CONFORMATIONAL DYNAMICS OF PROTEINS: INSIGHTS FROM STRUCTURAL AND COMPUTATIONAL STUDIES

By

Lin Liu

BS in Biological Science, University of Science and Technology of China, 2004

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This thesis was presented

by

Lin Liu

It was defended on

September 26th, 2011

and approved by

Dr. Gordon S. Rule, Professor, Department of Biological Sciences, CMU

Dr. Ronald Wetzel, Professor, Department of Structural Biology

Dr. Daniel M. Zuckerman, Associate Professor, Department of Computational and Systems Biology

Thesis Advisors: Dr. Ivet Bahar, Professor, Department of Computational and Systems Biology

and Dr. Angela M. Gronenborn, Professor, Department of Structural Biology

CONFORMATIONAL DYNAMICS OF PROTEINS:

INSIGHTS FROM STRUCTURAL AND COMPUTATIONAL STUDIES

Lin Liu, PhD

University of Pittsburgh, 2011

Proteins are not static; they undergo both random thermal fluctuations near a given equilibrium state, and transitions between different sub-states. These motions are usually intricately connected to the function of the protein. Therefore, understanding the dynamics of proteins is important to gain insights into the mechanisms of many biological phenomena. Only the combination of structure and dynamics does allow for describing a functional protein (or biological molecule) properly. Therefore, this thesis is centered on computational and structural studies of protein dynamics. I carried out full atomic simulations and coarse-grained analyses (using elastic network models) as computational approaches, and used NMR as well as X-ray crystallography on the experimental side. With regard to the understanding of the fluctuations accessible under equilibrium conditions, a detailed analysis of high-resolution structural data and computationally predicted dynamics was carried out for a designed sugar-binding protein. The mean-square deviations in the positions of residues derived from NMR models and those inferred from X-ray crystallographic B-factors for two different crystal forms were compared with the predictions based on the Gaussian network model (GNM) and the results from molecular dynamics (MD) simulations. The results highlighted the significance of considering ensembles of structures (or structural models) from experiments, in order to make an accurate

assessment of the fluctuation dynamics of proteins under equilibrium conditions. Moreover, we analyzed the amplitudes, correlation times, and directions of residue motions in multiple MD runs of durations varying in the range 1 ns – 400 ns. Our data show that the distribution of residue fluctuations is insensitive to the simulation length, while the amplitudes increase with simulation time with a power law. Another area of interest concerned the phenomenon of "domain swapping". We investigated the molecular basis of this unusual multimerization, using a broad range of approaches. A systematic analysis of a large set of domain-swapped structures was performed to this aim. Results suggest that almost any protein may be capable of undergoing domain swapping, and that domain swapping is solely a specialized form of oligomer assembly but is closely associated with the unfolding/folding process of proteins. We also use experimental ^{19}F-NMR to study the thermodynamic and kinetic properties in CV-N domain swapping. The activation energy barrier for the passage between monomeric anddomain-swapped dimeric form is of similar magnitude to that for complete unfolding of the protein, indicating that the overall unfolding of the polypeptide is required for domain swapping. Crystal structures of a domain-swapped trimer and a tetramer of CV-N provide further insights into the potential mechanics of CV-N domain swapping.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I gratefully appreciate my advisors Dr. Ivet Bahar and Dr. Angela Gronenborn for their directions throughout my graduate studies. What I've learned from them is not only the knowledge about computational and structural biology, but also the attitude of a good scientist. Under their guidance, I have worked on several research projects complied in the thesis.

I sincerely thank my dissertation committee members, Dr. Gordon Rule, Dr. Ronald Wetzel, and Dr. Daniel Zuckerman for the constructive suggestions and stimulating comments and discussions.

I thank the department of Structural Biology and the department of Computational and Systems Biology. The academic environment is rich and free, and people around are nice and helpful. I especially thank Dr. Leonardus Koharudin. Working with him is a pleasant experience, and his enthusiasm always inspirits me. I thank other lab members in Angela's group and Ivet's group. I extend my thanks to friends in Pittsburgh.

Finally, I thank the members of my family: my father Guorui Liu, my mother Xiuai Kou, and my husband Wei Wang for their love and support.

1.0 INTRODUCTION

1.1 CONFORMATIONAL DYNAMICS

A general way to investigate biological phenomena is to study an individual component from a living organism, such as a protein, the major constituent of cells. Proteins are polymers of covalently linked amino acids, with the amino acid sequence characteristic of each protein. The spatial arrangement of atoms in a protein is called its conformation. The most stable conformation under physiological conditions, known as the native state, is encoded by the protein's amino acid sequence, and is highly related to the protein's function. In a strict sense, the native state is an ensemble of fluctuating conformations, or microstates, narrowly distributed around a global energy minimum. At each instantaneous conformation, the interactions responsible for maintaining the arrangement of atoms in the neighborhood of the native energy minimum originate from various physicochemical effects: hydrophobic contacts, hydrogen bond formation, electrostatic interactions, disulfide bridges, and so on. Therefore, the protein is not static; it undergoes both thermal fluctuations near its equilibrium state and occasional transitions between sub-states, and thus samples multiple conformations. The conformational dynamics of the protein or the ability to sample various conformations usually assists in its chemical or biological activities (e.g. interacting with different substrates).[1] Therefore, it is necessary to examine the dynamics of a protein in addition to its static structure in order to gain a better understanding of its mechanisms of activities.

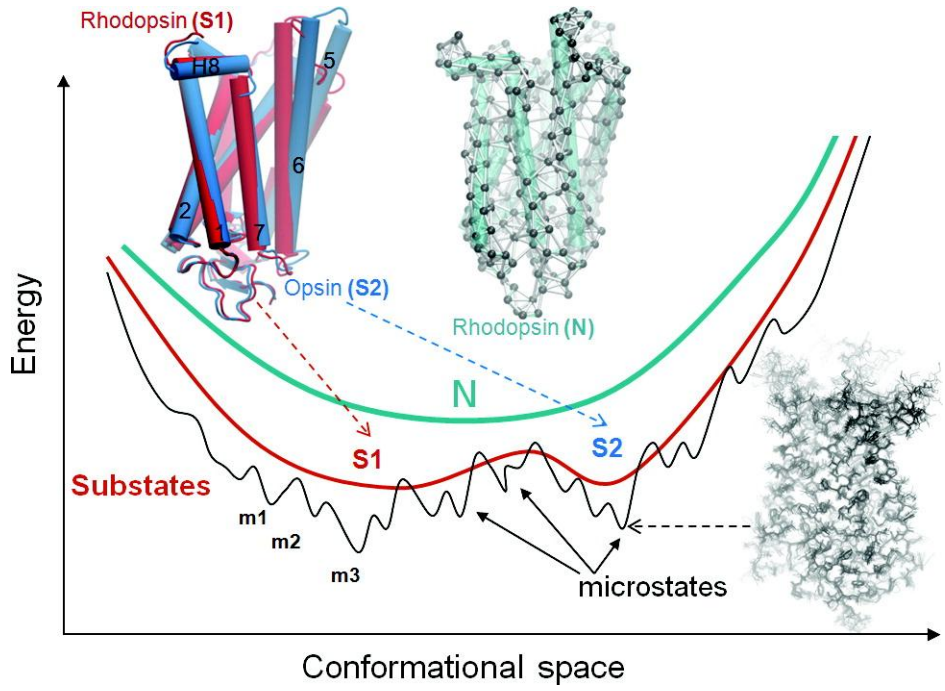

Figure 1.1 Schematic diagram of energy profile near native state conditions, modeled at different resolutions.

N denotes the native state, modeled at a coarse-grained scale as a single energy minimum. A more detailed examination of the structure and energetics reveals two or more sub-states (S1, S2, etc.), which in turn contain multiple microstates (m1, m2, etc.). Structural models corresponding to different hierarchical levels of resolution are shown: an elastic network model representation where the global energy minimum on a coarse-grained scale (N) is approximated by a harmonic potential along each mode direction; two sub-states S1 and S2 sampled by global motions near native state conditions; and an ensemble of conformers sampled by small fluctuations in the neighborhood of each substate. The diagrams have been constructed using the following rhodopsin structures deposited in the Protein Data Bank: 1U19 (N); 1U19 and 3CAP (S1 and S2); and 1F88, 1GZM, 1HZX, 1L9H, 1U19, 2G87, 2HPY, 2I35, 2I36, 2I37, 2J4Y, 2PED, 3C9L, and 3C9M (microstates). Figure is adopted from Bahar et al. *Chem. Rev.*, 2010, 110: 1463-1479.

Although the conformational space is vast, a folded protein is often confined to a significantly narrower distribution of conformations in the close neighborhood of its native state, compared to disordered polymers. It is possible to view these conformations as different sub-states (on a more global scale) or different microstates (at a higher resolution). Microstates

usually share the overall 'fold' and regular secondary structure, with variations in bond lengths, bond angles, dihedral angles, loop conformations, substructure packing, or even entire domain or subunit positions and orientations. Importantly, there is a dynamic equilibrium among these microstates, allowing for their continual interconversions and maintaining their probability distribution,[2] which could be altered by a change in the system (e.g., ligand binding or changes in external conditions).[3] Figure 1.1 illustrates the different hierarchical levels of structures, from native 'state', to sub-states, to microstates that coexist in a dynamic equilibrium.[2] It is clear that transitions between two or more microstates may be treated as the thermal motions around one state. 'Equilibrium motions' of a folded protein are referred to as all types of motions, including fluctuations between microstates or passages between sub-states, that are achieved while maintaining the fold and navigating within the global energy minimum corresponding to the native state.

1.2 STRUCTURAL AND COMPUTATIONAL METHODS

Structures deposited in Protein Data Bank (PDB)[12] have increased rapidly from 695 in 1991 to about 75,000 in 2011,[13] benefitting from the developments in multi-dimensional NMR analysis,[14] restrained refinement of structural models,[15] automated multiple wavelength anomalous diffraction (MAD) and multiple isomorphous replacement (MIR).[16] 99% structures deposited in the PDB are solved by one of the two classical methods: NMR spectroscopy and X-ray crystallography, indicating their dominant and important position in structural biology. Sometimes, one protein has more than one resolved structure, indicating its dynamic intermediates such as crystal structures of the cytochrome P450,[17] or structures resolved in the presence of different substrates /inhibitors, or under different conditions.

Protein dynamics became a major topic of investigation in many recent studies. A broad range of experimental techniques provides information on protein dynamics, including NMR relaxation measurements,[18, 19] Laue X-ray diffraction data,[20, 21] infrared and fluorescence spectroscopy,[22] and single-molecule studies,[23] although they inform about different aspects and time scales of protein dynamics. On the computational side, structure-based methods such as molecular dynamics (MD) simulations[24] and normal mode analysis (NMA) with elastic network models (ENMs)[25-28] have been broadly exploited in recent years, so as to gain insights into biomolecular systems dynamics at multiple scales. For example, the cyclophilin A catalysis dynamics has been investigated by NMR relaxation experiments;[29] its substrate binding dynamics has been observed by single-molecule FRET as well as MD simulations.[7] Many studies focus on principal components analysis (PCA)[30] of biomolecular experimental structures or simulation models, in order to extract information on dominant patterns, or cooperative events. One example is the recent ensemble study about ubiquitin,[8] whose conformational space built based on residual dipolar coupling measurements has been shown to share similarities with the conformational space deduced from PCA of different ubiquitin crystal complexes (resolved with different substrates). More details about structural and computational methods will be presented in the following chapters.

1.3 DOMAIN SWAPPING

Four levels of organization are usually used for describing protein structures: primary, secondary, tertiary, and quaternary structures. Primary structure is the description of all covalent bonds linking the consecutive amino acid residues in a polypeptide chain, and as such they essentially provide information on a one-dimensional sequence space; secondary structure refers

to the local, particularly stable, arrangements of residues forming structural patterns; tertiary structure refers to all aspects of the three-dimensional folding of a polypeptide such as the packing of secondary structural elements and their topological features; and quaternary structure describes the arrangement of two or more polypeptide subunits in space.

A small but growing subset 'domain-swapped oligomers', as originally coined by Eisenberg,[31] have received more and more attention in recent years, as a special type of quaternary structure. True domain-swapped structures require that both, monomeric and oligomeric states must be observed for the protein.[32] However, this stringent designation is not always adhered to in the literature. Sometimes, structures are called domain-swapped, even if no structure of the closed monomer has ever been observed or where only a homolog exhibits a closed monomer. In the first case, the protein is a 'candidate' for domain swapping, while in the second, the oligomers are classified as 'quasi-domain-swapped'.

In true domain-swapped structures, the exchanged subunit or domain in the oligomer is identical to the one in the corresponding monomer, exhibiting no differences in the ϕ, ψ dihedral angles on the backbone, except for the region that links the exchanging domains. This region is called the 'hinge-loop' and often adopts an extended conformation in the domain-swapped oligomer while it folds back on itself in the monomer. Although called 'domain swapping', the term 'domain' encompasses a variety of structural units: the largest may be an independently folded domain, while the smallest can be single secondary structure elements, such as a single β-strand or an isolated α-helix. The inter-molecular interfaces in the oligomer that possess identical intra-molecular counterparts in the monomer form are called the 'closed' or 'primary' interface while the newly created contact surfaces constitute the 'open' or 'secondary' interface. A

schematic representation of different domain swapping scenarios as well as the delineation of the different structural interfaces is provided in Figure 1.2.

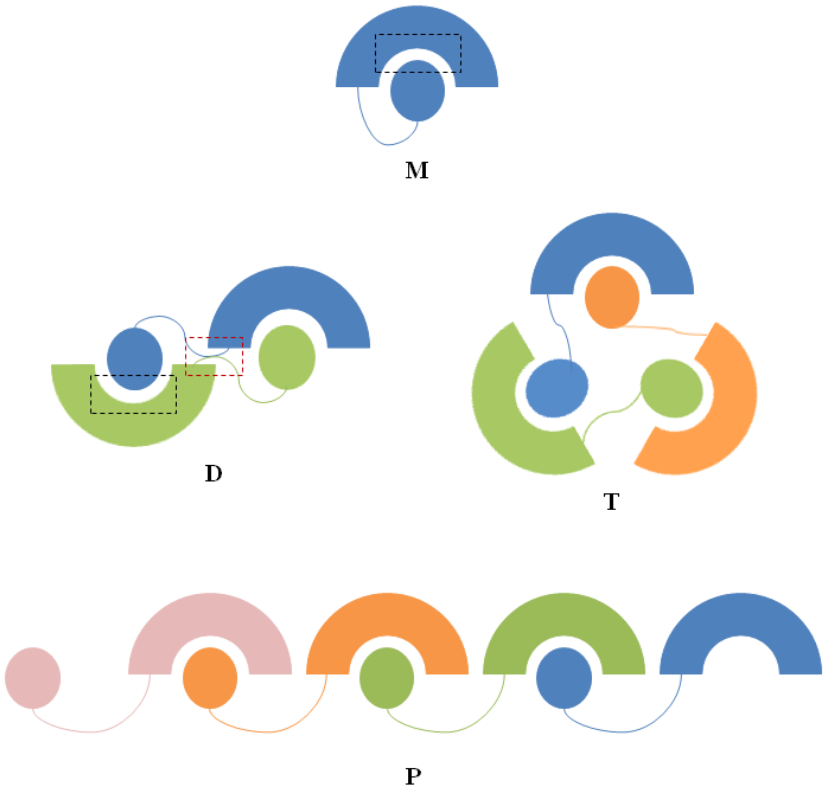

Figure 1.2 Schematic representation of domain-swapped structures and their pertinent features.

M, monomer; D, dimer; T, trimer; P, daisy chain-type multimer. Closed and open interfaces are boxed-in by black and red squares, respectively.

In this thesis, we consider mainly those proteins that contain swapped elements in their multimeric forms and for which a monomeric structure is seen for a mutant or close relative. We focus in particular on cyanovirin-N (CV-N),[33] a well-characterized protein with domain swapping abilities.

1.4 THE GOAL AND SPECIFIC SUBPROJECTS

Protein motions are usually intricately connected to the function of the protein. Therefore, understanding the dynamics of proteins is important to gain insights into the mechanisms of many biological phenomena. Only the combination of structure and dynamics does allow for describing a functional protein (or biological molecule) properly. For these reasons, I combined experimental and computational approaches in my work. My thesis is centered on computational and experimental studies of protein dynamics. I carried out full atomic (MD) simulations and coarse-grained analyses (using ENMs) as computational approaches, and used NMR spectroscopy as well as X-ray crystallography for dynamic study and structure determination on the experimental side.

With regard to the study of proteins' equilibrium dynamics (i.e., the fluctuations accessible under equilibrium conditions), I carried out the following two specific investigations reported in Chapters 2 and 3, respectively:

- A comparative analysis of the equilibrium dynamics of a designed protein inferred from NMR, X-ray, and computational studies.
- Extensive MD simulations of a CV-N to demonstrate that full atomic simulations provide insights into the mechanics, but not the time scales, of protein motions under equilibrium conditions.

Another area of investigation within the scope of my dissertation studies has been the phenomenon of "domain swapping". We investigated the molecular basis of this unusual multimerization, using a broad range of approaches. A systematic analysis of a large set of domain-swapped structures was performed to this aim, along with experimental studies of the

folding thermodynamics and structural properties of CV-N. The results reported in the respective Chapters 4 and 5 therefore include:

- Bioinformatics analysis of domain-swapped proteins.
- Elucidation of domain swapping thermodynamics with a ^{19}F-NMR study of CV-N, to show that domain swapping proceeds via complete unfolding.

Overall, both equilibrium and transition dynamics of proteins were studied in my thesis, using multiple biophysical and computational approaches. Moreover, two recently solved crystal structures of CV-N domain-swapped oligomers in my recent study enlighten our understanding about domain swapping. These results show that computational and experimental methods yield complementary results and are ideally used in combination for evaluating protein dynamics and gaining insights into the molecular basis of observed phenomena.

2.0 A COMPARATIVE ANALYSIS OF THE EQUILIBRIUM DYNAMICS OF A DESIGNED PROTEIN INFERRED FROM NMR, X-RAY AND COMPUTATIONAL STUDIES

The results presented in this chapter have been published in *Proteins*, 2009, 77: 927-39. Detailed analyses of high-resolution structural data and computationally predicted dynamics were carried out in this study for a designed sugar binding protein, LKAMG. The mean-square-deviations in the positions of residues derived from NMR models, and those inferred from X-ray crystallographic B-factors for two different crystal forms were compared with the predictions based on the Gaussian Network Model (GNM), and the results from MD simulations. The GNM systematically yielded a higher correlation than MD, with experimental data, suggesting that the lack of atomistic details in the coarse-grained GNM is more than compensated for by the mathematically exact evaluation of fluctuations using the native contacts topology. Evidence is provided that particular loop motions are curtailed by intermolecular contacts in the crystal environment causing a discrepancy between theory and experiments. Interestingly, the information conveyed by X-ray crystallography becomes more consistent with NMR models and computational predictions when ensembles of X-ray models are considered. Less precise (broadly distributed) ensembles indeed appear to describe the accessible conformational space under native state conditions better than B-factors. Our results highlight the importance of utilizing multiple conformations obtained by alternative experimental methods, and analyzing

results from both coarse-grained models and atomic simulations, for accurate assessment of motions accessible to proteins under native state conditions.

2.1 INTRODUCTION

Understanding structure and dynamics is essential for elucidating protein function that is governed by the complement of accessible energetically favored motions as seen for ligand/substrate binding in catalysis and protein-protein interactions in signaling and regulation.

It has long been appreciated that native proteins are not confined to a single, static conformation, but sample numerous sub-states under equilibrium conditions.[34-36] Similarly, the denatured state also consists of an ensemble of conformations. The main difference between the two states is simply that the native ensemble is narrow, confined to fluctuating conformations that maintain the native fold, whereas the denatured ensemble consists of a wide range of conformations. Both experiments and computations indicate that ensemble-based approaches provide superior information on the properties of a given molecule and the advantages of ensemble-based approaches have been demonstrated for NMR[37] and X-ray structure refinement.[38] Novel methods that simultaneously and synergistically determine structure and dynamics, called dynamic ensemble refinement,[8, 39] hold great promise for providing insight into equilibrium dynamics.

Focused efforts in developing and interpreting relaxation measurements, primarily by NMR spectroscopy, provide increased understanding of the temporal and spatial scales that are associated with the broad range of protein motions. Small-scale ($\leq$ 1.5 Å) motions, such as the small fluctuations in the positions of backbone and side chain atoms occur on femto- to picosecond time scales. These are accessible via NMR Lipari-Szabo order parameters (S^2)[19] or

short (< 1 ns) molecular dynamics (MD) simulations.[24] This fast motional regime is also reflected in the X-ray crystallographic temperature factors[40, 41] or can be studied using infrared or fluorescence correlation spectroscopy.[22] Mid-scale motions that take place over hundreds of pico- to nanoseconds or low microseconds may comprise loop or terminal-end fluctuations as well as peptide plane motions (change in dihedral angles) and other local dynamics information. This regime can be also be extracted via NMR Lipari-Szabo order parameters (S^2) as long as these motions are faster than the overall correlation time (τ_c). Computationally, this regime may be probed by performing long (10-100 ns) MD simulations.[42] This mid-scale range has also been evoked to contribute to the spread of conformers in NMR ensembles[43] or may be accessible from collections of X-ray structures of the same protein in different crystal isomorphs.[44] Slow motions are most frequently associated with large displacement (> 15 Å) of entire secondary structure elements, domains or subunits. If these occur on the micro- to millisecond timescale, they can be detected in the T_2 or $T_{1\rho}$ Carr–Purcell–Meiboom–Gill (*CPMG*) type NMR relaxation experiments.[18] Such motions may also be been studied in the crystal by Laue diffraction.[21] On the computational side, this regime is beyond the range accessible by MD.

