Published in final edited form as:
Proteins. 2012 February ; 80(2): 616–625. doi:10.1002/prot.23225.

Longer simulations sample larger subspaces of conformations while maintaining robust mechanisms of motion

Lin Liu[1,2], **Angela M. Gronenborn**[2], and **Ivet Bahar**[1,*]

[1]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213

[2]Department of Structural Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213

Abstract

Recent studies suggest that protein motions observed in molecular simulations are related to biochemical activities, although the computed time scales do not necessarily match those of the experimentally observed processes. The molecular origin of this conflicting observation is explored here for a test protein, cyanovirin-N (CV-N), through a series of molecular dynamics simulations that span a time range of three orders of magnitude up to 0.4 microseconds. Strikingly, increasing the simulation time leads to an approximately uniform amplification of the motional sizes, while maintaining the same conformational mechanics. Residue fluctuations exhibit amplitudes of 1–2 Å in the nanosecond simulations, while their average sizes increase by a factor of 4–5 in the microsecond regime. The mean-square displacements averaged over all residues (y) exhibit a power law dependence of the form $y \propto x^{0.26}$ on the simulation time (x). Essential dynamics analysis of the trajectories, on the other hand, demonstrates that CV-N has robust preferences to undergo specific types of motions that already can be detected at short simulation times, provided that multiple runs are performed and carefully analyzed.

Keywords

structure-encoded dynamics; molecular dynamics simulations; power law; global motions; equilibrium fluctuations of cyanovirin-N

INTRODUCTION

Native proteins are not static entities under physiological conditions. On the contrary, they undergo a broad range of motions around their native state structures, ranging from local conformational changes such as peptide bond re-orientations or amino acid side chain isomerization to global rearrangements involving entire domains or subunits. The type and size of these motions are governed by the free energy landscape near native state conditions.[1–3] In terms of functional relevance, many structural rearrangements, especially

[*]Correspondence to: Ivet Bahar, Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Suite 3064 Biomedical Science Tower 3, 3501 Fifth Ave., Pittsburgh, PA 15213, Voice: 412 648 3332 - Fax: 412 648 3163, bahar@pitt.edu.

those collectively involving large substructures, are necessary for proteins to carry out their chemical and biological activities.[1–5] Therefore, in order to understand protein function, it is necessary to also examine the *dynamics* of proteins and not only their atomic *structures*. In particular, the lowest frequency internal motions, or *global* motions, need to be evaluated since they usually relate to the molecules' biological functions.

Despite the complexity of protein motions, and contrary to expectations, experimental and computational studies suggest that dynamic features that are detected computationally or experimentally at short times may explain experimental data associated with much slower processes. A typical example is the dataset of order parameters derived by Palmer and coworkers for protein G binding domain 3,[6] based on two alternative datasets: NMR relaxation parameters for probing motions on the order of nanoseconds[7] and residual dipolar couplings (RDCs) that probe motions on the microsecond time scale.[8] Notably, the order parameter profiles extracted from these two datasets exhibit similar shapes,[6] and the most 'disordered' residues, associated with the minima in the order parameter profiles plotted as a function of residue number (Fig. 1A), become more pronounced in the longer-time events. In contrast, the *shape* of the profiles, i.e., the distribution of order parameters as a function of residue index, remains essentially unchanged, suggesting that events at short time scales and those at long time scales share common features. Another example that indicates similar behavior is an NMR study of ubiquitin in which RDC and spin-lattice relaxation experiments exhibit comparable profiles that also agree with the predictions of accelerated molecular dynamics (MD) simulations, except for the amplitudes of the motions at long times.[9–11] Likewise, results for GB1 from two different length MD runs (Fig. 1B) also demonstrate that the two simulations result in comparable order parameter profiles.[12] In addition, other observations indicate a correspondence between experiments and computations, such as the relationship between MD events and catalytic turnover times observed by Kern and coworkers for adenylate kinase, even though the MD events are several orders of magnitude faster than the experimental ones.[13] All these observations point to the existence of robust mechanism(s) of motions that underlie both short-time and long-time dynamics.

Atomic motions can be divided into three basic components: the time scale of the motion, its amplitude, and its direction. In the strictest sense, characterization of protein dynamics requires the collection of thousands of time-resolved data at multiple length and time scales.[5] As mentioned above, a broad range of experimental techniques provides information on protein dynamics, including NMR relaxation measurements,[14,15] Laue X-ray diffraction data,[16,17] infrared and fluorescence spectroscopy,[18] and single-molecule studies,[19] although they inform about different aspects and time scales of protein dynamics. On the computational side, structure-based methods such as MD simulations[20] and normal mode analysis (NMA) with elastic network models (ENMs)[21–24] have been exploited to gain insights into biomolecular systems dynamics. In particular, MD simulations are uniquely suited for examining time-resolved events in proteins at high resolution. Although extremely powerful, two shortcomings are inherent to MD simulations.[25] The first arises from sampling inefficiency, which becomes increasingly noticeable in large molecular systems.[25–27] Limitations of this nature can be alleviated to some extent by performing multiple independent runs for assessing convergence.[26,28–31] Second, the lengths of MD runs often remain below microseconds due to memory and computing time limitations.[25] Therefore, it still is an open issue whether functional motions at low frequencies can be inferred from relatively short MD runs.

The present study is carried out using cyanovirin-N (CV-N), a potent HIV-inactivating protein widely investigated in our previous studies,[32] to answer the following questions: (i) How similar are the residue fluctuation profiles for different lengths runs? (ii) Do top-ranking modes from a short simulation become high frequency modes with increasing simulation time?[33] (iii) Do short MD simulations provide insights into functional motions, i.e., to what extent are the directions of motions near the native state energy minimum at short simulation times preserved at longer times? (iv) Do simulations provide information on the absolute sizes of various mechanisms of motions?

The results obtained for CV-N in combination with data reported previously for other systems, suggest that the distribution (or *relative* size) of residue fluctuations along the polypeptide chain is a robust quantity under equilibrium conditions, predominantly defined by the 3-dimensional architecture in the native state, while their *absolute* size predicted by MD simulations change with simulation duration, in the time regime (< 400 ns) investigated. The ratios of the mean-square displacements, $y = <(\Delta \boldsymbol{R})^2>_{MDk}/<(\Delta \boldsymbol{R})^2>_{MDk'}$ observed in two MD runs k and k' of different durations, and that of the total simulation times, $x = t_{MDk}/t_{MDk'}$, are governed by a power law of the form $y = x^{0.26}$, similar to results reported by Scheraga and co-workers.[34,35] The decomposition of the trajectories into essential modes reveals that well-defined directions of the global motions, encoded by the native topology of inter-residue contacts, can be discerned even in short runs, as long as the region around the native state energy minimum is comprehensively sampled by multiple runs.

