Published in final edited form as:
Phys Biol. ; 9(1): . doi:10.1088/1478-3975/9/1/014001.

The Importance of Slow Motions for Protein Functional Loops

Aris Skliros[1], **Michael T. Zimmermann**[1], **Debkanta Chakraborty**[1], **Saras Saraswathi**[1,2], **Ataur R. Katebi**[1], **Sumudu P. Leelananda**[1], **Andrzej Kloczkowski**[1,2], and **Robert L. Jernigan**[1]

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

[2]Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital and Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH 43205

Abstract

Loops in proteins connect secondary structures such as alpha-helix and beta-sheet, are often on the surface, and may play a critical role in some functions of a protein. The mobility of loops is central for the motional freedom and flexibility requirements of active-site loops and may play a critical role for some functions. The structures and behaviors of loops have not been much studied in the context of the whole structure and its overall motions, and especially how these might be coupled. Here we investigate loop motions by using coarse-grained structures (C^{α} atoms only) to solve for the motions of the system by applying Lagrange equations with elastic network models to learn about which loops move in an independent fashion and which move in coordination with domain motions, faster and slower, respectively. The normal modes of the system are calculated using eigen-decomposition of the stiffness matrix. The contribution of individual modes and groups of modes are investigated for their effects on all residues in each loop by using Fourier analyses. Our results indicate overall that the motions of functional sets of loops behave in similar ways as the whole structure. But, overall only a relatively few loops move in coordination with the dominant slow modes of motion, and that these are often closely related to function.

Introduction

The importance of understanding loops in proteins

Protein motions are extremely important for their functioning, and it is now well established that domain motions are the dominant motions and that these motions are often relatable to their functions. So immediately there the question arises of whether loops are controlled in their motions by the domain motions and are slow, or whether they move independently and thus more rapidly. The first case, where the loops move together with the large domain motions, corresponds to the protein structure being strongly cooperative in its motions, and these motions will reflect a type of allostery. Distinguishing between these two extremes is the intention of the present study of the behavior of protein surface loops.

One category of loops whose function is clear are those large loops that cover binding sites, and are clearly important since they have to open in order to facilitate binding, and thus these motions are clearly critical for function. But, other loops may have functional roles such as chaperoning the transport of ligands from secondary binding sites towards their primary binding site. Such behaviors might reflect a more deterministic behavior than is usually observed in molecular simulations. New ways of investigating loop behaviors may assist in our understanding of such biased, non-random behavior in biology.

If loops that are functionally important move in a strongly coordinated way with the larger domain motions, i.e., the slowest motions, then it might even be possible to predict which loops are likely to be functional based on computations that identify them as moving more slowly. In addition there is the issue of how the loops move with respect to the domain to which they are attached. If they are fully coordinated in a positive way then they are moving as if the domain and the loop together were a rigid block. If they are moving in a strongly anti-correlated way, this would require articulated joints between the domain and the loop, but whether correlated or anti-correlated both could be motions effectively under the control of the domain motions.

First we must define loops for the purpose of this investigation. Proteins consist primarily of three types of secondary structure elements: α-helices, β–strands (forming either parallel or antiparallel β-sheets) and loops. Loop regions here are taken to be those conformationally irregular fragments of the chain, that connect between two secondary structure elements and lie upon the surface.

Loops are also quite variable in their lengths and sometimes there are even gaps in the loop regions of some reported protein structures, because of disorder. Godzik and collaborators (1) recently surveyed the PDB to find ordered-disordered pairs; residues of the same protein in two different crystals, where the atomic coordinates were resolved in one, but not in the other. They found that this type of disorder (sometimes relates to post-translational modification) is overrepresented in loop regions (46% of ordered-disordered pairs). While it might be tempting to interpret the missing parts of loops as moving independently, it seems likely that the details of the crystal packing would be likely to confound such a simple interpretation. Completing the structures of these missing regions in loops and learning about the range of loop conformations are important issues not yet a standard computation, that will not be considered here, even though these issues are important for evaluating the importance of the mobilities of all loops. Solving these issues would however result in an improved understanding of the functional roles played by loops.

Diverse approaches have been applied to fill in the missing information in loop regions. Joosten *et al.* (2) have combined structural and electron density information to find likely conformations of loop regions. Fiser *et al.* (3) presented an improved and automated modeling technique for loop predictions using spatial information and optimization of a pseudo-energy function. Felts *et al.* (4) predicted native conformations using Optimized Potentials for Liquid Simulations, all atom (OPLS-AA) force fields, and Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent models, in combination with torsion angle conformation based search. By understanding the relationship between the motions of loops and larger domains, it might even be possible to improve crystallographic refinements of loop regions. Importantly, a full comprehension would permit predictions of ensembles of loop conformations, rather than static structures.

Sellers *et al.* (5) predicted loops in inexact environments by examining how loop refinement accuracy is affected by errors in the surrounding elements such as backbone and side-chain positions. They used augmented loop prediction methods that optimize the conformations of the side chains simultaneously. This method helps to recover near-native conformations for many perturbed structures. Olson *et al.* (6) examined *ab initio* methods for predicting protein loops by using multi-scale conformational sampling. They used physical energy functions to score the models. Peng and Yang (7) developed a knowledge-based loop prediction method without the necessity of constructing hierarchically clustered length-dependent loop libraries. This method first predicts the local structure of the loop and then structurally aligns it against all possible motif templates. Zhu *et al.* (8) have developed an improved sampling algorithm and an energy model for protein loop prediction that yields a smaller root mean square deviation from the native structure. They discussed their results in the context of the accuracy of continuum solvation models. Xiang (9) discussed the advances in protein homology modeling and the contribution of loop structure predictions for the overall prediction of protein structures. According to Radivojac *et al.* (10), some of the intrinsically disordered regions of protein structures consist of long loop regions with functional roles. Since conformation and dynamics are intrinsically related, any improvements in understanding one will likely improve the understanding of the other. Thus, the methods proposed in this paper may help us to better understand the relationship between functional motions and loop conformations.

Protein loops play an important role in protein function since they are often exposed to the solvent environment and hence may readily interact with other molecules. It is widely understood that their structures are not random coils (even for longer loops), and thus have some defined characteristics (11). Conformations of loops play a significant role in protein docking (12) and in stabilizing active sites through loop-scaffold interactions (13). Smith *et al.* (12,13,14) investigated the idea of guiding protein-protein interactions through contacts between surface loops in proteins. Hence flexibility of protein loops and their dynamics are important factors for understanding protein functions, as demonstrated further by Yao *et al.* (15) who used sampling algorithms to explore conformations of flexible loops. Krieger *et al.* (16) have shown that folding mechanisms in proteins vary widely depending on native-state topology and details such as the relative contact order (RCO). This indicates that protein loops and their topologies might also play an important part in protein folding. Conformational evaluation of loops and their structural variability was studied by Li *et al.* (17) who indicated the importance of loop structures for protein design.

Hu *et al.* (18), demonstrated through high-resolution design of protein loops that small changes in protein energetics can perturb the structure of proteins. They studied longer loops adopting specific conformations with the Rosetta molecular modeling program to find low-energy sequence-structure pairs. Their results suggest that the high-resolution design of protein loops may become feasible.

We previously investigated (19) the fluctuation dynamics of the tubulin dimer to elucidate the functional motions that might relate to activities such as binding, polymerization and assembly and discovered that a loop that covers the GTP binding site moves in coordination with a large-scale rotation between the α and β subunits. This illustrated how loop motions can be controlled by the large domain motions and can be slow. Also we investigated (30) the enzyme triose phosphate isomerase and observed that its binding site loop opened and closed only in the intact dimer, and not in the monomer, in a slow motion coordinated with a large-scale domain-domain motion.

Espadaler *et al.* (20) developed ArchDB, an automated classification tool for the structures of protein loops that connect different supersecondary structures and play an important role in initiating and maintaining the overall functions of a protein. Oliva *et al.* (20,21), computationally derived an extensive characterization of loop conformations that could enhance model building by comparison studies. Groban *et al.* (22) illustrated phosphorylation driven changes in loop conformation using the activation loop in CDK2. Kolodny *et al.* (23) approached the ‘loop closure problem’ using inverse kinematics. They proposed an algorithm for generating conformations of candidate loops within gaps in protein structures to complete protein structures so that their biological functions can be determined. Gerstein and Chothia (24) demonstrated the significant mobility of surface loops that can move over a distance of 10 Å to cover the active site, and showed that this motion is propagated outwards towards other regions of protein structure that have no contact with the ligand. They suggested that the whole protein consists of several different shells of increasing mobility. Andrec *et al.* (25) have developed a novel approach for detecting statistically significant differences in the structures of loops between crystal and

NMR-determined structures. Their approach is based on structural superposition and the analysis of the distributions of atomic positions relative to a mean structure. Their studies indicate that physical factors and the environment play a role in determining protein conformations. Sudarsanan *et al.* (26) used information from the backbone conformation of dimers to develop an automated method for modeling the backbones of protein loops that obtains near-native loop conformations from an ensemble of sterically allowed conformations. Street *et al.* (27) investigated the physical-chemical determinants of the turn conformations in globular proteins, concluding, as have many others, that turns can be classified into a small number of discrete conformations. Kempf *et al.* (28) examined how the loops in triosephosphate isomerase facilitate substrate access and catalysis. They investigated the dynamic requirements for functional hinges and elucidated the important principle of motional freedom and flexibility requirements for active-site loops, which control the open and closed states of active sites. Their results demonstrate the importance of catalytic hinge design in proteins.