To overcome the limitations of MD simulations and predict the mechanisms of low frequency, or '*global'*, modes of motion, coarse-grained models and methods based on inter-residue contact topology have been proposed, such as the elastic network models (ENMs) introduced a decade ago.[26-28, 45] ENMs have been broadly used in normal mode analysis (NMA) of known structures,[46] and shown to yield results that correlate with those from principal component analysis of ensembles of structures.[47] Such large scale movements were evoked in a recent study of ubiquitin where an ensemble of conformations based on residual dipolar couplings was determined.[8] The ensemble covered a conformational space similar to that seen

for the X-ray structures of ubiquitin complexed with different substrates, and were consistent with structural changes along a well-defined principal direction of motion.[8]

ENMs have gained widespread use given their simplicity and ability to yield a unique, analytical solution for low frequency motions (e.g., cooperative domain movements), without requiring knowledge of detailed force fields or implementation of expensive energy minimization algorithms.[48, 49] Notably, *global modes* are insensitive to details of force field parameters or specific interactions at the atomic scale.[50, 51] They are uniquely defined by the native contact topology for a particular structure, and provide insights into the potentially functional motions intrinsically favored by the proteins' native structure.[5]

We previously investigated the correlation between (i) the mean-square (ms) deviations (MSDs) in atomic coordinates for NMR ensembles, (ii) the B-factors observed in X-ray crystallographic structures, and (iii) the equilibrium fluctuations in residue positions predicted by a simple ENM, the Gaussian Network Model (GNM),[26, 45] for a large set of proteins structurally characterized by both techniques.[52] GNM results exhibited then a better correlation with the NMR data than with X-ray data.[52] We suggested that the superior correlation with NMR data may arise from the larger spectrum of modes accessible in solution, which may be represented by the NMR ensemble, as opposed to the crystalline environment where the largest amplitude modes of motion may be suppressed by crystal contacts. Another study by Phillips and coworkers[53] demonstrated that the GNM results for B-factors outperform those predicted by models that attribute the observed mobilities exclusively to rigid-body motions.[54] More recent applications suggest that the ENM methodology provides a reasonable estimate of the anisotropic displacement parameters[55, 56] and can assist in the structural refinement of supramolecular complexes.[57]

Despite these practical successes there still remain a number of uncertainties about the origin of the agreement between the GNM results and experimental ensembles. In principle, the GNM exclusively depends on inter-residue contact topology. Thus, the results for a given protein are uniquely determined, irrespective of the experimental conditions. On the other hand, different crystal packing arrangements may result in disparaging B-factors for the same protein crystallized under varying conditions. Song and Jernigan pointed out that *selected* modes may be favored or suppressed, depending on different crystal packing geometries,[58] and Phillips and coworkers noted that crystal packing selects conformers from the ensemble of structures accessible in solution.[59] Furthermore, B-factors may contain contributions from rigid-body rotations of the molecules in the crystal environment. Hinsen recently showed that crystal packing considerably modifies the distributions of atomic fluctuations, and that thermal fluctuations are not necessarily the dominant contribution to the crystallographic Debye-Waller factors.[60] Therefore, the observed discrepancies between the GNM predictions and X-ray B-factors could arise from the packing of the protein in the crystal lattice, from static disorder, or approximations (such as the lack of amino acid specificity) inherent to the GNM method.

Comparing GNM, X-ray and NMR models the question arises why one observes better agreement between GNM and NMR RMSDs, compared to X-ray B-factors. The width of the distribution among the NMR models usually results from a combination of sparse data and motion of the polypeptide chain in solution. Furthermore, most methods for calculating NMR ensembles use Nuclear Overhauser effect (NOE) distances as the predominant constraints, which represent a similar contact topology inherent to the GNM analysis. Thus, the good agreement between NMR data and GNM predictions could be caused by the commonality in methodology and similar inherent assumptions in the two approaches.

To address these open questions, we undertook a comprehensive analysis for a designed sugar-binding protein, LKAMG, which we have structurally characterized by both NMR and X-ray crystallography (Koharudin et al., submitted). We simultaneously analyzed the ensemble of NMR models and the X-ray models obtained from two crystal forms, as well as computational data from both the GNM analysis and full atomic MD simulations, for a rigorous assessment of the origins of similarities and differences between the experimental and computational data. Our results show that ensembles, NMR or X-ray, agree well with GNM predictions. The noted consistency of MD and GNM results point to the dominance of inter-residue contact topology (basic ingredient of the GNM) in equilibrium dynamics, even if a detailed force field with non-linear and specific interactions is used, as in MD simulations. Interestingly, our data suggest that less precise ensembles appear to describe the accessible conformational space under native state conditions better than tight ensembles.

2.2 MATERIALS AND METHODS

2.2.1 Materials

We used two sets of independently determined structures of LKAMG, a cyanovirin-N homolog (CVNH) chimera, as our defined model system, determined by NMR spectroscopy and X-ray crystallography. LKAMG is a small protein of 107 residues. It is monomeric both in solution and in the crystalline state. LKAMG crystals were obtained in two different space groups, P_{21} and P_{212121}, designated as X1 and X2 throughout this manuscript. The NMR structure was solved using commonly used methodology[15] and a final ensemble comprising 100 conformers with the lowest energy was selected from the calculated 4000 structural models. The backbone RMSD of the NMR ensemble with respect to the mean was 0.23 ± 0.04 Å and the lowest energy model is

designated as N1. Details about the design, expression, and structural characterization of the protein by both methods are given in an accompanying manuscript (Koharudin et al., submitted). We have additionally constructed a homology model of LKAMG, called H1, using a cyanovirin-N structure (PDB ID: 2EZM)[33] as template in MODELLER 8v2.[61] The sequence identity between LKAMG and its template was 29%. H1 has been adopted as the starting structure in MD simulations.

The structure of LKAMG is displayed in Figure 2.1A. The protein has a pseudo-symmetric architecture comprised of two domains, and closely resembles other members of the CVNH family. Each domain is composed of a three-stranded β-sheet on top of which resides a β-hairpin (β-strands are colored yellow). The two domains are connected by short helical turns (red). In addition, three loops (residues 25-29, 68-73, 81-87; colored purple) protrude out from the core structure. A superposition of the X-ray models X1 and X2 (blue and green), the NMR conformer N1 (magenta) and the homology model H1 (gray) is displayed in Figure 2.1B. Table 2.1 lists the root-mean-square differences (RMSDs) in the backbone atom coordinates of these models. The RMSDs vary from 0.36 Å (between X1 and X2) to 2.01 Å (between N1 and H1).

Table 2.1 Backbone RMSD (Å) between different LKAMG structural models.[a]

	X2	**N1**	**H1**
X1	0.36	0.99	1.69
X2	-	0.96	1.78
N1	-	-	2.01

[a] X1 and X2 are the P_{21} and P_{212121} crystal structures, respectively; N1 is the lowest energy conformer in the NMR solution structure ensemble; H1 is the homology model.

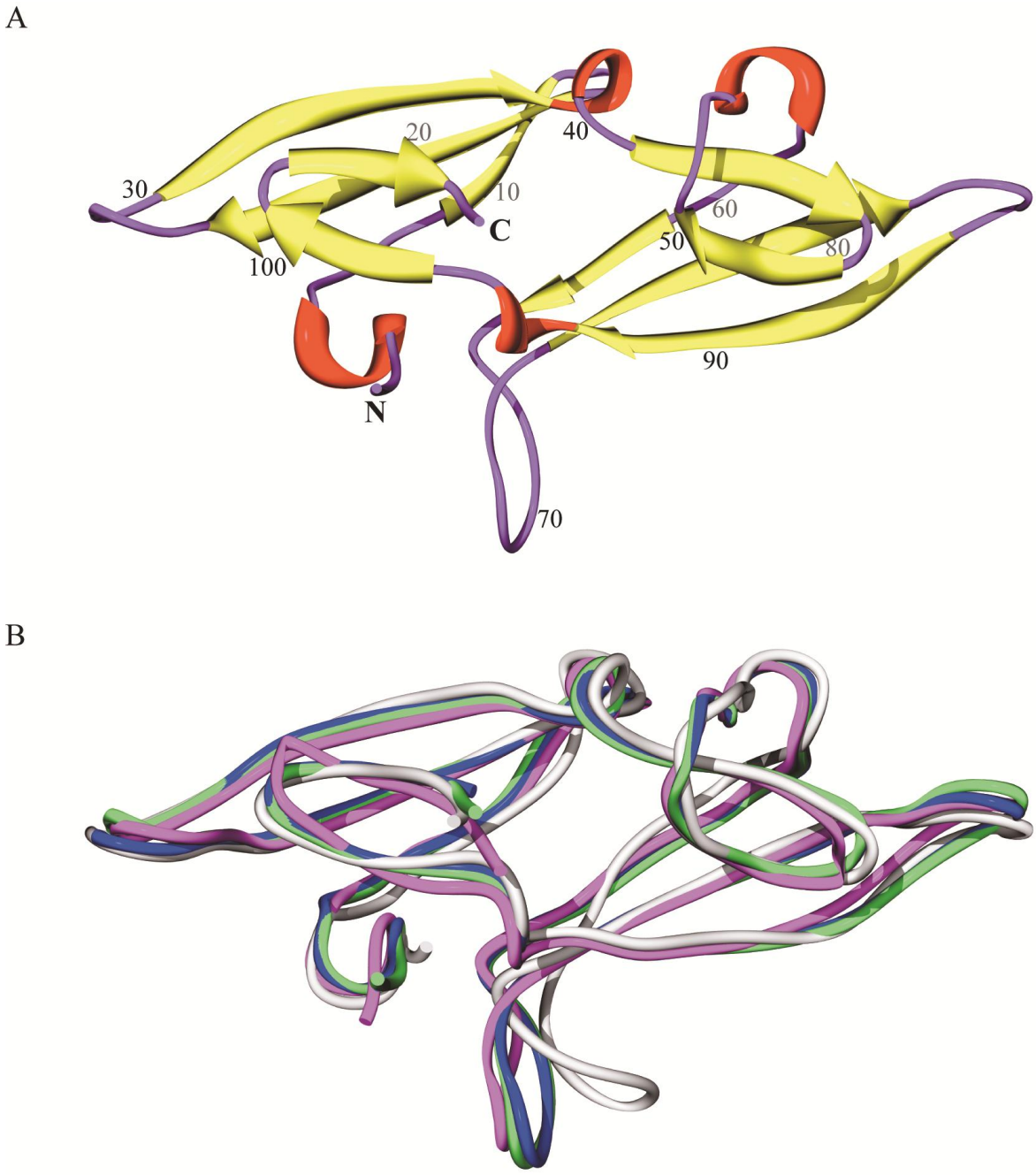

Figure 2.1 Structure of the designed protein, LKAMG.

(A) Ribbon representation, color-coded according to secondary structure; β-strands are shown in yellow, helical turns in red, and loops and chain termini in purple. Amino acid sequence positions are labeled at every 10th residue. (B) Best-fit superposition of four different structural models for LKAMG in modified ribbon representation; the X-ray models X1 and X2 are shown in blue and green, respectively, the lowest energy conformer of the NMR ensemble N1 in magenta, and the homology model H1 in gray.

2.2.2 RMSD calculation for the ensemble of NMR models

The RMSD in the position of residue *i* is calculated as:

$$\langle(\Delta \boldsymbol{R}_i)^2\rangle^{\frac{1}{2}} = \sqrt{\frac{\sum_{k=1}^{m}|\boldsymbol{r}_{i,k}-\bar{\boldsymbol{r}}_i|^2}{m}} \qquad (2.1)$$

where $\bar{\boldsymbol{r}}_i$ designates the position of that particular residue averaged over all optimally superimposed models (*m* of them). The MSDs in residue positions, $< (\Delta \boldsymbol{R}_i)^2 >$ as a function of residue index *i* are referred to as the *fluctuations profile* in residue positions.

2.2.3 Generation of NMR-like ensembles from the X-ray models

NMR-like ensembles were created using inter-proton distance constraints with commonly employed methodology.[15] In order to extract inter-proton distances from X-ray models, hydrogen atoms were added using REDUCE.[62] In this manner, standardized geometry and optimized orientations for OH, SH, NH_3^+, Met methyls, Asn and Gln sidechain amino groups, and His rings were created. Since we use high resolution X-ray models (1.56 Å and 1.36 Å for the P_{21} and P_{212121} data, respectively), one-cycle of refinement in the presence of the added hydrogen atoms was carried out using PHENIX.[63] The resulting models exhibit R and R_{free} values of 0.1607 and 0.2018 for the P_{21} and 0.1669 and 0.1972 for the P_{212121} structures, respectively. Inter-protons distances shorter than or equal to 5 Å were then extracted using MOLMOL[64] and a total of 3972 and 3982 inter-protons distances were generated for the P_{21} and P_{212121} structures, respectively. Note that a total of 2756 inter-proton distances were used for calculating the NMR ensemble. Therefore, an equal number of constraints is used in the pseudo-X-ray ensemble with ~ 70% of the complete constraints set. In order to mimic the structure calculation methodology by NMR, we classified these distances according to three NOE classes (strong, medium, and weak) and added distance corrections to the upper bounds to allow for some distance variability. The upper bound was set to 3.0, 4.0, and 6.0 Å for any extracted distances that were less or equal to 2.5 Å

(strong NOE), less or equal to 3.5 Å but more than 2.5 Å (medium NOE), and less or equal to 5.0 Å but more than 3.5 Å (weak NOE), respectively. This correction reflects distance allowances of 0.5, 0.5, and 1.0 Å for the short, medium, and long distances. From the total set of distance constraints, we randomly removed 20% or 50% of the data, yielding the 80% or 50% distance sets. Inter-proton distances were measured including exchangeable hydrogens, some of which may not be observable in the experimental setting due to fast exchange with solvent. No intra-residue proton distances, however, were included. Note, removal of the exchangeable hydrogens from the lists did not affect the generated NMR-like ensembles in any significant manner (data not shown).

2.2.4 Fluctuations and collective modes predicted by the Gaussian Network Model

In the GNM, the structure is modeled as a 3-dimensional elastic network of *n* nodes. The position of each node is determined by the α-carbons. The network topology is described by a $N \times N$ Kirchhoff matrix $\mathbf{\Gamma}$

$$\mathbf{\Gamma}_{ij} = \begin{cases} -1 & if \;\; i \neq j \; and \; r_{ij} \leq r_c \\ 0 & if \;\; i \neq j \; and \; r_{ij} > r_c \\ -\sum_{i,i\neq j} \mathbf{\Gamma}_{ij} & if \;\; i = j \end{cases} \qquad (2.2)$$

where r_c is the cutoff distance that defines pairs of residues to be connected in the network. r_{ij} is the equilibrium distance between residue *i* and residue *j*, calculated using the Protein Data Bank (PDB)[65] coordinates. The cross-correlations between the fluctuations $\Delta \boldsymbol{R}_i$ and $\Delta \boldsymbol{R}_j$ of the nodes *i* and *j* are given by[45]

$$\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle = \frac{3k_B T}{\gamma} \, [\mathbf{\Gamma}^{-1}]_{ij} \qquad (2.3)$$

where k_B is the Boltzmann constant, *T* is the absolute temperature and γ is a uniform spring constant. The inverse of $\mathbf{\Gamma}$ is expressed in terms of the nonzero eigenvalues λ_k ($1 \leq k \leq N-1$) and corresponding eigenvectors $\boldsymbol{u}_k$ of $\mathbf{\Gamma}$ as[26]

$$\mathbf{\Gamma}^{-1} = \sum_{k=1}^{N-1} \lambda_k^{-1} \boldsymbol{u}_k \boldsymbol{u}_k^T \quad (2.4)$$

which permits us to express the ms fluctuations of a given residue as a sum over the contributions of all modes

$$\langle (\Delta R_i)^2 \rangle = \sum_{k=1}^{N-1} \frac{3k_B T}{\gamma} (\lambda_k^{-1} u_k u_k^T)_{ii} \quad (2.5)$$

Here the subscript *ii* designates the i^{th} diagonal element of the matrix enclosed in parenthesis. The X-ray crystallographic B-factors are compared with the theoretical predictions using

$$B_i \equiv \frac{8\pi^2}{3} \langle (\Delta \boldsymbol{R}_i)^2 \rangle = \sum_{k=1}^{N-1} \frac{8\pi^2 k_B T}{\gamma} (\lambda_k^{-1} \boldsymbol{u}_k \boldsymbol{u}_k^T)_{ii} \quad (2.6)$$

The GNM predictions for (i) NMR model N1, (ii) the mean structure of NMR ensemble, or (iii) those averaged over all models in the NMR ensemble were found to be almost identical (correlation coefficients above 0.95); hence we use NMR model N1 as a representative model for the NMR ensemble.

2.2.5 Comparison of MD essential modes with GNM global modes

The MD simulations were performed using NAMD[66] with the Charmm22 force field[67]. Three runs were performed with explicit water for a total duration of 10 ns, each, at constant temperature (298 K) and pressure (1 atm). Instantaneous conformations were saved every 1ps excluding the first 1.5 ns portion of the trajectories (Figure 2.2A). The resulting M snapshots where organized in the fluctuation trajectory matrix

$$\Delta R = \begin{bmatrix} \Delta R_1(t_1) & \Delta R_1(t_2) & \cdots & \Delta R_1(t_M) \\ \Delta R_2(t_1) & \Delta R_2(t_2) & \cdots & \Delta R_2(t_M) \\ \Delta R_3(t_1) & \Delta R_3(t_2) & \cdots & \Delta R_3(t_M) \\ \cdots & \cdots & \cdots & \cdots \\ \Delta R_N(t_1) & \Delta R_N(t_2) & \cdots & \Delta R_N(t_M) \end{bmatrix}_{3N \times M} \quad (2.7)$$

$\Delta \boldsymbol{R}_i(t_j)$ is the 3-dimensional vector representing the departure of the i^{th} α-carbon from its mean position, at the j^{th} snapshot. Multiplication of $\Delta \boldsymbol{R}$ by its transpose yields the $3N \times 3N$

covariance matrix $\boldsymbol{A}$. $\boldsymbol{A}$ can be viewed as an $N \times N$ supermatrix, the ij^{th} 'element' of which is the 3× 3 matrix

$$\boldsymbol{A_{ij}} = M \begin{bmatrix} \langle \Delta \boldsymbol{X_i} \Delta \boldsymbol{X_j} \rangle & \langle \Delta \boldsymbol{X_i} \Delta \boldsymbol{Y_j} \rangle & \langle \Delta \boldsymbol{X_i} \Delta \boldsymbol{Z_j} \rangle \\ \langle \Delta \boldsymbol{Y_i} \Delta \boldsymbol{X_j} \rangle & \langle \Delta \boldsymbol{Y_i} \Delta \boldsymbol{Y_j} \rangle & \langle \Delta \boldsymbol{Y_i} \Delta \boldsymbol{Z_j} \rangle \\ \langle \Delta \boldsymbol{Z_i} \Delta \boldsymbol{X_j} \rangle & \langle \Delta \boldsymbol{Z_i} \Delta \boldsymbol{Y_j} \rangle & \langle \Delta \boldsymbol{Z_i} \Delta \boldsymbol{Z_j} \rangle \end{bmatrix}_{3\times 3} \tag{2.8}$$

The cross-correlation between the fluctuations of residues *i* and *j* is found from the trace of $\boldsymbol{A_{ij}}$ as

$$\langle \Delta \boldsymbol{R_i} \cdot \Delta \boldsymbol{R_j} \rangle = (1/\mathrm{M}) tr\left[\boldsymbol{A}_{ij}\right] \tag{2.9}$$

These cross-correlations may be conveniently organized in an *N* x *N* covariance matrix $\boldsymbol{C}$, the diagonal elements of which are simply the ms fluctuations of residues. $\boldsymbol{C}$ may be expressed in terms of its eigenvalues (s_l) and eigenvectors ($\boldsymbol{q}_l$) as

$$\boldsymbol{C} = \sum_l s_l \boldsymbol{q}_l \boldsymbol{q}_l^T \tag{2.10}$$

The eigenvalues serve as weights for square displacements induced by different modes. Trajectories along the essential modes 1, 2, 4 and 16 of an MD run are illustrated in the Figure 2.2B.

$\boldsymbol{C}$ is the counterpart of $\boldsymbol{\Gamma^{-1}}$. Likewise, s_l is the counterpart of $(\frac{3k_BT}{\gamma})\lambda_k^{-1}$, and $\boldsymbol{q}_l$ is the counterpart of $\boldsymbol{u}_k$. Therefore the eigenvalues extracted from MD can be directly compared to the reciprocal eigenvalues from the GNM. Likewise, the top-ranking eigenvectors (corresponding to the lowest frequency, or global, modes) may be directly compared. The cumulative square correlation $\{\sigma^2(\mathrm{k})\}_{ltot}$ between a given GNM mode (e.g., $\boldsymbol{u}_k$) and an ensemble of l_{tot} MD modes is evaluated from

$$\{\sigma^2(\mathrm{k})\}_{ltot} = \Sigma_l \cos^2(\boldsymbol{u}_k, \boldsymbol{q}_l) \tag{2.11}$$

where the summation is performed for $1 \leq l \leq l_{tot}$.