MATERIALS AND METHODS

MD simulations

The starting CV-N structure (PDB ID: 2EZM)[32] is highly anisotropic, occupying a volume of about 30 × 52 × 27 Å^3. The inset of Fig. 2 shows the CV-N structure. We used a simulation box of size 40 × 62 × 37 Å^3, which ensured a minimal water layer thickness of 5 Å for all surface residues. This thickness has been verified in our earlier simulations,[36] and shown in previous work,[37] to satisfactorily solvate the protein. The resulting system consisted of 8,159 atoms, including 2,216 TOP3P water molecules. NAMD[38] with the Charmm22 force field[39] was used with a 2 fs time step. After energy minimization and equilibration, multiple independent runs were performed at constant temperature (298 K) and pressure (1 atm).

Principal component analyses (PCAs) of MD trajectories and NMR models

The instantaneous position $\boldsymbol{R}_i(t)$ of each residue i is defined by the coordinates of its α-carbon atoms, which are organized into a 3n-dimensional vector of instantaneous configurations, $\boldsymbol{R}(t)$, for the protein of n residues. The configuration vector definition applies to snapshots recorded at fixed intervals from MD runs, as well as the models in the NMR structure ensemble (where t is replaced by the model index). The global changes in configuration originating from the collective fluctuations sampled in each MD run, or associated with the structural deviations observed in NMR ensemble, are identified by the same procedure, described here for MD snapshots. First, the instantaneous fluctuation $\Delta\boldsymbol{R}_i(t) = \boldsymbol{R}_i(t) - \langle\boldsymbol{R}_i\rangle$ from mean position $\langle\boldsymbol{R}_i\rangle$ is evaluated for each residue, for each recorded time t (a total of m snapshots or models). This is performed after optimal superimposition of the configuration onto the starting structure so as to eliminate the rigid-body translations and rotations. The superimposition is achieved by least squares fitting to backbone heavy atoms. Second, the fluctuation vectors $\Delta\boldsymbol{R}_i(t)$ $(1 \leq i \leq n)$ are organized in a trajectory matrix $\mathbf{A}$ of dimension $3n \times m$, for a set of m snapshots. Multiplication of $\mathbf{A}$ by its transpose and division by m yields the $3n \times 3n$ covariance matrix $\mathbf{C}$ for each run (or for the NMR ensemble). $\mathbf{C}$ may be expressed as an $n \times n$ supermatrix, the element $\mathbf{C}_{ij}$ of which is a 3×3 matrix of the form

$$\mathbf{C}_{ij} = \begin{bmatrix} \langle\Delta X_i \Delta X_j\rangle & \langle\Delta X_i \Delta Y_j\rangle & \langle\Delta X_i \Delta Z_j\rangle \\ \langle\Delta Y_i \Delta X_j\rangle & \langle\Delta Y_i \Delta Y_j\rangle & \langle\Delta Y_i \Delta Z_j\rangle \\ \langle\Delta Z_i \Delta X_j\rangle & \langle\Delta Z_i \Delta Y_j\rangle & \langle\Delta Z_i \Delta Z_j\rangle \end{bmatrix} \quad (1)$$

Here, $\langle\Delta\boldsymbol{X}_i \Delta\boldsymbol{Y}_j\rangle$ represents the cross-correlation between the X-component of $\Delta\boldsymbol{R}_i$ for residue i and the Y-component of $\Delta\boldsymbol{R}_j$ for residue j, averaged over all m snapshots. Third, the eigenvalue decomposition of $\mathbf{C}$ produces $3n - 6$ nonzero eigenvalues and the corresponding eigenvectors. The eigenvectors define the directions of motions, and the eigenvalues scale the squared amplitudes of fluctuations. The Gaussian Network Model (GNM)[21,22] and anisotropic network model (ANM)[40,41] analyses also lend themselves to a series of eigenmodes, as described in the Supporting Information.

RESULTS AND DISCUSSION

The distribution of residue fluctuations is insensitive to the duration of simulations

We selected CV-N as our model system, based on its small size (n = 101 residues), its considerable thermodynamic stability and the large body of prior data available in our laboratory.[42–45] We compared the dynamic information retrieved from 1 ns to 400 ns MD runs. As depicted in the Fig. 2 inset, CV-N has a compact, pseudo-symmetric fold and is made up of two domains. Residues 1-39 and 91-101 form domain A (green), and 40-90, domain B (blue). The two domains share 32% sequence identity and are connected by short helical linkers. Each domain is composed of a triple-stranded β-sheet with a β-hairpin packed on top. There are two carbohydrate-binding sites located at distal positions (shown in red), one in each domain.[46] The two binding sites exhibit distinct affinities and specificities for high-mannose sugars.[47] The rotational correlation time τ_c of CV-N has been measured to be 4.5 ns.[48] Our simulations thus permit us to investigate both the sub-τ_c and supra-τ_c dynamics of CV-N under native state conditions.

Fig. 2 presents the results from a series of fifty-eight runs, adding up to a total simulation time of 2 microseconds. Multiple trajectories were generated for each simulation time (t_{MDk} = 1, 5, 25, 100 and 400 ns, also called the *time window*) to reduce inaccuracies arising from inadequate sampling of sub-states near the native state, especially for the short runs. The curves in Fig. 2 represent the mean-square-fluctuations (MSFs) in residue positions, $<(\Delta \boldsymbol{R}_i)^2>$ for residue $1 \leq i \leq n$, for each time window in the range 1 to 400 ns, averaged over all runs of a given duration. Residue positions are those of the α-carbons.

As can be appreciated, the family of curves shown in Fig. 2 exhibits a striking similarity between the shapes of the residue fluctuation profiles for the different time windows. Essentially, all peaks/maxima that are noted at short time scales (e.g., 1–5 ns simulations) are amplified at longer times, with small changes in the relative sizes of the residue excursions. In principle, one might expect to detect new modes of motion at longer times, possibly changing the MSF profiles. However, only slight variations can be discerned in the profiles, such as the emergence of a peak near the helical hairpin loop around residues 65-67 in domain B in the longer time windows. Indeed, most features are robustly maintained: the loop regions usually tend to have high fluctuations, while secondary structure elements exhibit more restricted motions. Interestingly, an asymmetry in residue fluctuations can be seen, with residues in domain B exhibiting larger motions than those in domain A, consistently noted in all simulations.

A quantitative measure of the degree of similarity between these MSF profiles is provided by the correlation coefficients listed in Table I. The correlation coefficient between the MSFs for the 1 ns runs and the 400 ns runs is 0.83. Thus increasing the time window of observation by 4–5 orders of magnitude essentially leaves the shape of the fluctuation profile unchanged. A recent study of MDM2 dynamics also showed that the correlations between dihedral angle motions were conserved while the motional amplitudes changed upon binding the p53-peptide ligand,[49] which also supports the view that the conformational mechanisms are robustly maintained while the sizes of motions differ.

