IOWA STATE UNIVERSITY
Digital Repository

Biochemistry, Biophysics and Molecular Biology Publications

Biochemistry, Biophysics and Molecular Biology

2012

Protein Loop Dynamics Are Complex and Depend on the Motions of the Whole Protein

Michael T. Zimmermann
Iowa State University

Robert L. Jernigan
Iowa State University, jernigan@iastate.edu

Follow this and additional works at: http://lib.dr.iastate.edu/bbmb_ag_pubs


Part of the Biochemistry Commons, Bioinformatics Commons, Biophysics Commons, Molecular Biology Commons, and the Structural Biology Commons

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/bbmb_ag_pubs/161. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.

This Article is brought to you for free and open access by the Biochemistry, Biophysics and Molecular Biology at Iowa State University Digital Repository. It has been accepted for inclusion in Biochemistry, Biophysics and Molecular Biology Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Protein Loop Dynamics Are Complex and Depend on the Motions of the Whole Protein

Abstract
We investigate the relationship between the motions of the same peptide loop segment incorporated within a protein structure and motions of free or end-constrained peptides. As a reference point we also compare against alanine chains having the same length as the loop. Both the analysis of atomic molecular dynamics trajectories and structure-based elastic network models, reveal no general dependence on loop length or on the number of solvent exposed residues. Rather, the whole structure affects the motions in complex ways that depend strongly and specifically on the tertiary structure of the whole protein. Both the Elastic Network Models and Molecular Dynamics confirm the differences in loop dynamics between the free and structured contexts; there is strong agreement between the behaviors observed from molecular dynamics and the elastic network models. There is no apparent simple relationship between loop mobility and its size, exposure, or position within a loop. Free peptides do not behave the same as the loops in the proteins. Surface loops do not behave as if they were random coils, and the tertiary structure has a critical influence upon the apparent motions. This strongly implies that entropy evaluation of protein loops requires knowledge of the motions of the entire protein structure.

Keywords
protein dynamics, protein loops, molecular dynamics, elastic network models, correlated motions

Disciplines
Biochemistry | Bioinformatics | Biophysics | Molecular Biology | Structural Biology

Comments
This article is published as Zimmermann, Michael T., and Robert L. Jernigan. "Protein loop dynamics are complex and depend on the motions of the whole protein." Entropy 14, no. 4 (2012): 687-700. doi: 10.3390/e14040687. Posted with permission.

Creative Commons License

This work is licensed under a Creative Commons Attribution 3.0 License.

This article is available at Iowa State University Digital Repository: http://lib.dr.iastate.edu/bbmb_ag_pubs/161

Entropy **2012**, *14*, 687-700; doi:10.3390/e14040687

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Protein Loop Dynamics Are Complex and Depend on the Motions of the Whole Protein

Michael T. Zimmermann [1,2] and Robert L. Jernigan [1,2,*]

[1] L. H. Baker Center for Bioinformatics and Biologic Statistics, Iowa State University, Ames, IA 50011, USA

[2] Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA 50011, USA

* Author to whom correspondence should be addressed; E-Mail: jernigan@iastate.edu; Tel.: +1-515-294-7278; Fax: +1-515-294-3841.

Received: 16 February 2012; in revised form: 27 March 2012 / Accepted: 29 March 2012 / Published: 10 April 2012

Abstract: We investigate the relationship between the motions of the same peptide loop segment incorporated within a protein structure and motions of free or end-constrained peptides. As a reference point we also compare against alanine chains having the same length as the loop. Both the analysis of atomic molecular dynamics trajectories and structure-based elastic network models, reveal no general dependence on loop length or on the number of solvent exposed residues. Rather, the whole structure affects the motions in complex ways that depend strongly and specifically on the tertiary structure of the whole protein. Both the Elastic Network Models and Molecular Dynamics confirm the differences in loop dynamics between the free and structured contexts; there is strong agreement between the behaviors observed from molecular dynamics and the elastic network models. There is no apparent simple relationship between loop mobility and its size, exposure, or position within a loop. Free peptides do not behave the same as the loops in the proteins. Surface loops do not behave as if they were random coils, and the tertiary structure has a critical influence upon the apparent motions. This strongly implies that entropy evaluation of protein loops requires knowledge of the motions of the entire protein structure.

Keywords: protein dynamics; protein loops; molecular dynamics; elastic network models; correlated motions

1. Introduction

A longstanding point of view has been that the dynamics of protein loops might be modeled as if they were polymers capable of randomly sampling their various degrees of freedom. This view has its roots in theoretical polymer physics. Contrary evidence has been presented in studies using Elastic Network Models (ENMs) where the loops are observed to move in strong correlation with the large domains of the structures. From these we have even suggested that the functional loops move with the slow domain motions, and not with any significant independence. However, this remains unproven.

The polymer physics viewpoint would treat loops as random Gaussian chains. This approach has been used to treat the statistical distribution of covalently linked rings in condensation polymers given by Jacobson and Stockmayer [1]. The occurrence of rings diminishes for longer chains because of the conformational entropy that grows rapidly with increases in the lengths of the chains. Flory [2] gave an explanation for the formation of the small rings or coils based on various statistical parameters. In this present work, we aim to see whether this random point of view has any validity for protein loops when tested against atomic Molecular Dynamics (MD), and then we compare the MD dynamical freedom, representing the entropies of the loops, to see which extreme viewpoint is more likely, either the random viewpoint or the controlled behavior from the elastic models. In the present study we consider a small set of diverse protein structures, to investigate the behavior of their protein loops, and show that their motional behaviours are far from random, show a high level of complexity and a strong dependence upon the tertiary structure.

