phosphate isomerase loop number 3 ($n = 5$), (d) protease loop number 7 ($n = 6$), and (e) myoglobin loop number 4 ($n = 5$). We see that for these cases the local approximation based on eq 2 provides an excellent result that very well approximates the whole structure result based on the GNM.

In addition to covariances of the instantaneous fluctuations, it is interesting to analyze correlations among them defined by

$$\text{corr} = \frac{\langle(\Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j)\rangle}{\sqrt{\langle(\Delta\mathbf{R}_i)^2\rangle\langle(\Delta\mathbf{R}_j)^2\rangle}} \qquad (6)$$

Figure 4 shows the correlations obtained by the GNM and the polymer elastic theory for the loops analyzed earlier in Figure 3.

Figure 5 shows the values of covariances computed for four loops of reverse transcriptase with asymmetry in the connectivities ϕ_1, ϕ_2 of the terminal junctions for loops: (a) loop number 4 ($n = 7$, $\phi_1 = 4$, $\phi_2 = 11$), (b) number 5 ($n = 5$, $\phi_1 = 8$, $\phi_2 = 3$), (c) number 11 ($n = 6$, $\phi_1 = 5$, $\phi_2 = 10$), and (d) number 14 ($n = 4$, $\phi_1 = 5$, $\phi_2 = 13$). We see that

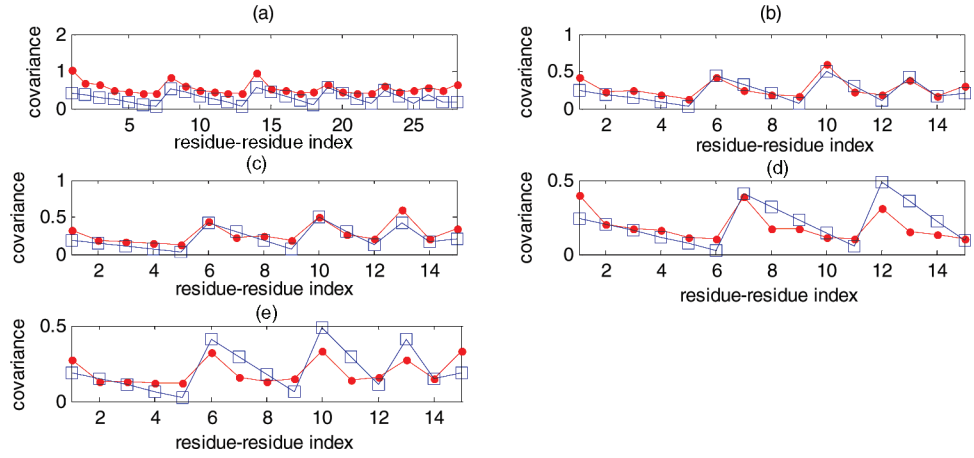

Figure 3. Covariances of instantaneous fluctuations calculated by using theory of rubberlike elasticity (eq 2) (squares) and GNM (eq 4) (dots) for the following individual loops: (a) reverse transcriptase loop no. 4, (b) tubulin loop no. 4, (c) triose phosphate isomerase loop no. 3, (d) protease loop no. 7, and (e) myoglobin loop no. 4. The abscissa shows the index for pairs of residues.