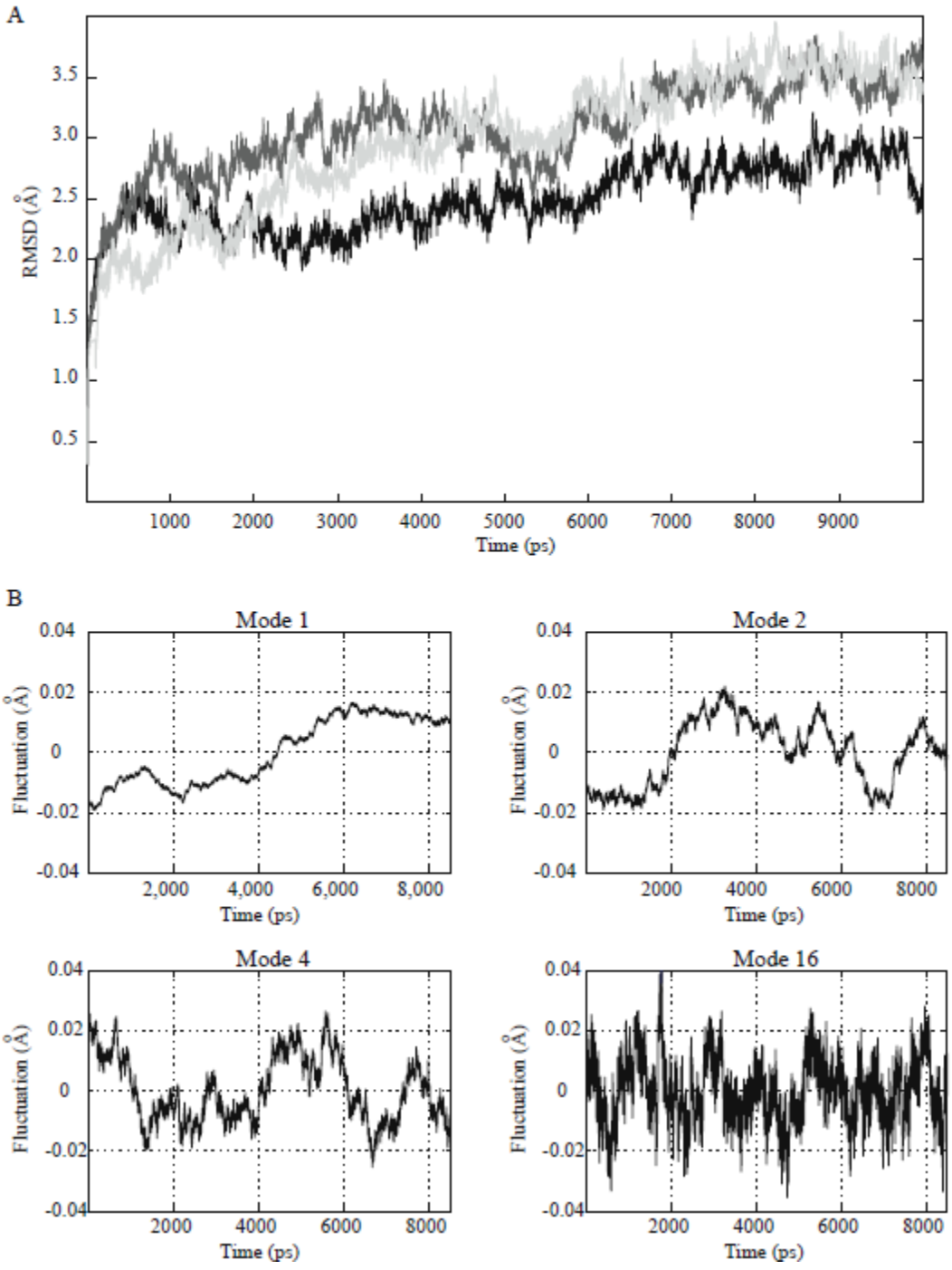

Figure 2.2 Analysis of MD trajectories.

(A) Time evolution of average RMSD (with respect to the starting conformation) in Cα-coordinates for three runs MD1 (black), MD2 (dark gray) and MD3 (light gray). (B) Motions along essential modes, illustrated for modes 1, 2, 4 and 16 evaluated for MD1, after excluding the equilibration period of 1500 ps.

2.3 RESULTS AND DISCUSSION

2.3.1 Comparison of the two computational approaches

The main impetus for our current study was to uncover any reasons that cause the better agreement between predicted equilibrium dynamics by GNM and the NMR RMSDs compared to X-ray B-factors. In order to exclude any potential errors that may arise from neglecting nonlinear effects in the GNM, we first compared the results predicted by the GNM with those obtained by MD simulations.

Figure 2.3 compares the MSDs, $\langle(\Delta \boldsymbol{R}_i)^2\rangle$, $1 \le i \le N$, extracted from experimental data (NMR ensemble and X-ray crystallographic B-factors) with the square fluctuations computed by MD and GNM. The MSDs refer to the positions of the α-carbons with respect to their mean positions. The MSD profiles based on NMR, X1 and X2 data, designated as $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{NMR}$, $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{X1}$, and $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{X2}$, are colored magenta, blue, and green, respectively (top panel). $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{GNM\text{-}N1}$, $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{GNM\text{-}X1}$ and $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{GNM\text{-}X2}$ are their counterparts predicted by the GNM, using the NMR model N1 and the two crystal structures X1 and X2, respectively, as input (middle panel). $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{MD1}$, $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{MD2}$ and $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{MD3}$ are the square fluctuations profiles observed in three independent MD runs (bottom panel). The correlation coefficients between these profiles are summarized in Table 2.2.

As can be appreciated from the results presented in Table 2.2, GNM predictions for different models (N1, X1 or X2) are highly correlated, also reflected by the very similar profiles in Figure 2.3 (middle panel). The pairwise correlations between $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{GNM\text{-}N1}$, $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{GNM\text{-}X1}$ and $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{GNM\text{-}X2}$ are all equal to or higher than 0.95. Such close agreement is not surprising since GNM results are primarily defined by the coarse-grained distribution of inter-residue contacts (C^α-C^α pairs within an interaction cutoff distance of r_c = 7 Å). The three models N1,

X1 and X2, which differ in their backbone coordinates by less than 1 Å (Table 2.1), are expected to exhibit very similar contact topologies.

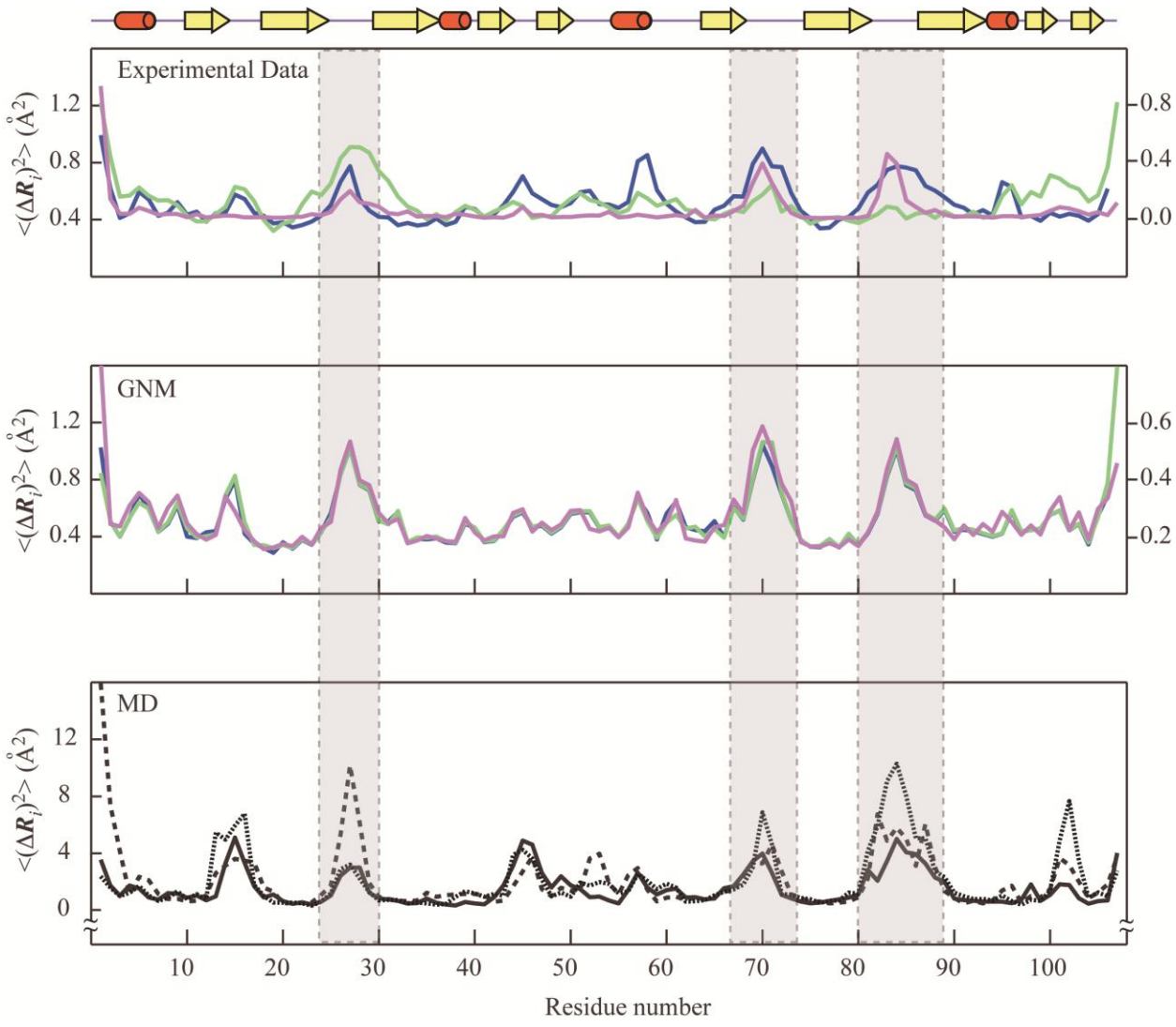

Figure 2.3 Mean-square fluctuations profiles of LKAMG from experimental data and computations.

The fluctuations in the positions of the residues, $\langle(\Delta \boldsymbol{R}_i)^2\rangle$, are plotted as a function of residue position along the polypeptide chain, $1 \le i \le N$. The upper panel displays the MSDs from experimental data, $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{NMR}$, $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{X1}$ and $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{X2}$ colored magenta, blue and green, respectively. Crystallographic fluctuations are extracted from the B-factors, using $B_i = (8\pi^2/3)\ \langle(\Delta \boldsymbol{R}_i)^2\rangle$, The left and right ordinates correspond to NMR and X-ray data, respectively. The middle panel displays the square fluctuations

predicted by the GNM for the different structural models, $<(\Delta \boldsymbol{R}_i)^2>_{GNM\text{-}N1}$ (magenta,), $<(\Delta \boldsymbol{R}_i)^2>_{GNM\text{-}X1}$ (blue), and $<(\Delta \boldsymbol{R}_i)^2>_{GNM\text{-}X2}$ (green). The lower panel shows the results from three MD runs, $<(\Delta \boldsymbol{R}_i)^2>_{MD1}$ (solid black), $<(\Delta \boldsymbol{R}_i)^2>_{MD2}$ (dotted black), and $<(\Delta \boldsymbol{R}_i)^2>_{MD3}$ (dashed black). A schematic representation of the LKAMG secondary structure is displayed on top. The three loop regions are indicated by the gray columns.

Table 2.2 Correlation coefficients for mean-square fluctuations (MSFs) and MSDs in residue positions observed in experiments and computations.[a]

$<(\Delta R_i)^2>$	GNM (N1)	X1	GNM(X1)	X2	GNM(X2)	MD1	MD2	MD3
NMR	**0.80**	0.64	0.77	0.31	0.78	**0.54**	**0.60**	**0.65**
GNM (N1)	-	0.76	0.95	0.50	0.95	0.62	0.61	0.58
X1	-	-	**0.76/0.72[b]**	0.25	0.76	**0.69**	**0.62**	**0.60**
GNM(X1)	-	-	-	0.52	0.99	0.69	0.65	0.61
X2	-	-	-	-	**0.51/0.69[b]**	**0.18**	**0.49**	**0.13**
GNM(X2)	-	-	-	-	-	0.68	0.66	0.63
MD1	-	-	-	-	-	-	0.64	0.79
MD2	-	-	-	-	-	-	-	0.65

[a] Computational results were obtained by GNM predictions for the NMR model N1, and the X-ray models X1 and X2, as well as by MD simulations MD1-3. Boldface entries refer to correlations between experimental data the corresponding computational predictions.

[b] 0.76 and 0.51 are the correlation coefficients based on the GNM predictions for the isolated protein, and 0.72 and 0.69 are their counterpart for the protein in the lattice (see Figure 2.7).

The correlations between GNM and MD profiles, on the other hand, vary from 0.58 (between $<(\Delta \boldsymbol{R}_i)^2>_{GNM\text{-}N1}$ and $<(\Delta \boldsymbol{R}_i)^2>_{MD3}$) to 0.69 (between $<(\Delta \boldsymbol{R}_i)^2>_{GNM\text{-}X1}$ and $<(\Delta \boldsymbol{R}_i)^2>_{MD1}$) and are mainly influenced by the particular trajectories (MD1, MD2, and MD3). This level of agreement is reasonable, given the fundamentally different assumptions and methodologies inherent to the two types of computations: GNM is a low resolution approach, based exclusively on inter-residue contact topology; MD includes full atomic details with elaborate force fields. Notably, GNM yields consistent solutions for the fluctuations behavior of LKAMG and results

are obtained within seconds. MD runs, on the other hand, take weeks, and the results suffer from sampling inaccuracies, as evidenced by the correlations of 0.64, 0.65 and 0.79 between pairs of MD runs. What is even more striking is that GNM results consistently agree better with experimental data, either NMR or X-ray, than MD results, as will be discussed below.

Two further comparative analyses of the two sets of computational results were carried out focusing on (i) their level of agreement with experimental data, and (ii) the spectra of modes predicted in each case.

2.3.2 Comparison of computational and experimental data

As shown in Table 2.2, the fluctuations profiles predicted by the GNM for the models N1 ($<(\Delta \boldsymbol{R}_i)^2>_{GNM\text{-}N1}$), X1 ($<(\Delta \boldsymbol{R}_i)^2>_{GNM\text{-}X1}$) and X2 ($<(\Delta \boldsymbol{R}_i)^2>_{GNM\text{-}X2}$) yielded respective correlation coefficients of 0.80, 0.76 and 0.51 with their experimental counterparts, $<(\Delta \boldsymbol{R}_i)^2>_{NMR}$, $<(\Delta \boldsymbol{R}_i)^2>_{X1}$ and $<(\Delta \boldsymbol{R}_i)^2>_{X2}$, respectively. For the MD trajectories, on the other hand, respective correlation coefficients of {0.54, 0.69 and 0.18} were found between $<(\Delta \boldsymbol{R}_i)^2>_{MD1}$ (from MD1) and {$<(\Delta \boldsymbol{R}_i)^2>_{NMR}$, $<(\Delta \boldsymbol{R}_i)^2>_{X1}$ and $<(\Delta \boldsymbol{R}_i)^2>_{X2}$} and their counterparts for MD2 and MD3 were {0.60, 0.62 and 0.49} and {0.65, 0.60, and 0.13}, respectively. These entries are listed in boldface in the Table 2.2.

These results clearly show that the fluctuations profiles predicted by the GNM exhibit higher correlation with experimental data compared to those obtained by MD. It is also interesting to note that the correlation between the results from the three different MD runs is 0.69 ± 0.08, indicating that the results from MD simulations are not as robust as those from GNM, despite the fact that all MD runs were performed with the same starting structure (H1) while GNM calculations, almost identically reproduced, were performed using different structural models.

Additionally, it can also be easily noticed that, in both cases, the computational results obtained for X2 exhibit poorer agreement with the experimental data, compared to those obtained for N1 and X1. It is worth noting, however, that, even for the X2 data, GNM systematically yielded a higher correlation than MD, suggesting that the lack of atomistic details in the GNM is more than compensated for by the mathematically exact evaluation of fluctuations using the complete, collective coupling of all residues. A detailed analysis pertaining to the comparison of the X2 data with computational predictions is discussed below.

2.3.3 Comparison of essential modes from MD and GNM

To provide a more in-depth analysis of the GNM and MD results, we decomposed the predictions into the contributions of the underlying modes and compared both methods' individual (top-ranking) modes. In doing so, we verified that the motions in different time regimes predicted by GNM are comparable to those sampled by MD simulations. There is, however, no one-to-one correspondence between pairs of modes.

GNM equilibrium fluctuations result from the superposition of N-1 normal modes for a protein of N residues. On the other hand, the essential dynamics analysis of a MD trajectory yields 3N-6 modes (unless the number of snapshots M is smaller than 3N-6). As described in the Methods, the $3N$ x $3N$ covariance matrix derived from a given MD trajectory may be conveniently organized into an N x N covariance matrix **C** of residue fluctuations, the eigenvalues and eigenvectors of which can be directly compared to those predicted by the GNM. We focused on the top-ranking modes at the low frequency end of the spectrum. These modes, also referred to as the global or *essential* modes,[68] define those motions that contribute the most to the observed dynamics, and are usually relevant to functional changes in conformation.[5, 49]

The correlation coefficients between the top-ranking modes extracted from MD and those predicted by the GNM are displayed in Figure 2.4A. Two pertinent observations emerge: (i) a given MD mode can be correlated with more than one GNM mode. For example, the MD mode 3 exhibits a correlation of 0.5 or higher with both modes 1 and 2 predicted by the GNM (see the corresponding brown and red boxes in Figure 2.4A), and (ii) the order of the modes in the two methods differ (for example, the 4th MD mode is highly correlated with the 5th GNM mode; i.e., the red boxes are not necessarily clustered along the diagonal). This analysis shows that it is hard, if not impossible, to identify a unique counterpart of each GNM mode in MD, or vice versa, probably due to different types and scales of movements represented by these modes. Yet, similarities between preferred modes of motions could be detected by consolidating the results using subsets of modes. We examined to this aim the combined contributions of the first 10 MD modes in relation to the individual GNM modes k in the range $k \leq 10$. The cumulative correlation cosine (squared) $\{\sigma^2(k)\}_{10}$ between the set of 10 MD modes and the k^{th} GNM mode (Eq. 2.11 in Methods) is shown in Figure 2.4B. The result for the first GNM mode is 0.95, shown by the magenta bar at $k = 1$, i.e., the combined first 10 MD modes $[\{\sigma^2(k)\}_{10}]^{½}$ overlap by 97% with the 1st GNM mode. The overlap with the 2nd GNM mode is equally high and only gradually decreases with mode number, remaining above 0.75 for 5 out of 10 GNM modes. Note that the 10 MD modes represent only a small fraction (less than ten percent) of the entire set of modes retrieved by decomposing the MD covariance matrix **C**. However, their weighted contribution amounts to 98% while that of first 10 GNM modes represents 47% of the predicted motions.

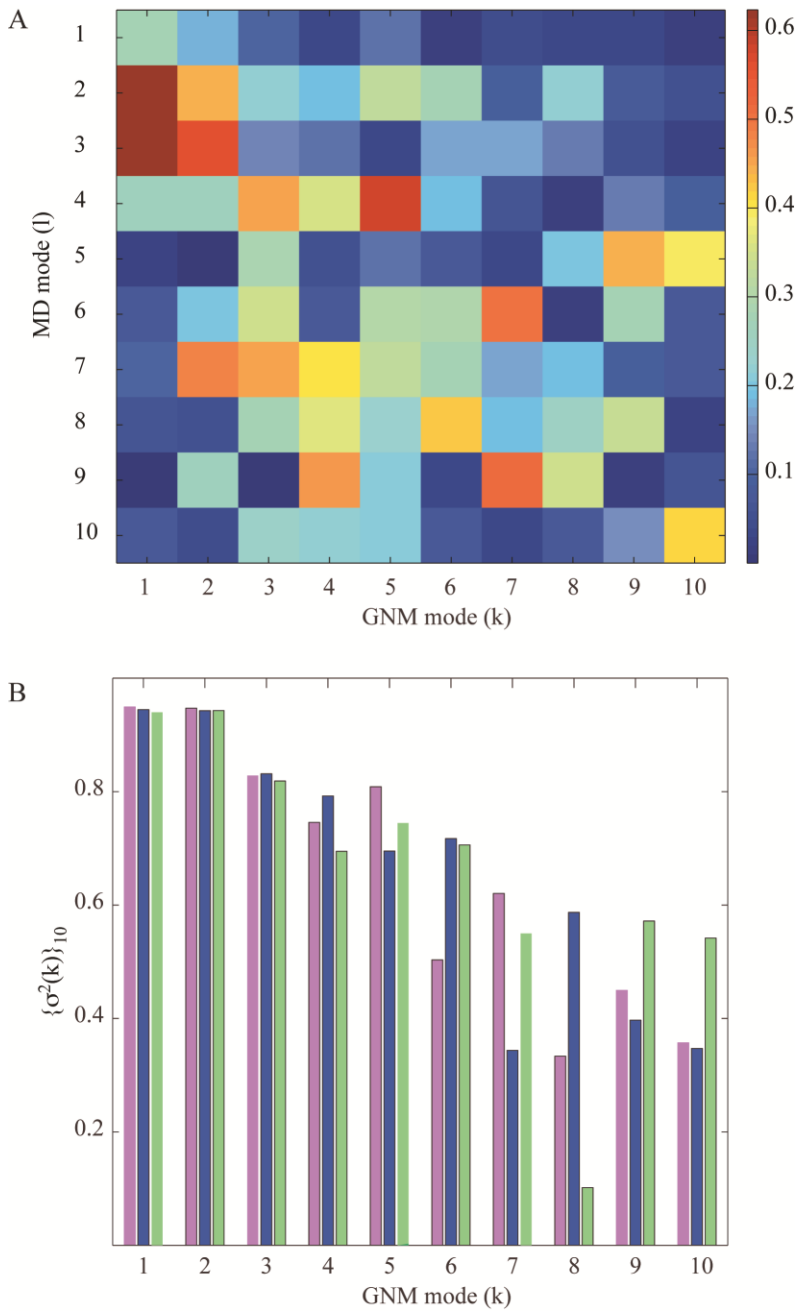

Figure 2.4 Correlation map for essential modes predicted by the GNM and derived from MD.

(A) Correlations, $[\boldsymbol{q}_l . \boldsymbol{u}_k]$, between the essential modes $\boldsymbol{q}_l$ ($1 \leq l \leq 10$) retrieved from MD1 and those ($\boldsymbol{u}_k$, $1 \leq k \leq 10$) predicted by the GNM. (B) Cumulative correlations (sum over cosines squared; see Eq. 2.12) for the first ten essential MD modes and individual GNM modes predicted for the models N1 (magenta), X1 (blue) and X2 (green). See Figure 2.5 for a more extensive comparison of the mode spectra obtained by MD and GNM.