The loop regions are often thought to be conformationally less regular fragments of the chain which connect between two secondary structure elements, *i.e.*, alpha helix and beta strands and and also to be more generally exposed at the surface. They have quite variable lengths in their different occurrences. Loops exposed on the surface often play a vital role in protein functions, primarily because they have a greater chance of interacting with the solvent and other molecules. Multiple experimentally determined structures often show the apparent restricted motion of protein loops [3], such as those pairs of structures corresponding to the trajectory between an ‘open’ and a ‘closed’ state. But, these pairs of structures are extremely limited, and in some cases other important intermediate states may to exist. The general results for loops from elastic network studies support the point of view that they are not just random coils moving randomly, but instead often possess well defined characteristics showing limited motions coupled with the large domain motions.

Two recent reviews [4,5] discussed the importance of modelling of loops and their entropic contributions to investigate protein folding pathways. The relative disorder in the folded and unfolded ensembles was quantified as an entropic difference playing an important role in the folding process. Others have realized the importance of loop entropies in the ligand binding process [6]. Successes in RNA structure prediction based on secondary structure considerations have lead to many papers that consider the entropies of nucleic acid stem loops [7]. There have been many recent papers that have devised new methods for sampling conformations to improve entropy evaluations [8–10]. In one of our recent studies [11], we showed that the loops and the coordination of their motions with the entire structure are critical for the functional purposes, and that the funcitonal loops tend to move in coordination with the dominant slow modes of motions of the protein structures; whereas the functionally unimportant loops moved more independently.

In this paper, we investigate whether there is any plausible relationship between the loop motions and the characteristics of the loop such as its length and surface exposure. We present a detailed analysis of the dynamic trajectories based on atomic Molecular Dynamics and also show that these motions closely resemble those computed with ENM, specifically the Anisotropic Network Models (ANM) [12]. In the following work we find that there is no simple relationship between the length of the loop, its flexibility and function. The loops behave not as a random coil, but instead in ways that relate inherently to the topology of the specific protein and its tertiary structure.

2. Results and Discussion

The relative mobilities of each residue are important and are used frequently for studying protein motions; these can serve as an approximate measure of entropy. We compute Mean Square Fluctuations (MSFs) from the molecular dynamics trajectories and compare them to the X-ray crystallographic temperature factors, as well as the MSFs calculated from the ANM using a cutoff of 12 Å. Further details of the structures used and the molecular dynamics trajectories are given in Table 1. Figure 1 illustrates these mobilities for 1flh, while the other three proteins have their mobilities shown in Supplemental Figure S2. It is apparent that all three metrics show the same regions of the structure to be the most mobile, but the relative magnitudes of the motion differ somewhat among the X-ray B factors, the atomic MD and the ANM. While both the ANM and MD simulations correlate well with the X-ray temperature factors, they typically correlate even better with one another. The crystallographic B factors can be affected by the intermolecular interactions within the crystalline state or other contributions from the crystal environment, and these may account for some of the differences.

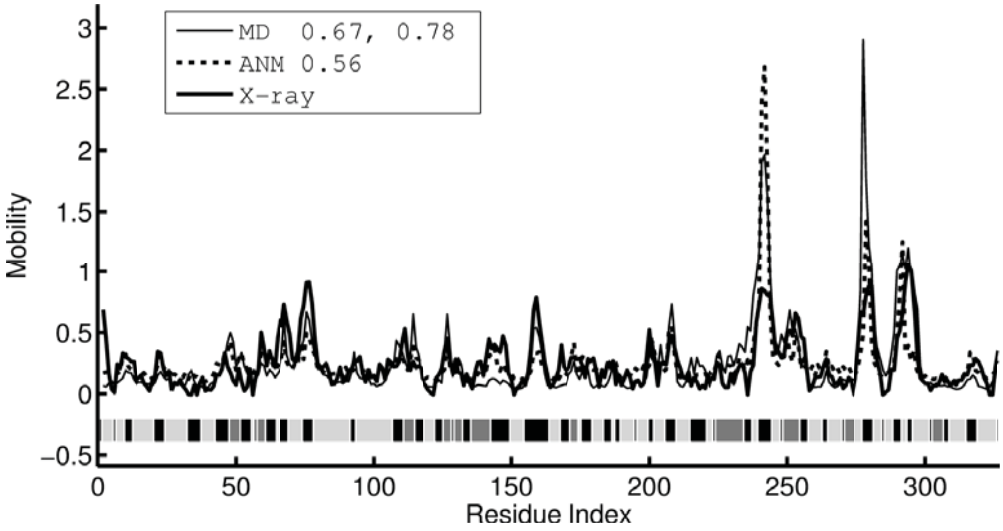
Table 1. Protein structures used in the analyses. MD trajectories for the homologs of the four proteins used in this study were downloaded from the MoDEL [13] database and are listed here. For two structures, a full match was not found. Rather, simulations consisting of only one domain were available.

Protein	PDB structure used	Number of Residues	MoDEL Homolog	% Sequence Identity	Number of Matching Residues	Subunits Included
protease	1j71	338	1flh	30	330	All
myoglobin (MB)	2v1k	153	1gjn	100	153	All
triosephosphate isomerase (TIM)	1wyi	496	1amk	52	239	1 of 2 Subunits
reverse transcriptase (RT)	1dlo	971	1o1w	100	138	RNase Domain

To confirm these similarities in motions, we show in Figure 2 a direct comparison of the motions of several loops in the structures calculated with two independent approaches for four protein structures. The computed mean square fluctuations of the loops from MD and ANM are colored spectrally (from blue to red), with blue indicating parts having the smallest fluctuations, and red the ones with the highest mobilities. There are strong similarities between the relative mobilities of the various parts of

the structures as can be seen in Figures 1 and 2 and Supplemental Figure S2. Thus, the ENM is capable of sampling the overall equilibrium fluctuations similarly to MD.