In the present study we will investigate loop motions with elastic network models. We are interested in analyzing loop motions to see if they move independently or in coordination with large domain motions. We thus are able to identify the local motions of loops that make the largest contribution to the overall domain motions. The focus is on the dynamics of all the surface loops present in five diverse proteins: reverse transcriptase, triosephosphate isomerase, tubulin, protease, and myoglobin. Each of these proteins is distinct from the others in its topology and function, thus providing a small but diverse test set. The loops present in these proteins are known to have diverse functional behaviors. The choice of reverse transcriptase was based upon the importance of the loop motions that provide access to the polymerase site, specifically related to how the fingers and the thumbs move to open and close this site, as shown by Bahar *et al.* (29). Including triosephosphate isomerase was motivated by our previous observation of the importance of the loop moving over the active binding site (30). The loop covering the GTP binding site in tubulin was also shown previously to be coordinated with a slow motion of the protein (19). We have previously shown that this motion occurs together with the dominant motion, which is a rotation between the two subunits. The flaps of the protease are well known loops that regulate access to its binding site. The behaviors of the loops in these five proteins are examined in detail below and our findings suggest that functional loops behave in coordinated ways with the rest of the structure, rather than as random motions.

Materials and methods

Normal Mode Analysis

To study the kinematics of residues constituting loops we use **NMA** (normal mode analysis) on the coarse-grained elastic network models of structures. The structures are represented by C^{α} atom coordinates only. Harmonic springs are used to connect the C^{α} atoms in order to represent the protein structure as an *elastic network*. The Gaussian Network Model (**GNM**) is one of the simplest of elastic network models, originally applied to protein dynamics by Bahar *et al.* (31) and Haliloglu *et al.* (31,32) who applied the approach of Tirion *et al.* (33) in a coarse-grained way to both bonded and non-bonded contacts in proteins and represent their interactions with a single universal spring parameter. This model has its deep origins in the rubber like elasticity theory of Flory, James and Guth, James and Guth, Kloczkowski *et al.*, Skliros *et al.* (34,35,36,37,38). Each normal mode corresponds to a different frequency of oscillation. Extensive applications of NMA to biological and chemical systems have been discussed in Cui *et al.*, Jernigan and Kloczkowski, Sen *et al.* (39,40,41). The Anisotropic Network Model (ANM) developed by Atilgan *et al.* (42), can be used to compute the *directions of motions* of all points in the structure with the coarse-grained elastic network model, whereas the original GNM provided only the amplitudes of motion. We employ the ANM model throughout our following analyses.

Kinematics of Proteins

Our method of solving the kinematics of proteins in the coarse-grained representation is based on Lagrange's equation for the potential and kinetic energy of the system, as described by Kim *et al.*, Kim *et al.* a, Kim *et al.* b, Schuyler and Chirikjian, Schuyler and Chirikjian (43,44,45,46,47). As a first step, a rigid body translation and rotation of the structure is performed, so that the center of mass lies at the origin of the coordinate system and the moment of inertia tensor is diagonal. This procedure is described in detail in Supporting Information Section A.

To solve for the kinematics of the protein we define the coordinates as

$$\begin{gathered}\vec{R}_i(t)=\vec{R}_i(0)+\overrightarrow{\Delta R}_i(t) \Rightarrow \overrightarrow{\Delta R}_i(t)=\vec{R}_i(t)-\vec{R}_i(0) \\ \overrightarrow{\Delta R}_i(t)=[\ \ \Delta x_i(t),\quad \Delta y_i(t),\quad \Delta z_i(t)\ \ ]\end{gathered} \qquad (1)$$

where $\vec{R_i}(t)$ and $\vec{R_i}(0)$ are the instantaneous and the starting position vectors for the i^{th} point and $\overrightarrow{\Delta R}_i(t)$ is the displacement vector. The potential energy of the system can be written as (details shown in Supporting Information Section B)

$$V=\frac{1}{2}\Delta\vec{R}^T(t)K\Delta\vec{R}(t)$$

where K is the matrix of the order $3N\times 3N$ which depends on spring constants and initial position vectors of all points in a structure.

If we take $\overrightarrow{\Delta R}(t)=\begin{bmatrix}\Delta R_1(t) & \ldots & \Delta R_N(t)\end{bmatrix}$ for each time t then we find the solution for the elastic model for all values of i is given by

$$\overrightarrow{\Delta R}(t)=\sum_{i=1}^{3N}[\frac{1}{\sqrt{\lambda_i}}\sin\left(t\sqrt{\lambda_i}\right)e_ie_i^T\,\overrightarrow{\Delta R}(0)+\cos\left(t\sqrt{\lambda_i}\right)e_ie_i^T\,\overrightarrow{\Delta R}(0)] \qquad (2)$$

where λ_i and e_i ($i = 1,\ldots, 3N$) are the eigenvalues (square of angular frequencies) and eigenvectors (normal modes) of the system. For more details see Supporting Information Section C. For evaluation of the motions of loops we select the components of the $3N$-dimensional vector $\overrightarrow{\Delta R}(t)$ that correspond to the coordinates of the residues in the loop. We then study their time evolution by solving Eq. (2).

Identifying the dominant normal modes by Fourier Analysis

The essence of equation (2) is that it calculates the displacement of each coordinate from the equilibrium position at any given time t. We set the initial conditions of the fluctuations in such a way that the initial moment of inertia of the system is zero. We want information from time-dependent displacements to reconstruct the signal. The solution comes from the Nyquist-Shannon theorem, Shannon (48), which states that if the signal $x(t)$ has no angular frequencies higher than Ω_0, it is completely determined by giving the ordinates as a series of points separated by time intervals $\frac{\pi}{\Omega_0}$. For the current case, we know that the maximum angular frequency of the system is the square root of the maximum eigenvalue, $\Omega_0 = \sqrt{\lambda_{\max}}$. This corresponds to selecting a sampling period of $T_s \leq \frac{2\pi}{2\Omega_0}$ or $\Omega_s \geq 2\Omega_0$. Furthermore we see from Eq. (2) that the motions of residues can be expressed as a combination of sinusoidal functions, making them periodic.

The maximum period of the system that defines the final time in our calculations is $T_{\max} = \frac{2\pi}{\sqrt{\lambda_{\min}}}$, $\lambda_{\min} \neq 0$. For each element of $\overrightarrow{\Delta R}(t)$ we calculate its time evolution, following from Eq. (2), at time intervals $t = 0, T_s, 2T_s, \ldots, sT_s$, with $s = \lceil \frac{T_{\max}}{T_s} \rceil$ up to $T_{\max}$. To each of the $3N$ coordinates we can assign an s-length discrete time signal, called $H(n) = \Delta R(nT_s)$ that is periodic with the period $T_{\max}$.

The Discrete Fourier Transform DFT of this signal is given by:

$$F_H(k) = \sum_{n=0}^{s-1} H(n) e^{-2\pi jkn/s}, \qquad k = 0, \ldots, s-1 \quad (3)$$

To calculate all the s-entries of that signal we require s^2 multiplications and $s(s-1)$ additions. The Fast Fourier Transform (FFT), as proposed by Cooley and Tukey and Singleton (49,50), significantly reduces the computational cost.

In order to recover $H(n)$ from $F_H(k)$, we apply the inverse Discrete Fourier Transform defined as:

$$H(n) = \sum_{k=0}^{s-1} F_H(k) e^{2\pi jkn/s}, n = 0, \ldots, s-1. \quad (4)$$

FFT also applies to the inverse Discrete Fourier Transform, and the interested reader might refer to the Digital Signal Processing literature such as given in Antoniou, ElAli, Hayes (51,52,53). Eqs (9–10 in (52)), imply that the magnitude of $F_H(k)$ is symmetrical about the point $k = \frac{s}{2}$, thus $|F_H(k)| = |F_H(s-k)|$, and $\angle F_H(k) = -\angle F_H(s-k)$ for the phase of the signal. The lowest frequencies of the signal are located at the ends of $F_H(k)$ whereas the highest frequencies are located in the middle. It is a symmetric signal with $F_H(p-K/2) = F_H(p+K/2)$, where $k, p \in [1, 2, \ldots, K]$. It was also noted in reference (52) that the distances between the successive values of k in $F_H(k)$ are given by the angular frequency resolution $\frac{\Omega_s}{s}$. The correspondence between indices k of $F_H(k)$ and the eigenvalues of the system can be specified as given in Supporting Information Section D.