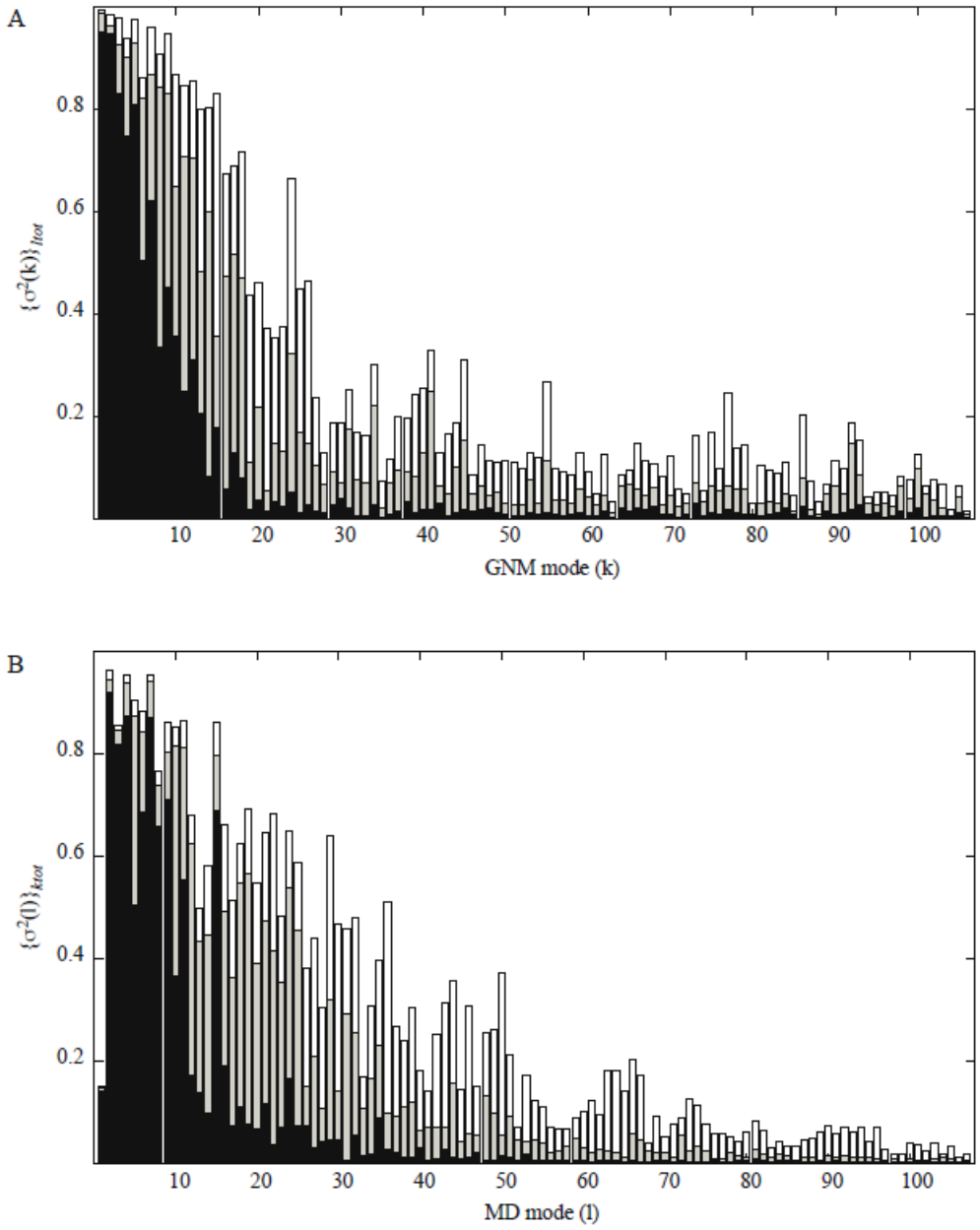

Figure 2.5 Cumulative correlations between mode spectra obtained from GNM and MD.

(A) Cumulative squared cosines $\{\sigma^2(k)\}_{ltot}$ between l_{tot} essential modes from MD simulations and each GNM mode (k) for l_{tot} = 10 (black), 20 (gray) and 30 (white). (B) $\{\sigma^2(l)\}_{ktot}$ between top-ranking k_{tot} = 10 (black), 20 (gray) and 30 (white) GNM modes with the MD modes (l) listed along the abscissa. Note, the dominant contribution of the slowest modes to the low frequency end of the spectrum, in each case, followed by the larger contribution of intermediate frequency, and then higher frequency modes, indicate the consistency between the two sets of mode spectra.

The above numbers refer to GNM calculations performed with N1 as the model. Similar results were obtained using the X-ray models X1 and X2, shown by blue and green bars in Figure 2.4B. The correlation falls below 0.1 beyond the 20th GNM mode. The dependence of $\{\sigma^2(k)\}_{ltot}$ on k, for l_{tot} = 10, 20 and 30 MD modes is provided in Figure 2.5 panel A for all GNM modes $1 \leq k \leq N-1$; and panel B in the same figure displays the joint contribution $\{\sigma^2(l)\}_{ktot}$ of k_{tot} = 10, 20 and 30 GNM modes to the l^{th} MD mode. Interestingly, there is a hierarchical influence of relatively higher frequency MD modes on the higher GNM modes, confirming consistency between the two spectra of modes. MD simulations and GNM predictions are thus comparable with regard to the dominant, usually biologically relevant, low frequency modes. The differences between the MD and GNM fluctuation profiles mainly originate from higher frequency modes that are known to be noisy.

2.3.4 The close relationship between NMR and GNM - is the agreement simply based on the similarity in methodology?

The above analysis indicates that MSDs predicted by the GNM consistently exhibit a better correlation with experimental data than MD results, and that the level of correlation between the fluctuations predicted by GNM and the MSD extracted from the NMR ensemble is higher than that between GNM and X-ray B-factors. NMR ensembles calculations use NOE distances as the predominant constraints, and GNM analysis is also based on knowledge of inter-residue contact topology. In order to critically evaluate whether the good correlation between the distribution of the conformers in an NMR ensemble and GNM-predicted fluctuations arises mainly from the similarity in the methodologies for NMR structure determination/refinement and for GNM calculations, we analyzed six differently calculated ensembles of structures that were derived from the X-ray models X1 and X2. For each crystal structure, three ensembles of 30 conformers

each were generated, using the standard constraints-based NMR structure determination procedure. As constraints, 100%, 80%, and 50% of all possible inter-proton distance constraints were used. These ensembles are designated as X1- and X2-ensembles. The final 30 conformer ensembles exhibit backbone RMSD values of 0.32 ± 0.07 Å, 0.33 ± 0.06 Å, and 0.43 ± 0.06 Å when 100%, 80%, and 50% of constraints were used, respectively. The corresponding values for the X2-ensembles are 0.29 ± 0.06 Å, 0.37 ± 0.06 Å, and 0.40 ± 0.07 Å, respectively. As expected, there is a correlation between the ensemble precision and the number of constraints used to generate these ensembles, i.e. the ensemble RMSDs increase with decreasing number of constraints.[15]

Using these so-called pseudo X-ray ensembles, we compared their MSDs with the ms fluctuations predicted by GNM and with the MSDs extracted from NMR data (N1) (Figure 2.6). Figure 2.6A displays the correlation coefficients between the MSDs in α-carbon coordinates $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{ensemble}$ for each X-ray ensemble and the fluctuations predicted by the GNM for the single crystal structures X1 (blue) and X2 (green); and Figure 2.6B displays the correlation coefficients between the MSDs $\langle(\Delta \boldsymbol{R}_i)^2\rangle_{ensemble}$ for each X-ray ensemble and the MSDs extracted from the original NMR data (N1). For comparative purposes, the correlations between the experimental B-factors and their GNM counterparts (Table 2.2 and Figure 2.6A) and between the experimental B-factors and the experimental MSD from the NMR ensemble (Table 2.2 and Figure 2.6B) are displayed by the light-colored bars on the panels.

If only methodological similarities between NMR structure determination and GNM would play a role in their better correlation, we would expect that decreasing the number of constraints used for generating the pseudo ensembles would increase the correlation between these pseudo ensemble MSDs and the predicted GNM fluctuation for both X1 and X2 pseudo

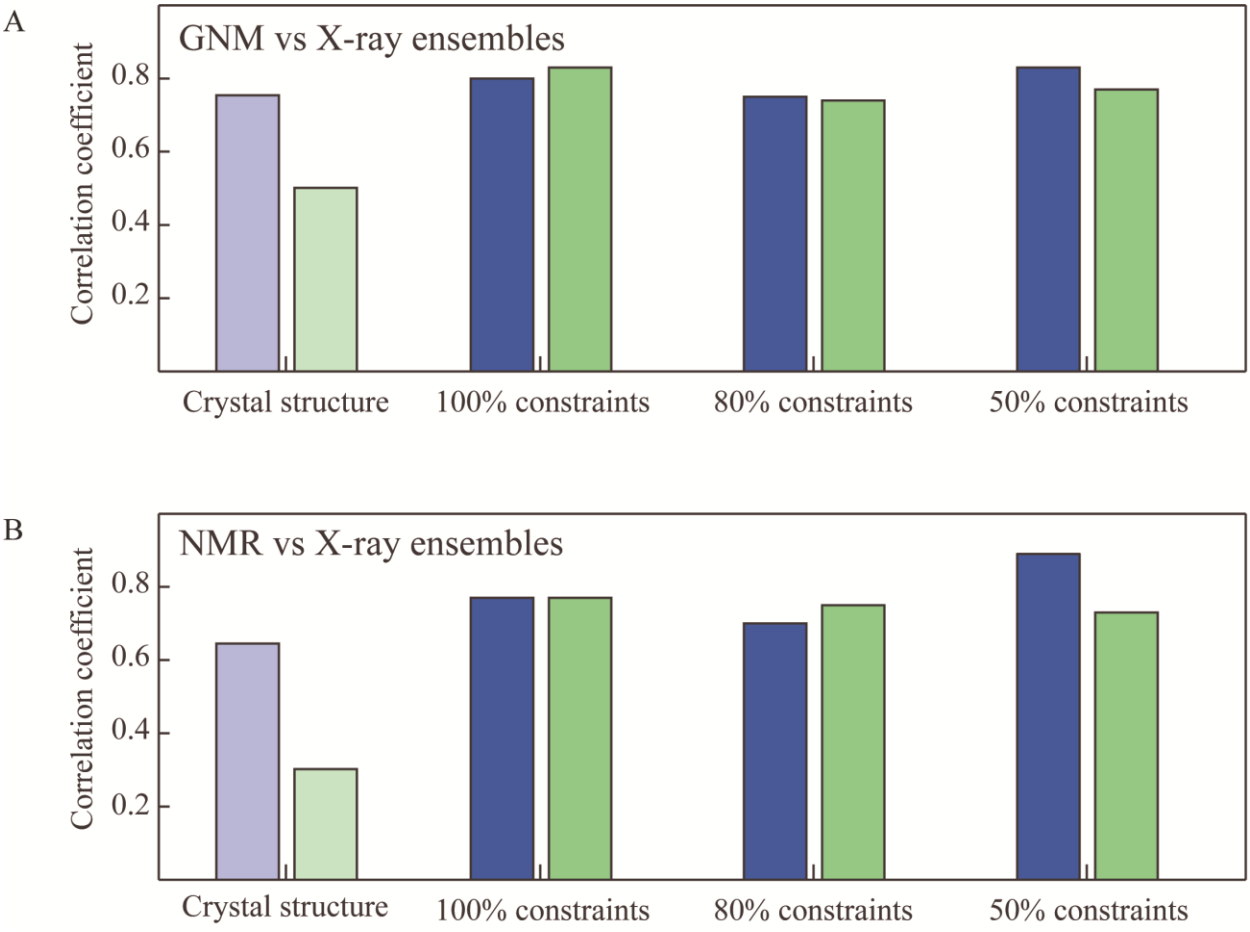

Figure 2.6 Correlations between residue fluctuations from theoretical predictions and inferred from pseudo X-ray ensembles (panel A) and NMR experiments (panel B).

Results for pseudo X-ray ensembles X1 and X2 are shown in blue and green bars. Theoretical data in panel A refers to GNM results obtained for the original crystal structures (X1 or X2). Experimental data in panel B refers to the RMSDs in Cα-positions between the models in the solution NMR ensemble. Results are displayed for three pseudo-X-ray ensembles, generated using 100%, 80% and 50% of the total constraints set. The light-colored bars on the left refer to the comparison of the original structures' B-factors with GNM theory (A) and NMRexperiments (B).

ensembles. Since the ensemble precision would be loosened with decreasing number of constraints, it might be mimicking the GNM methodology of using C^{α}-C^{α} distances of 7 Å. As can be appreciated from Figure 2.6A, we did not observe this effect. For X1 pseudo ensembles,

the correlations of the ensemble MSDs with the GNM predicted fluctuations were 0.80, 0.75, and 0.83 for the 100%, 80%, and 50% constraints employed, respectively. Therefore, the degree of correlation is very similar, irrespective of how many constraints were employed. For the X2 ensembles, the correlations are 0.83, 0.74, and 0.77 for the 100%, 80%, and 50% constraints set, respectively, similar to what is observed for the X1 ensembles. Therefore, our data show that the methodological similarity between NMR structure determination and GNM analysis is not a major factor causing good agreement between GNM predictions and NMR ensemble data.

Most importantly, we also noticed that the X-ray ensembles' MSDs are in better agreement with the equilibrium fluctuations inferred from GNM than the sole use of X-ray crystallographic B-factors. This is especially true for the X2 pseudo ensembles. As can be appreciated from Figure 2.6A and 2.6B, the correlations among the MSD profiles of both X1 and X2 ensembles agree equally well with their GNM predictions (panel A) and with the experimental NMR data (panel B), across the three different constraint sets. While there seems to be no noticeable change comparing the pseudo X1 ensembles and their GNM predictions versus the X1 B-factors and the GNM prediction, a large improvement in the correlations was seen in the X2 case.

Similar behavior was noted in the comparison of the pseudo X1 and X2 ensemble MSDs with the experimental NMR MSD. For X1, no significant differences in correlation were observed for all three ensemble MSDs and the corresponding experimental NMR MSD (0.77, 0.70, and 0.89 for 100%, 80%, and 50%, respectively) versus the correlation between the X1 B-factors and the NMR MSD (0.64). In contrast, a large improvement in the correlation between the pseudo X2 ensembles and the experimental NMR MSDs (0.77, 0.75, and 0.73 for 100%, 80%, and 50%, respectively) was noted, compared to the poor correlation of 0.31 between the B-

factors and the NMR MSDs. Based on these findings, we conclude that it is preferable to consider the MSDs obtained from an ensemble of conformers, rather than solely considering the B-factors from a single crystal structure, for assessing the equilibrium fluctuation behavior of residues.

2.3.5 Interactions between neighboring molecules affect the dynamics in the crystal lattice

One may ask why improvements in the correlation were only found for X2 and not for X1? In order to answer this question, we analyzed the distinctive behavior of the X1- and X2-models and how it may relate to influencing B-factors. As pointed out previously, crystal packing can influence residue motions since the interactions between one molecule and its neighbors can dampen equilibrium motions.[5, 58, 60] The fluctuations accessible in the crystal environment may therefore deviate from those observed in solution (or under physiological conditions), depending on the extent of intermolecular contacts in a given crystal lattice.[53, 58, 60] Motions in the crystal will also deviate from those calculated by the GNM, since the GNM, by definition, predicts the 'intrinsic' dynamics in the absence of intermolecular interactions. It therefore is critical in any comparative assessment of the equilibrium dynamics to consider the isolated molecule and that in the crystal environment and elucidate any biases induced by crystal contacts.

We therefore carried out additional GNM calculations that took into account inter-molecular contacts between adjacent proteins in the crystal lattices, including all immediate neighbors in the crystal lattice (Figure 2.7, panels C and D). The resulting MSD profiles for the crystal forms X1 and X2 are shown by the dashed gray curves in Figure 2.7 along with the experimental data ($<(\Delta \boldsymbol{R}_i)^2>_{X1}$, blue, and $<(\Delta \boldsymbol{R}_i)^2>_{X2}$, green, by solid curves. For comparative purposes, we also display the GNM predictions for the isolated protein (dotted blue and green curves). No significant differences are observed for the two sets of GNM results for X1

(correlation coefficients of 0.76 and 0.72 for the isolated and lattice embedded chain, respectively), while for X2 an increase from 0.51 to 0.69 is noted.

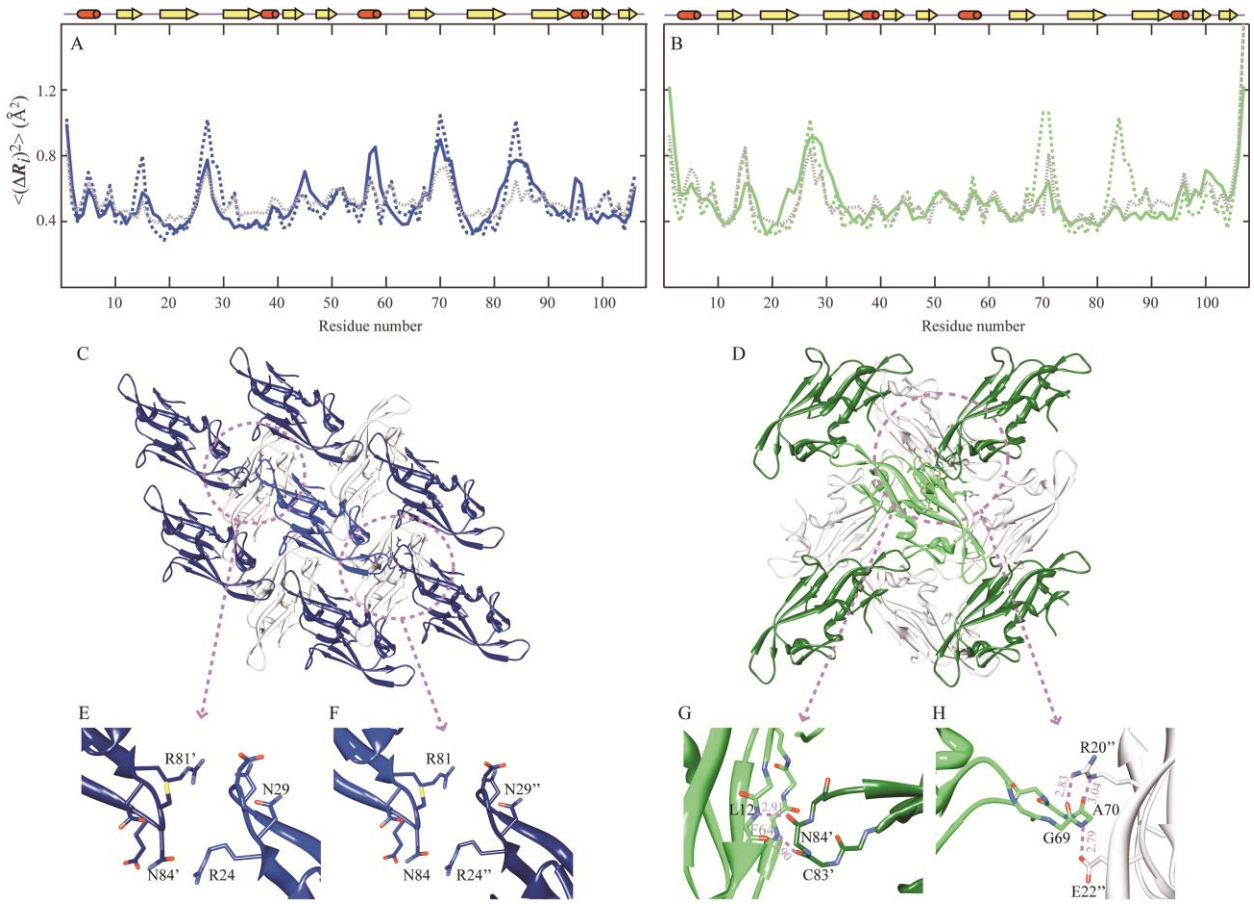

Figure 2.7 Comparison of theoretical and experimental residue fluctuations based on crystal packing of LKAMG in two different lattices.

Panels A and B refer to the crystal structures X1 and X2, respectively. Mean-square fluctuations of residues predicted by the GNM for the isolated protein (dashed blue in panel A, dashed green in panel B) and those in the crystal lattice (dashed gray in both panels) are compared with those inferred from X-ray crystallographic B-factors (solid blue and green in the respective panels). (C) and (D) ribbon diagram of LKAMG surrounded by its first neighbors in the respective P_{21} (X1) and P_{212121} (X2) crystal forms. The total number of surrounding molecules is 14 and 12 in the respective crystals. Four symmetrically related molecules on the upper plane are not displayed in each diagram for clarity. Encircled regions are enlarged in panels E, F, G, and H. Panels (E) and (F) highlights the inter-molecular contacts in X1, (G) and (F) those in X2.

Closer examination of the fluctuations profiles reveals that the three loops comprising residues 25-29, 68-73 and 81-87 are predicted by the GNM to be the most mobile regions. In the case of X1, these regions, indeed, exhibit relatively high B-factors. For X2, on the other hand, motions in loops 68-73 and 81-87 are dampened as evidenced by the experimentally observed smaller B-factors. The GNM calculations performed in the presence of neighboring molecules in the crystal lattice unambiguously reveal that the observed deviations are related to crystal packing. Note, a total of 15 (X1) or 13 (X2) molecules, including the central molecule of interest, were considered in the GNM predictions, and the fluctuations profiles for the central molecules are shown in the figure.

The different behavior of the GNM predictions for the two X-ray models in the context of their crystal neighbors is related to the different arrangement of individual proteins in the two different crystal space groups, P_{21} and P_{212121}. In the X1 structure, one molecule is surrounded by 14 neighbors (Figure 2.7C) and the loop comprising residues 25-29 of the central molecule is in close contact with the 81-87 loop in the translationally related neighboring molecule (Figure 2.7E). In the contact region, the side chains of Arg24 and Asn29 of one molecule engage in electrostatic interactions with Asn84 and Arg81 of the neighboring molecule. Another intermolecular interaction involves the 68-73 and 94-96 loops (not shown). Clearly, such crystal contacts will influence the observed fluctuations, causing slight suppressions in GNM-predicted motions, compared to those obtained for the isolated protein.