Figure 1. Comparison of relative mobility of the residues in 1flh for three different metrics. The relative mobilities of each residue are frequently used for studying protein motions. For the uropepsin structure 1flh, we compute the mean square fluctuations (MSFs) from the molecular dynamics trajectory and compare these to the X-ray crystallography temperature factors for 1fly, and the computed MSF for the ANM using a cutoff of 12 Å and all normal modes. Both the ANM and X-ray data has been scaled to fit the MSFs computed from MD for purposes of visualization. Correlation coefficients are shown in the key; the MD and ANM have correlations of 0.67 and 0.56 with the temperature factors, respectively, and exhibit a higher correlation of 0.78 with each other. SSEs are indicated below the curves; black indicates loops (L), dark gray helices (H), and light gray extended strands (E).



How does the motion of individual secondary structure elements (SSEs) compare with each other and coordinate with that of the rest of the structure? In Figure 3, we report the computed time-averaged dot product between SSEs using Equation (2). The secondary structure is taken for the initial conformation from DSSP [14] and the individual SSEs are identified by contiguous blocks of secondary structure. The values are derived from the trajectories downloaded from the MoDEL database [13]. It is evident from the results shown in Figure 3 that few of the loops exhibit correlated motions with one another, but that many loops exhibit highly correlated motions with their sequence proximal SSEs of other types. To explore the loop length dependence, we report the average MSF for each different length of loop (see Figure 3) and find no apparent dependence on loop size. Rather, the pattern of mobility appears to be dependent upon the given protein structure in a complex way.

It has often been postulated that the center of a loop may be less restrained than the flanking parts, because it might be less restricted in its motions by the covalent bonds connecting it to the rest of the globular protein structure. For loops longer than three residues, we divide them into three approximately equal sections; the N-terminal end, the center, and the C-terminal end. The residue-averaged MSF from the MD simulations of each segment is plotted in Figure 4 for 1flh and in the

Supplemental Figure S1 for the other three proteins. Interestingly, for 1fly, there is a bias for the C-terminal end to have lower mobility. This is echoed in 1amk, but is strongly reversed in 1o1w where the C-terminal loop residues are on average twice as mobile. 1gjn shows little bias. Thus, it is not possible to state that residue mobility is a simple function of location within the loop.

Figure 2. Comparison of loop fluctuations calculated with MD (left side) and with the ANM (right side) for four proteins: Protease (1flh), Myoglobin (1gjn), TIM (1amk) and RT (1o1w) shown in the diagrams from top to bottom respectively. Coloring is spectral with blue being having the smallest motions and red the largest motions.