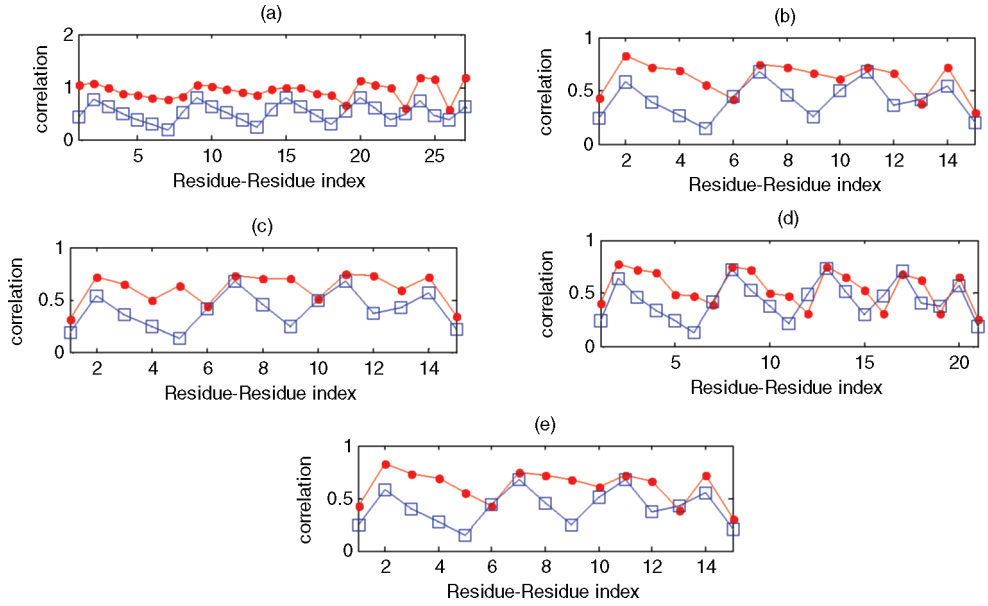

Figure 4. Correlations of instantaneous fluctuations computed by using the theory of rubberlike elasticity (eq 2) (squares) and the GNM (eq 4) (dots) for the following individual loops: (a) reverse transcriptase loop no. 4, (b) tubulin loop no. 4, (c) triose phosphate isomerase loop no. 3, (d) protease loop no. 7, and (e) myoglobin loop no. 4. The abscissa shows the index for pairs of residues.

the local approximations based on the polymer network model are more successful for longer loops than for short ones.

It is also relevant to examine the relationships between the fluctuations of the loop residues computed from polymer rubberlike elasticity model, and from GNM with experimental *B*-factors. Figure 6 shows plots of the mean square fluctuations of loop residues obtained from these three different sources for all 16 loops of the protease. For purposes of comparison, all three quantities are normalized. (For example $B_i^{\text{norm}} = (B_i - B_{\min})/(B_{\max} - B_{\min})$, where i is the residue index for the particular loop. The same normalization is carried for the fluctuations computed from GNM and from polymer phantom network model.) We see that for some of these cases the local approximation based on our model approximates the GNM results very well. The main reasons for the good or bad approximation by the local method to the GNM results are the connectivities of the residues of the loops.

Another crucial issue is whether the covariances of instantaneous fluctuations decay similarly with respect to the sequence distance between the residues. To address this problem, we have plotted in Figure 7 the covariances of the first residue in each loop with respect to the other residues of the same loop as a function of the sequence distance between these two residues.

Figures 3−7 indicate that the covariances of instantaneous fluctuations obtained both by considering the loops as individual entities (theoretical model) and as a whole structure (GNM) are closely similar. We computed the correlations of these covariances for all loops for all the proteins studied here. Our computations indicate that the