What distinguishes the different MSFs is their absolute size. The longer the simulation, the further the displacement of a residue from its mean position is. The increase in fluctuations is also evident from the root-mean-square-deviation (RMSD) profiles provided in Fig. S1. The RMSD from the original structure remains around 3.7 Å, which may be viewed as an indication of sampling the native state energy minimum even though this state may comprise narrowly distributed microstates that differ in their local conformers. But the fluctuations around the average RMSD increase with increasing simulation time, consistent with the observed dependence of $<(\Delta \boldsymbol{R}_i)^2>$ on the duration of the simulation. In order to uncover whether and what kind of dependency exists between the MSFs and the simulation time, we analyzed the data further (below).

The increase in residue MSFs with simulation duration obeys a power law

First, we consider two sets of trajectories, corresponding to two simulation times, e.g., $t_{MD1} = 1$ ns and $t_{MD2} = 5$ ns. Fig. 3A displays the $<(\Delta \boldsymbol{R}_i)^2>$ values of residues $2 \leq i \leq 101$ for these two time windows: the abscissa represents the MSFs observed in MD1, and the ordinate, that in MD2. Linear regression of the data yields a correlation coefficient R^2 of 0.95, the slope of which, 1.34 in the present case, represents the average ratio of residue MSFs observed in MD2 to those in MD1. In other words, increasing the simulation time by a factor of 5 increases the residue MSFs by 34%, on average. Panel B represents a similar plot for two other time windows, t_{MD3} = 25 ns and t_{MD5} = 400 ns, which, in turn, yields a slope of 2.14, i.e., increasing the simulation time by a factor of 16 enhances the square displacements by a factor of 2.14.

Repeating the same analysis for all pairwise combinations of simulation times, t_{MDk} for k = 1–5 (5!/3!2! = 10 of them), yields the *master curve* displayed in Fig. 3C. The data points show the enhancements in the MSFs accompanying the increases in the simulations, also listed in Table SI, for each pairwise combination. In other words, the ratio of MSFs for each pair of MD runs is plotted against the ratio of simulation times in Fig. 3C. Each point represents the average behavior of *all* residues, averaged over multiple runs, i.e., the resulting dependence represents the outcome from the *complete* dataset of trajectories with a cumulative simulation time of 2 μs. Note that the scales of both, abscissa and ordinate, is logarithmic and a linear relationship on such a log-log plot indicates a power law of the form $y \sim x^{\alpha}$. The value of the exponent is evaluated from the slope of the best fit and is 0.26. Thus, the overall dependence is

[illegible] (2)

The subscript i in $<(\Delta \boldsymbol{R}_i)^2>$ has been removed since the MSFs refer to averages over all residues. Eq. (2) conveys two messages: (i) the MSFs observed in MD simulations depend on the duration of the simulations, and (ii) the dependence obeys a power law, with exponent 0.26. While this dependence seems small, it maps to displacements of $<(\Delta \boldsymbol{R})^2>_{MD1}$ = 0.5 Å^2 for t_{MD1} = 1 ns, and $<(\Delta \boldsymbol{R})^2>_{MD5}$ = 2.6 Å^2 for t_{MD5} = 400 ns. Thus, the square amplitudes of motions are enhanced by a factor of ~ 5 in long simulations. The major difference between short and long runs appears to be the larger excursions undertaken by the molecule around the native state energy minimum in longer runs, while the preferred directions of motions exhibit little, if any, changes.

The power law observed in present simulations (Eq. (2)) applies to CV-N equilibrium dynamics near its native state; it cannot be extended to larger scale transitions, such as those occurring during unfolding events. Evidently, the shape of the native state energy minimum defines the maximal size of fluctuations accessible to a given protein under native state conditions, and those beyond a certain range inevitably fall into new energy minima, including the unfolded state; and fluctuations in the unfolded state are limited by chain connectivity or covalent bonds. Such structural changes involving partial or complete unfolding events are beyond the range of current equilibrium simulations which maintain the native fold. The increase in the motional amplitudes simply reflects the sampling of a broader range of the global energy basin with increasing time window (up to 400 ns), and suggests that the observed MSFs simply reflect the portion of the global energy basin that is being accessed in a given run.

We further analyzed the behavior of each residue. Calculations yielded a range from 0.13 to 0.46, for the exponent α, depending on residue position/conformation (see Fig. 4). Larger exponents indicate a more pronounced dependence of the fluctuation sizes on the simulation time, i.e., residues with larger exponents enjoy larger conformational freedom. Examining the exponents with respect to secondary structure elements clearly indicated that loop residues possess larger exponents than their neighbors located in helices and β-strands. Another interesting observation is the distinctive distributions of exponents in the two structurally similar, but distinct, domains of CV-N. Fig. 4 suggests that in some cases it may be possible to use the exponent of individual residues or substructures to gain information on intrinsic dynamics, or conformational flexibility, which, in turn, may inform on functional properties.

The above power law relationship suggests that there may be a time-dependent conformational drift throughout our simulations, even though we are exploring the neighborhood of the native state energy minimum. The deviation of the time-dependence of observed motion from that of a classical Brownian motion (where the exponent α is unity) might be attributed not only the subdiffusive motion which has been suggested to originate from the trapping in a local minimum/sub-state of the native state in the energy landscape[35,50] and from the sampling of infrequent and large jumps between such local minima,[51] but also to the bounded motion of the protein constrained by native contact topology in addition to covalent bonds.

Comparison of essential modes extracted from different MD runs

Towards gaining a better understanding of the physical basis of the comparable RMSF profiles observed at different time scales (Fig. 1) and identifying the shared mechanisms that underlie the observed similarity across different time windows, we examined the principal motional modes sampled in simulations of different lengths. To this end, we decomposed the CV-N motions in each MD trajectory into a series of collective modes, each ranked by their weights. We focused on the top-ranking modes (global modes), also called essential modes, since these are usually the most collective modes and numerous applications have shown their relevance to biological function.[1]

We considered two most extreme runs: the 1 ns and 400 ns simulations. The global (lowest frequency) mode obtained from two such runs is illustrated in Fig. 5, panels A and B. Strikingly, although one might expect that the longer simulations probe more collective motions that only emerge at longer time scales, the global motional behavior is remarkably similar in the two runs. The correlation coefficient between the two modes is 0.77, suggesting that the global modes at either short or long times share robust features that are uniquely defined by the structure, and can be extracted to a good approximation from short runs.