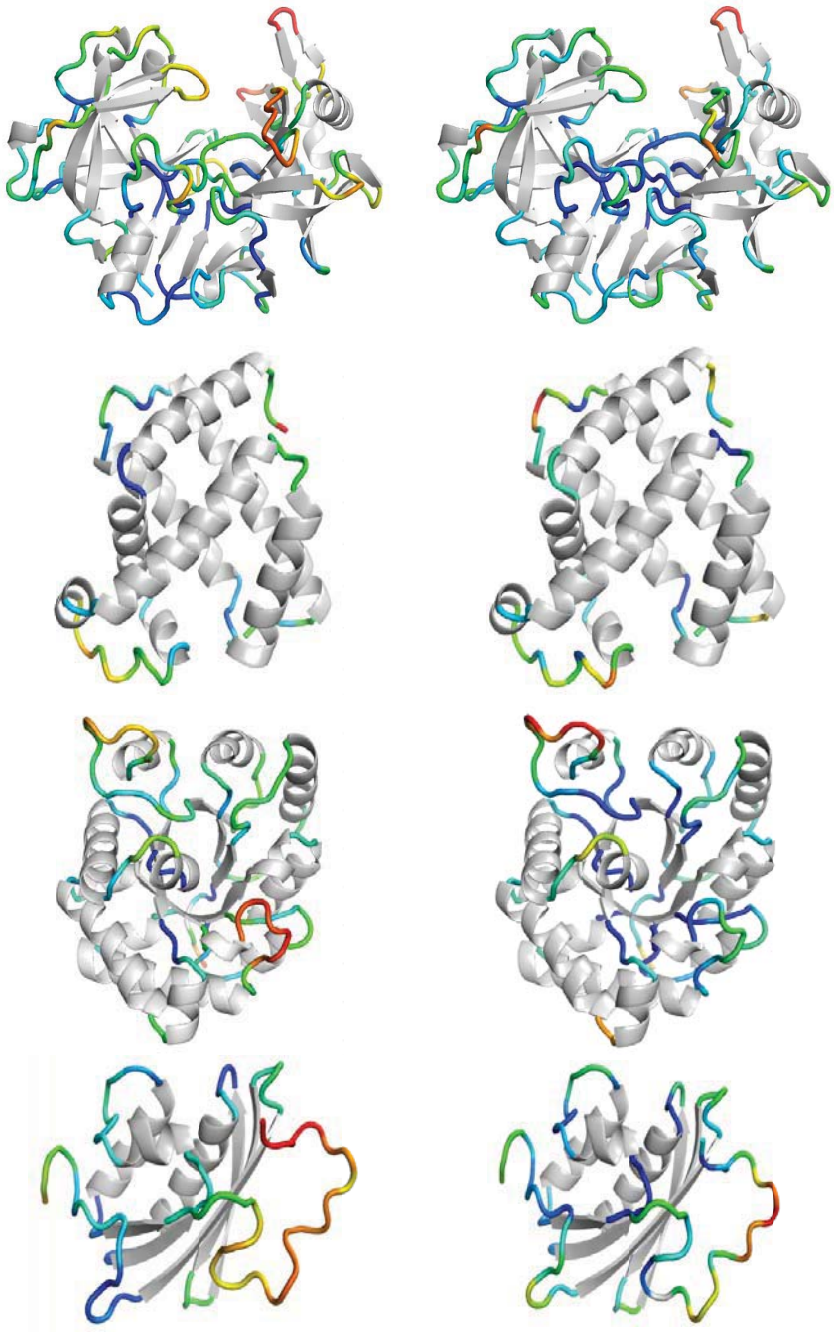


Figure 3. Overlaps between directions of motion of secondary structure elements and the relationship between loop length and residue mobility from MD simulations. For each secondary structure element (SSE), we compute the time averaged dot product between the direction of motion of its residues and the residues of other SSEs using Equation (1). Proteins studied are identified in Table 1 and are (**A**) 1flh, (**B**) 1gjn, (**C**) 1amk, and (**D**) 1o1w. Each of the small blocks in these images is a pairwise comparison between two SSEs. The matrices have been sorted so that SSEs of the same type (strand, E; helix, H; loop, L) are grouped together and within each group they appear in sequence order. It is evident that few of the loops exhibit correlated motions with one another, but that many loops exhibit highly correlated motions with their sequence proximal SSEs of other types. This is most evident in the regions of modest to high correlation running roughly parallel to the main diagonal. To explore the loop length dependence, we report on the right side the average MSF for each different length of loop. Bars represent the range of MSF for residues within loops of a certain length; the mean is shown as a diamond, and the individual residue values as dots. It is difficult to draw any general conclusions about the dependences of the mobilities on the lengths of the loops.

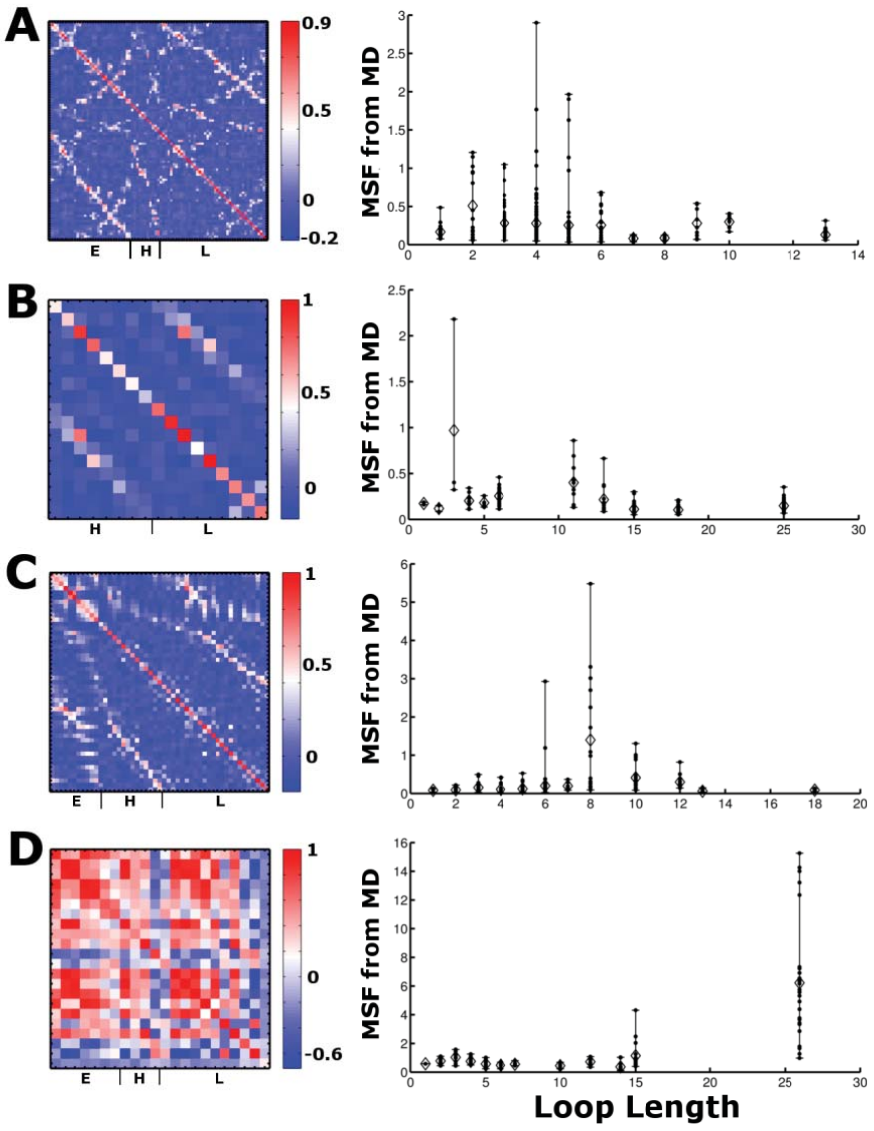
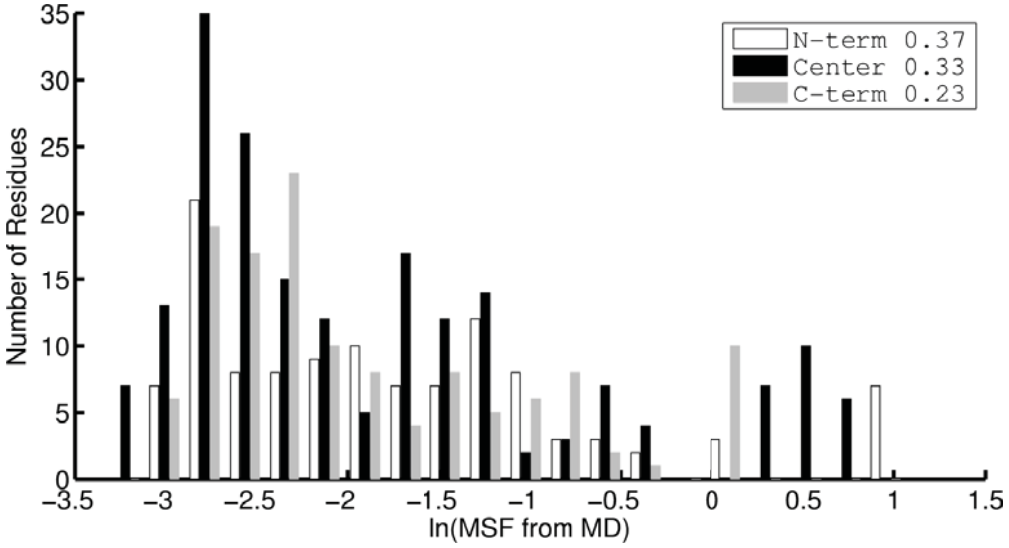