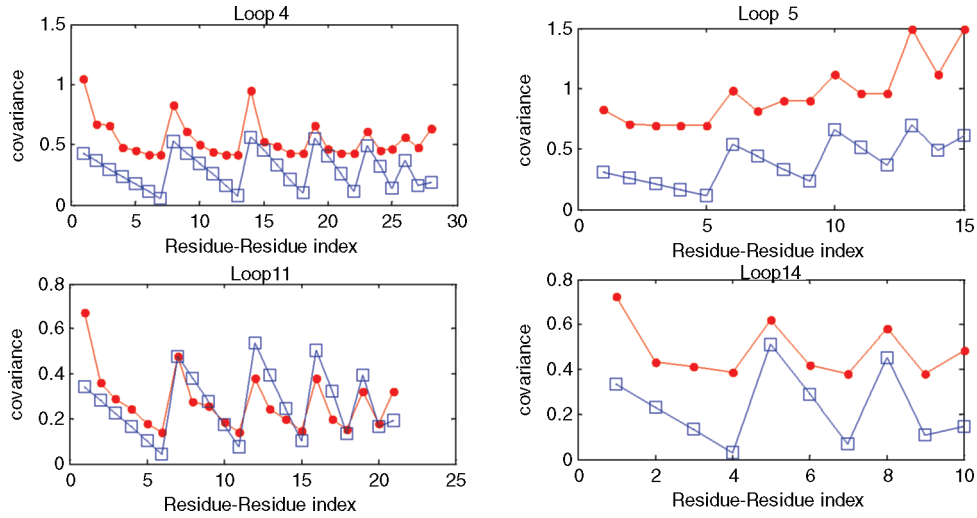

Figure 5. Covariances of the instantaneous fluctuations calculated by using the theory of rubber elasticity (eq 2) (squares) and GNM (eq 4) (dots) for loops no. 4, 5, 11, and 14 of reverse transcriptase. The abscissa shows the index for pairs of residues, with indexing described in the text.

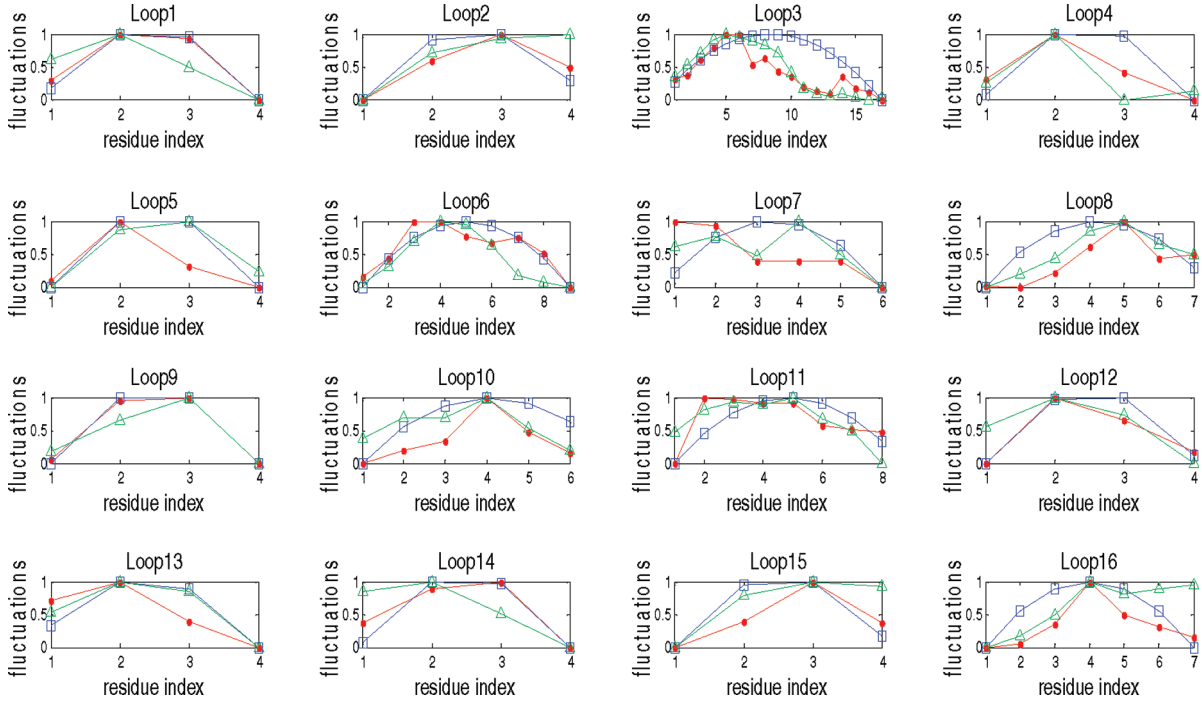

Figure 6. Fluctuations of the residues for all loops of protease. Fluctuations computed from polymer theory of rubberlike elasticity (squares) and from GNM (dots) are compared with *B*-factors (triangles). All fluctuations have been normalized. The abscissa is the residue index in the loop.