To evaluate the impact of the lowest frequency motions on the system, we first identify the proper k indices in $F_H(k)$. Then we set the value of $F_H(k)$ for the k's that do not belong in that range to zero which leads to the new FFT $F'_H(k)$. Then we take the inverse DFT (Discrete Fourier Transform) of $F'_H(k)$, thus obtaining a new discrete time signal $H'(n)$ that depends only on the lowest normal modes of the system. Finally we compute the Pearson correlation between $H'(n)$ and $H(n)$ (46). The higher the value of the correlation, the greater the impact of the lowest normal modes is on the motions of the system.

Computing Changes in Internal Distances

We also consider the changes in the internal locations of the structure points with ANM. This is the change in internal distance, computed as

$$<(\Delta R_i - \Delta R_j)^2> = <\Delta R_i{}^2> + <\Delta R_j{}^2> - 2 <\Delta R_i \cdot \Delta R_j> \quad (5)$$

These values are obtained directly from the inverse of the Hessian matrix from which the normal modes are derived.

$$<(\Delta R_i - \Delta R_j)^2> = \Gamma^{-1}{}_{ii} + \Gamma^{-1}{}_{jj} - 2\Gamma^{-1}{}_{ij} \quad (6)$$

where Γ is the matrix of second derivatives of the potential energy (Hessian) for the structure for which the normal modes (e_i) are computed. Since there are six zero eigenvalues in ANM corresponding to the rigid body motions, Γ is not invertible. Thus, instead we compute its pseudo-inverse: $\Gamma^{-1} = \sum_{i=7}^{3N} \frac{1}{\lambda_i} e_i e_i^T$.

Loop Identification

In our study we first identify the surface loops on the proteins. We identify these loops in proteins by excluding all residues identified by DSSP (54) as H, G, I, or E, corresponding to standard α, 3_{10}, and π-helices, and β-strands, respectively. We retain isolated beta-bridges and hydrogen bonded turns to prevent short loops interrupted by these elements from being discarded. We focus on surface loops by also rejecting any residue having surface exposure less than 5% in an extended A-X-A chain using NACCESS program (55) to compute relative solvent accessibility. We also set the requirement that the length of a loop must be four or more residues. Visual inspection of the 5 protein structures studied in this work show that this selection appears reasonable. The identity of all loops studied here is given in the Supporting Information Section E. The functional loops are defined as those loops which contain one or more functional sites. Information about the functional parts of the proteins and the functional loops is provided in the Supporting Information Section F. Functional information is derived from the NCBI Protein database and manually related to the corresponding protein structures.

Results and Discussion

The main purpose of this study is to answer three major questions. (i) Do protein residues move overall independently, or do they move in coordination with the entire structure? (ii) Do loops in proteins move along with the whole structure or do they exhibit a different behavior? (iii) Do the functional loops move independently or in coordination with the slow motions, and do they move like non-functional loops? We attempt to answer these questions by investigating five different proteins in terms of function and topology, which are given in Table 1. To address the first question, that is whether proteins residues move individually or collectively we employ the Anisotropic Network Model (ANM). This model depends on the whole structure of the protein through a connectivity matrix, dependent on topology. In this study, we find that there is a close correspondence between the behavior of loop motions and the entire structure as observed in the protein reverse transcriptase (19). In Figure 1, we show the correlation between the motions obtained for the first 6 normal modes (the slowest motions) for all residues in comparison with the loop residues for four proteins studied in this work. (For myoglobin see Supporting Information G.) Residue indices are sorted according to the increasing values of correlations. The first six modes are the slow, collective, low frequency motions of the structure. We see that in this case, the overall motions of the loops do not differ much from the motions of the whole structure.

For each coordinate of each residue, we calculate the displacement from the initial position at several time instances. We thus construct a discrete time signal H. The details of how we obtain this signal are explained in the Methods section. In the computations given below, H is the kinematic response of each coordinate of each residue based on all normal modes whereas H' is the similar kinematic response based on a subset of the lowest normal modes. Thus H is the full FFT signal while H' is the FFT signal corresponding to the low frequency normal modes. High correlations between these two will indicate a dominance of the low frequency motions. In Figure 1 these correlations are shown, and we can see that when all protein residues are considered that they move with the global (collective) domain motions since the percentage of motion represented by the six lowest normal modes is always above 50%.

We have also computed the mean correlations $< \rho_{H,H'} >$ between H and H' averaged over the residues within the loops, for all loops in a given protein. Similarly, we compare correlations computed by using all normal modes in Eq. (2) with those obtained by using only the slowest modes (details are given in the Methods section). Results are shown in Figure 2 for loops belonging to chains A and B of reverse transcriptase and of tubulin (for triosephosphate isomerase, protease and myoglobin see Supporting Information G). Similarly as in Figure 1, the loop indices are sorted according to ascending values of the correlations. Circles identify functionally important loops.

From Figure 2 we see that the mean impact of the first 6 normal modes on the loops ranges between 65–99%, a slightly larger range of correlations than the impact of the first 6 normal modes on all residues of the proteins. Thus from Figure 2 we conclude that protein loops move as a part of a domain to a somewhat greater extent than all other parts of the protein structures. The slightly larger correlations may be attributed simply to the loops being investigated residing on the outside of the structures

For the loops of reverse transcriptase we know that loops with indices 2,4,5,6,8,14,18,22,24,27,28 on chain A and 8 on chain B control access to the catalytic residues or contain binding residues. We see from Figure 2 that motions of these functional loops do not differ significantly from motions of other, non-functional loops. Likewise for tubulin (Figure 2), the loops with index numbers: 4, 6, 9 from chain A and 5, 8, 9 from chain B responsible for regulation of the interactions with other tubulin dimers do not differ in behavior much from the average behavior of the loops of tubulin. We also randomly generated 100,000 partitions of the loops into two groups (data not shown) where the smaller group was the minimum of 15 or half of the loops. The most significant partitions (determined by the amount of difference in average mean squared fluctuation) were either trivially different from one another or corresponded to groups of loops that were farthest from the protein's center of mass.

In Supporting Information F we show the location of functional loops on each structure. Hence, the answer to the third question could be that functional loops do not move in a more coordinated way with the rest of the structure than regular loops, although some individual loops may do so (see Fig. 2).

Normal mode calculations are often performed to elucidate which residues or atoms in a molecular structure are the most mobile. Active site residues are often held relatively rigid. Two supporting cases here are reverse transcriptase and protease where the catalytic residues are within a cleft where they are held relatively rigid. It is the movement of the surrounding structural elements that regulate access to these catalytic residues that facilitate access to the protein active site. However, another quantity which may be informative is the internal mean square distance changes described by Eq. 5. Internal mean square distance changes can be calculated directly from the Hessian matrix used to generate the normal modes in ANM using Eq. 6. We have employed (57) ANM models built with uniform springs with cutoffs ranging from 10–15 Å, as well as with springs having inverse square dependences on distance and obtained similar results. The mean square internal distance fluctuations, $(\Delta R_i - \Delta R_j)^2 >$, describe the change within a structure; how the normal modes stretch, compress, or otherwise rearrange the internal structure locally. If this change in internal distance is zero for a given (i, j) pair, it means that the two points move together fully rigidly (the distance between them does not change). We have analyzed the present five protein structures (data not shown) and concluded that the areas of a protein with the smallest internal mean square distance changes are the cores, and as one moves further away from the stable cores the internal distance fluctuations increase. Figure 3 shows the mean internal RMSD for each loop of reverse transcriptase and tubulin. (For triosephosphate isomerase, protease and myoglobin see Supporting Information G). We see that those loops that are functional do not have lower or higher RMSDs than nonfunctional loops. Hence the nonfunctional loops do not differ in the internal conformational behavior from the nonfunctional ones.

It is also of interest to locate the loops on protein structures that have the highest correlations according to Figure 2. These loops are highlighted in Figure 3 for the HIV-1 reverse transcriptase structure and are mostly associated with the areas surrounding catalytic residues (see Supplemental Section H for the other four proteins). In Figure 4 we show a zoomed in view of the polymerase active site where a large loop hangs over the opening between the thumb and fingers. This loop may act to regulate substrate access to the catalytic residues and influence binding on the interior of the thumb and fingers (white arrow in the lower part of Figure 4) domains. Yellow surfaces correspond to experimentally verified nucleotide binding residues. Another loop with a very high correlation coefficient is marked with a solid black arrow in part C that may also interact with bound substrate. It is likely that many of these loops owe their high correlation to the large hinge motion through the middle of the structure that is seen in the dominant mode of motion.

For the sake of finding the amount of correlation or anti-correlation among the residues which correspond to only functional and non functional loops, we have reduced the correlation map for ANM from all residue set up to loop residues only. We have analyzed these correlation maps for all these functional and non-functional loops for all five proteins for only first six normal modes which correspond to the global motions of these proteins as shown in Figure 5 (for protease) and in the Supporting Information I (for myoglobin, triosephosphate isomerase, tubulin and reverse transcriptase). These correlation plots for only loop residues exhibit a significant amount of correlations mainly among the functional loops. Also in some cases, there is an extent of anti-correlation among certain functional loops in particular modes, which again explains a particular functionality for that protein.