Figure 4. Distributions of the loop fluctuations based on sequence location within the loops of 1flh, from molecular dynamics trajectories downloaded from the MoDEL database. It has often been postulated that residues at the center of a loop would likely be less restrained than their flanking segments, as they might be less restricted in their motion by the covalent bonds attaching the segment to the rest of the structure. For loops longer than three residues in 1flh, we divide them into three approximately equal sections; the N-terminal end, center, and C-terminal end. The residue-averaged MSF of each section is plotted in a 20 bin histogram. To normalize the plot, a natural log transform was used. However, analysis was performed on the untransformed data. The untransformed mean within each group is given in the key.

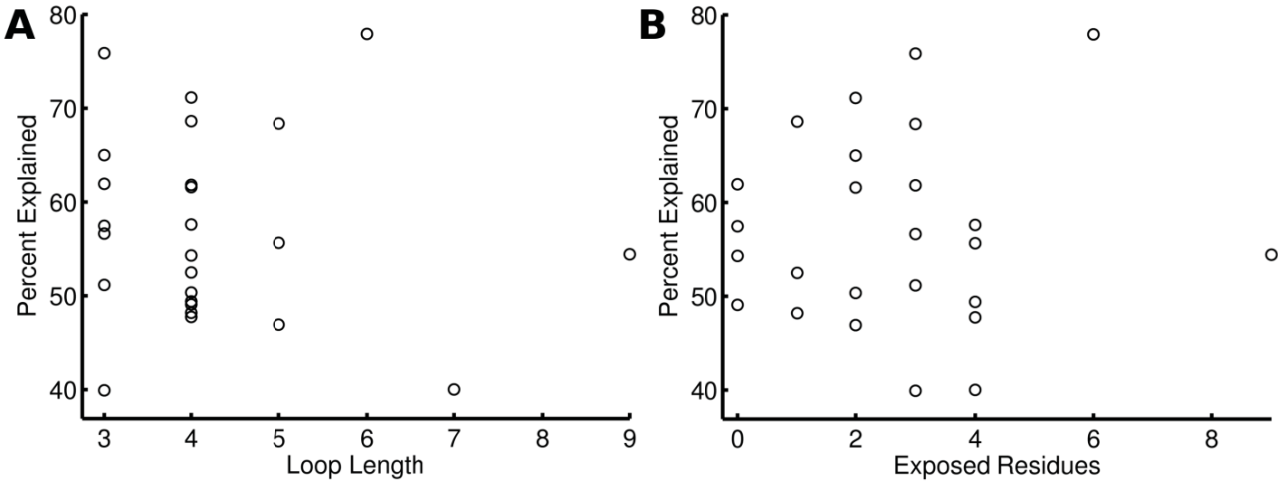


We calculate the Principal Components (PCs) for each loop to determine whether there is any apparent relationship between the magnitudes of motions, loop length, and solvent exposure. The percent of variance explained by the first PC alone is plotted in Figure 5 according to the loop length and the number of solvent exposed residues in the loop and shows that even the relatively long loops that are more solvent exposed exhibit quite cohesive motions within the context of the complete protein. This is demonstrated by a single linear deformation vector (the first PC) capturing more than half, and up to 87%, of the conformational variation of loops, regardless of length or surface exposure. Results are shown in Figure 5 for 1flh and in Supplemental Figure S3 for the other three proteins.

The percent of variance captured by the first PCs is taken as a measure of the cohesiveness or simplicity of the motions, and these are given in Table 2 for four selected loops of length 4, 7, 9, and 15 residues and in five different contexts. Simpler sets of motions are indicative of correlated motions within the loop. That is, a loop's motions are less random if the percentage of variance captured in the first modes is high. Motions of long surface exposed loops are considerably less random within protein structures compared to free peptides. Considering the first 5 PCs, 9 out of the 25 loops in 1flh have 95% or more of their motions captured, 21 at 90% of the variance, and all loops have 80% or more captured. Comparing the percentages of variance captured in Table 2, it is evident that the motions of free peptides behave very differently from the motions of the same fragment within the context of

tertiary structures. See Supplemental Information for results for all loops within each structure as well as for the sequences for the 1flh fragments that were simulated.

Figure 5. Loop motions sampled from MD trajectories of the protease, human uropepsin, are cohesive across lengths and solvent accessibility. We calculate Principal Components (PCs) for each loop in 1flh to determine whether any relationship exists between the motions and the loop length, and solvent exposure. The percentage of variance captured by the first PC is plotted according to the (**A**) loop length and the (**B**) number of solvent exposed residues in the loop. The majority of values are above 50%, so that even the relatively long loops that are solvent exposed exhibit cohesive motions. The single linear deformation vector (the first PC) is capturing more than half, and up to 78%, of the conformational variation. The extreme scatter seen here means that there is no simple relationship between the cohesiveness and either the number of residues in the loop or their solvent exposure.