In X2, each individual molecule is surrounded by 12 neighbors (Figure 2.7D) and the 81-87 loop makes intimate backbone contacts with residues in β-strands 1 (7-13) and 6 (59-68) of the neighboring molecule. In particular, the backbone atom Cys83-O forms a hydrogen bond with Phe64-N, and Asn84-O with Leu12-N (Figure 2.7G). In addition, a number of side chain-

backbone interactions are observed, including Ala70-N and Glu22-Oε, Ala70-O and Arg20-Nε, and Gly69-O and Arg20-Nη. Clearly, such intimate interactions exert a significant effect of the fluctuations profile, and the experimentally observed suppression of residue motions in the crystal structure is reproduced by the GNM calculations performed for the X2 lattice.

2.4 CONCLUSION

The current work extends our previous analysis of NMR and X-ray structure parameters and GNM predictions.[47] In order to uncover the origin of the correlation between NMR data and computations, we undertook here a detailed analysis for a specific protein. We chose LKAMG, given its small size, high thermodynamic stability and its multiple structures solved in our laboratory to high resolution. We applied multiple experimental and computational methods to examine its structure and dynamics, allowing us to assess the limitations inherent to the different methodologies, and reconciling the apparent disparate data derived using different methodologies.

We previously suggested that the lower correlation between X-ray crystallographic B-factors and GNM results may be caused by the inaccessibility of large-scale motions in the crystal lattice, while solution NMR ensembles may inherently contain such motional characteristics.[47] Although compelling, the validity of this conjecture, and/or the contribution of other effects, had to be established. The present study provides data to that effect. Furthermore, in view of potential errors due to lack of specificity and nonlinear effects in the GNM predictions, we also compared the GNM results with MD simulations that use realistic force fields.

Our results show that the fluctuations profiles predicted by the GNM and observed in MD simulations exhibit a correlation of 0.64 ± 0.04 (comparable to the correlation between the

individual MD runs), despite their fundamental differences in terms of the underlying model (e.g., all atoms *vs.* only α-carbons, specific nonlinear potentials *vs.* nonspecific, linear potentials) and method (simulations *vs.* unique analytical solution). Strikingly, GNM exhibits even higher correlation than MD with the experimental data, suggesting that the improved accuracy of the mathematically 'exact' GNM method that takes into account the entire network of structural interactions more than counterbalances the lack of precision/specificity in the model. An important feature of elastic network models is their ability to capture the cohesiveness and cooperativity in the structures overall. This cohesiveness accounted for by the network connectivity appears to play a dominant role in defining the accessible motions. Since GNM results can be generated extremely rapidly, our data suggest that they can be securely and effectively used to assess the equilibrium dynamics of proteins. The relatively good correlation between the GNM results obtained for different conformers (N1, X1, X2) also support the notion that GNM is relatively insensitive to atomic details.

An interesting finding pertains to the crystallographic data. The GNM predictions did not exhibit comparable correlations with the B-factor of the two crystal structures, although both X-ray structures are of the same protein and were solved to similar resolution: the correlation with X2 B-factors was distinctively lower than that with X1 B-factors (Table 2.2). Likewise, all MD runs yielded poorer correlation with X2 data, pointing to an inherent feature of the X2 data. We therefore generated NMR-like ensembles of conformers, called X2-ensembles, using different sets of distance constraints extracted from the X-ray model. Three sets with 50-100% of the complete distance constraints were considered. The resulting MSD profiles exhibited distinctively better agreement with both GNM predictions and NMR data. This suggests that the inferior behavior observed for the X2 predictions originates from incomplete coverage of the

accessible conformational space. Examination of the crystal contacts in the X2 structure substantiates this conclusion and GNM calculations in the presence of crystallographic neighbors confirmed that the origin of discrepancy between theory/computations and experiments lies in crystal contacts.

Our results also lend credence to the view that ensembles of conformers, rather than unique structures, allow computational methods to assess equilibrium dynamics more accurately.[8, 38, 69, 70] Not surprisingly, higher accuracy comes at the expense of lower precision, paralleling the lack of precision in coarse-grained analytical approaches such as GNM compared to MD simulations. However, useful information on structural dynamics, otherwise inaccessible, can be extracted in this fashion.

3.0 MOLECULAR SIMULATIONS PROVIDE INSIGHTS INTO THE MECHANICS, BUT NOT THE TIME SCALES, OF PROTEIN MOTIONS UNDER EQUILIBRIUM CONDITIONS

This chapter is based on a recent study that has been submitted for publication in *Proteins*, and has been recently accepted for publication with minor revision. 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 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). The effective correlation times, on the other hand, tend to increase linearly with the total length of the simulations. Our results demonstrate that proteins possess 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. In contrast, experimental relaxation time scale and absolute size of the motions cannot be extracted unambiguously from current state-of-the-art atomic simulations in the submicroseconds regime.

3.1 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.[3, 35, 71] In terms of functional relevance, many structural rearrangements, especially those collectively involving large substructures, are necessary for proteins to carry out their chemical and biological activities.[1, 3, 71, 72] 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 can be 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 (GB3),[73] based on two alternative datasets: NMR relaxation parameters for probing motions on the order of nanoseconds[74] and residual dipolar couplings (RDCs) that probe motions on the microsecond time scale.[75] Notably, the order parameter profiles extracted from these two datasets exhibit similar shapes,[73] and the most 'disordered' residues, associated with the minima in the order parameter profiles plotted as a function of residue number (Figure 3.1A), become even more pronounced in the longer-time events. In

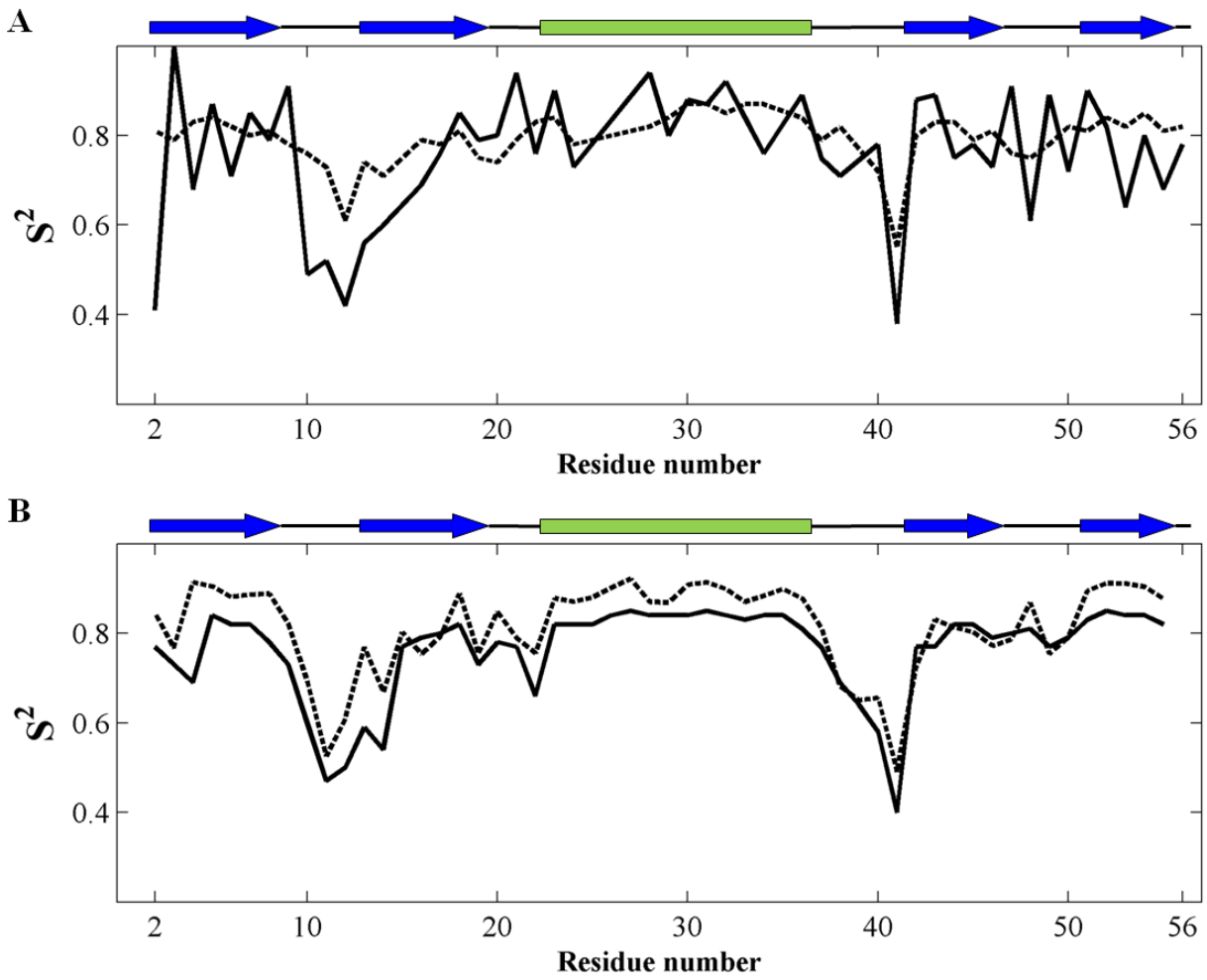

Figure 3.1 Experimental and computational literature data exhibit similar motional behavior for short and long times.

(A) Order parameters S^2 of GB3 extracted from NMR data: spin-relaxation (7), dashed black; and RDC (8), solid black. (B) Order parameters S^2 of GB1 extracted from MD simulations: 10 ns MD simulation (12), dashed black; and 175 ns MD simulation (12), solid black. Secondary structure elements are depicted at the top of each panel.

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.[76-78] Likewise, results for GB1 from two different length MD runs (Figure 3.1B**)** also demonstrate that the two simulations result in comparable order parameter profiles.[79] 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.[80] All these observations point to the existence of robust mechanism(s) of motions that dominate 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.[1] As mentioned above, a broad range of experimental techniques provides information on protein dynamics, including NMR relaxation measurements,[18, 19] Laue X-ray diffraction data,[20, 21] infrared and fluorescence spectroscopy,[22] and single-molecule studies,[23] although they inform about different aspects and time scales of protein dynamics. On the computational side, structure-based methods such as MD simulations[24] and normal mode analysis (NMA) with elastic network models (ENMs)[26-28, 45] 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.[81] The first arises from sampling inefficiency, which becomes increasingly noticeable in large molecular system.[81-83] Limitations of this nature can be alleviated to some extent by performing multiple independent runs for assessing convergence.[82, 84] Second,

the lengths of MD runs often remain below microseconds due to memory and computing time limitations.[81] Therefore, it still is an open issue whether functional motions at low frequencies can be inferred from relatively short MD runs. The present study was carried out 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?[85] (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 time scales and sizes of various mechanisms of motions?

Our results in combination with data reported previously for other systems, suggest that the distribution (or *relative* size) of residue fluctuations along the polypeptide chain, or the conformational mechanics, is a robust quantity under equilibrium conditions, predominantly defined by the 3-dimensional architecture in the native state, while their *absolute* size and effective correlation times predicted by MD simulations change with simulation duration, in the time regime (< 400 ns) investigated. The ratios for the observed mean-square displacements, $y = \langle(\Delta \boldsymbol{R})^2\rangle_{MDk} / \langle(\Delta \boldsymbol{R})^2\rangle_{MDk'}$ observed in two MD runs k and k' of different durations, and for the total simulation time, $x = t_{MDk} / t_{MDk'}$, are governed by a power law of the form $y = x^{0.26}$, similar to results reported by Scheraga and co-workers.[86, 87] The decomposition of the trajectories into essential modes revealed 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.

3.2 MATERIALS AND METHODS

3.2.1 MD simulations

The starting structure (PDB ID: 2EZM)[33] is highly anisotropic, occupying a volume of about 30 × 52 × 27 Å^3. We adopted 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,[4] and shown in previous work,[88] to satisfactorily solvate the protein. The resulting system consisted of 8,159 atoms, including 2,216 TOP3P water molecules. NAMD[66] with the Charmm22 force field[67] was used with a 2 fs time step. After energy minimization and equilibration, multiple independent runs were performed at constant temperature (298K) and pressure (1 atm).

3.2.2 Principal component analysis (PCA) 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 each snapshot from MD runs or each model in the NMR structure ensemble (where t is replaced by the model index). In order to identify global changes in configuration originating from the collective fluctuations sampled in each MD run, or associated with the structural deviations observed in NMR ensemble, the following steps are taken. 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 $\Delta\boldsymbol{R}_i(t)$ ($1 \le i \le n$) are organized in a trajectory matrix $\mathbf{A}$ of dimension 3n x m,

for a set of *m* snapshots. Multiplication of A by its transpose and division by *m* yields the $3n \times 3n$ covariance matrix **C** for each run (or for the NMR ensemble). **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

$$\mathbf{C}_{ij} = \begin{bmatrix} \langle \Delta X_i \Delta X_j \rangle & \langle \Delta X_i \Delta Y_j \rangle & \langle \Delta X_i \Delta Z_j \rangle \\ \langle \Delta Y_i \Delta X_j \rangle & \langle \Delta Y_i \Delta Y_j \rangle & \langle \Delta Y_i \Delta Z_j \rangle \\ \langle \Delta Z_i \Delta X_j \rangle & \langle \Delta Z_i \Delta Y_j \rangle & \langle \Delta Z_i \Delta Z_j \rangle \end{bmatrix} \quad (3.1)$$

Here, $<\Delta \boldsymbol{X}_i \Delta \boldsymbol{Y}_j>$ 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 **C** is performed, which produces $3n - 6$ nonzero eigenvalues and the corresponding eigenvectors. The eigenvectors define the directions of motions and the eigenvalues scale with the amplitudes.

3.2.3 GNM and ANM

The Gaussian Network Model (GNM)[26, 45] and anisotropic network model (ANM)[56, 89] analyses also lend themselves to a series of eigenmodes. In the GNM and ANM, the atomic structure could be simplified to a three-dimensional elastic network of *n* nodes (defined by positions of α-carbons), where *n* is the residue number. By assuming that the fluctuations of nodes are isotropic and Gaussian distributed, the Kirchhoff matrix is used to describe the connectivity of the network as below:

$$\mathbf{\Gamma}_{ij} = \begin{cases} -1 & if \ \ i \neq j \ and \ R_{ij} \leq r_c \\ 0 & if \ \ i \neq j \ and \ R_{ij} > r_c \\ -\sum_{i,i \neq j} \mathbf{\Gamma}_{ij} & if \ \ i = j \end{cases} \quad (3.2)$$

Here r_c is the cutoff distance that defines pairs of residues to be connected in the network. R_{ij} is the distance between node *i* and node *j*, $1 \leq i,j \leq n$. The inverse of $\mathbf{\Gamma}$ can be expressed in terms of the non-zero eigenvalues λ_k ($1 \leq k \leq n$-1) and corresponding eigenvectors $\boldsymbol{u}_k$ of $\mathbf{\Gamma}$ as $\Gamma^{-1} = \sum_{k=1}^{n-1} \lambda_k^{-1} \boldsymbol{u}_k \boldsymbol{u}_k^{\boldsymbol{T}}$, and the MSF of a given residue is the sum over the contributions of all modes

$\langle(\Delta R_i)^2\rangle = \sum_{k=1}^{n-1} \frac{3k_B T}{\gamma} (\lambda_k^{-1} u_k u_k^T)_{ii}$ where k_B is the Boltzmann constant, T is the absolute temperature, γ is a uniform spring constant, and the subscript ii designates the i^{th} diagonal element of the matrix enclosed in parenthesis.

For ANM, the $n \times n$ Hessian matrix is used, with each element $\mathbf{H}_{ij}$ is a 3×3 matrix that holds the anisotropic information:

$$\mathbf{H}_{ij} = \frac{\gamma \boldsymbol{r}_{ij}}{R_{ij}^2} \begin{bmatrix} X_{ij}X_{ij} & X_{ij}Y_{ij} & X_{ij}Z_{ij} \\ Y_{ij}X_{ij} & Y_{ij}Y_{ij} & Y_{ij}Z_{ij} \\ Z_{ij}X_{ij} & Z_{ij}Y_{ij} & Z_{ij}Z_{ij} \end{bmatrix} \quad (3.3)$$

The decomposition of **H** produces 3*n*-6 eigenvectors and their respective non-zero eigenvalues, where the eigenvectors describe the vibrational directions and the relative amplitudes of different modes.

3.3 RESULTS AND DISCUSSION

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

In our study we compared the dynamic information retrieved from 1 ns to 400 ns MD runs for the protein cyanovirin-N (CV-N).[33] 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.[90-93] CV-N's high stability at room temperature makes it a good candidate for performing extended simulations without the risk of significant structural changes or large conformational drift.[94] As depicted in the Figure 3.2 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

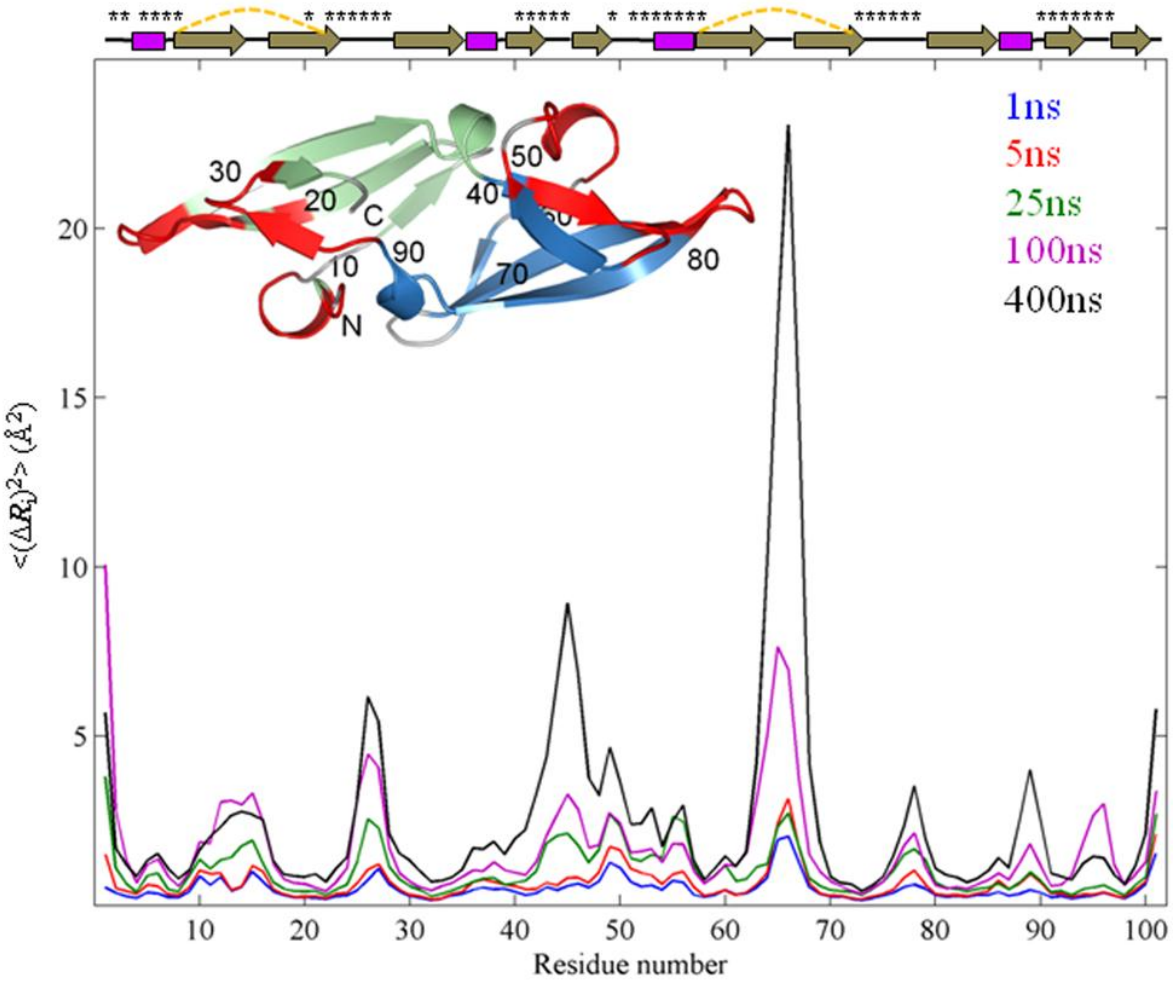

Figure 3.2 Mean-square-fluctuation profiles of CV-N from simulations with different durations.

The MSFs $<(\Delta \boldsymbol{R}_i)^2>$ in the residue positions are plotted along the polypeptide chain of CV-N. Averages 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.

positions (shown in red), one in each domain.[95] The two binding sites exhibit distinct affinities and specificities for high-mannose sugars.[96] The rotational correlation time τ_c of CV-N has been measured to be 4.5 ns.[97] Our simulations thus permit us to investigate both the sub-τ_c and supra-τ_c dynamics of CV-N under native state conditions.

Figure 3.2 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 Figure 3.2 represent the mean-square-fluctuations (MSFs) in residue positions, $\langle(\Delta \boldsymbol{R_i})^2\rangle$ 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 Figure 3.2 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 minimal changes in the relative sizes of the residue excursions. In principle, one might expect to detect new motional modes 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 3.1. The correlation coefficient between the MSFs for the 1 ns and the 400 ns runs is 0.83. Thus increasing the time window of observation by 4-5 orders of magnitude essentially leaves 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,[98] which also supports the view that the conformational mechanics are robustly maintained while the sizes of motions differ.

Table 3.1 Correlation coefficients between the MSFs of CV-N Residues observed in MD simulations[a] and those predicted by the GNM

cc of MSFs	$1ns_{avg}$	$5ns_{avg}$	$25ns_{avg}$	$100ns_{avg}$	$400ns_{avg}$
$5ns_{avg}$	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).