average correlation of covariances (averaged over all loops in a given protein) is the largest (0.70) for triose phosphate isomerase, 0.66 for reverse transcriptase, and the smallest (0.38) for myoglobin, which has only four very short loops. All profiles shown in these figures show a close resemblance between the behavior of our new loop modeling and the whole protein modeling with the Gaussian network model. This is at first surprising, since previous results with the Gaussian network model indicated that the whole structure was needed in order to compute the motions of any part. What we are seeing here is that the individual loops and their simplified representations are generally sufficient to compute the relative mobilities for the individual parts of the loop. Information contained in Figures 3−7 for additional loops of the proteins are provided in Appendix D in the Supporting Information. Appendix E in the Supporting Information gives the correlations of the covariances for every loop of each protein we have studied as a function of the sum of the functionalities of the two terminal junctions in each loop. We have not noticed any apparent relationship between these two quantities. Appendix F in the Supporting Information shows results of computations of covariances of instantaneous fluctuations of residues belonging to helices for the proteins of study, similarly as was done earlier for loops. Our computations show that polymer network approximation does not work as well for helices as for the loops. Appendix G in the Supporting Information shows the computed correlation coefficients between the predicted

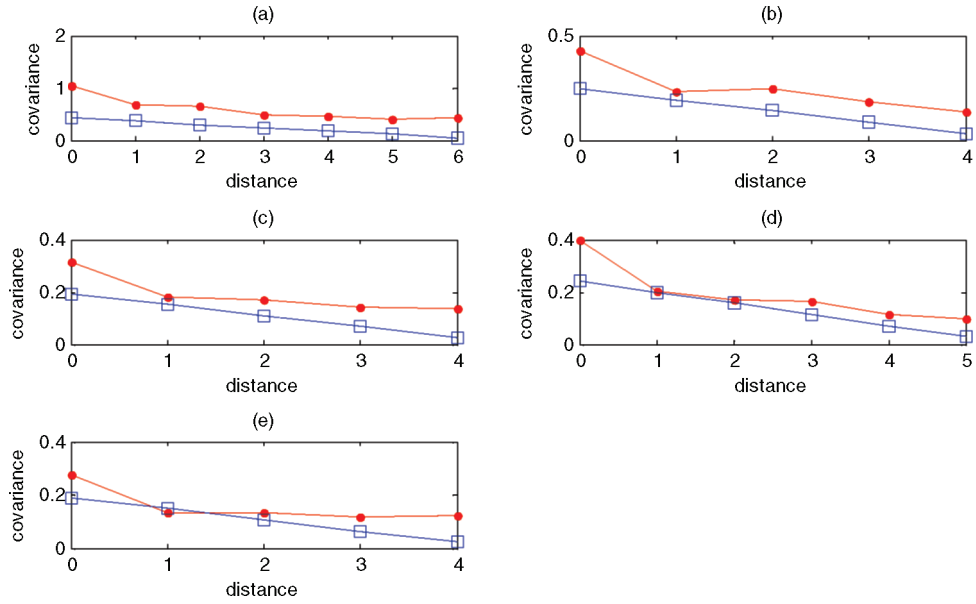

Figure 7. Covariances of instantaneous fluctuations of the first residue of the loop with other residues in the same loop as a function of the sequence distance between them for (a) reverse transcriptase loop no. 4, (b) tubulin loop no. 4, (c) triose phosphate isomerase loop no. 3, (d) protease loop no. 7, and (e) myoglobin loop no. 4. Results computed from polymer theory of rubberlike elasticity (squares) and from GNM (dots) are compared.

amplitudes of fluctuations of loop residues (using both the present analytical model and GNM theory) and the experimental *B*-factors.