Because the percentage of variance captured is a coarse measure of the randomness of a trajectory, the directions of motion encoded by the PCs are directly compared by using dot products. Data is visualized in Supplemental Figure S4 and shows that the ensembles are sampled differently in the excised loops and the whole structure. At first glance, it might appear that the motion space is similar and that only the order of the PCs has been shuffled in one ensemble relative to the other. This, however, is not the case since each PC has a weight factor—the percentage of variance captured. For example, the direction of the 1st PC of the 4 amino acid loop in 1flh has an overlap of 0.77 with the free peptide's 9th PC, indicating a strong directional agreement, but the shift from PC1 to PC9 indicates that the direction of motion that is most dominant in the full structure contributes only weakly to the conformational ensemble of the free peptide. In comparing one ensemble against another it is important to account for not only the direction of motion in each PC, but also its relative importance (percent of variance captured) by each. Two ensembles may have a significant one-to-one relationship among their PCs, but if the order of the PCs is quite different, then the ensembles are sampling the motion space in dissimilar ways. Thus, these can possibly be completely different ensembles.

Table 2. Analysis of simulations for isolated loop peptides. The loop dynamics are simulated in one of five contexts: (1) '1flh' indicating the fragment's motions in the full structure were extracted; (2) 'Free' indicating that the loop was simulated in isolation; (3) 'ALA' for poly-alanine of the corresponding length; (4) 'Free-EC' where the free peptide is simulated, but the ends are constrained; and (5) 'ALA-EC' for end-constrained poly-alanine. Four loop fragments of 1flh are extracted, MD is performed using each fragment, and the trajectory is analyzed with PCA using the C^{α} coordinates. Poly-alanine chains of the same length are also simulated to test any effects of the specific side chain interactions. The percent of variance captured by the first 5 PCs is shown for each trajectory. The smaller three loops are surface exposed loops, while the 15 residue fragment is a buried strand that connects two surface exposed loops with each of the three SSEs containing 5 residues. See Supplemental Information for the sequences and location of each fragment. Overall, certain trends are evident: Both the free and alanine behaviors are similar, with the shorter segments showing greater cohesion in their motions, with the longer fragments having a small fraction of their motions captured by the first PCs. Within the context of the protein these segments consistently show less cohesion for the motions of the longer segments, but nonetheless generally a greater cohesion than for the excised segments and alanine segments. One result seems to be readily comprehended—that these segments have significantly less freedom when they are attached to the remainder of the protein, as can be seen from the numbers in the second to last column. The last column shows the WRMSIP [Equation (3)] between each loop trajectory and the trajectory extracted from the full structure.

Length	Context	PC1	PC2	PC3	PC4	PC5	Σ(PC1-5)	WRMSIP
4	1flh	49.1	29.3	18.7	1.5	0.9	99.4	1.00
4	Free	65.6	23.9	7.9	1.5	0.8	99.7	0.16
4	Free-EC	42.5	38.4	16.1	1.4	1.0	99.4	0.21
4	ALA	68.5	17.0	10.7	2.8	0.8	99.8	0.16
4	ALA-EC	58.3	32.3	7.5	1.2	0.5	99.8	0.23
7	1flh	40.0	22.1	17.0	12.0	3.8	95.0	1.00
7	Free	31.7	22.1	10.5	9.3	5.6	79.2	0.41
7	Free-EC	22.5	19.1	13.0	10.1	9.0	73.8	0.30
7	ALA	28.5	14.5	13.9	9.1	7.2	73.2	0.27
7	ALA-EC	18.7	17.2	12.4	10.7	9.5	68.6	0.34
9	1flh	54.4	28.2	7.0	5.0	3.0	97.6	1.00
9	Free	26.0	15.3	14.4	6.5	5.9	68.1	0.28
9	Free-EC	18.5	15.3	12.9	9.3	8.0	64.1	0.22
9	ALA	26.4	13.0	9.4	8.8	7.4	65.0	0.35
9	ALA-EC	20.9	17.9	9.7	8.6	7.7	64.8	0.27
15	1flh	43.0	26.2	11.2	5.8	3.5	89.6	1.00
15	Free	23.7	14.4	11.7	7.3	6.1	63.1	0.52
15	Free-EC	15.3	14.6	11.1	9.0	6.6	56.5	0.41
15	ALA	23.6	12.1	10.4	6.5	5.7	58.3	0.51
15	ALA-EC	15.2	12.9	9.9	7.2	6.8	52.1	0.36

To capture the relationship between the directions (the PCs) and the weights (the percentage of variance captured), we compute a weighted root mean squared inner product according to Equation (3) and present the results in Table 2. Supplementary Table S1 compares the trajectories of free loops with the corresponding end-constrained trajectory. We find that the motions of free loops or end-constrained loops are considerably different from the motions realized within the protein structures.

Given the high agreement between mobilities calculated from ENM and MD, it appears feasible to compute conformational entropy using the ENM from a representative structure. The extent of motion is related to entropy because it is an indication of the number of accessible microstates that the system can occupy. These fluctuation-based entropies could be used in many contexts, and we have begun to use them in combination with knowledge-based energy potentials for refinement of protein tertiary structure predictions and similarly as the basis for selection of native-like docking poses. Our proof-of-concept paper for this approach is given in the reference Zimmermann *et. al.* [15]. An extended study with positive results across three datasets of commonly used docking benchmarks is to appear in *The Journal of Physical Chemistry* [16].

3. Methods

3.1. Anisotropic Network Model

The Anisotropic Network Model (ANM), which has been extensively developed and summarized in depth elsewhere [12,17] and was based on the original concept from Tirion [18], is utilized here to compute coarse-grained dynamics based for a structure. Such structures can be taken from many sources, including X-ray crystallography, NMR, Electron Microscopy, and pre-equilibrated Molecular Dynamics (MD) conformations. The ANM model assumes that the structure represents a minimum energy conformation and that all deviations from this conformation have an energetic cost. The motions of the structure that have the least energetic cost are favored and dominate the computed motions. Such motions also are collective, involving internal motion throughout the bulk of the structure. These have been extensively applied and are often found to represent the large scale domain motions better than atomic molecular dynamics simulations because they do not require the long computed trajectories from molecular dynamics. They have been shown to provide an efficient sampling of the motions of proteins structures, and as a result should also be useful for evaluating entropies.