In Figure 5 the functional and non-functional loops are shown in blue and red, respectively. The total number of loops of protease is 18 of which only 9 (with loop indices 1,2,5,6,7,10,11,14 & 15) are functional. For protease, the total number of residues is 338, but as we have excluded the residues belonging to other secondary structure elements (alpha helices and beta sheets) and considered only the residues belonging to the loops, the number of residues is much lower than 338. We have used a white demarcation line between the two adjacent loops to clearly distinguish the correlations and anti-correlations in these loop residues. The correlation ranges from +1 (shown in green) to −1 (black). Also there have been cases where we have found no correlation which has been marked by white color.

We have also calculated the percentage of total number of functional loops which move in correlation for all these five proteins under the first six normal modes. This enables us to address in a more informative way a question whether the functional loops behave in a more coordinated way with the slow motions, or if their behavior is independent of the global motion. Figure 6 shows the results for all the five proteins. We have found that approximately 40% to 70% of the total number of functional loops move in coordination for majority of these proteins in most of the different lowest six normal modes which is again a significant manifestation of dynamical cooperativity considering that these functional loops are not adjacent and some of them are really far apart from each other.

Conclusion and Outlook

In the present paper, we have considered the motions of the surface loops in five proteins. By applying a novel method that combines ANM and FFT we are able to identify which normal modes have the largest impact on the motions of individual loops. We observe a broad range of behaviors, with some loops moving in the slowest modes, which implies that their motion is strongly linked with the global, collective motions of the structure and others moving with the fast modes.

We know that loops are parts of protein structure that are likely to be more susceptible to the influence of the external environment. Environmentally influenced changes in loop structures or dynamics may lead to radical structural changes of the whole protein. ***The reverse may hold also because of the protein’s cohesiveness, so that external influences changing loop conformations could also push the large domains into different positions, leading to allosteric transmission***.

Despite the evident successes of normal mode analysis and elastic network models in explaining functional protein motions, ligand binding, allosteric effects, etc. there are certain limitations that are intrinsic characteristics of these models. One basic assumption is that the potential is harmonic, i.e. it is a simple quadratic function with a single minimum. This significantly simplifies the mathematics of the problem, allows for the analytical computation of Gaussian integrals for the elastic network models and the reduction of the problem to a simple eigen-analysis problem of the Kirchhoff connectivity matrix. However we know well that the actual protein energy landscape is much more complex, with many local minima surrounding any global minimum, and that some of these minima originate in the details of the side chain atoms. In the present case, we have not included these details in our coarse-grained model. So, accounting for all details of the energy landscape and dynamics will require those further details. Nonetheless these simple coarse-grained models have advantages in filtering out some of the high frequency motions, which is both an advantage and a disadvantage. The advantage is that the system behaves dynamically in a simpler way, and the disadvantage is that all details cannot be seen in the computations.

Anharmonic potentials may be required that have more energy minima, and would need at least a quartic function of the distance to obtain a more realistic potential energy function for the elastic network models describing transitions between two states, but unfortunately that leads immediately to a major complication in the mathematics of the problem, and its becomes analytically insoluble. Another major limitation of the elastic network models is their usual independence of the type of amino acids, whereas, it is well known that certain single amino acid substitutions can have large effects on protein dynamics and allostery, without significant changes to the protein structure. This problem has been addressed few years ago by Erman [58], who assumed that the spring constants depend on the type of amino acids. Future extensions of the elastic network models using mixed coarse-grained models may bring a better explanation of the single amino acid substitution effects on protein dynamics.”

Prediction of loop motions with and without external environmental influences can lead to a better understanding of the functions of loops and their mechanisms. More specifically, we can potentially identify the mechanics of the hyper-variable loops of antibodies and how they may move in response to the presence of a specific antigen. We can also try to understand the mechanism of motions at the polymerase sites, the mechanism of GTP binding sites in tubulin, and the loop at the active site in triose phosphate isomerase. For instance in the introduction we referred to Keskin *et al.* who found that boundary regions of collective motions seem to act as linkages in secondary structures elements. The loops of tubulin act as these linkages, since they are dominated by low normal modes that move loops with the whole domain. Our study also confirms the finding of Gerstein *et al.* (24) that the whole protein consists of different shell regions of increasing mobility. Since most protein residues' motions are dominated by the lowest frequencies, this implies that the protein residues form clusters of rigid bodies. Another important issue is in understanding how the binding site of proteases open and close. We would like to answer the following questions: What is the mechanism for this allosteric transition? What are the roles of loops, and how do the structures of loops change during this and other transitions? Here we have made a first computational step in this direction by demonstrating that the slow motions control the loops that are most pertinent to the principle function. Our future computations will focus on the dynamical behavior of loops under certain environmental conditions and the transmission of any induced changes through the structure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge the financial support provided by the National Institutes of Health through grants R01GM081680, R01GM072014, and R01GM073095.

References

1. Zhang Y, Stec B, Godzik A. Between order and disorder in protein structures: analysis of “dual personality” fragments in proteins. Structure. 2007; 15:1141–1147. [PubMed: 17850753]
2. Joosten K, Cohen SX, Emsley P, Mooij W, Lamzin VS, Perrakis A. A knowledge-driven approach for crystallographic protein model completion. Acta Crystallogr D Biol Crystallogr. 2008; 64:416–424. [PubMed: 18391408]
3. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Prot Science. 2000; 9:1753–1773.
4. Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA, Levy RM. Prediction of Protein Loop Conformations using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. J Chem Theory Comput. 2008; 4:855–868. [PubMed: 18787648]
5. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP. Toward better refinement of comparative models: Predicting loops in inexact environments. Proteins. 2008; 72:959–971. [PubMed: 18300241]
6. Olson MA, Feig M, Brooks CL. Prediction of protein loop conformations using multiscale Modeling methods with physical energy scoring functions. J Comp Chem. 2008; 29:820–831. [PubMed: 17876760]
7. Peng HP, Yang AS. Modeling protein loops with knowledge-based prediction of sequence-structure alignment. Bioinformatics. 2007; 23:2836–2842. [PubMed: 17827204]
8. Zhu K, Pincus DL, Zhao SW, Friesner RA. Long loop prediction using the protein local optimization program. Proteinss. 2006; 65:438–452.
9. Xiang Z. Advances in homology protein structure modeling. Curr Protein Pept Sci. 2006; 7:217–227. [PubMed: 16787261]

10. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uverskyn VN, Dunker AK. Intrinsic disorder and functional proteomics 19. Biophys J. 2007; 92:1439–1456. [PubMed: 17158572]
11. Panchenko AR, Madej T. Structural similarity of loops in protein families: toward the understanding of protein evolution. BMC Evol Biol. 2005; 5:10. [PubMed: 15691378]
12. 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; 180:605–613. [PubMed: 9457864]
13. Li C, Banfield MJ, Dennison C. Engineering copper sites in proteins: Loops confer native structures and properties to chimeric cupredoxins. J Am Chem Soc. 2007; 129:709–718. [PubMed: 17227035]
14. Smith JW, Tachias K, Madison EL. Protein loop grafting to construct a variant of tissue-type plasminogen activator that binds platelet integrin alpha(IIb)beta(3). J Biol Chem. 1995; 270:30486–30490. [PubMed: 8530479]
15. Yao P, Dhanik A, Marz N, Propper R, Kou C, Liu GF, van den Bedem H, Latombe JC, Halperin-Landsberg I, Altman RB. Efficient Algorithms to Explore Conformation Spaces of Flexible Protein Loops. IEEE-ACM Trans Comp Biol Bioinformatics. 2008; 5:534–545.
16. Krieger, Florian; Fierz, Beat; Axthelm, Fabian; Joder, Karin; Meyer, Dominique; Kiefhaber, Thomas. Intrachain diffusion in a protein loop fragment from carp parvalbumin. Chemical Physics. 2010; 307:209–215.
17. Li WZ, Liu ZJ, Lai LH. Protein loops on structurally similar scaffolds: Database and conformational analysis. Biopolymers. 1999; 49:481–495. [PubMed: 10193195]
18. Hu, Xiaozhen; Wang, Huanchen; Ke, Hengming; Kuhlman, Brian. High-resolution design of a protein loop. PNAS. Nov 6.2007 104:17668–17673. [PubMed: 17971437]
19. Keskin O, Durell SR, Bahar I, Jernigan RL, Covell DG. Relating molecular flexibility to function: A case study of tubulin. Biophys J. 2002; 83:663–680. [PubMed: 12124255]
20. Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJE, Oliva B. ArchDB: automated protein loop classification as a tool for structural genomics. Nucl Acids Res. 2004; 32:D185–D188. [PubMed: 14681390]
21. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE. An automated classification of the structure of protein loops. J Mol Biol. 1997; 266:814–830. [PubMed: 9102471]
22. Groban ES, Narayanan A, Jacobson MP. Conformational changes in protein loops and helices induced by post-translational phosphorylation. Plos Comp Biol. 2006; 2:238–250.
23. Kolodny R, Guibas L, Levitt M, Koehl P. Inverse kinematics in biology: The protein loop closure problem. Int J Robotics Res. 2005; 24:151–163.
24. Gerstein M, Chothia C. Analysis of Protein Loop Closure - 2 Types of Hinges Produce One Motion in Lactate-Dehydrogenase. J Mol Biol. 1991; 220:133–149. [PubMed: 2067013]
25. Andrec M, Snyder DA, Zhou ZY, Young J, Montellone GT, Levy RM. A large data set comparison of protein structures determined by crystallography and NMR: Statistical test for structural differences and the effect of crystal packing. Proteins. 2007; 69:449–465. [PubMed: 17623851]
26. Sudarsanam S, Dubose RF, March CJ, Srinivasan S. Modeling Protein Loops Using A Phi-I+1, Psi-I Dimer Database. Prot Science. 1995; 4:1412–1420.
27. Street TO, Fitzkee NC, Perskie LL, Rose GD. Physical-chemical determinants of turn conformations in globular proteins. Prot Science. 2007; 16:1720–1727.
28. Kempf JG, Jung JY, Ragain C, Sampson NS, Loria JP. Dynamic requirements for a functional protein hinge. J Mol Biol. 2007; 368:131–149. [PubMed: 17336327]
29. Bahar I, Erman B, Jernigan RL, Atilgan AR, Covell DG. Collective motions in HIV-1 reverse transcriptase: Examination of flexibility and enzyme function. J Mol Biol. 1999; 285:1023–1037. [PubMed: 9887265]
30. Kurkcuoglu O, Jernigan RL, Doruker P. Loop motions of triosephosphate isomerase observed with elastic networks. Biochemistry. 2006; 45:1173–1182. [PubMed: 16430213]
31. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding Des. 1997; 2:173–181.