Discussion. We have presented a comparison of the loop motions using both the GNM and a new approach based on the theory of rubberlike elasticity of polymer networks. For the latter, the loop is modeled by assuming that the two ends have functionalities ϕ_1 and ϕ_2 connecting them with the remainder of the protein structure, whereas the intermediate residues of the loop uniformly have the functionality of two, connecting them only to their sequence neighbors. We then calculated the mean square fluctuations (variances) and covariances for the loop residues by using formulas derived analytically by us for a treelike polymer network with alternating functionalities. We then applied this new approach to all external loops in five different proteins (reverse transcriptase, triose phosphate isomerase, tubulin, protease, and myoglobin) and compared analytical results for the mean square fluctuations and the covariances with GNM computations based on the coordinates for all residues of the loops. For each loop we have plotted the covariances between instantaneous fluctuations of pairs of residues of the loop, and the mean square fluctuations of individual loop residues. We have also compared the covariance of instantaneous fluctuations between two residues of the loop as a function of the distance between them. The comparisons between these two models show that the local approximation of the loops that describes the motion of the loop residues based on polymer treelike network topology independently from the rest of the protein structure closely approximates the results obtained from GNM where the whole protein structure is taken into account.

Loop mobilities have long been considered to be important for function. Loops on the surfaces of proteins are often the most uncertain parts of structures. It is quite likely that many loops are artificially immobilized and compressed against the body of the protein in the crystal environment. What has been done in the present work is to develop and compare two simple models for loop mobilities. These two quite different approaches yield similar results. Our finding, that *computed loop motions are similar whether treated as independent or in the context of the whole protein structure*, implies that the loop motions are occurring along relatively well-defined pathways, which likely could be mapped out in future computations. This approach may be used not only for loops of proteins but also for the large loops frequently occurring in nucleic acid stem−loop structures, as well as the much larger loops originating in double-stranded DNA, when multiple subunit proteins bind at widely separated positions. This critical problem for transcription regulation requires coarse graining of the structure since the loops can be thousands of base pairs in length. The more difficult problem may be how to introduce additional interactions within these DNA loops when supercoiling is introduced.

Utility of This Approach for Describing the Motions of Loops with Unknown Structures. Often in protein structures loops are missing in the crystal structure. The approach described in this paper could be used to predict probably structures for these missing loops since the only requirement is knowing the connectivity numbers for the two ends and the number of residues in the loop.

In the present work, we have considered only the relative magnitudes of these motions, which compare favorably. In the future, it will be essential to develop a vector version of the present loop modeling to define actual pathways, to compare with the ANM and the anisotropic temperature factors. These pathways will be presented in future work, where we also investigate the effects of interactions between loop residues and the body of the protein. The results from the present work indicate that our new approach based on

polymer elasticity theory provides a close approximation to the GNM model that includes the effects of the whole structure.

The polymer rubberlike elasticity model only predicts a relatively simple pattern of fluctuations of residues across a loop with a convex shape, and with the residues close the center of the loops having a higher amplitude. However, according to this model the central residue of the loop does not always have the maximum amplitude of fluctuations. We have also an effect from the functionality of junctions on both ends of the loop. The maximum is shifted toward the junction which having lower functionality. This effect is also observed for experimental *B*-factors for loop residues, although sometimes there are exceptions to this rule, due to significant interactions of loop residues with the remainder of the protein. Our predictions are limited somewhat by the simplicity of the theoretical model, and by the assumption that the motions of loops are unobstructed by the remaining part of the protein, but nevertheless they do enable predicting the basic features of these motions.

Acknowledgment. We are pleased to acknowledge the financial support provided by the National Institutes of Health through grants R01GM081680, R01GM072014, and R01GM073095.