To validate these findings, the first two modes of two 400 ns simulations were compared with the global modes extracted from all other shorter simulations. The results of this analysis are presented in Table SII. Thirty-two of all fifty-six short (≤ 100 ns) simulations yielded global motions similar to those in the first 400 ns simulation, with similarity defined as a correlation coefficient > 0.6 between the two modes. A very similar result was obtained, performing the analysis for the second 400 ns simulation. Even though not all the different length simulations in our dataset converged completely, a large fraction of them share the low frequency motions with the longest runs.

The correlation between the global mechanisms of motions sampled in runs of different durations becomes even more apparent when we combine trajectories from all individual runs with the same duration, and compare the principal modes of motions computed for each combined trajectory. The results, presented in Table II, further confirm that the directions (not the size) of the global motions are effectively conserved across runs of various lengths, although the orders of the modes (shown in parentheses) may shift in some cases. This analysis also underscores the importance of carrying out multiple simulations and subjecting the compiled data to mode decomposition in order to detect the 'consensus' global modes and extract information on collective mechanics.[27,52]

Given that the top-ranking modes of motions from long simulations can be extracted to a good approximation from short simulations (provided that multiple runs are combined), insights into biological motions of low frequencies may be gained via multiple short simulations. The explanation for such unexpected behavior may lie in the nature of the folding energy landscape. The energy space may be described in terms of an orthogonal basis set, with each basis vector defining a different mode of motion. If the global modes of motion in long and short simulations, respectively, display the same patterns, this suggests that the molecule tends to move along the same direction or to sample the same subspace, in both cases, although the amplitudes of the displacements differ.

Both ENM and NMR results are consistent with the MD simulation results

As further verification of the relevance of our findings to CV-N dynamics, we performed the GNM[21,22] analysis of the PDB structure 2EZM, the NMA[53,54] of the same structure using the ANM,[40] and the PCA of the NMR ensemble of 40 structural models for CV-N.[32] ANM modes have been observed in previous studies to correlate with the structural dynamics intrinsically accessible to enzymes[1,55,56] and with the microseconds dynamics of G-protein coupled receptors.[27] The distribution of NMR models also provides information on structural variabilities, which may be compared to those observed in MD runs.[36,57,58]

The correlations between the distribution of MSFs predicted by the GNM, $<(\Delta \boldsymbol{R_i})^2>_{\mathrm{GNM}}$, and those observed in different MD runs are presented in Table I. The correlations vary from 0.60 (with $<(\Delta \boldsymbol{R_i})^2>_{\mathrm{100ns,avg}}$) to 0.74 (with $<(\Delta \boldsymbol{R_i})^2>_{\mathrm{25ns,avg}}$). Here the subscript designates that the MSFs refer to the averages over multiple MD runs of a given duration (e.g., 12 runs of 25 ns each, or eight runs of 100 ns, etc). These results are consistent with our previous findings where correlations of 0.64 ± 0.04 were obtained[36] between GNM-predicted MSFs and the MSFs inferred from multiple 10 ns MD simulations. The results presented in Fig. 5C, Table II row 5 and Table SIII further show that the global modes predicted by the ANM correlate with the global modes derived from MD simulations, irrespective of the length of the simulation.

Tables II also displays the correlations between the principal modes of structural deviations inferred from NMR models (last row) and global modes observed in MD simulations. The correlations between the NMR principal modes and MD global modes, 0.55 ± 0.06, are not as high as those among MD runs with different lengths, 0.68 ± 0.11 (Table II). This may be attributed to the fact that there are only 40 models in the NMR ensemble, which may not provide a complete description of the accessible reconfigurations. The level of agreement appears to decrease with increasing simulation duration, presumably be due to the inadequate sampling of the accessible (larger) conformational subspace by fewer independent runs. Finally, we examined the subspace spanned by the first ix global modes extracted from MD simulations, the NMR ensemble, and ANM predictions. As observed in Table SIV, there are large overlaps between these subspaces, with an average RMSIP (root mean-square inner product)[59] of 0.59 ± 0.11. The observed subspace similarities at the low frequency region suggest the global motions observed in MD simulations and those predicted by the ANM share robust features uniquely encoded by the equilibrium structure.

The conformational dynamics usually consists of a continuous spectrum of motions, with varying frequencies and amplitudes. As such, it can hardly be divided into two distinctive groups, fast and slow. However, in the literature, for simplicity, two time regimes have been defined, sub-τ_c and supra-τ_c, to describe fast and slow motions, respectively. τ_c is the correlation time deduced from T_1/T_2 ratio measured by NMR spectroscopy. [12, 60] In the case of CV-N, the experimentally measured τ_c is 4.5 ns.[48] Therefore, the time scale of present simulations includes motions in the 'fast' regime, as well as 'slow' regime. The frequency range of slow motions varies by two orders of magnitude up to 0.4 microseconds time scale. The conclusions drawn therefore apply to this time regime. Yet, it is worth noting that the most cooperative (global) modes of internal motions derived from short and long simulations share close similarities (compare, for example, panels A and B in Fig. 5). Furthermore, these global motions exhibit reasonable agreement with the results from ANM calculations, and NMR data, which also supports the robustness of the results from simulations and the fact that these robust modes are uniquely defined by the architecture.

CONCLUSION

In the present work, we have analyzed the amplitudes and directions of residue motions in multiple MD runs of durations varying in the range 1 ns – 400 ns. The simulation conditions were identical in all runs, except for the lengths of the simulations. Our data show that the distribution of residue fluctuations, or the MSF profile, is insensitive to the simulation length, while the amplitudes increase with simulation time. The square amplitudes exhibit a power law dependence on the simulation time, while the correlation times are linearly dependent. These findings suggest that the types of motions, but not their absolute size scales, can be accurately extracted from MD runs in the observed time regime, which includes both sub- and supra-τ_c motions up to hundreds of nanoseconds. Accurate assessment of the distribution of residues, however, and extraction of the global modes that predominantly define the common features of the RMSFs observed at various time windows, require performing multiple simulations and decomposing (by PCA) the trajectories to identify consensus modes.

The present study also explains why and how simulations that sample several order of magnitude faster events may provide insights into the conformational mechanics of much slower processes. Our in-depth examination of the spectra of essential modes retrieved from the different simulations suggests that highly robust and usually functional modes that persist (or fully evolve) at longer times can be discerned even in short simulations provided that the dominant modes are extracted by a PCA of the combination of multiple trajectories.

The motions are robustly defined by the shape of the native state energy minimum, which apparently governs protein fluctuations not only in the close neighborhood but also during relatively large excursions away from the minimum. The fact that the GNM and ANM results are consistent with MD simulation results also points to the dominance of shape of the energy landscape near the native state minimum in defining the accessible routes/modes of reconfiguration.