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 Figure 3.3. The RMSD remains around 3.7Å, 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).

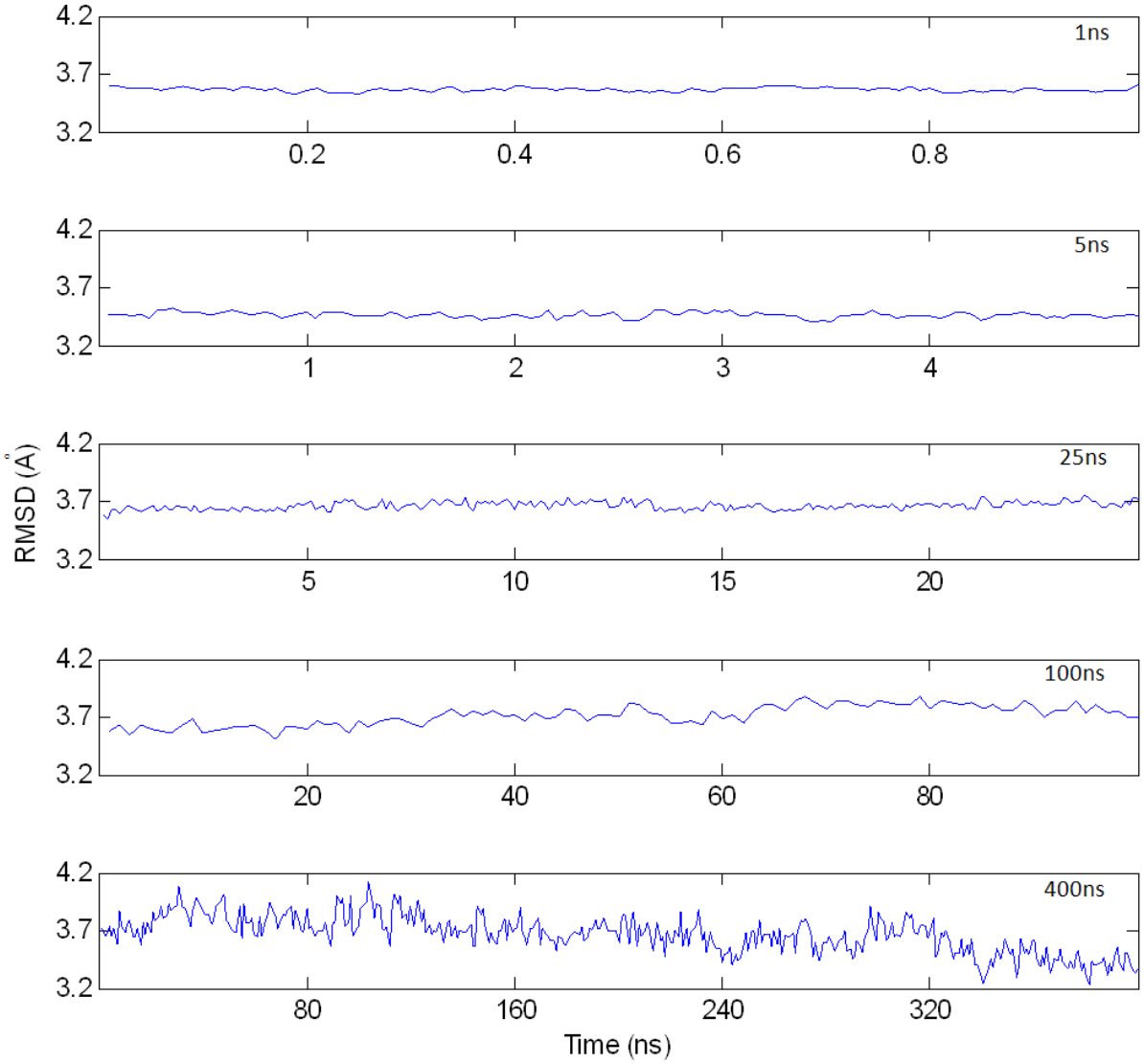

Figure 3.3 RMSD profiles for several simulation times.

3.3.2 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. Figure 3.4A 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 Figure 3.4C. The data points show the enhancements in the MSFs accompanying the increases in the simulations, also listed in Table 3.2, 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 Figure 3.4C. 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 can be extracted from the slope of the best fit and is 0.26. Thus, the overall dependence is

$$<(\Delta \boldsymbol{R})^2>_{MDk} / <(\Delta \boldsymbol{R})^2>_{MDk'} = (t_{MDk} / t_{MDk'})^{0.26} \qquad (3.4)$$

The subscript i in $<(\Delta \boldsymbol{R}_i)^2>$ has been removed since the MSFs refer to averages over all residues.

Equation 3.4 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.

Table 3.2 Scaling factors for MSFs between different MD runs[a]

Run1\Run2	***1ns$_{avg}$***	***5ns$_{avg}$***	***25ns$_{avg}$***	***100ns$_{avg}$***
5ns$_{avg}$	1.34			
25ns$_{avg}$	1.93	1.42		
100ns$_{avg}$	3.31	2.41	1.56	
400ns$_{avg}$	4.76	3.54	2.14	1.44

[a] See Figure 3.4C for the corresponding plot.

We note that the power law observed in present simulations (Eq. 3.4) applies to CV-N equilibrium dynamics near its native state, and 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 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.

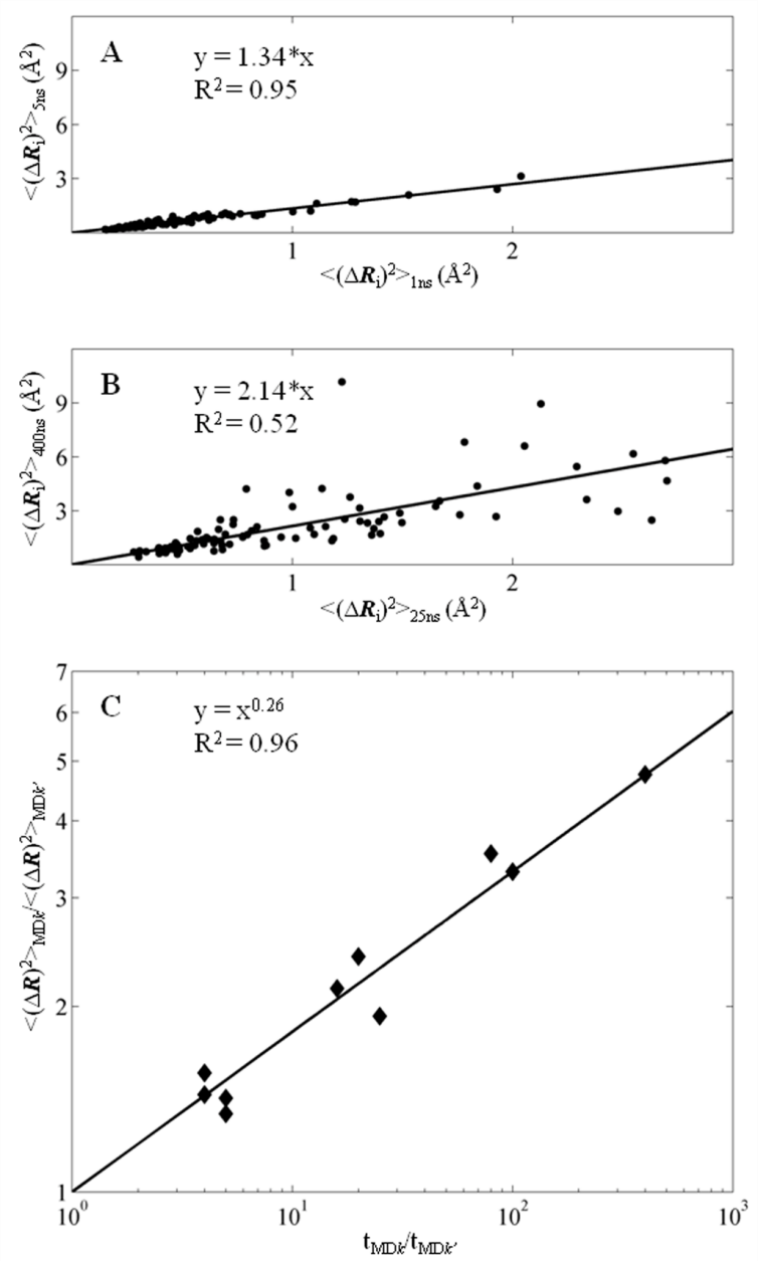

Figure 3.4 The magnitude of the fluctuations increases with increasing simulation time.

(A) and (B) Comparison of the mean-square fluctuations for different simulations. (A) $<(\Delta \boldsymbol{R}_i)^2>$ of residue *i* in the 5 ns simulation (y axis) is plotted against $<(\Delta \boldsymbol{R}_i)^2>$ of the same residue in the 1 ns simulation (x axis). (*B*) $<(\Delta \boldsymbol{R}_i)^2>$ of residue *i* in the 400 ns simulation (y axis) versus $<(\Delta \boldsymbol{R}_i)^2>$ 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.

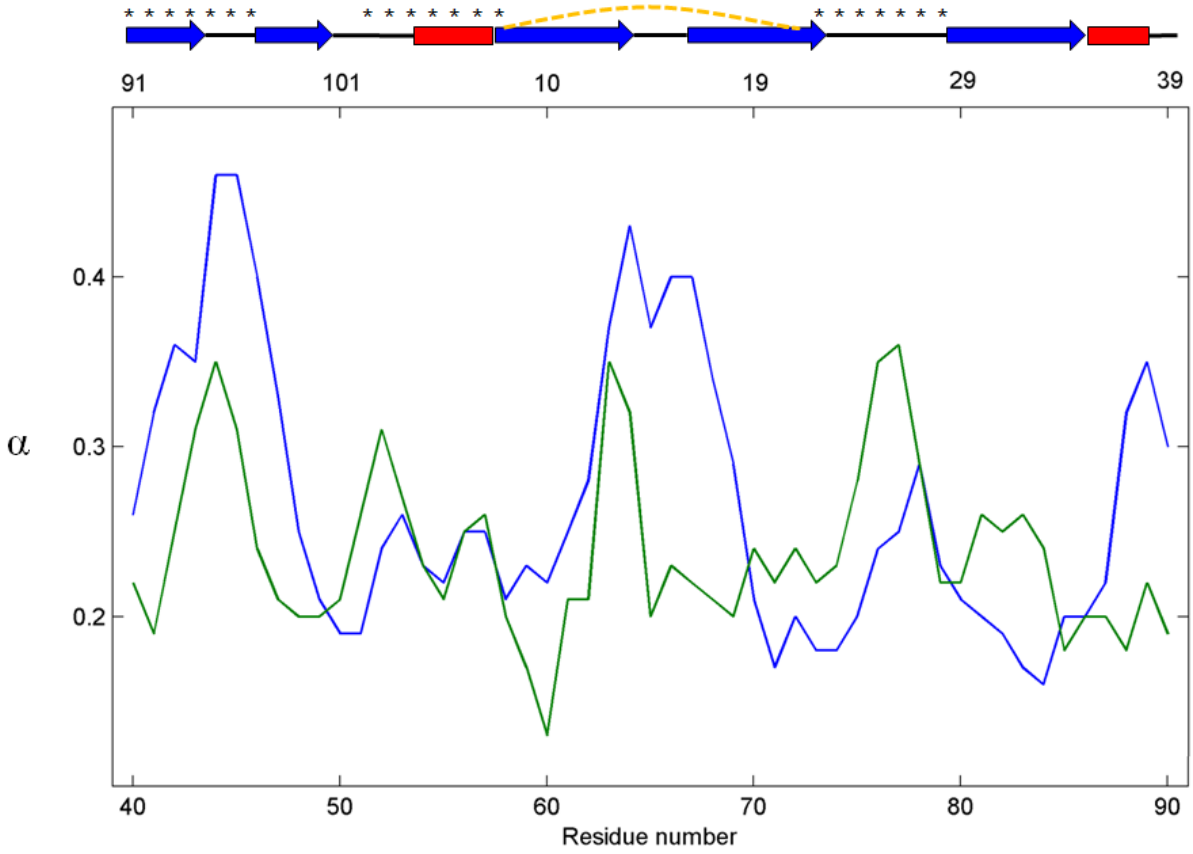

Figure 3.5 Power law exponents for the fluctuation size of CV-N residues as a function of simulation time.

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.

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 Figure 3.5). 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 finding is the observation that the two structurally similar, but distinct, domains of CV-N exhibit distinctive distributions of exponents. This 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[87, 99] and from the sampling of infrequent and large jumps between such local minima,[100] but also the bounded motion constrained by native contact topology in addition to covalent bonds.

3.3.3 Longer simulations yield larger correlation times

Next, we explored the time scales of observed motions. To this end, we evaluated the autocorrelation time τ_i for each residue in each run and averaged the results over all residues, and all runs of equal length to extract an *effective correlation time* for the protein for each simulation length. The autocorrelation time for residue *i* is obtained from the time decay of the time-delayed autorrelation function $<\Delta \boldsymbol{R}_i(t) \cdot \Delta \boldsymbol{R}_i(t+\Delta t)>$. Figure 3.6 illustrates the time decay of the autocorrelation function for Gln78, based on 1 ns, 5 ns, and 100 ns runs, on the respective panels A-C. The function decays exponentially at short Δt, and fluctuates before leveling off to zero (indicating the loss of any correlation). The correlation times extracted for Gln78 by fitting the

early portion (20%) of the profiles to a single exponential are 0.013 ns for the 1 ns simulation, 0.097 ns for the 5 ns simulation, and 3.64 ns for the 100 ns simulation, i.e., longer simulations yield larger correlation times. We further checked if the evaluated correlation times were stable by comparing the results from fitting the first 5%, 20%, and 60% portions of the autocorrelation decay curves (Figure 3.6C), to find out that the fluctuation of the observed time was acceptable. Calculations were repeated for all residues and all runs to obtain highly robust values for the effective correlation time of the protein for each simulation length. The resulting average scaling factors, by evaluating the ratios of effective correlation times for all pairs of simulation lengths, are listed in Table 3.3 and plotted in Figure 3.6D. Paralleling the increase in the motional amplitudes with increasing simulation time, the correlation times also increase. However, this increase exhibits a near linear dependence. Least square fitting to the results shown in Figure 3.6D yields an exponent of 1.1. Notably, the correlation times (either τ_e extracted from the original model-free approach, or the fast and slow dynamics correlation times, τ_f and τ_s based on the extended model-free approach with four parameters) reported by Bui et al. for their MD simulation of GB1[79] also exhibited a dependence on simulation time, with the correlation time of the 175 ns simulation being 7.8 or 17.5 times larger than that of the 10 ns simulation (using original or extended model-free approach), suggesting that it is not possible to make an unambiguous assessment of the absolute time scale of configurational relaxation motions based on the correlation times observed in MD simulations up to hundreds of nanoseconds.

In principle, the representation of the time-delayed autocorrelation functions' decay by a single exponential is an approximation that overlooks the multitude of motions/modes effectively controlling the dynamics. However, performing this analysis for each individual residue yields a

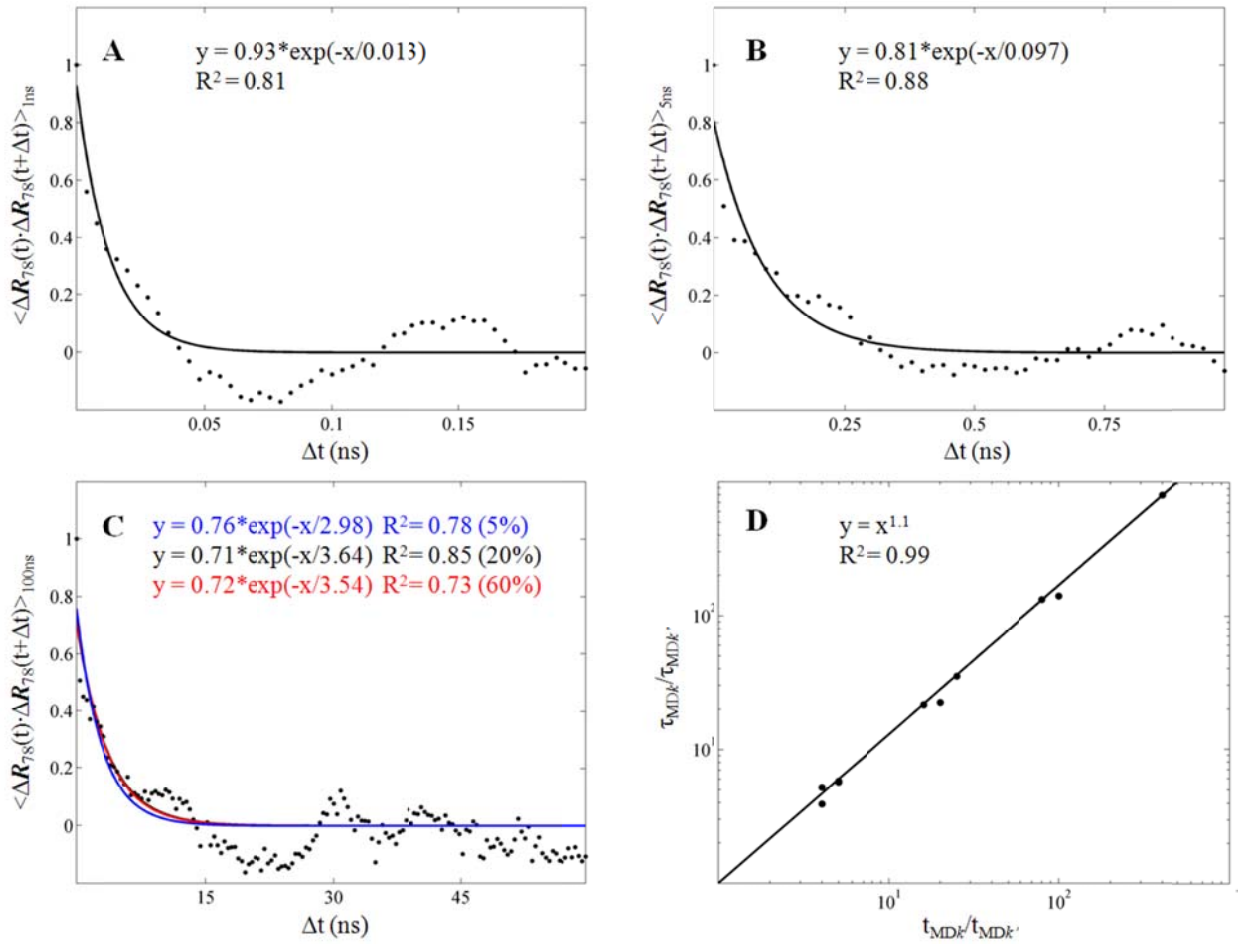

Figure 3.6 The autocorrelation time τ increases with the simulation duration.

(A) The first 20% time-delayed autocorrelations for the fluctuations of residue Gln78 in a 1 ns simulation are plotted against the delay time Δt. (B) The first 20% time-delayed autocorrelations for the fluctuations of residue Gln78 in a 5 ns simulation are plotted against the delay time Δt. (C) The first 60% time-delayed autocorrelations for the fluctuations of residue Gln78 in a 100 ns simulation are plotted against the delay time Δt. Single exponential fitting the first 5%, 20%, and 60% of the fluctuation correlations in the 100 ns simulation are colored blue, black, and red, respectively. (D) The ratios of τ among different simulations are plotted against the corresponding ratios of simulation durations.

range of correlation times in accord with the multiplicity of operating modes which differentially affect the individual residues. This permits us to include, albeit indirectly, the contributions from

different modes in the evaluation of the effective correlation times. Application of the same procedure to all runs, and then taking averages over all runs of a given length, provided us with a consistent metric, called effective correlation times.

Table 3.3 Scaling factors for autocorrelation time (τ) between different MD simulations[a]

Run1\Run2	***1ns$_{avg}$***	***5ns$_{avg}$***	***25ns$_{avg}$***	***100ns$_{avg}$***
5ns$_{avg}$	5.8			
25ns$_{avg}$	35.3	5.7		
100ns$_{avg}$	140.4	22.4	3.9	
400ns$_{avg}$	797.2	132.5	21.6	5.2

[a] See Figure 3.6D for the corresponding plot.

3.3.4 Comparison of essential modes extracted from different MD runs

As a further test, we examined the principal motional modes inferred from simulations of different lengths. To this end, we decomposed the CV-N motions that were sampled in each MD run into a series of collective modes, each ranked by their weights. We next focused on the top-ranking modes, also called global or essential modes, since these are usually the most collective modes and numerous applications have shown their relevance to biological function.[71]

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 Figure 3.7, 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.

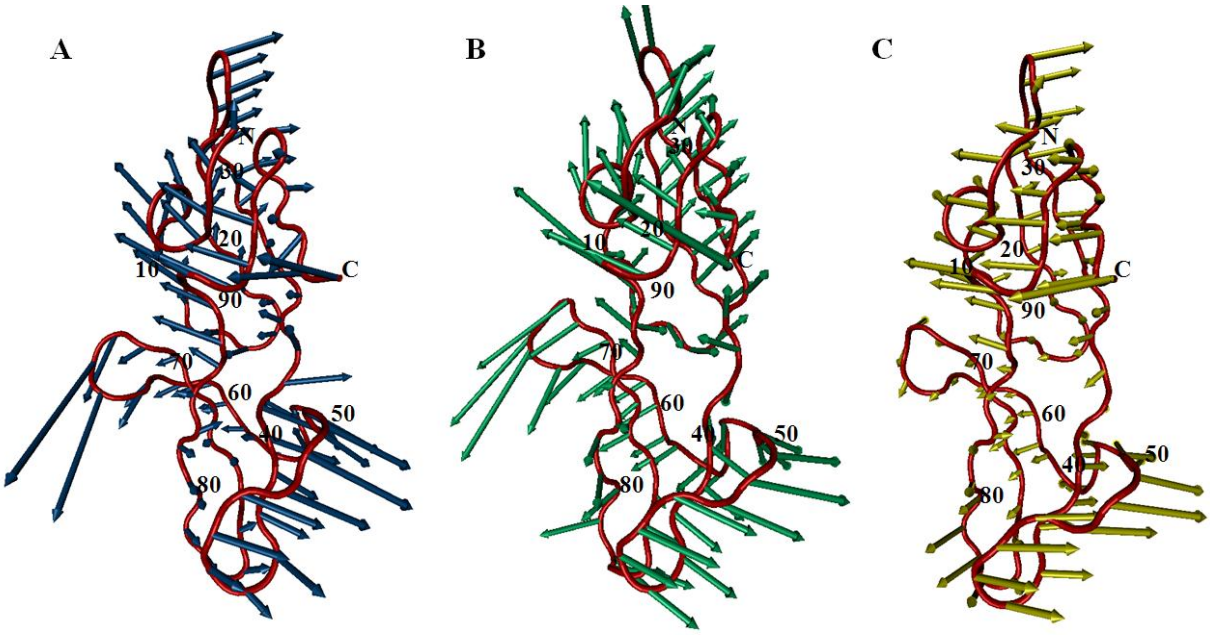

Figure 3.7 The shared global mode between theory and simulations.