32. Haliloglu T, Bahar I, Erman B. Gaussian dynamics of folded proteins. Phys Rev Lett. 1997; 79:3090–3093.

33. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett. 1996; 77:1905–1908. [PubMed: 10063201]

34. Flory PJ. Statistical Thermodynamics of Random Networks. Proceedings of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences. 1976; 351:351–380.

35. James HM, Guth E. Theory of the elastic properties of rubber. J Chem Phys. 1943; 11:455–481.

36. James HM, Guth E. Statistical Thermodynamics of Rubber Elasticity. J Chem Phys. 1953; 21:1039–1049.

37. Kloczkowski A, Mark JE, Erman B. Chain Dimensions and Fluctuations in Random Elastomeric Networks. 1. Phantom Gaussian Networks in the Undeformed State. Macromolecules. 1989; 22:1423–1432.

38. Skliros A, Mark JE, Kloczkowski A. Chain dimensions and fluctuations in elastomeric networks in which the junctions alternate regularly in their functionality. J Chem Phys. 2009; 130:064905. [PubMed: 19222296]

39. Cui, Q.; Bahar, I. Normal Modes Analysis: Theory and applications to biological and chemical systems. 2006.

40. Jernigan RL, Kloczkowski A. Packing regularities in biological structures relate to their dynamics. Methods Mol Biol. 2007; 350:251–276. [PubMed: 16957327]

41. Sen TZ, Feng YP, Garcia JV, Kloczkowski A, Jernigan RL. The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models. J Chem Thy Comp. 2006; 2:696–704.

42. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J. 2001; 80:505–515. [PubMed: 11159421]

43. Kim MK, Jernigan RL, Chirikjian GS. Efficient generation of feasible pathways for protein conformational transitions. Biophys J. 2002; 83:1620–1630. [PubMed: 12202386]

44. Kim MK, Li W, Shapiro BA, Chirikjian GS. A comparison between elastic network interpolation and MD simulation of 16S ribosomal RNA. J Biomol Struct Dyn. 2003; 21:395–405. [PubMed: 14616035]

45. Kim MK, Jernigan RL, Chirikjian GS. An elastic network model of HK97 capsid maturation. J Struct Biol. 2003; 143:107–117. [PubMed: 12972347]

46. Schuyler AD, Chirikjian GS. Normal mode analysis of proteins: a comparison of rigid cluster modes with C-alpha coarse graining. J Mol Graphics Model. 2004; 22:183–193.

47. Schuyler AD, Chirikjian GS. Efficient determination of low-frequency normal modes of large protein structures by cluster-NMA. J Mol Graphics Model. 2005; 24:46–58.

48. Shannon CE. Communication in the Presence of Noise. Proc Inst Radio Eng. 1949; 37:10–21.

49. Cooley JW, Tukey JW. An Algorithm for Machine Calculation of Complex Fourier Series. Math Comput. 1965; 19:297–301.

50. Singleton RC. An Algorithm for Computing Mixed Radix Fast Fourier Transform. IEEE Trans Audio Electroacoustics. 1969; AU17:93–103.

51. Antoniou, A. Digital Signal Processing. 2006.

52. ElAli, TS. Discrete Systems and Digital Signal Processing with MATLAB. 2004.

53. Hayes, MH. Digital Signal Processing. 1999.

54. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]

55. Hubbard, SJ.; Thornton, JM. NACCESS. University College London; 1993.

56. Seckler JM, Howard KJ, Barkley MD, Wintrode PL. Solution structural dynamics of HIV-1 reverse transcriptase heterodimer. Biochemistry. 2009; 48:7646–7655. [PubMed: 19594135]

57. Yang L, Song G, Jernigan RL. Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci USA. 2009; 106:12347–12352. [PubMed: 19617554]

58. Erman B. The Gaussian network model: Precise prediction of residue fluctuations and application to binding problems. Biophys J. 2006; 91:3589–3599. [PubMed: 16935951]

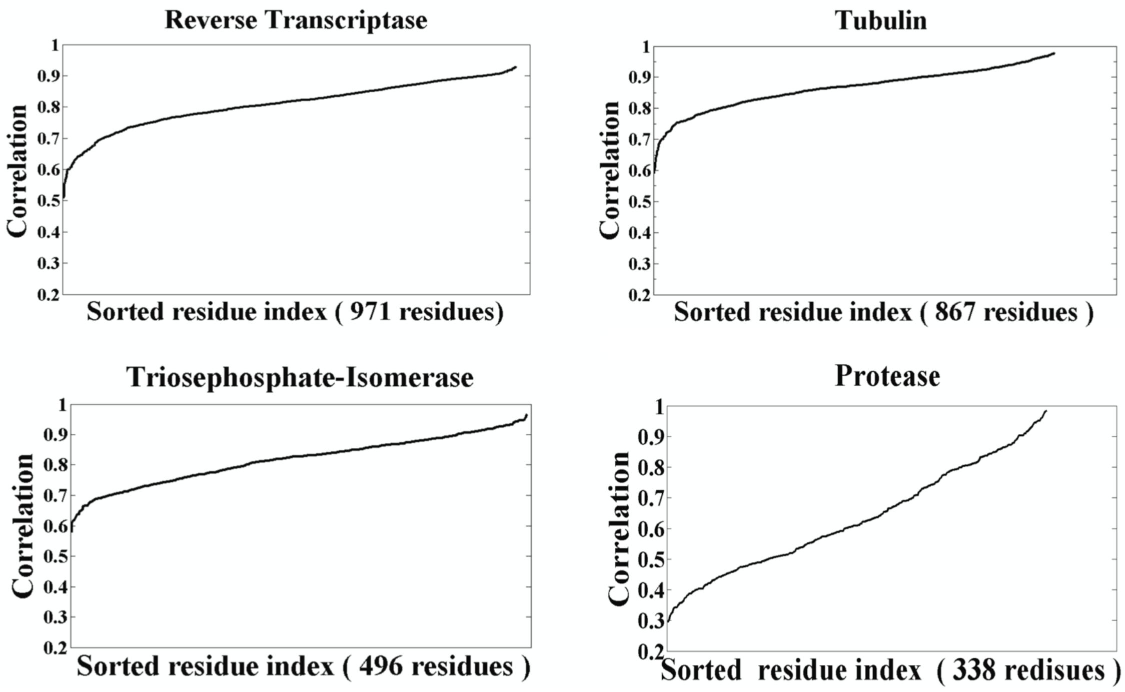


Figure 1.

Impact of the motions of the first 6 normal modes on the overall motion, for all residues of reverse transcriptase, tubulin, triosephosphate-isomerase and protease (myoglobin in Supplemental Information G).