The observations made here for CV-N are consistent with different levels of coverage of the native state energy well, shorter simulations covering the bottom only, while longer simulations reaching distant locations while remaining in the same well. CV-N's high stability at room temperature[61] made it a good candidate for performing extended simulations without the risk of significant structural changes or large conformational drift. In principle, the shape of the energy landscape would affect the observed shifts in fluctuation profiles. One might expect to see a broad range of excursions under native state conditions when that the structure's stability is mainly entropic in nature, i.e., when the structure has access to multiple microstates while maintaining its native fold. In such cases, the observed size of fluctuations would exhibit a relatively more pronounced dependence on the simulation duration. This behavior is evidenced by the higher exponent observed in loop regions that enjoy multiple conformations. It remains to be seen if the residue fluctuation profiles of proteins predominantly stabilized by enthalpic interactions (native state characterized by a deep, but not necessarily broad energy, minimum), exhibit any shifts similar to those observed here for CV-N prior to the onset of their unfolding.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health Grants 5R01LM007994 (IB), and GM080642 (AMG).

References

1. Bahar I, Lezon TR, Yang LW, Eyal E. Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys. 2010; 39:23–42. [PubMed: 20192781]
2. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature. 2007; 450:964–972. [PubMed: 18075575]
3. Smock RG, Gierasch LM. Sending signals dynamically. Science. 2009; 324:198–203. [PubMed: 19359576]
4. Cavanagh J, Venters RA. Protein dynamic studies move to a new time slot. Nat Struct Biol. 2001; 8:912–914. [PubMed: 11685230]
5. Falke JJ. Enzymology. A moving story. Science. 2002; 295:1480–1481. [PubMed: 11859184]
6. Trbovic N, Kim B, Friesner RA, Palmer AG III. Structural analysis of protein dynamics by MD simulations and NMR spin-relaxation. Proteins. 2008; 71:684–694. [PubMed: 17975832]
7. Hall JB, Fushman D. Variability of the ^{15}N chemical shielding tensors in the B3 domain of protein G from ^{15}N relaxation measurements at several fields. Implications for backbone order parameters. J Am Chem Soc. 2006; 128:7855–7870. [PubMed: 16771499]

8. Bouvignies G, Bernado P, Meier S, Cho K, Grzesiek S, Bruschweiler R, Blackledge M. Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings. Proc Natl Acad Sci U S A. 2005; 102:13885–13890. [PubMed: 16172390]

9. Briggman KB, Tolman JR. De novo determination of bond orientations and order parameters from residual dipolar couplings with high accuracy. J Am Chem Soc. 2003; 125:10164–10165. [PubMed: 12926926]

10. Lakomek NA, Walter KF, Fares C, Lange OF, de Groot BL, Grubmuller H, Bruschweiler R, Munk A, Becker S, Meiler J, Griesinger C. Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics. J Biomol NMR. 2008; 41:139–155. [PubMed: 18523727]

11. Markwick PR, Bouvignies G, Salmon L, McCammon JA, Nilges M, Blackledge M. Toward a unified representation of protein structural dynamics in solution. J Am Chem Soc. 2009; 131:16968–16975. [PubMed: 19919148]

12. Bui JM, Gsponer J, Vendruscolo M, Dobson CM. Analysis of sub-τ_c and supra-τ_c motions in protein Gβ1 using molecular dynamics simulations. Biophys J. 2009; 97:2513–2520. [PubMed: 19883594]

13. Henzler-Wildman KA, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, Pozharski E, Wilson MA, Petsko GA, Karplus M, Hubner CG, Kern D. Intrinsic motions along an enzymatic reaction trajectory. Nature. 2007; 450:838–844. [PubMed: 18026086]

14. Akke M, Palmer AG. Monitoring macromolecular motions on microsecond to millisecond time scales by $R_{1\rho}$-R_1 constant relaxation time NMR spectroscopy. J Am Chem Soc. 1996; 118:911–912.

15. Lipari G, Szabo A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. J Am Chem Soc. 1982; 104:4559–4570.

16. Schotte F, Soman J, Olson JS, Wulff M, Anfinrud PA. Picosecond time-resolved X-ray crystallography: probing protein function in real time. J Struct Biol. 2004; 147:235–246. [PubMed: 15450293]

17. Srajer V, Teng T, Ursby T, Pradervand C, Ren Z, Adachi S, Schildkamp W, Bourgeois D, Wulff M, Moffat K. Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography. Science. 1996; 274:1726–1729. [PubMed: 8939867]

18. Kolano C, Helbing J, Kozinski M, Sander W, Hamm P. Watching hydrogen-bond dynamics in a beta-turn by transient two-dimensional infrared spectroscopy. Nature. 2006; 444:469–472. [PubMed: 17122853]

19. Michalet X, Weiss S, Jager M. Single-molecule fluorescence studies of protein folding and conformational dynamics. Chem Rev. 2006; 106:1785–1813. [PubMed: 16683755]

20. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nat Struct Biol. 2002; 9:646–652. [PubMed: 12198485]

21. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des. 1997; 2:173–181. [PubMed: 9218955]

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

23. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins. 1998; 33:417–429. [PubMed: 9829700]

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

25. Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips GN Jr. A sampling problem in molecular dynamics simulations of macromolecules. Proc Natl Acad Sci U S A. 1995; 92:3288–3292. [PubMed: 7724554]

26. Caves LS, Evanseck JD, Karplus M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. Protein Sci. 1998; 7:649–666. [PubMed: 9541397]

27. Romo TD, Grossfield A. Validating and improving elastic network models with molecular dynamics simulations. Proteins. 2011; 79:23–34. [PubMed: 20872850]

28. Smith LJ, Daura X, van Gunsteren WF. Assessing equilibration and convergence in biomolecular simulations. Proteins. 2002; 48:487–496. [PubMed: 12112673]

29. Huang X, Bowman GR, Bacallado S, Pande VS. Rapid equilibrium sampling initiated from nonequilibrium data. Proc Natl Acad Sci U S A. 2009; 106:19765–19769. [PubMed: 19805023]

30. Morcos F, Chatterjee S, McClendon CL, Brenner PR, Lopez-Rendon R, Zintsmaster J, Ercsey-Ravasz M, Sweet CR, Jacobson MP, Peng JW, Izaguirre JA. Modeling conformational ensembles of slow functional motions in Pin1-WW. PLoS Comput Biol. 2010; 6:e1001015. [PubMed: 21152000]

31. Fersht AR. On the simulation of protein folding by short time scale molecular dynamics and distributed computing. Proc Natl Acad Sci U S A. 2002; 99:14122–14125. [PubMed: 12388785]

32. Bewley CA, Gustafson KR, Boyd MR, Covell DG, Bax A, Clore GM, Gronenborn AM. Solution structure of cyanovirin-N, a potent HIV-inactivating protein. Nat Struct Biol. 1998; 5:571–578. [PubMed: 9665171]

33. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal component analysis and long time protein dynamics. J Phys Chem. 1996; 100:2567–2572.