3.2. Analysis of Molecular Dynamics Trajectories

Numerous molecular dynamics trajectories are available to download from the MoDEL database [13]. They are distributed from the database in a compressed form that captures the motion apparent in the first Principal Components [19], or PCs, of the simulation such that at least 90% of the variation is captured. In this way, much of the random noise is filtered out, as in Essential Dynamics [20]. Four different proteins were chosen based their diversity in size and function: aspartic protease, myoglobin, triosphosphate isomerase, and reverse transcriptase. Trajectories of these structures or their close homologues were downloaded, and these are listed in Table 1. For two structures, simulations of only one domain from the full structure were available. In the case of triosphosphate isomerase, the protein acts as a homodimer, but the simulation was performed on the monomer. For reverse transcriptase, the

RNase H domain has been simulated. All trajectories used were simulated for 10ns using Amber 8.0 software [21], the Amber99 force field, and TIP3P explicit solvation. To ensure that we are using properly equilibrated data, the first 5ns of the trajectories were discarded. The final 5ns were used in analysis and the first of these frames used in ANM generation.

Because we seek to analyze the results on the residue level, the atomic trajectories are first reduced to only the C^{α} atom positions prior to analysis. The covariance between each atom pair is then quantified with the normalized time averaged dot product of the changes in position [22]:

$$C(i,j) = \frac{< \Delta R_i \cdot \Delta R_j >}{< \Delta R_i^{1/2} >< \Delta R_j^{1/2} >} \quad (1)$$

where ΔR_i is the displacement vector of atom i between consecutive time steps and <> denotes time (ensemble) averaging. In this study, we focus on the dynamics of loops and their relationship to the remainder of the structure. Therefore, it is of interest to generalize Equation (1) to an even more coarse level. Secondary Structure Elements (SSEs) are identified from DSSP [14] and are defined as a segment of sequence with the same secondary structure. To investigate the correlation between motions of pairs of secondary structure segments, including individual loops, the time averaged dot product between two SSEs is defined in Equation (2) and is the average of the covariance of the individual atoms within each of the two SSEs:

$$SSE_a \cdot SSE_b = \frac{\sum_{i \in SSE_a} \sum_{j \in SSE_b} C(i,j)}{n(SSE_a)n(SSE_b)} \quad (2)$$

where $n(SSE_a)$ is the number of residues in SSE a.

In order to capture the difference in sampling between two trajectories we compute the weighted root mean squared inner product (WRMSIP) between the first I PCs from the first trajectory and the first J PCs from the second. First, we compute the relative weight of the i^{th} PC, w_i. For each pair of PCs between the trajectories we compute the ratio of their weights: $r_{ij} = \frac{\min(w_i^1, w_j^2)}{\max(w_i^1, w_j^2)}$, where the superscript denotes which trajectory it is from. We then weight the pairwise inner products:

$$WRMSIP = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} r_{ij}\left|PC_i^1 \cdot PC_j^2\right| \quad (3)$$

where PC_i^1 is the i^{th} principal component from the first trajectory, PC_j^2 is the j^{th} from the second. The dot product accounts for the agreement in direction, while the weight accounts for the extent of sampling in that direction. RMSIP has been used in many studies and is explained well in Leo-Macias *et al.* [23], however it does not capture the PC weights as does our modification. Values approaching 1 indicate that the ensembles are identical, while smaller numbers indicate reduced coverage. All calculations presented here use $I = J = 10$ so that we consider the bulk of the important motions.

It should be noted that our WRMSIP counts differences in a nonlinear way. For example, say we consider the first three PCs from two trajectories where the directions of the PCs are identical and they have weights (percent of variance) of $w^1 = [0.5, 0.3, 0.2]$ and $w^2 = [0.5, 0.4, 0.1]$. The WRMSIP

would be 0.75. While ten percent of the variance has shifted from PC3 in the first trajectory to PC2 in the second, the WRMSIP decreases by more than 0.1.

3.3. Comparison of Structural Loops and Free Peptides

The complexity of loop motions within the MD simulations is quantified by Principal Component Analysis (PCA). PCA transforms the input trajectory data into a new coordinate system where the PCs form the basis. The first PC captures the largest fraction of the variance, the second captures the largest part of the remaining variance, and so on. If loop motions are highly random in nature, then individual PCs can be anticipated to capture only a fraction of the total variance. More correlated motions will be more concisely captured by a smaller set of PCs. This will allow us to distinguish between loop motions with either highly diverse motions or those having more internally correlated directions of motion. We perform PCA for each loop in each structure individually in order to determine the cohesiveness of its motions.

As control cases, we also perform 10 ns atomic MD simulations of representative loops extracted from the protease 1flh as free peptides of lengths 4, 7, 9, and 15 residues using the CHARMM27+CMAP [24,25] force field. While small differences in long timescale dynamics (hundreds of ns) have been observed between different force fields, overall their agreement is quite high [26,27]. It is possible that the specific side chains present in these loops will affect the types or extent of motions sampled. Thus, a second set of control simulations is performed by using the same parameters for poly-alanine chains of the same lengths.

It is possible that the differences in motions between the free peptides and the peptides within their structural context are due to end-constraints. That is, when the peptide is within a protein structure, its ends are not free to move, since they are constrained by the flanking structure. To test this effect, we perform a second type of control simulation where the N- and C-terminal C^{α} atoms are harmonically constrained by a force of 5 kcal/mol. These are also presented in Table 2 and labeled by "–EC" for End-Constrained.