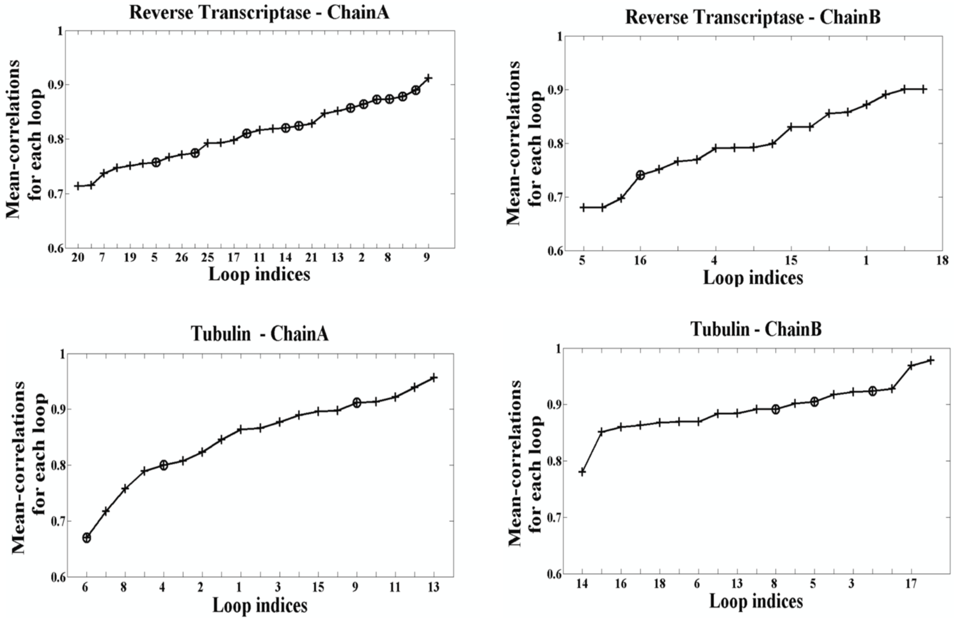
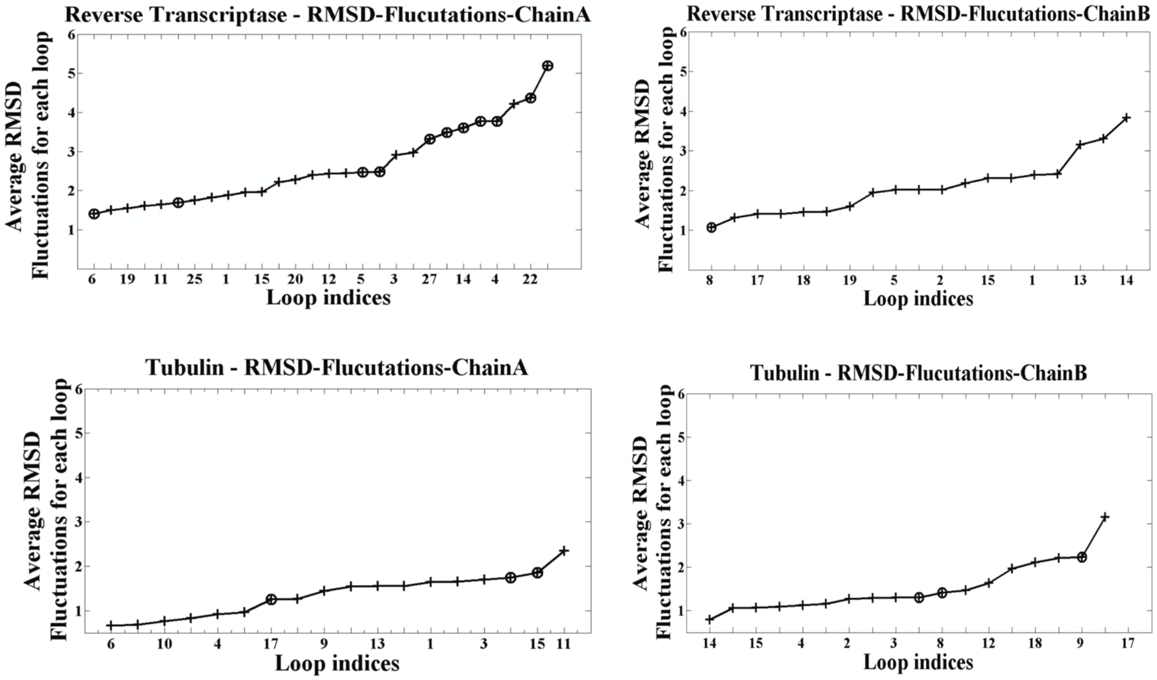


Figure 2.

Mean correlations of the motions derived from the first six normal modes with the total motions for each of the loops of reverse transcriptase and tubulin. The functional loops are denoted by circles.



Reverse Transcriptase - RMSD-Flucutations-ChainA

Reverse Transcriptase - RMSD-Flucutations-ChainB

Tubulin - RMSD-Flucutations-ChainA

Tubulin - RMSD-Flucutations-ChainB

Figure 3.

Mean RMSD for the loops of chain A and chain B of the reverse transcriptase and chain A and chain B of tubulin. Functional loops are indicated by circles.

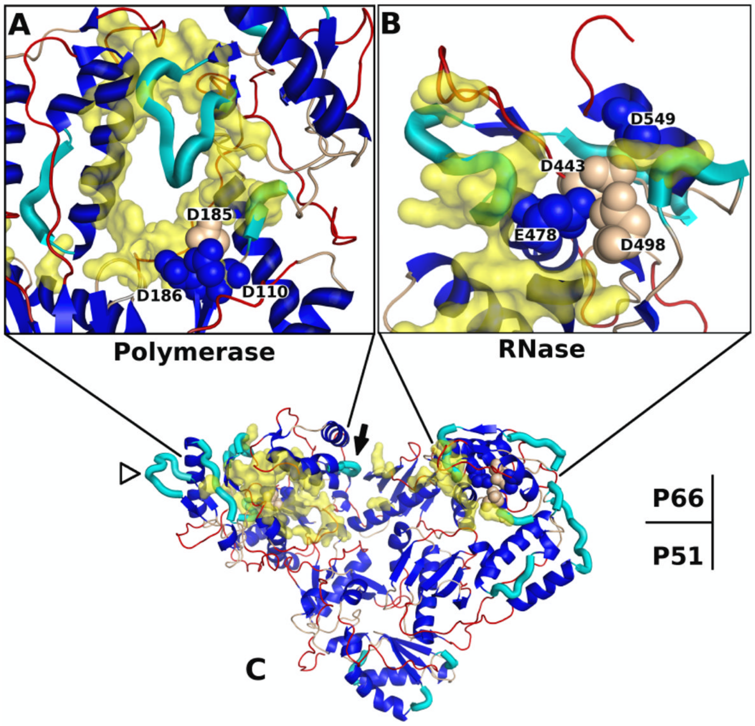
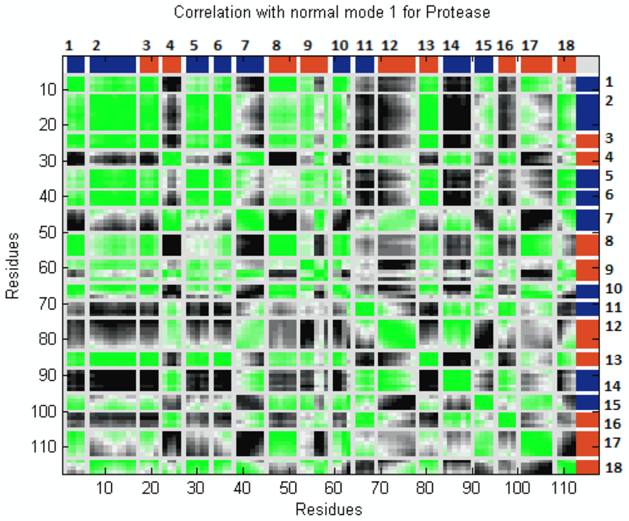
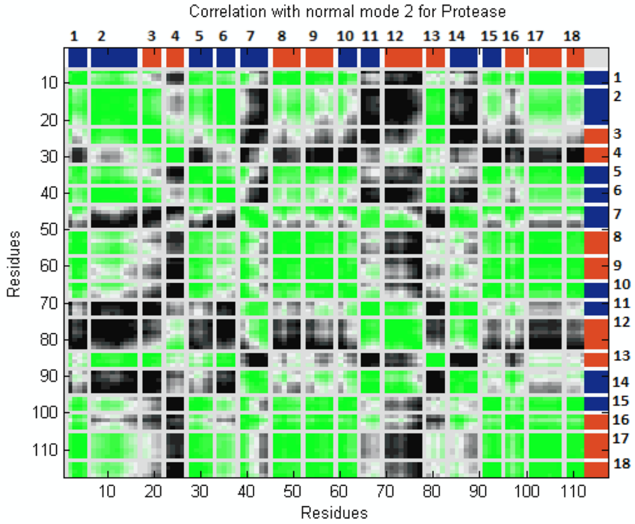
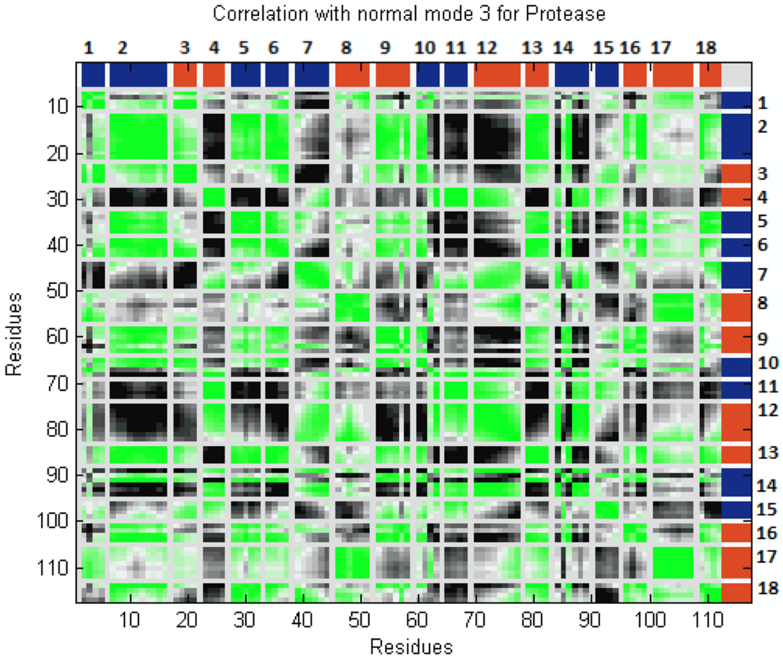