34. Cote Y, Senet P, Delarue P, Maisuradze GG, Scheraga HA. Nonexponential decay of internal rotational correlation functions of native proteins and self-similar structural fluctuations. Proc Natl Acad Sci U S A. 2010; 107:19844–19849. [PubMed: 21045133]

35. Senet P, Maisuradze GG, Foulie C, Delarue P, Scheraga HA. How main-chains of proteins explore the free-energy landscape in native states. Proc Natl Acad Sci U S A. 2008; 105:19708–19713. [PubMed: 19073932]

36. Liu L, Koharudin LM, Gronenborn AM, Bahar I. A comparative analysis of the equilibrium dynamics of a designed protein inferred from NMR, X-ray, and computations. Proteins. 2009; 77:927–939. [PubMed: 19688820]

37. de Souza ON, Ornstein RL. Effect of periodic box size on aqueous molecular dynamics simulation of a DNA dodecamer with particle-mesh Ewald method. Biophys J. 1997; 72:2395–2397. [PubMed: 9168016]

38. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem. 2005; 26:1781–1802. [PubMed: 16222654]

39. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B. 1998; 102:3586–3616. [PubMed: 24889800]

40. 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]

41. Eyal E, Chennubhotla C, Yang LW, Bahar I. Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. Bioinformatics. 2007; 23:i175–i184. [PubMed: 17646294]

42. Barrientos LG, Gronenborn AM. The highly specific carbohydrate-binding protein cyanovirin-N: structure, anti-HIV/Ebola activity and possibilities for therapy. Mini Rev Med Chem. 2005; 5:21–31. [PubMed: 15638789]

43. Matei E, Furey W, Gronenborn AM. Solution and crystal structures of a sugar binding site mutant of cyanovirin-N: no evidence of domain swapping. Structure. 2008; 16:1183–1194. [PubMed: 18682220]

44. Matei E, Zheng A, Furey W, Rose J, Aiken C, Gronenborn AM. Anti-HIV activity of defective cyanovirin-N mutants is restored by dimerization. J Biol Chem. 2010; 285:13057–13065. [PubMed: 20147291]

45. Sandstrom C, Hakkarainen B, Matei E, Glinchert A, Lahmann M, Oscarson S, Kenne L, Gronenborn AM. Atomic mapping of the sugar interactions in one-site and two-site mutants of cyanovirin-N by NMR spectroscopy. Biochemistry. 2008; 47:3625–3635. [PubMed: 18311923]

46. Shenoy SR, Barrientos LG, Ratner DM, O'Keefe BR, Seeberger PH, Gronenborn AM, Boyd MR. Multisite and multivalent binding between cyanovirin-N and branched oligomannosides: calorimetric and NMR characterization. Chem Biol. 2002; 9:1109–1118. [PubMed: 12401495]

47. Barrientos LG, Matei E, Lasala F, Delgado R, Gronenborn AM. Dissecting carbohydrate-Cyanovirin-N binding by structure-guided mutagenesis: functional implications for viral entry inhibition. Protein Eng Des Sel. 2006; 19:525–535. [PubMed: 17012344]

48. Kelley BS, Chang LC, Bewley CA. Engineering an obligate domain-swapped dimer of cyanovirin-N with enhanced anti-HIV activity. J Am Chem Soc. 2002; 124:3210–3211. [PubMed: 11916396]

49. Li DW, Showalter SA, Bruschweiler R. Entropy localization in proteins. J Phys Chem B. 2010; 114:16036–16044. [PubMed: 21077640]

50. Luo G, Andricioaei I, Xie XS, Karplus M. Dynamic distance disorder in proteins is caused by trapping. J Phys Chem B. 2006; 110:9363–9367. [PubMed: 16686476]

51. Wong IY, Gardel ML, Reichman DR, Weeks ER, Valentine MT, Bausch AR, Weitz DA. Anomalous diffusion probes microstructure dynamics of entangled F-actin networks. Phys Rev Lett. 2004; 92:178101. [PubMed: 15169197]

52. Roy J, Laughton CA. Long-timescale molecular-dynamics simulations of the major urinary protein provide atomistic interpretations of the unusual thermodynamics of ligand binding. Biophys J. 2010; 99:218–226. [PubMed: 20655850]

53. Cui, Q.; Bahar, I. Normal mode analysis: theory and applications to biological and chemical systems. London: CRC Press; 2006.

54. Kitao A, Go N. Investigating protein dynamics in collective coordinate space. Curr Opin Struct Biol. 1999; 9:164–169. [PubMed: 10322205]

55. Bakan A, Bahar I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. Proc Natl Acad Sci U S A. 2009; 106:14349–14354. [PubMed: 19706521]

56. May A, Zacharias M. Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. J Med Chem. 2008; 51:3499–3506. [PubMed: 18517186]

57. Abseher R, Horstink L, Hilbers CW, Nilges M. Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. Proteins. 1998; 31:370–382. [PubMed: 9626697]

58. Yang LW, Eyal E, Chennubhotla C, Jee J, Gronenborn AM, Bahar I. Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. Structure. 2007; 15:741–749. [PubMed: 17562320]

59. Amadei A, Ceruso MA, Di NA. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. Proteins. 1999; 36:419–424. [PubMed: 10450083]

60. Lange OF, Lakomek NA, Fares C, Schroder GF, Walter KF, Becker S, Meiler J, Grubmuller H, Griesinger C, de Groot BL. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science. 2008; 320:1471–1475. [PubMed: 18556554]

61. Barrientos LG, Lasala F, Delgado R, Sanchez A, Gronenborn AM. Flipping the switch from monomeric to dimeric CV-N has little effect on antiviral activity. Structure. 2004; 12:1799–1807. [PubMed: 15458629]