As the percent of variance captured is only an approximate measure of the conciseness of a set of motions and does not compare the directions of motion between two datasets, we also compute the dot products between the directions indicated in essential dynamics of loops within the protease 1flh and the corresponding excised free peptides. To further capture the agreement between the trajectories including the percent of variance that each PC captures, WRMSIP is calculated according to Equation (3).

4. Conclusions

In this work we investigate the relationships between loop motions and the motions of protein structures. There exists no apparent dependence on loop length or the number of solvent exposed residues within a loop. Rather, the nature of the tertiary structure is a dominant influence over these motions, and prevents the development of any general rules. Many loops have high agreement in their direction of motion with the secondary structures they connect—helices or strands. Due to the cohesive nature of the ANM, it could be argued that sets of cohesive motions derived for protein loops may be an artifact of the model, and not the genuine behavior of protein structures. However, in the present

study we have presented an analysis of atomic MD trajectories, which might be expected to enable a greater extent of randomness for the molecular interactions, and these studies also confirm the large difference in loop dynamics between the free and structured contexts, as well as the lack of any general relationship between loop size, exposure, and mobility. We have shown that free peptides behave extremely differently from loops within proteins. Therefore, surface loops do not behave as if they were random coils, but the tertiary structure has a critical impact upon the realized motions.

Acknowledgements

This work has been support by NIH grant R01GM072014. We thank Debkanta Chakraborty for assistance with the figures.

References

1. Jacobson, H.; Stockmayer, W.H. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* **1950**, *18*, 1600–1606.
2. Flory, P.J.; Semlyen, J.A. Macrocyclization equilibrium constants and the statistical configuration of poly (dimethylsiloxane) chains. *J. Am. Chem. Soc.* **1966**, *88*, 3209–3212.
3. Hu, X.; Stebbins, C.E. Dynamics of the WPD loop of the Yersinia protein tyrosine phosphatase. *Biophys. J.* **2006**, *91*, 948–956.
4. Chirikjian, G.S. Modeling loop entropy. *Methods Enzymol*. **2011**, *487*, 99–132.
5. Zhou, H. Loops, linkages, rings, catenanes, cages, and crowders: Entropy-based strategies for stabilising proteins. *Acc. Chem. Res.* **2004**, *37*, 123–130.
6. Baron, R.; McCammon, J.A. (Thermo)dynamic role of receptor flexibility, entropy, and motional correlation in protein-ligand binding. *Chemphyschem* **2008**, *9*, 983–988.
7. Srinivasan, J.; Miller, J.; Kollman, P.A.; Case, D.A. Continuum solvent studies of the stability of RNA hairpin loops and helices. *J. Biomol. Struct. Dyn.* **1998**, *16*, 671–682.
8. Mihailescu, M.; Meirovitch, H. Entropy and free energy of a mobile loop based on the crystal structures of the free and bound proteins. *Entropy* **2010**, *12*, 1946–1974.
9. General, I.J.; Meirovitch, H. Relative stability of the open and closed conformations of the active site loop of streptavidin. *J. Chem. Phys.* **2011**, *134*, 025104.
10. General, I.J.; Dragomirova, R.; Meirovitch, H. New method for calculating the absolute free energy of binding: The effect of a mobile loop on the avidin/biotin complex. *J. Phys. Chem. B* **2011**, *115*, 168–175.
11. Skliros, A.; Zimmermann, M.T.; Chakraborty, D.; Saraswathi, S.; Katebi, A.R.; Leelananda, S.P.; Kloczkowski, A.; Jernigan, R.L. The importance of slow motions for protein functional loops. *Phys. Biol.* **2012**, *9*, 014001.
12. 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*, 505–515.
13. Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Perez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; *et al*. MoDEL (Molecular Dynamics Extended Library): A database of atomistic molecular dynamics trajectories. *Structure* **2010**, *18*, 1399–1409.

14. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
15. Zimmermann, M.T.; Leelananda, S.P.; Gniewek, P.; Feng, Y.; Jernigan, R.L.; Kloczkowski, A. Free energies of coarse-grained proteins by integrating multibody statistical contact potentials with entropies from elastic network models. *J. Struct. Funct. Genomics* **2011**, *12*, 137–147.
16. Zimmermann, M.T.; Leelananda, S.P.; Kloczkowski, A.; Jernigan, R.L. Combining statistical potentials with dynamics based entropies improves selection from protein decoys and docking poses. *J. Phys. Chem.* **2012**, in press.
17. Zimmermann, M.T.; Kloczkowski, A.; Jernigan, R.L. MAVENs: Motion analysis and visualization of elastic networks and structural ensembles. *BMC Bioinformatics* **2011**, *12*, 264.
18. Tirion, M.M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.
19. Jolliffe, I.T. *Principal Component Analysis*; Springer-Verlag: New York, NY, USA, 2002.
20. Hayward, S.; de Groot, B.L. Normal modes and essential dynamics. In *Molecular Modeling of Proteins*; Humana Press: Totowa, NJ, USA, 2008; p. 89–106.
21. Case, D.A.; Cheatham, T.E., III.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
22. Ichiye, T.; Karplus, M. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins* **1991**, *11*, 205–217.
23. Leo-Macias, A.; Lopez-Romero, P.; Lupyan, D.; Zerbino, D.; Ortiz, A.R. An analysis of core deformations in protein superfamilies. *Biophys. J.* **2005**, *88*, 1291–1299.
24. MacKerell, A.D., Jr.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2000**, *56*, 257–265.
25. MacKerell, A.D. Jr.; Feig, M.; Brooks, C.L., III. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.
26. Cerutti, D.S.; Freddolino, P.L.; Duke, R.E., Jr.; Case, D.A. Simulations of a protein crystal with a high resolution X-ray structure: Evaluation of force fields and water models. *J. Phys. Chem. B* **2010**, *114*, 12811–12824.
27. Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Perez, A.; Camps, J.; Hospital, A.; Gelpi, J.L.; Orozco, M. A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 796–801.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).