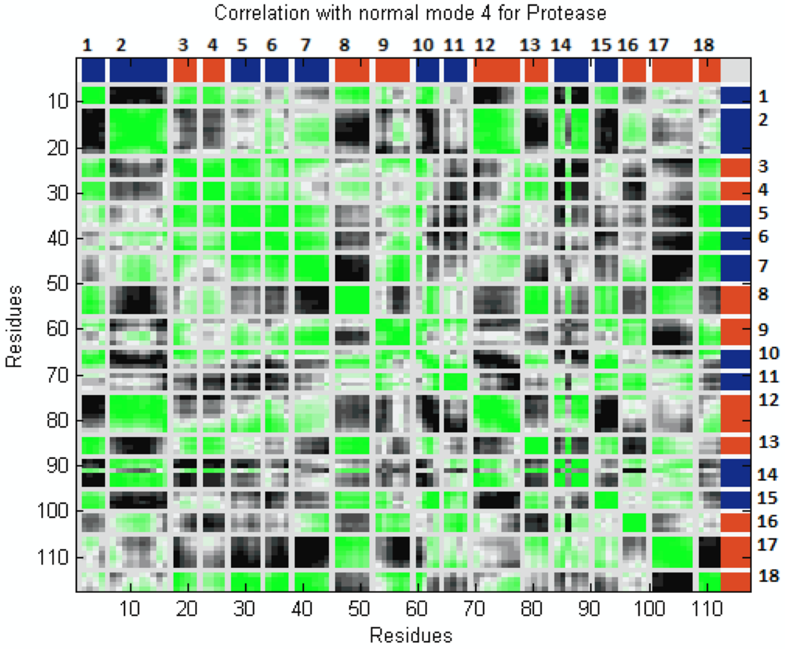
Figure 4.

We highlight with thick cyan tubes the surface loops of HIV-1 reverse transcriptase that have the highest average correlation (coefficient > 0.87) of motion between the first six modes and all modes. Catalytic residues of the polymerase and RNase domains are labeled and shown as spheres. Yellow molecular surfaces are shown for residues experimentally determined to contact the nucleotide template. The other surface loops (see Methods) are colored red with other loops in tan. (A) Zoomed and rotated to show the polymerase catalytic domain with the fingers on the right and thumb on the left. (B) RNase catalytic domain. (C) The P66 and P51 dimer is shown. The white arrow head points to the polymerase finger domain which contains three cyan loops. The filled black arrow points to a loop which is likely to interact with the nucleotide chain in the dominant modes of motion. See Supplemental Information H for similar figures for the other four proteins.


Correlation with normal mode 1 for Protease
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Residues
10 20 30 40 50 60 70 80 90 100 110
Residues


Correlation with normal mode 2 for Protease
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Residues
10 20 30 40 50 60 70 80 90 100 110
Residues


Correlation with normal mode 3 for Protease
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Residues
10 20 30 40 50 60 70 80 90 100 110
Residues


Correlation with normal mode 4 for Protease
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Residues
10 20 30 40 50 60 70 80 90 100 110
Residues

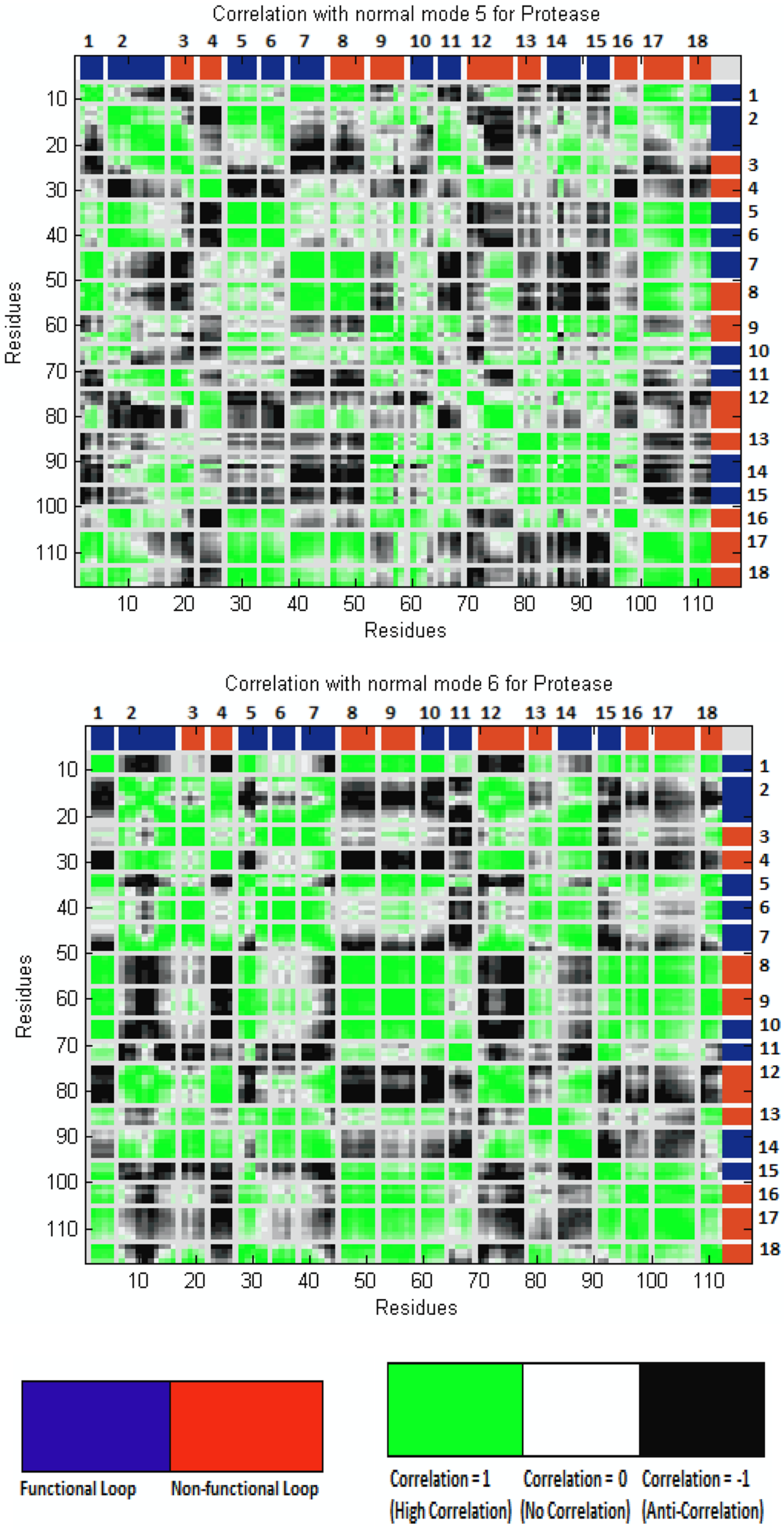
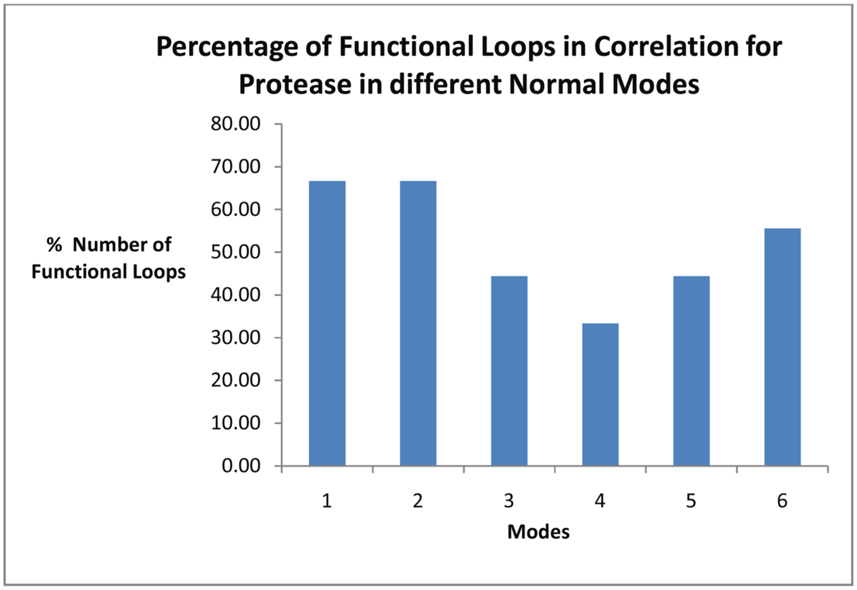
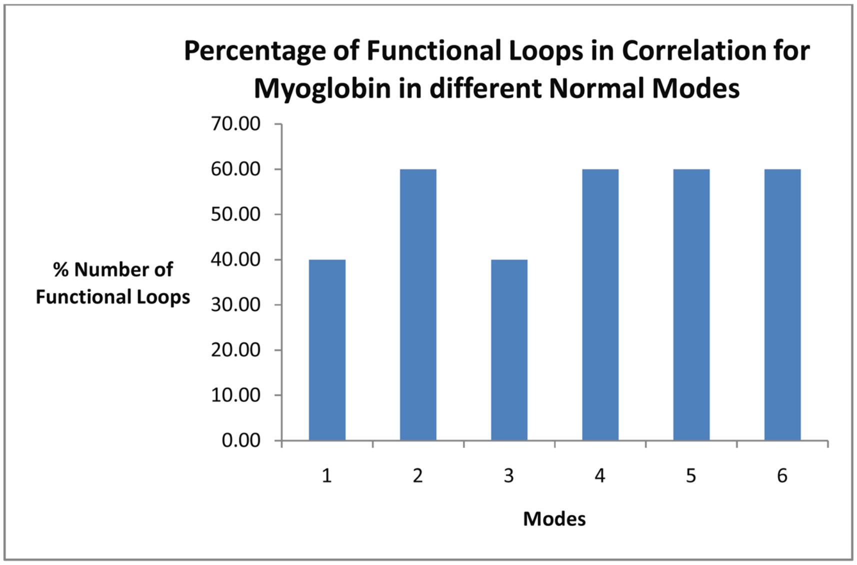
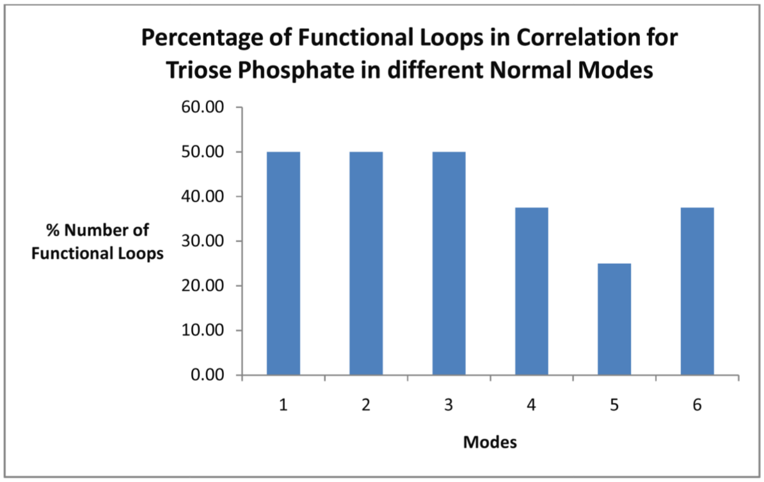