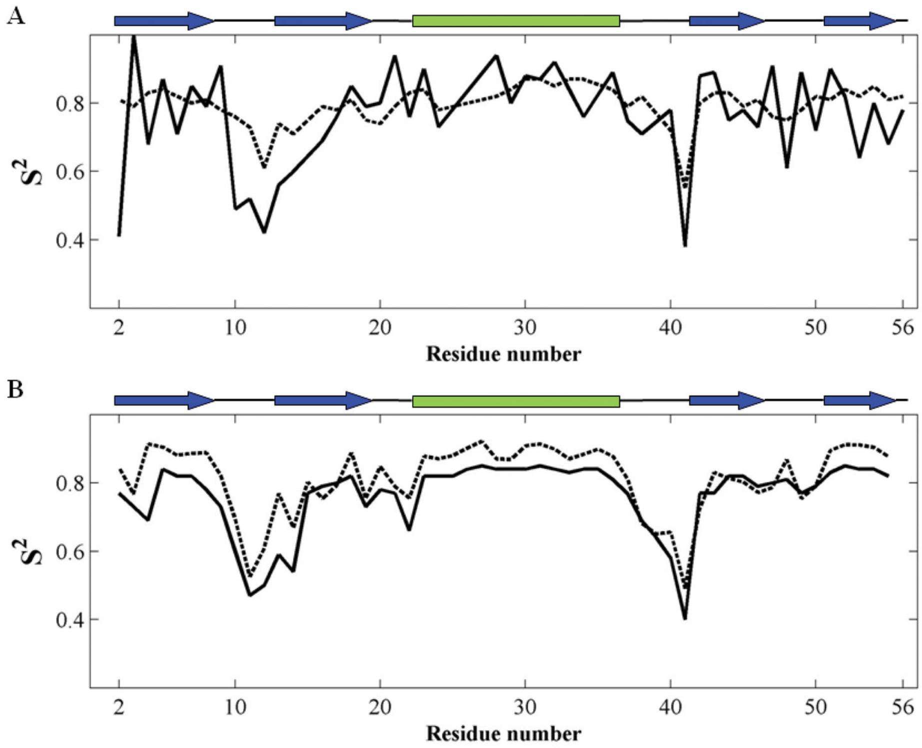

Figure 1.

Experimental and computational literature data exhibit similar motional behavior for short and long times. (**A**) Order parameters S^2 of G protein B 3 extracted from NMR data: spin-relaxation,[7] dashed black; and RDC,[8] solid black. (**B**) Order parameters S^2 of GB1 extracted from MD trajectories[12] of 10 ns (dashed black) and 175 ns (solid black). Secondary structure elements are depicted at the top of each panel.

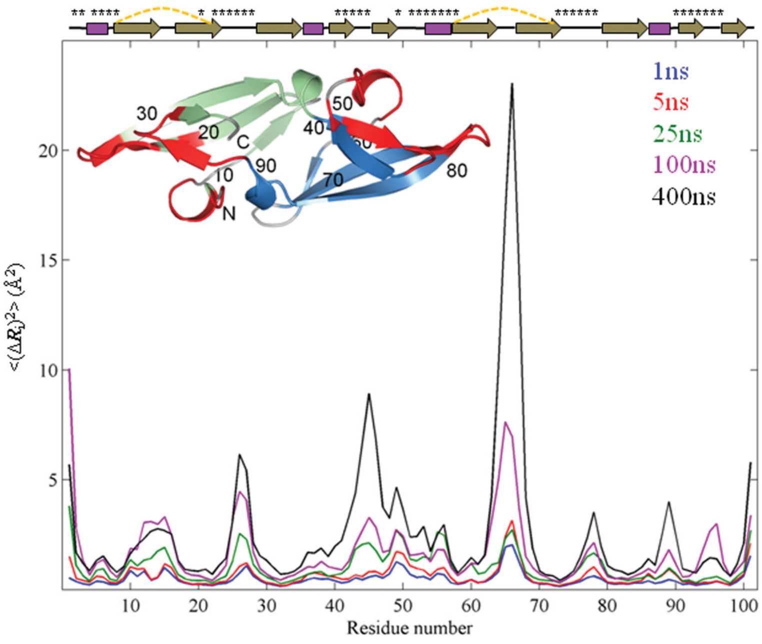

Figure 2.

Mean-square-fluctuation profiles of CV-N from simulations with different durations. The MSFs $<(\Delta \boldsymbol{R}_i)^2>$ of residues averaged over twenty independent 1 ns, sixteen 5 ns, twelve 25 ns, eight 100 ns and two 400 ns runs are shown in blue, red, green, magenta, and black, respectively. Secondary structure elements of the protein are depicted at the top with disulfide bonds represented by dashed yellow lines and residues in the sugar binding sites labeled by asterisks. The inset shows the CV-N structure in ribbon representation. Domains A and B are colored green and blue, respectively, and the two sugar binding sites are colored red. Amino acid sequence positions are labeled for every 10th residue.

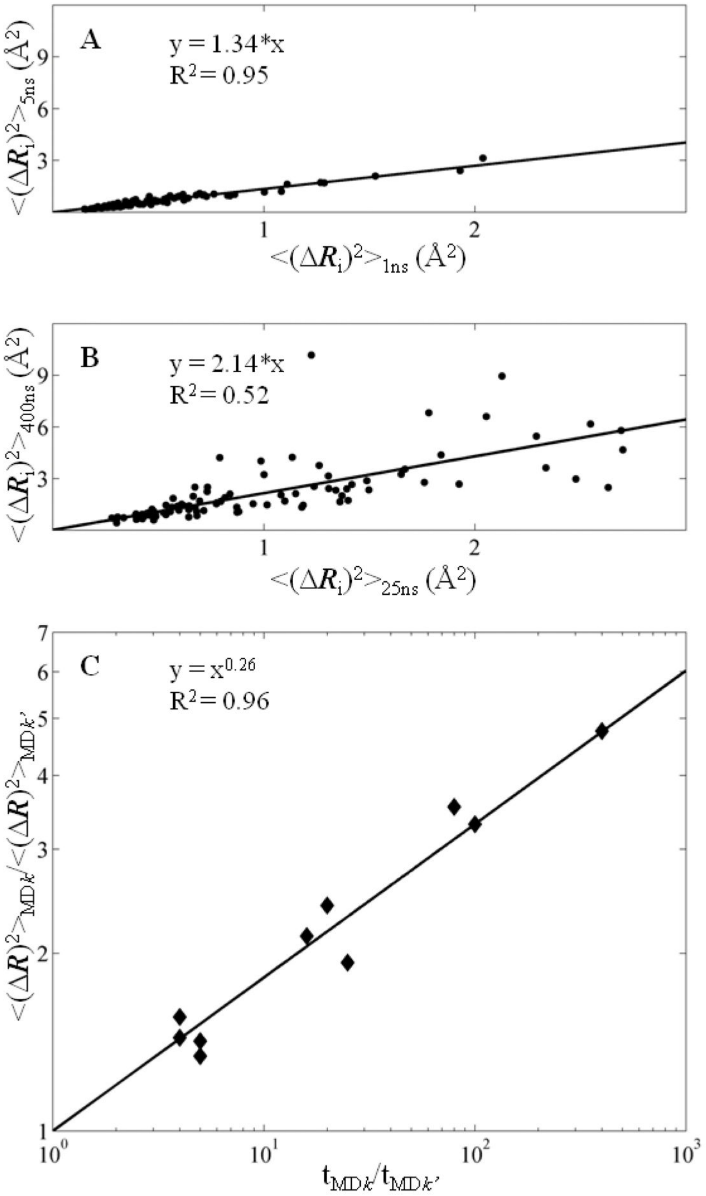

Figure 3.