Supporting Information Available: In Appendix A we provide a synopsis of the theory of random polymer networks. In the Appendix B we present a summary of the theory of random polymer networks with alternating functionality. In Appendix C we show tables listing the loops and the loop residues for all the proteins of this study. In Appendix D we provide information (supplementary to that contained in Figures 3−7) for more loops for all the proteins of study. In Appendix E we show the calculated correlation of covariances of instantaneous fluctuations for every loop of each protein as a function of the sum of the functionalities of their terminal junctions. In Appendix F we have calculated the covariances of instantaneous fluctuations of residues belonging to helices for all proteins studied, as we did earlier for the loops. Appendix G lists the computed correlation coefficients between the predicted amplitudes of fluctuations (using both the present analytical model and GNM theory) and the experimental *B*-factors for protein loops.

This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Des.* **1997**, *2* (3), 173–181.

(2) Kundu, S.; Melton, J. S.; Sorensen, D. C.; Phillips, G. N. Dynamics of proteins in crystals: Comparison of experiment with simple models. *Biophys. J.* **2002**, *83* (2), 723–732.

(3) Sen, T. Z.; Feng, Y. P.; Garcia, J. V.; Kloczkowski, A.; Jernigan, R. L. The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models. *J. Chem. Theory Comput.* **2006**, *2* (3), 696–704.

(4) Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R. L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* **2008**, *16* (2), 321–330.

(5) Yang, L.; Song, G.; Jernigan, R. L. How well can we understand large-scale protein motions using normal modes of elastic network models. *Biophys. J.* **2007**, *93* (3), 920–929.

(6) Lu, M. Y.; Ma, J. P. The role of shape in determining molecular motions. *Biophys. J.* **2005**, *89* (4), 2395–2401.

(7) Fiser, A.; Do, R. K. G.; Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9* (9), 1753–1773.

(8) Espadaler, J.; Fernandez-Fuentes, N.; Hermoso, A.; Querol, E.; Aviles, F. X.; Sternberg, M. J. E.; Oliva, B. ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res.* **2004**, *32*, D185–D188.

(9) Oliva, B.; Bates, P. A.; Querol, E.; Aviles, F. X.; Sternberg, M. J. E. An automated classification of the structure of protein loops. *J. Mol. Biol.* **1997**, *266* (4), 814–830.

(10) Panchenko, A. R.; Madej, T. Structural similarity of loops in protein families: toward the understanding of protein evolution. *BMC Evolut. Biol.* **2005**, *5*, Art. No. 10.

(11) Groban, E. S.; Narayanan, A.; Jacobson, M. P. Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PloS Comput. Biol.* **2006**, *2* (4), 238–250.

(12) Hu, X. Z.; Wang, H. C.; Ke, H. M.; Kuhlman, B. High-resolution design of a protein loop. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (45), 17668–17673.

(13) Bos, C.; Lorenzen, D.; Braun, V. Specific in vivo labeling of cell surface-exposed protein loops: Reactive cysteines in the predicted gating loop mark a ferrichrome binding site and a ligand-induced conformational change of the Escherichia coli FhuA protein. *J. Bacteriol.* **1998**, *180* (3), 605–613.

(14) Li, C.; Banfield, M. J.; Dennison, C. Engineering copper sites in proteins: Loops confer native structures and properties to chimeric cupredoxins. *J. Am. Chem. Soc.* **2007**, *129*, 709–718.

(15) Smith, J. W.; Tachias, K.; Madison, E. L. Protein loop grafting to construct a variant of tissue-type plasminogen activator that binds platelet integrin alpha(IIb)beta(3). *J. Biol. Chem.* **1995**, *270* (51), 30486–30490.

(16) Sudarsanam, S.; Dubose, R. F.; March, C. J.; Srinivasan, S. Modeling Protein Loops Using A Phi-I+1, Psi-I Dimer Database. *Protein Sci.* **1995**, *4* (7), 1412–1420.

(17) vanVlijmen, H. W. T.; Karplus, M. PDB-based protein loop prediction: Parameters for selection and methods for optimization. *J. Mol. Biol.* **1997**, *267* (4), 975–1001.

(18) Kolodny, R.; Guibas, L.; Levitt, M.; Koehl, P. Inverse kinematics in biology: The protein loop closure problem. *Int. J. Robotics Res.* **2005**, *24* (2−3), 151–163.