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 10^{th} residue.

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 3.5. 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. As a further analysis, we combined trajectories from all individual runs with the

same duration, and compared the principal modes of motions computed for different time scales. The results compiled in Table 3.4 also confirm that the directions (not the size) of the global motions are reproducible and conserved across runs of various lengths, although the orders of the modes may shift in some cases. It thus is important to carry out multiple simulations and subject the compiled data to mode decomposition in order to detect the 'consensus' global modes and extract information on collective mechanics.[101]

Table 3.4 Shared global modes between MD simulations[a], NMR structural ensemble, and ANM predictions[b]

cc of Modes	***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.

[b] 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).

Given that the top-ranking modes of long simulations can be extracted to a good approximation from short simulations, 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

Table 3.5 Shared modes between 400 ns simulations and shorter simulations

Runs	400ns-01				Runs	400ns-01			
	Mode 1		Mode 2			Mode 1		Mode 2	
1ns-02	1	0.61	-	-	**5ns-01**	-	-	2	0.63
1ns-03	3	0.61	-	-	**5ns-02**	-	-	3	0.73
1ns-05	1	0.77	-	-	**5ns-03**	-	-	3	0.66
1ns-08	-	-	2	0.70	**5ns-05**	1	0.62	2	0.64
1ns-10	2	0.68	1	0.72	**5ns-06**	1	0.63	2	0.64
1ns-11	1	0.60	1	0.62	**5ns-07**	2	0.67	-	-
1ns-12	-	-	3	0.72	**5ns-09**	3	0.71	2	0.71
1ns-14	1	0.73	-	-	**5ns-10**	1	0.71	3	0.71
1ns-16	5	0.68	1	0.61	**5ns-11**	-	-	2	0.72
1ns-18	2	0.60	-	-	**5ns-12**	1	0.77	3	0.75
$1ns_{comb}$	-	-	2	0.71	**$5ns_{comb}$**	2	0.76	4	0.64
25ns-01	1	0.78	-	-	**100ns-01**	-	-	4	0.63
25ns-03	-	-	3	0.65	**100ns-03**	3	0.62	-	-
25ns-04	6	0.61	-	-	**100ns-04**	-	-	5	0.61
25ns-05	1	0.61	2	0.61	**100ns-05**	-	-	1	0.65
25ns-06	3	0.60	-	-	**100ns-06**	-	-	2	0.63
25ns-07	1	0.68	3	0.75	**100ns-08**	2	0.60	-	-
$25ns_{comb}$	2	0.61	-	-	**$100ns_{comb}$**	3	0.80	-	-

Runs	400ns-02				Runs	400ns-02			
	Mode 1		Mode 2			Mode 1		Mode 2	
1ns-01	1	0.62	-	-	**5ns-14**	-	-	1	0.64
1ns-05	-	-	1	0.62	**5ns-15**	-	-	1	0.62
1ns-12	-	-	1	0.61	**5ns-16**	-	-	1	0.72
1ns-15	5	0.64	-	-	**25ns-01**	2	0.62	4	0.64
1ns-17	2	0.60	3	0.64	**25ns-03**	-	-	1	0.76
$1ns_{comb}$	-	-	1	0.72	**25ns-04**	-	-	1	0.75
5ns-02	-	-	1	0.63	**25ns-05**	-	-	1	0.68
5ns-04	2	0.65	3	0.62	**$25ns_{comb}$**	-	-	2	0.88
5ns-05	2	0.60	1	0.73	**100ns-01**	4	0.65	-	-
5ns-08	-	-	1	0.60	**100ns-02**	-	-	3	0.69
5ns-09	-	-	1	0.67	**100ns-03**	2	0.62	-	-
5ns-11	2	0.69	1	0.69	**100ns-04**	-	-	3	0.66

[a] The independent runs are indexed by the duration of the simulation, followed by the simulation number. For example, there are twenty 1 ns simulations, the first indicated as 1ns-01 and the last as 1ns-20. Likewise, we have sixteen 5 ns runs, twelve 25 ns, etc. Only those runs that exhibit shared modes are listed. The results from combing the multiple individual runs of the same duration are highlighted by gray.

[b] The upper and lower parts of the table refer to two independent runs of 400 ns each. The correlating global mode (mode 1-6) is listed provided that a correlation cosine of 0.6 or more is observed with the mode 1 or 2 of the 400 ns simulation.

suggests that the molecule tends to move along the same direction, or samples the same subspace, in both cases, although the amplitudes of the displacements differ. All the observations made here 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.

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

Table 3.6 Shared modes between ANM prediction and different MD simulations

ANM / MD	Mode 1		Mode 2		ANM / MD	Mode 1		Mode 2	
1ns-01	3	0.76	-	-	**5ns-12**	-	-	1	0.70
1ns-03	-	-	3	0.73	**5ns-13**	4	0.65	3	0.60
1ns-05	3	0.72	1	0.69	**5ns-15**	-	-	2	0.68
1ns-07	2	0.63	-	-	**5ns-16**	-	-	4	0.64
1ns-08	2	0.67	1	0.77	**5ns$_{comb}$**	-	-	2	0.60
1ns-09	4	0.64	2	0.81	**25ns-01**	4	0.63	1	0.68
1ns-10	-	-	2	0.71	**25ns-02**	4	0.66	-	-
1ns-14	-	-	1	0.67	**25ns-07**	-	-	1	0.75
1ns-15	2	0.68	1	0.61	**25ns-10**	-	-	3	0.63
1ns-16	6	0.60	-	-	**25ns$_{comb}$**	-	-	3	0.61
1ns-17	3	0.66	-	-	**100ns-01**	-	-	3	0.60
1ns-20	1	0.67	-	-	**100ns-04**	2	0.66	4	0.69
5ns-02	-	-	2	0.65	**100ns-05**	-	-	3	0.61
5ns-05	5	0.66	-	-	**100ns-06**	-	-	6	0.62
5ns-07	-	-	2	0.75	**100ns$_{comb}$**	-	-	3	0.60
5ns-08	6	0.68	-	-	**400ns-01**	-	-	1	0.64
5ns-09	-	-	3	0.64	**400ns-02**	2	0.61	-	-
5ns-10	-	-	1	0.72					

[a] Same indexing as Table 3.5 is adopted to label the runs.

[b] The correlating global mode (mode 1-6) is listed provided that a correlation cosine of 0.6 or more is observed with the mode 1 or 2 predicted by the ANM.

As further verification of the relevance of our findings to CV-N dynamics, we performed the GNM[26, 45] analysis of the PDB structure 2EZM, the NMA[46, 102] of the same structure using the ANM,[89] and the PCA of the NMR ensemble of 40 structural models for CV-N.[33] ANM modes have been observed in previous studies to correlate with the structural dynamics intrinsically accessible to enzymes[71, 103, 104] and with the microseconds dynamics of G-protein coupled receptors.[83] The distribution of NMR models also provides information on structural variabilities, which may be compared to those observed in MD runs.[4, 52, 105]

The correlations between the distribution of MSFs predicted by the GNM, $<(\Delta \boldsymbol{R_i})^2>_{GNM}$, and those observed in different MD runs are presented in Table 3.1. The correlations vary from 0.60 (with $<(\Delta \boldsymbol{R_i})^2>_{100ns,avg}$) to 0.74 (with $<(\Delta \boldsymbol{R_i})^2>_{25ns,avg}$). 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[4] between GNM-predicted MSFs and the MSFs inferred from multiple 10 ns MD simulations. The results presented in Figure 3.7C, Table 3.4, and Table 3.6 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, again suggesting the global motions observed in MD simulations and those predicted by coarse-grained models such as the ANM share robust features uniquely encoded by the equilibrium structure. Table 3.4 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 3.4), presumably due to the fact that there are only 40 models in the NMR ensemble, which may provide an incomplete description of the accessible reconfigurations. The level of agreement appears to decrease with increasing simulation duration, which may be due to the inadequate sampling of the accessible (larger) conformational subspace by fewer independent runs. The above results emphasize the importance of performing a sufficient number of independent runs in order to ensure complete coverage and adequate sampling of accessible conformers.

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.[8, 79] In the case of CV-N, the experimentally measured τ_c is 4.5 ns.[97] 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 Figure 3.7). Furthermore, they exhibit reasonable agreement with the results from ANM calculations, and NMR data, which also supports the robustness of the results from simulations.

3.4 CONCLUSION

In the present work, we have analyzed amplitudes, correlation times, 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 and correlation times 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 time and length 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

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. We suggest that performing multiple simulations should be considered as a key strategy for identifying

consensus modes and that PCA may help test the convergence and conservation of collective motions in a given protein.

4.0 BIOINFORMATIC ANALYSIS OF DOMAIN-SWAPPED PROTEINS

Work discussed in this chapter has been accepted for publication as a chapter in *Comprehensive Biophysics*, 2012. Among thousands of homo-oligomeric protein structures, there is a small but growing subset of 'domain-swapped' proteins. The term 'domain swapping', originally coined by D. Eisenberg, describes a scenario in which two or more polypeptide chains exchange identical units for oligomerization.[12] This type of assembly could play a role in disease-related aggregation and amyloid formation or as a specific mechanism for regulating function, and hence it is important to understand how proteins perform domain swapping. Although a lot of effort has been directed towards analyzing domain swapping, no unifying molecular mechanism of domain swapping has emerged to date. We compiled all domain-swapped protein structures in the PDB, performed a detailed examination of the common/different features of the chains in our collection and summarized ideas about putative mechanisms. Results from this analysis, for instance with respect to chain lengths, structural classification or amino acid composition, did not reveal any special properties associated with domain-swapped proteins or the exchanged domains. The diversity of sequences and architectures suggests that almost any protein may be capable of undergoing domain swapping and that domain swapping maybe solely a specialized form of oligomer assembly. On the other side, structure-based computational analysis, i.e., GNM, on the monomeric conformations of our collection suggested that native contact and topology information alone is not sufficient for uncovering hinge residues in our diverse set of domain-swapped proteins.

4.1 INTRODUCTION

It is generally accepted as a central truth in biochemistry that the amino acid sequence of a protein encodes all necessary information for the chain in a given environment to fold into a single, well-defined stable structure.[106] For most proteins, this structure is under physiological conditions the native, functional state. Under certain circumstances, however, proteins may be able to fold into distinctly different structures, and over the past few years, increasing numbers of alternative folds are being discovered. Lymphotactin[107] and Mad2 (the mitotic arrest deficiency 2 protein)[108] are extreme examples of this type.

The most common alternative structures comprise different multimeric assemblies of identical polypeptide chains. Multimers are endowed with structural and functional advantages, such as improved stability and control over the accessibility and specificity of active sites, explaining why oligomerization is favored during protein evolution.[109] Special cases of multimers are the so-called morpheeins, homo-oligomeric proteins that can switch their structure between functionally distinct alternate quaternary states. The prototypical example of a morpheein is the enzyme porphobilinogen synthase (PBGS) which exists in an equilibrium between an octamer, a hexamer, and two dimer conformations.[110] Another special case of oligomerization has been described as '3D domain swapping'.[32] A 'domain-swapped' structure contains two or more polypeptide chains that exchange identical units. The exchanged portion may consist of a single secondary structure element or an entire globular domain. If exchange is reciprocal between two monomers, dimers are formed, or, if more chains are involved, oligomers ensue.

Folding into the native state is driven by a combination of entropic and enthalpic forces that result in burial of hydrophobic residues in the interior and exposure of polar residues on the

surface of the protein. This network of defined attractive and repulsive forces arranges the chain in well-defined, secondary structure elements. In multimers, each single polypeptide chain usually adopts the same conformation, although the assembly of individual chains in the oligomer can vary. Often, small changes in protein composition or environment can tip the balance from one arrangement to the next, with some proteins coexisting in more than one oligomeric state. A classic example of alternate oligomers is the Bence-Jones protein, characterized by X-ray diffraction more than 40 years ago. This protein exists in the crystal in three quaternary structures[111] that vary in their domain interactions.

By Jul 2010, PDB[112] contains 28723 homo-oligomeric protein structures. The most commonly found assembly patterns are 'side by side' and 'head-to-tail', but domain-swapped structures are becoming a sizeable fraction. In the current review not all oligomeric structures are considered; we solely concern ourselves with proteins for which domain swapping has been described.

The term 'three-dimensional (3D) domain swapping', or simply domain swapping, was originally coined by Eisenberg and colleagues for describing the X-ray structure of a diphtheria toxin (DT) dimer in 1994.[113] However, already in 1962 a report was published describing the exchange of an N-terminal fragment for bovine pancreatic ribonuclease A (RNase A) upon dimerization.[114] The first protein X-ray structures that contained domain-swapped elements were determined in the early 1980s,[115-117] with more and more structures of domain-swapped multimers following suit (Figure 4.1).

In the following, we will report on domain-swapped structures we derived from the PDB, summarize ideas about putative mechanisms for this type of oligomerization, and describe a few

examples in detail for which domain swapping may be important for regulating function or triggering disease.

Figure 4.1 Growth in domain-swapped structures deposited in the PDB.

Proteins with identical sequences for monomeric and oligomeric states are shown in red; proteins that share > 90% sequence identity between the monomer and oligomer are shown in green, and proteins for which swapped structures have been described without monomeric counterparts are shown in blue.

4.2 GENERAL ASPECTS

4.2.1 Dataset of domain-swapped proteins

Currently, more than 100 domain-swapped structures are deposited in the PDB, with 38 examples for which both monomeric and oligomeric structures are available (Table 4.1 and 4.2). These 38 proteins are non-related and exhibit < 20% pairwise sequence identity. Among them, 19 cases exist with identical sequences for monomeric and oligomeric states, thus they are

examples of true domain swapping. The other 19 share > 90% sequence identity between the monomer and oligomer polypeptide, some involving single amino acid mutations. Not surprisingly, structures for most domain-swapped oligomeric proteins have been determined by X-ray crystallography.

Analysis of the chain lengths, structural classification or amino acid composition, does not reveal any special properties associated with domain-swapped proteins. In our dataset, the shortest protein is the immunoglobulin binding domain B1 of streptococcal protein G (GB1)[118] which comprises only 56 residues, and the longest one is DT[119] with 535 amino acids. The ratio of all α proteins, all β proteins and mixed α/β proteins for domain-swapped proteins is 2:2:5, identical to the ratio reported for all structures in Structural Classification of Proteins (SCOP),[120] and there appears to exist no specific amino acid requirements for domain-swapped proteins, compared to overall protein space.

Similar findings hold when examining only the exchanged domains. They exhibit different sizes, ranging from a few residues to more than 100 amino acids. Single α-helix or β-strand can be swapped, bundles of α-helices or β-hairpins are found exchanged and even mixed α-helix and β-strand elements can serve as the swapped domain, without any discernable sequence signature among them.[32] Although, the exchanging unit can be located anywhere in the sequence, it is often found at one of the two termini. Human antithrombin III is an example in which the exchanged domain resides in the middle of the protein; this kind of exchange has also been termed 'hairpin insertion'.[121] An example in which almost one half of the entire polypeptide chain is exchanged is cyanovirin-N (CV-N).[33]

Taken together, the above analysis reveals that proteins found in domain-swapped structures display the same diversity as any protein in the PDB. This suggests that almost any

protein may be capable of undergoing domain swapping and that domain swapping is solely a specialized form of oligomer assembly.

Table 4.1 Proteins for which monomeric and swapped oligomeric structures are available for the identical polypeptide sequence.

Protein	PDB ID Monomer[a]	PDB ID Oligomer	Polypeptide length[b]	Hinge location[c]	exchanged element(s)	References
Syntaxin TLG1	2C5K	2C5J	95	65-69	helix	[122]
VAMP-7	2VX8	2VX8	169	40-45[e]	helix	[123]
spo0A	1QMP	1DZ3	130	107	helix	[124, 125]
Barnase	1BRN[d]	1YVS	110	37-41	helices	[126, 127]
FOXP2	2A07	2A07	93	538, 544	helices	[128]
Bcl2-L-1	1R2D	2B48	218	158-159	helices	[129, 130]
trpR	1P6Z[d]	1MI7	107	64-67, 76-78	helices	[131, 132]
CD47	2JJS	2VSC	127	101-102	β-strand	[133]
DAP-150	2HKQ	2HKN	97	37-40	β-strand	[134]
LB1	1K50	1K50	63	52-56	β-strand	[135]
cspB	1C9O	2HAX	66	37	β-strands	[136, 137]
CV-N	2EZM	3EZM	101	50-54	β-strands	[33, 138]
ATIII	1ATH	2ZNH	432	338-339, 390-406[e]	β-strands	[121, 139]
RNase A N-swap C-swap	5RSA	1A2W 1F0V	124	19-20 112	helix β-strand	[140-142]
ASP1	3BFB	3CYZ	119	13	β-strand	[143, 144]
yopH	1M0V	1K46	136	28-29	mixed	[145, 146]
Cystatin-A	1DVC	1N9J	98	48-50	mixed	[147, 148]
ptsH	1Y51	1Y50	88	54	mixed	[149]
DT	1MDT	1DDT	535	379-386	domain	[113, 119]

[a] Some structures are not available as isolated monomers.

[b] Sequence information was obtained from the FASTA file in the PDB. Coordinate information may be not available for all residues in the PDB file.

[c] Hinge residues are numbered according to the monomer PDB file; these numbers may differ between monomer and dimer.

[d] No monomeric structure is available. The comparison is carried out for the monomer unit in a non-swapped dimer or oligomer (see text for details).

[e] The protein contains a cleaved peptide bond in the hinge region or has no coordinate information in the PDB file.

Table 4.2 Proteins for which monomeric and swapped oligomeric structures are available for closely related polypetide sequences[a].

Protein	PDB ID Monomer[b]	PDB ID Oligomer	Polypeptide length[c]	mutation; extension[d]	Hinge location[e]	exchanged element(s)	References
TRX	2O7K	3DIE	107	1; 1	27-30	mixed	[150, 151]
CABP	1N65	1HT9	75	1; 1	42-45	helices	[152, 153]
CD2	1T6W	1CDC	99	3; 0	45-46	β-	[154, 155]
Rab27b	2ZET	2IF0	203	0; 3	43, 77	β-strands	[156, 157]
GRB2	1BM2	1FYR	117	1; 2	121-122	mixed	[158, 159]
GB1	1GB1	1Q10	56	4; 0	38-41	β-strands	[118, 160]
OBP	2HLV	1OBP	160	4; 0	121-122	mixed	[161, 162]
PrP^{C}	2W9E	1I4M	113	0; 5	190-197	helix	[163, 164]
HasA	1YBJ	2CN4	178	0; 5	48-50	mixed	[165, 166]
iNOS	1M8D[f]	1QOM	434	0; 6	104	mixed	[167, 168]
TNase	1SNC	1SND	149	6; 0	112-120[g]	helix	[169, 170]
GR	3BQD	3H52	255	7; 0	547-552	mixed	[171, 172]
Trk-A	1WWW	1WWA	101	0; 8	297	β-strand	[173, 174]
IL-10	1LK3	1ILK	160	6; 3	107-114	helices	[175, 176]
HDGF	1RI0	2NLU	110	0; 10	34-41	β-strands	[177, 178]
CA-CTD	2KOD[f]	2ONT	70	2; 12	177	helix	[179, 180]
EMMPRIN	3B5H[f]	3I84	184	>15	93-94	β-strand	[181, 182]
RGS7	2D9J	2A72	139	>15	100	helix	[183, 184]
afaD	2IXQ	2AXW	142	>15	116-130	β-strand	[185, 186]

[a] The monomeric and swapped oligomeric structures for each pair are in the same entry in Uniprot.[187]

[b] Some structures are not available as isolated monomers.

[c] Sequence information was obtained from the FASTA file in the PDB. Coordinate information may be not available for all residues in the PDB file.
[d] Sequence information was obtained from the FASTA file in the PDB. The polypeptide lengths in the pairs are different. Some such cases, for instance HasA is indeed a *bona fide* example of domain swapping.
[e] Hinge residues are numbered according to the monomer PDB file; these numbers may differ between monomer and dimer.
[f] No monomeric structure is available. The comparison is carried out for the monomer unit in a non-swapped dimer or oligomer (see text for details).
[g] The protein contains a cleaved peptide bond in the hinge region or has no coordinate information in the PDB file.

4.2.2 Mechanistic considerations

Comparison between the closed conformation of the monomeric polypeptide chain and the open conformation of the same chain in the domain-swapped dimer implies that the observed large conformational differences most likely require some kind of un/refolding. Intra-molecular interactions involving hydrophobic contacts, hydrogen-bonding, electrostatic interactions, and even disulfide bridge interactions[163, 188, 189] at the closed interface in the monomer are exchanged to inter-molecular interactions. Naturally, such breaking and reforming of contacts requires energy, the activation energy for 3D domain swapping.[190] In order to overcome the activation barrier between the monomer and dimer, changes in environment, in particular conditions that favor unfolding, may play a role.