The magnitude of the fluctuations increases with increasing simulation time. (**A**) and (**B**) Comparison of MSFs for different simulations. (**A**) $\langle(\Delta \boldsymbol{R}_i)^2\rangle$ of residue *i* in the 5 ns simulation (y axis) is plotted against $\langle(\Delta \boldsymbol{R}_i)^2\rangle$ of the same residue in the 1 ns simulation (x axis). (**B**) $\langle(\Delta \boldsymbol{R}_i)^2\rangle$ of residue *i* in the 400 ns simulation (y axis) versus $\langle(\Delta \boldsymbol{R}_i)^2\rangle$ of the same residue in the 25 ns simulation (x axis). (**C**) The relationship between MSF and simulation time is a power function, with exponent 0.26. The MSF scaling factors for different simulations are plotted against the corresponding ratios of simulation lengths.

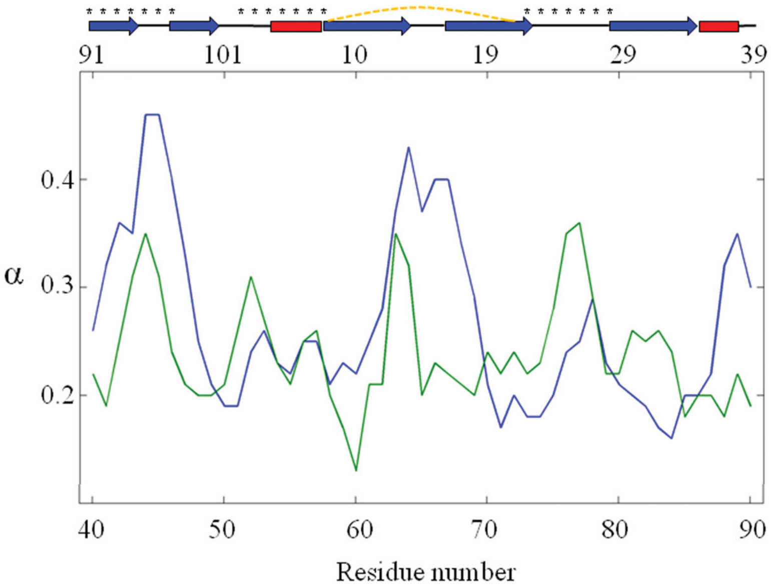

Figure 4.

Power law exponents for CV-N residues. The results are shown on domain A (green), and domain B (blue). The upper abscissa displays residue positions in domain A, and the lower abscissa, the residue positions in domain B. The secondary structures with disulfide bonds (dashed yellow lines) are represented on the top, and residues comprising the binding sites are labeled by asterisks.

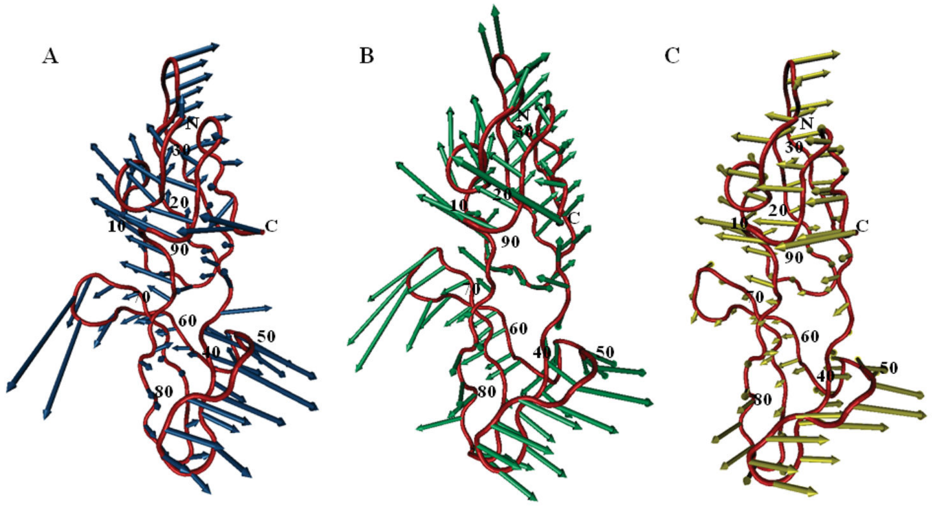

Figure 5.

Shared global mode between theory and simulations. The CV-N backbone structure is shown in tube representation (red) with the directions of the global motion for the 1 ns simulation (**A**) and the 400 ns simulation (**B**), or the second mode predicted by the ANM (**C**) depicted by blue, green, and yellow arrows, respectively. The correlation coefficients between pairs of modes displayed are 0.77 (blue/green), 0.69 (blue/yellow), and 0.64 (green/ yellow). Primary sequence positions are labeled for every 10th residue.

Table I

Correlation Coefficients Between the MSFs of CV-N Residues Observed in MD Simulations[a] and Those Predicted by the GNM

	$1ns_{avg}$	**$5ns_{avg}$**	**$25ns_{avg}$**	**$100ns_{avg}$**	**$400ns_{avg}$**
$5ns_{avg}$	0.96				
$25ns_{avg}$	0.76	0.80			
$100ns_{avg}$	0.71	0.76	0.79		
$400ns_{avg}$	0.83	0.83	0.63	0.77	
GNM	0.71	0.70	0.74	0.60	0.67

[a] Averages over multiple runs (see the text).

Table II

Correlation Coefficients Between the Global Modes from MD Simulations,[a] NMR Structural Ensemble, and ANM Predictions

	1ns$_{comb}$	**5ns$_{comb}$**	**25ns$_{comb}$**	**100ns$_{comb}$**	**400ns$_{comb}$**
5ns$_{comb}$	0.80 (1,1)				
25ns$_{comb}$	0.75 (2,1)	0.64 (2,1)			
100ns$_{comb}$	0.58 (5,3)	0.57 (2,1)	0.84 (1,1)		
400ns$_{comb}$	0.59 (1,4)	0.67 (1,2)	0.80 (4,4)	0.59 (1,3)	
ANM	0.57 (1,1)	0.60 (2,2)	0.61 (2,3)	0.60 (2,3)	0.58 (1,4)
NMR	0.60 (2,1)	0.63 (2,1)	0.57 (2,2)	0.56 (2,2)	0.47 (2,2)

[a] The MD global modes refer to the combination of multiple trajectories of a given simulation length. Entries in parentheses represent the mode numbers, e.g. the 2nd mode of the combined 25 ns simulations (total of 12 runs) displays a correlation coefficient of 0.75 with the 1st mode of the combined 1 ns simulations (20 runs).