(19) Gerstein, M.; Chothia, C. Analysis of Protein Loop Closure - 2 Types of Hinges Produce One Motion in Lactate-Dehydrogenase. *J. Mol. Biol.* **1991**, *220* (1), 133–149.

(20) Krieger, F.; Fierz, B.; Axthelm, F.; Joder, K.; Meyer, D.; Kiefhaber, T. Intrachain diffusion in a protein loop fragment from carp parvalbumin. *Chem. Phys.* **2004**, *307* (2−3), 209–215.

(21) Li, W. Z.; Liu, Z. J.; Lai, L. H. Protein loops on structurally similar scaffolds: Database and conformational analysis. *Biopolymers* **1999**, *49* (6), 481–495.

(22) Burke, D. F.; Deane, C. M. Improved protein loop prediction from sequence alone. *Protein Eng.* **2001**, *14* (7), 473–478.

(23) James, H. M.; Guth, E. Theory of the Increase in Rigidity of Rubber During Cure. *J. Chem. Phys.* **1947**, *15* (9), 669–683.

(24) James, H. M. Statistical Properties of Networks of Flexible Chains. *J. Chem. Phys.* **1947**, *15* (9), 651–668.

(25) James, H. M.; Guth, E. Simple Presentation of Network Theory of Rubber, with A Discussion of Other Theories. *J. Polym. Sci.* **1949**, *4* (2), 153–182.

(26) James, H. M.; Guth, E. Statistical Thermodynamics of Rubber Elasticity. *J. Chem. Phys.* **1953**, *21* (6), 1039–1049.

(27) Flory, P. J. Statistical Thermodynamics of Random Networks. *Proc. R. Soc. London Ser. A—Math. Phys. Eng. Sci.* **1976**, *351* (1666), 351–380.

(28) Kloczkowski, A.; Mark, J. E.; Erman, B. Chain Dimensions and Fluctuations in Random Elastomeric Networks 0.1. Phantom Gaussian Networks in the Undeformed State. *Macromolecules* **1989**, *22*, 1423–1432.

(29) Kloczkowski, A.; Mark, J. E.; Frisch, H. L. The relaxation spectrum for Gaussian Networks. *Macromolecules* **1990**, *23*, 3481–3490.

(30) Treloar, L. R. G. The Elasticity of A Network of Long-Chain Molecules 0.3. *Trans. Faraday Soc.* **1946**, *42* (1−2), 83–94.

(31) Treloar, L. R. G. The Statistical Length of Long-Chain Molecules. *Trans. Faraday Soc.* **1946**, *42* (1−2), 77–82.

(32) Kloczkowski, A.; Mark, J. E.; Erman, B. Fluctuations, Correlations, and Small-Angle Neutron-Scattering from End-Linked Gaussian Chains in Regular Bimodal Networks. *Macromolecules* **1991**, *24*, 3266–3275.

(33) Kloczkowski, A.; Mark, J. E.; Erman, B. A Diffused-Constraint Theory for the Elasticity of Amorphous Polymer Networks 0.1. Fundamentals and Stress-Strain Isotherms in Elongation. *Macromolecules* **1995**, *28*, 5089–5096.

(34) Bahar, I.; Rader, A. J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* **2005**, *15* (5), 586–592.

(35) Doruker, P.; Jernigan, R. L.; Bahar, I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.* **2002**, *23* (1), 119–127.

(36) Kurkcuoglu, O.; Jernigan, R. L.; Doruker, P. Collective dynamics of large proteins from mixed coarse-grained elastic network model. *QSAR Comb. Sci.* **2005**, *24* (4), 443–448.

(37) Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y. H. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Struct., Funct., Genet.* **2000**, *41* (1), 1–7.

(38) Tama, F.; Sanejouand, Y. H. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **2001**, *14* (1), 1–6.

(39) Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **1996**, *77* (9), 1905–1908.

(40) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **2001**, *80* (1), 505–515.