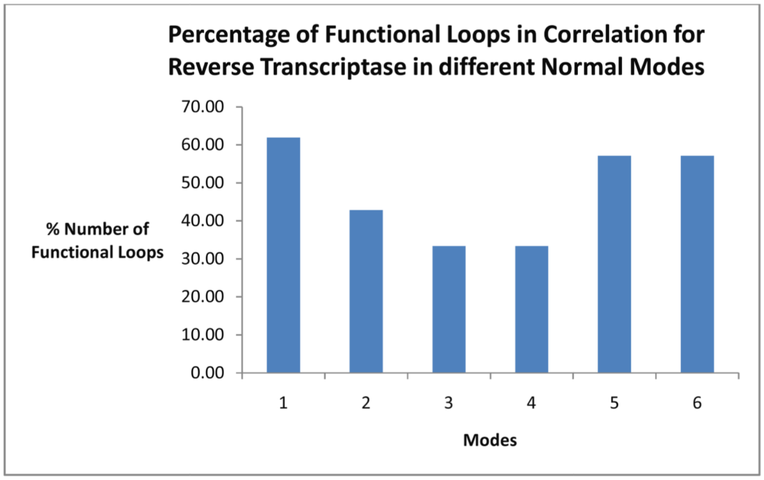
Figure 5.

Correlation Plot for first six normal modes for the functional and non-functional loops of protease. The functional loops are marked in blue, while non-functional ones are shown in red color. We use green, white and black colors to show the transition from the highest positive correlation (green) to anti-correlation (black). Numbers (1 to 18) indicate loop indices for protease.


Percentage of Functional Loops in Correlation for Protease in different Normal Modes
% Number of Functional Loops
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1
2
3
4
5
6
Modes


Percentage of Functional Loops in Correlation for Myoglobin in different Normal Modes
% Number of Functional Loops
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1
2
3
4
5
6
Modes


Percentage of Functional Loops in Correlation for Triose Phosphate in different Normal Modes
% Number of Functional Loops
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1
2
3
4
5
6
Modes


Percentage of Functional Loops in Correlation for Reverse Transcriptase in different Normal Modes
% Number of Functional Loops
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1
2
3
4
5
6
Modes

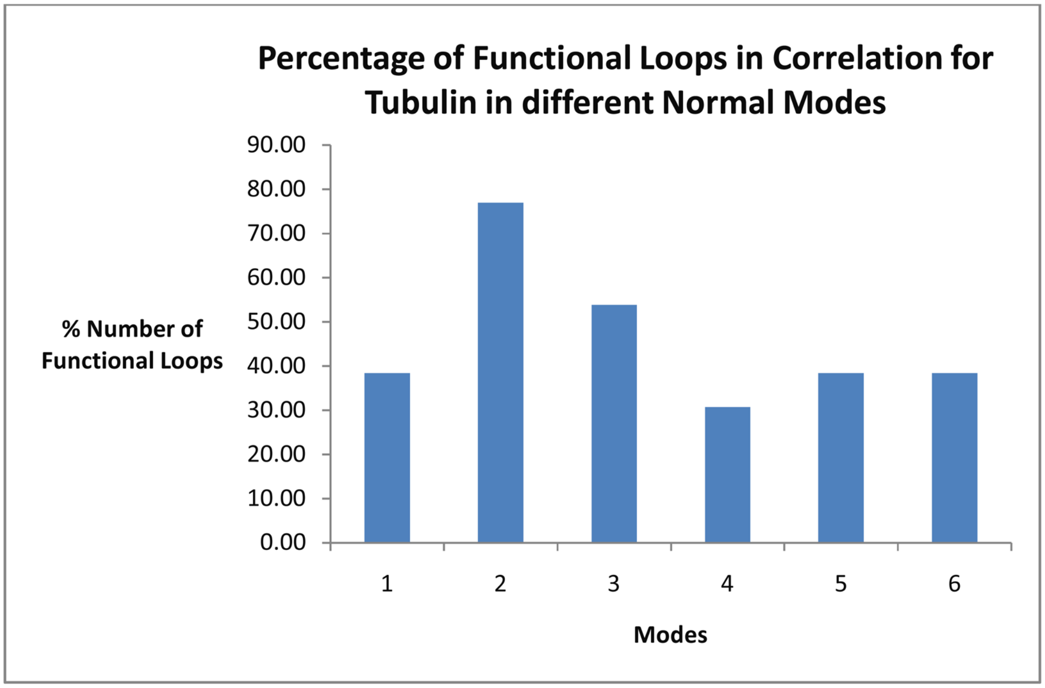


Figure 6.

Percentage number of Functional Loops moving in correlation for five proteins under first six normal modes.

Table 1

The proteins used in this study

Name	PDB ID	State	Residues	# Loops
Tubulin	1TUB	Heterodimer	867	36
HIV-1 reverse transcriptase	1DLO	Heterodimer	971	47
Triosephosphate isomerase	1WYI	Homodimer	496	20
Protease	1J71	Monomer	338	18
Myoglobin	2V1K	Monomer	153	5