(41) Song, G.; Jernigan, R. L. vGNM: A better model for understanding the dynamics of proteins in crystals. *J. Mol. Biol.* **2007**, *369* (3), 880–893.

(42) Skliros, A.; Mark, J. E.; Kloczkowski, A. Chain Dimensions and Fluctuations in Elastomeric Networks in which the Junctions Alternate Regularly in their Functionality. *J. Chem. Phys.* **2009**, *130*, 064905.

(43) Kuhn, W. Relationship between molecular size, static molecular shape and elastic properties of high polymer materials. *Kolloid-Z.* **1936**, *76*, 258.

(44) Jensen, J. H.; Gordon, M. S. An approximate formula for the intermolecular Pauli repulsion between closed shell molecules. II. Application to the effective fragment potential method. *J. Chem. Phys.* **1998**, *108* (12), 4772–4782.

(45) Skliros, A.; Mark, J. E.; Kloczkowski, A. Small-Angle Neutron Scattering from Elastomeric Networks in which the Junctions Alternate Regularly in their Functionality. *Macromol. Theory Simul.* **2009**, *18* (9), 537–544.

(46) Erman, B.; Kloczkowski, A.; Mark, J. E. Chain Dimensions and Fluctuations in Random Elastomeric Networks 0.2. Dependence of Chain Dimensions and Fluctuations on Macroscopic Strain. *Macromolecules* **1989**, *22*, 1432–1437.

(47) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **1997**, *79* (16), 3090–3093.

(48) Doruker, P.; Jernigan, R. L. Functional motions can be extracted from on-lattice construction of protein structures. *Proteins: Struct., Funct., Genet.* **2003**, *53* (2), 174–181.

(49) Yang, L. W.; Eyal, E.; Chennubhotla, C.; Jee, J.; Gronenborn, A. M.; Bahar, I. Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure* **2007**, *15* (6), 741–749.

(50) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **1997**, *79* (16), 3090–3093.

(51) Keskin, O.; Durell, S. R.; Bahar, I.; Jernigan, R. L.; Covell, D. G. Relating molecular flexibility to function: A case study of tubulin. *Biophys. J.* **2002**, *83* (2), 663–680.

(52) Keskin, O.; Bahar, I.; Flatow, D.; Covell, D. G.; Jernigan, R. L. Molecular mechanisms of chaperonin GroEL-GroES function. *Biochemistry* **2002**, *41*, 491–501.

(53) Navizet, I.; Lavery, R.; Jernigan, R. L. Myosin flexibility: Structural domains and collective vibrations. *Proteins: Struct., Funct., Genet.* **2004**, *54* (3), 384–393.

(54) Wang, Y. M.; Rader, A. J.; Bahar, I.; Jernigan, R. L. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.* **2004**, *147* (3), 302–314.

(55) Wang, Y. M.; Jernigan, R. L. Comparison of tRNA motions in the free and ribosomal bound structures. *Biophys. J.* **2005**, *89* (5), 3399–3409.

(56) Yan, A. M.; Wang, Y. M.; Kloczkowski, A.; Jernigan, R. L. Effects of Protein Subunits Removal on the Computed Motions of Partial 30S Structures of the Ribosome. *J. Chem. Theory Comput.* **2008**, *4* (10), 1757–1767.

(57) Kurkcuoglu, O.; Doruker, P.; Sen, T. Z.; Kloczkowski, A.; Jernigan, R. L. The ribosome structure controls and directs mRNA entry, translocation and exit dynamics. *Phys. Biol.* **2008**, *5* (4), XXX.

(58) Kundu, S.; Jernigan, R. L. Molecular mechanism of domain swapping in proteins: An analysis of slower motions. *Biophys. J.* **2004**, *86* (6), 3846–3854.

(59) Feng, Y. P.; Yang, L.; Kloczkowski, A.; Jernigan, R. L. The energy profiles of atomic conformational transition intermediates of adenylate kinase. *Proteins: Struct., Funct., Bioinf.* **2009**, *77* (3), 551–558.

CT1001413