For proteins capable of domain swapping, folding from the unfolded polypeptide chain can lead, in principle, to either the closed monomer or the domain-swapped dimer. Partitioning between the two products is determined by their free energy difference. This difference is naturally very small, given that all interactions within the two structures are extremely similar;

only the hinge-loop conformation is distinct. Therefore, any free energy difference needs to be traced to the hinge-loop, which can either introduce or relieve strain during monomer-dimer interconversion.

4.2.2.1 The hinge-loop The hinge-loop is the only region of the protein that adopts a different conformation in monomeric and domain-swapped structures. Therefore, sequences and secondary structures have received considerable attention in the search for local signals that could cause or influence domain swapping.

Several studies show that altering the length of the hinge-loop can switch the domain swapping propensity of a protein. Intuitively, one would expect that long loops preferentially result in monomers and short ones in dimer structures: a short loop will make it difficult for the polypeptide to fold back on itself, and in turn allow the swapped portion of the chain to find partners more easily. This clearly is the case in staphylococcal nuclease.[169] The only sequence difference between the monomer and domain-swapped dimer is the loop length, with the monomer loop containing 6 more residues than the hinge in the dimer. Loop residue deletion has also been used in some designed proteins. An elegant example illustrating the importance of loop length is provided by two different three helix bundles that were engineered in the Eisenberg laboratory.[191] Loop deletion in one of these caused the formation of a domain-swapped dimer whereas loop deletion in the other resulted in fibril formation. On the other hand, Perutz and colleagues found that adding a stretch of polyglutamines into the active site loop of Chymotrypsin Inhibitor 2 caused domain swapping and higher order oligomer formation.[192] Indeed, in this case, oligomerization increased with increasing loop lengths. Therefore, a universal statement regarding the influence of hinge-loop length cannot be made at present.

Not every amino acid in the hinge-loop region has to change conformation. Sometimes it is one or two residues for which the alternative conformation is observed. These could be the key hinge amino acids and only their backbone phi and psi angles may have to change between monomer and dimer conformations. In our dataset, alanine and glycine are the most frequent amino acids in these key hinge positions, with their occurrence being much higher than commonly found. Glycine can adopt phi and psi angles in all four quadrants of the Ramachandran plot, due to the lack of a side chain; therefore it is possible to accommodate a glycine in any kind of turn, even quite sharp ones, that are sterically forbidden for other residues. For the cold shock protein cspB,[136] a flip in the backbone of G37 ($\Delta\varphi \approx 180^{o}$) is observed between monomer and domain-swapped dimer. Similarly, the small alanine residue is also more tolerant in terms of steric effects and in the N-terminal swapped dimer of RNase A only two adjacent alanines change their conformation compared to the monomer structure.

In the middle of hinge-loop sequences one also finds conserved prolines.[193] Since proline residues are thought to impart rigidity to the polypeptide backbone, Rousseau and colleagues suggested for the cyclin-dependent kinase regulatory subunit suc1[194] that the proline-caused strain in the hinge-loop influences domain swapping. Indeed, replacement of the first proline in the hinge with an alanine stabilized the monomer form, whereas the same substitution of the second proline stabilized the dimer form. The authors suggest that tension in the hinge-loop in the monomer caused it to behave like a loaded molecular spring which is released when the alternative conformation is adopted in the dimer.[194] Unlike in suc1, mutation of the single proline in the hinge-loop of CV-N to glycine, substantially stabilized both states of the protein, with greater stabilization of the monomer compared to the dimer.[195] Furthermore, adding a second proline residue by mutating a neighboring amino acid causes the domain-swapped dimer to

become the thermodynamically most stable state.[195] Similarly, the change of alanine in the hinge-loop of the FOXP2 to proline prevented the formation of the swapped dimer.[128] This suggests that the addition or deletion of prolines creates no uniform outcome and that each protein may have its unique signature of hinge-loop residues.

Aside from glycine, alanine and proline, other amino acids in the hinge-loops could also play a role in stabilizing particular secondary structure elements in the swapped domains. For example, a hinge-loop could be a coil in the monomer form, but become embedded into a long β-strand or an α-helix. This could stabilize the dimeric forms of these proteins, given the higher degree of secondary structure and the elimination of a flexible hinge region.

For a region in the protein to function as a hinge-loop, it needs to be pliable enough to adopt different conformations. RNase nicely illustrates this point. RNase A,[140] bovine seminal ribonuclease (BS-RNase)[196] and a human pancreatic ribonuclease (hRNase) chimera[197] share > 60% sequence identity and all three proteins undergo domain swapping of their N-terminal helices, albeit with different relative orientations of the helix and different conformations in the three hinge-loops. As an aside, RNase A is also one of the rare examples that can swap either N- or C-terminal parts, with C-terminal strand exchange resulting in a domain-swapped dimer[141] or cyclic swapped trimer (see detailed discussion below).[198]

Overall, the combined results obtained for hinge-loop properties provide useful hints with respect to domain swapping. However, no clear, predictive rules have emerged yet.

4.2.2.2 Mutations promoting domain swapping outside of the hinge-loop Several examples exist where residue changes in other parts of the protein, not the hinge-loop, are associated with domain swapping. A prime example is GB1. Compared to wild type monomeric GB1, the domain-swapped dimer comprises four mutations: L5V, F30V, Y33F, and A34F, none

of which is located in the hinge region.[160] A theoretical analysis of the quadruple mutant and wild type GB1 from Wodak's group[199] suggested different effects caused by each change: L5V introduces general destabilization due to unfavorable interactions with its surrounding residues, F30V induces local strain due to a clash with its own backbone, and A34F not only destabilizes the monomer conformation by forcing W43 to adopt a strained side chain conformation, and therefore disrupts the hydrophobic core of GB1, but also stabilizes the swapped dimer by tightly packing its side chains from both subunits against each other in the dimer core. The importance of the individual mutated residues (L5V/F30V/Y33F/A34F) in the integrity of the domain-swapped structure was also investigated by modeling and mutagenesis.[160] Inspection of the dimer structure suggested that the shorter mutant side chains of the L5V and F30V variants could easily be accommodated within the core, although possibly causing some destabilization of the structure. Indeed, each change is tolerated within the wild type structure. The Y33F mutation represents a conservative change and either side chain can substitute for the other in the respective cores. The position of F34 in the domain-swapped dimer appeared to be most crucial. This was verified experimentally, since reverting F34 in the amino acid sequence of the domain-swapped dimer mutant back to the wild type alanine residue resulted in a monomeric protein with a very similar structure as wild type GB1.[160]

In the T-cell surface antigen CD2, the propensity for dimer formation could be modulated by mutations in the new interface that is created by domain swapping.[200] In addition, a R87A mutation that destabilizes the monomer, simultaneously increased dimer formation. However, as with the majority of other proteins, the hinge residues in CD2 were still the most crucial amino acids with respect to domain swapping.[200]

In summary, residues distant from the hinge region can shift the relative stabilities of monomer and domain-swapped dimer and thereby modulate domain swapping properties. However, compared to the amino acids in the hinge-loop region, they appear to play only a secondary role.

4.2.2.3 Stability and folding of the monomer Despite substantial efforts, no compelling proposal for a generally applicable and unified molecular mechanism of domain swapping has emerged to date.[32, 201-204]

Eisenberg and colleagues suggested a free energy diagram involving pathways for domain swapping based on their studies on DT.[190] In their scenario, the 'open monomer' conformation retains the native fold of other parts of the 'closed monomer', and only interactions at the closed interface are disrupted during unfolding of the monomer. Such partial unfolding scheme may be at play in multi-domain proteins in which separate, independently folding domains are exchanged. However, the existence of a stable 'open monomer' is unlikely for most domain-swapped proteins in which only a few secondary structural elements are exchanged. These isolated structural elements will be unstable and therefore complete un/refolding is more likely to be at play in these cases.

In RNase A more than one portion of the chain can exchange, creating different oligomers (Figure 4.2). Two different domain-swapped dimers and two domain-swapped trimers are formed in different relative proportions.[198] Among the two dimers, the C-terminal swapped dimer is the major form, suggesting that it is more stable. For the trimers, only the crystal structure of the cyclic C-terminal swapped form has been solved. Biochemical studies suggested that the second, uncharacterized trimer may be a linear trimer in which one RNase A molecule swaps its N-terminal helix with a neighboring RNase A molecule at one end and its C-terminal

strand at the other end.[198] In this kind of trimer, both types of exchange occur simultaneously at very distant sites in the same protein molecule, supporting the notion that the closed monomers may fully unfold and refold to form these various forms of domain-swapped oligomers.

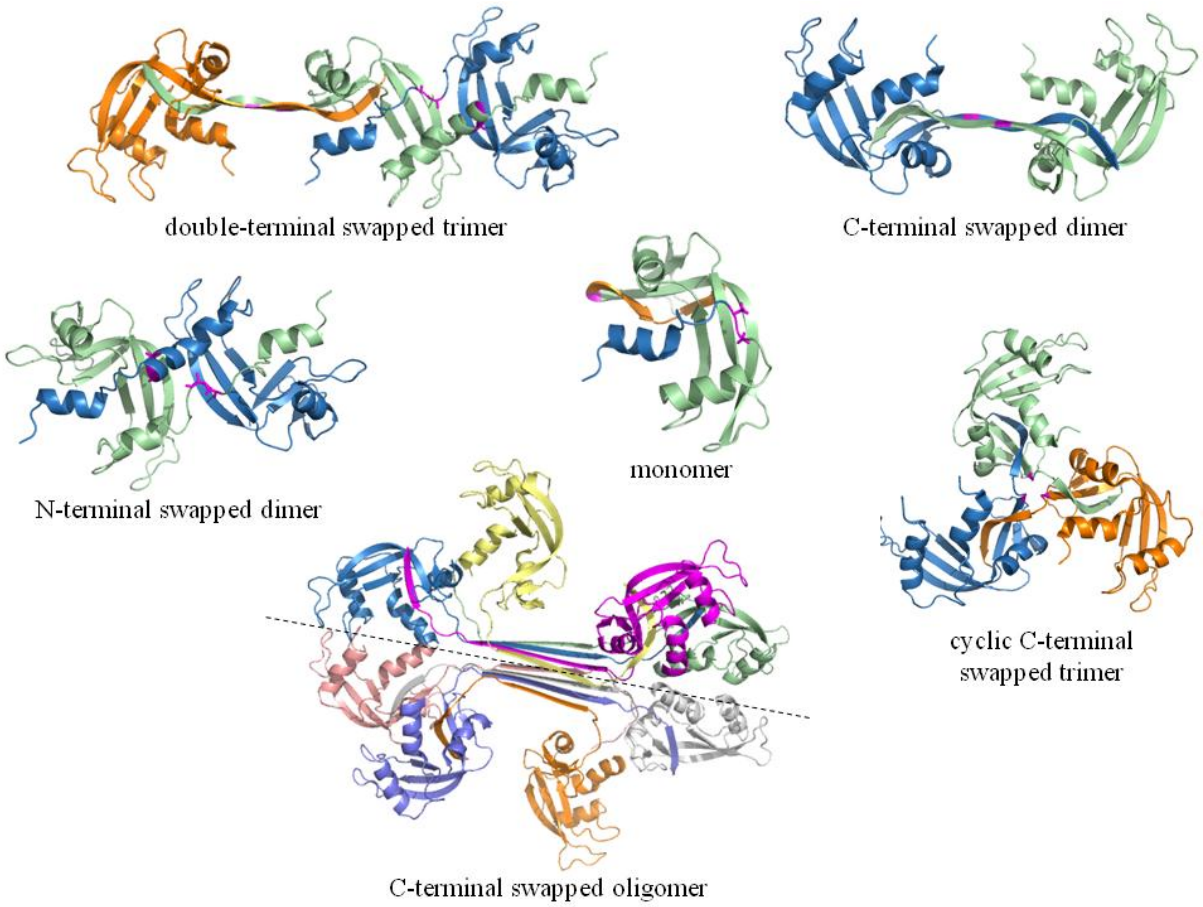

Figure 4.2 Structures of RNase A.

In the monomer, the two secondary structure elements involved in exchange are colored blue and orange. In the dimers and trimers, the individual polypeptide chains are colored green, blue and orange respectively. Hinge residues are shown with their side chains in stick representation and colored in magenta.

In the cyclin-dependent kinase regulatory subunit Cks1, exchange of the last β-strand β4, is involved in dimer formation.[205] NMR studies indicated that β4 in free monomeric Cks1

exhibits conformational heterogeneity.[206] This motion is abrogated by binding of Cdk2 to Cks1, resulting in a more homogeneous conformation of Cks1. Since Cdk2 binds to one face of the Cks1 β-sheet, the flexibility of β4 is reduced, preventing domain swapping. Interestingly, the binding of Cdk2 increases the binding affinity of Cks1 for phosphopeptides that bind to the other face of the β-sheet.[206] Therefore, configurational entropy not only influences ligand binding of Cks1 but also domain swapping.

4.2.3 Theoretical and computational explorations

A number of computational approaches for deciphering the basic events in protein folding and assembly are available, using reduced models and detailed atomistic simulations. Several groups are applying these methodologies to domain swapping. Movement of the polypeptide chain by Brownian motion through a funneled energy landscape with structure formation dominated by native stability[207] is the most elegant and widely accepted protein folding concept. This concept has also been applied to protein associations in domain-swapped multimers. In particular, Onuchic and Wolynes[208] have used a symmetrized Go-type potential to simulate domain swapping in MD simulations. For the epidermal growth factor receptor kinase substrate 8 (Eps8) SH3 dimer, they discovered a frustrated hinge region and suggested the following most favorable path for domain swapping: native monomers → partially folded monomers → unfolded monomers → open-end domain-swapped dimers → domain-swapped dimers. The authors suggested that the overall monomeric topology, rather than local signals in the hinge region, determines where in the polypeptide chain domain swapping will occur.[208] Although plausible, it appears at odds with some experimental results. For instance, in GB1 and LB1 (Protein L B1 domain, see below), proteins with identical monomeric topologies, different domain-swapped dimers are observed, clearly at odds with expectations if topology plays the

dominant role. Proteins with intrinsic symmetry of the sequence and/or structure are 'highly frustrated' in the language of these authors and in their simulations multi-mode domain swapping was observed and necessitated the inclusion of inter- or intra-molecular disulfide bonds.[209] Two proteins that fall into the 'highly frustrated' category are the human prion protein (PrP^C) and CV-N. However, at least for CV-N, the presence of disulfide bonds is not necessary for domain swapping since several homologs of CV-N with varying numbers of disulfide bonds appear to lack domain swapping[92, 210] and no differences in disulfides were noted for the monomers or domain-swapped dimers.

Coarse-grained MD simulations for several known domain-swapped proteins were also performed by Ding *et al*[211] who found that starting from monomeric conformations sometimes domain-swapped dimers formed. Based on native contact changes and topology maps, a web server for predicting the hinge region of domain-swapped proteins[211] was created. Testing the predictive value with the current set of 38 proteins resulted in correct predictions for only ~1/3 of the proteins in this set.

Analyzing large-scale domain motions of DT via Gaussian Network Models (GNM), Kundu and Jernigan[212] uncovered the major hinge in this protein based on the observed slower modes in GNM. The direction of the motion of the swapped domain about the hinge was predicted using the ANM.[212] However, it appears that DT is a special case among the domain-swapped proteins, given its multiple domain structure and the fact that a true folded domain undergoes the exchange and not single secondary structural elements.

We performed GNM analysis on the monomeric conformations of all 38 domain-swapped proteins (Table 4.1 and 4.2) in order to uncover any motions that may induce domain swapping. Initially, the domain-swapped structures were not used and were simply employed as

controls in this analysis. For each protein, hinge residues were defined by comparing backbone dihedral angles for the experimentally determined monomer and dimer structures (dihedral angle changes > 60^{o} at the open interface). The motional behavior for all residues via the first slow modes from GNM were examined. Disappointingly, GNM did not successfully distinguish hinge residues for our diverse set of domain-swapped proteins. Investigating the behavior of every residue we found that the hinge residues are neither the most mobile nor the most rigid ones in some proteins. For that matter, taking the picture of a hinge literally, the actual hinge usually stays fixed with the two objects that are connected by the hinge changing their relative positions. This would translate to relative rigidity of hinge residues and mobility at the edge of the hinge. On the other hand, hinge residues are often located in loops that naturally are more mobile than the cores of proteins, thereby allowing conformational changes to occur more easily.

A quite different mechanism of domain swapping has been proposed by the Wodak group, involving a progressive and reversible transformation between monomer and dimer.[213] This process, starts from either end of the polypeptide chain and intra-molecular contacts are traded for equivalent inter-molecular ones, with the total number of native contacts remaining essentially constant. In this manner more and more of the monomer chains are substituted for each other, until a stable state is reached. Exchange initiated at one end, such as the C-terminus, and did not involve unfolding. Conformational changes within the individual monomers and the binding between them were tightly coupled and the total number of native contacts was maximized. In this process, a large number of hinge conformations and association modes are sampled by the intermediates, suggesting that the exchange reaction is nonspecific and amino acid sequence only plays a minor role. However, so far, no experimental evidence exists for such a mechanism and it remains highly speculative.

4.3 INSRUCTIVE EXAMPLES AND BIOLOGICAL IMPLICATIONS

Is domain swapping an *in vitro* curiosity or does it serve a biological function? A number of results suggest that this type of oligomerization could be exploited in biology. One possible role for domain swapping could be to regulate protein function by modulating the populations of active molecules or the availability of functional sites. In addition, domain swapping could play a role in the allosteric regulation and signal transduction. Furthermore, in protein oligomerization scenarios, possible cytotoxic aggregation could be inhibited by domain-swapped dimerization. Finally, domain swapping is an efficient means for supramolecular structural organization of oligomers, such as seen in viral capsid structures. Therefore, although domain swapping may be involved in misfolding, aggregation, and amyloid formation of many proteins,[204, 214] this may not be the only function it serves.

Below we will discuss several notable examples of domain-swapped proteins in more detail. These are not stringent examples as defined above and for the associated proteins a stably folded monomeric structure may not be available.

4.3.1 RNase A

RNase A is the classic example of a protein engaged in domain swapping. Dimerization involving exchange of the N-terminus was proposed in 1962 prior to any structural information by Crestfield, Stein, and Moore to explain its behavior under acidic conditions.[10] The first X-ray structure for a domain-swapped RNase A dimer was solved in the late nineties by Eisenberg,[140] and the Eisenberg laboratory subsequently identified more domain-swapped dimers, trimers, and multimers (Figure 4.2).[141, 198] Because of its versatility, RNase A is frequently portrayed as the prototypical domain-swapped protein and with its different oligomeric states it beautifully illustrates the remarkable options of domain swapping modes.

Different folding conditions result in different types of RNase A oligomerization. Dimers are found at pH 6.5 and 37 ℃, close to the physiological conditions. However, the dissociation constant for the dimer under these conditions is ~2 mM, about 20-fold greater than the concentration of RNase A in the bovine pancreas. Polyethylene glycol (PEG) 10,000 stabilizes the RNase A minor trimer under crystallization conditions at pH 3.5.[198] Interestingly, RNase A oligomers exhibit higher enzyme activity on double-strand RNA than the monomer[215] and this is easily explained by the spatial arrangement of amino acids from different subunits that create the active site. Indeed, catalytic histidines are contributed by the N-terminal α-helix and the C-terminal β-strand, respectively.[216]

In one of the trimer forms of RNase A, both N- and C-terminal units are exchanged, resulting in a linear arrangement.[198] In the other trimer that only exhibits swapping of the C-terminal strand, a cyclic structure is formed. Therefore, for proteins that can swap two different domains, a variety of assembled oligomeric structures can be formed and models for such trimers, tetramers, and other oligomers have been proposed for RNase A.[217]

Although wild type RNase A does not form fibrils, a variant with a polyglutamine insertion in its hinge-loop (RNase A Q_{10}) forms amyloids *in vitro*.[218] A model for the RNase A Q_{10} fibrils was proposed in which the Q_{10} containing hinge-loops residues form β-strands that arrange into two β-sheets. The individual domains in this model keep their native fold and are involved in 'runaway' domain swapping.[218] In addition to the linear-type arrangements, simultaneous exchange of two different domains allows the formation of branched aggregates, possibly explaining the observation of some nonfibrillar aggregates.

4.3.2 B1 domain

GB1 is a small, 56 residue, stable, single domain protein. It comprises a four-stranded β-sheet with a single α-helix packed on top of it.[118] This protein exhibits astounding structural variability. A number of surprising structural variants were obtained in a large mutagenesis study involving a library of randomized hydrophobic core residues. Among the alternative structures was a domain-swapped dimer in which one hairpin was exchanged between the subunits.[160] The dimeric structure comprises an eight-stranded β-sheet made from four adjacent hairpins, resulting in two extensive new interfaces (Figure 4.3). The two α-helices are anti-parallel and cross at their C-termini. Half of the dimer, composed of the first β-hairpin and the α-helix from one polypeptide chain and the second β-hairpin from the other chain, is essentially identical to the monomer structure. The dimer dissociates into partially folded, monomeric species at low micromolar protein concentrations. The monomer is not a native, stable structure, but is a partially folded protein with extensive motions on the micro- to millisecond timescale. Despite these conformational fluctuations, the overall architecture of the monomer resembles that of wild type GB1. Thus, for this variant, dimerization via domain swapping stabilizes the molten, monomeric hydrophobic core.[219]

Structural comparison between the domain-swapped dimer and the wild type monomer suggested that the F34 side chain was the pivot for the monomer-dimer switch. Indeed, changing this residue back to the wild type alanine resulted in a wild type-like monomer structure. Interestingly, changing A34 to phenylalanine in the wild type sequence did not induce domain swapping, but resulted in a side-by-side dimer.[220]

GB1 variants are also capable of fibril formation, especially those sequences that are prone to domain swapping. Mutants that fold into the stable, wild type GB1 structure or variants

that exist as a highly destabilized, fluctuating ensemble of random, folded and partially folded structures under the same experimental conditions do not easily fibrillize. A left-handed helical ribbon model for the fibril was built, based on experimental disulfide cross-linking results, containing the swapped dimer structure as the smallest unit.[221]

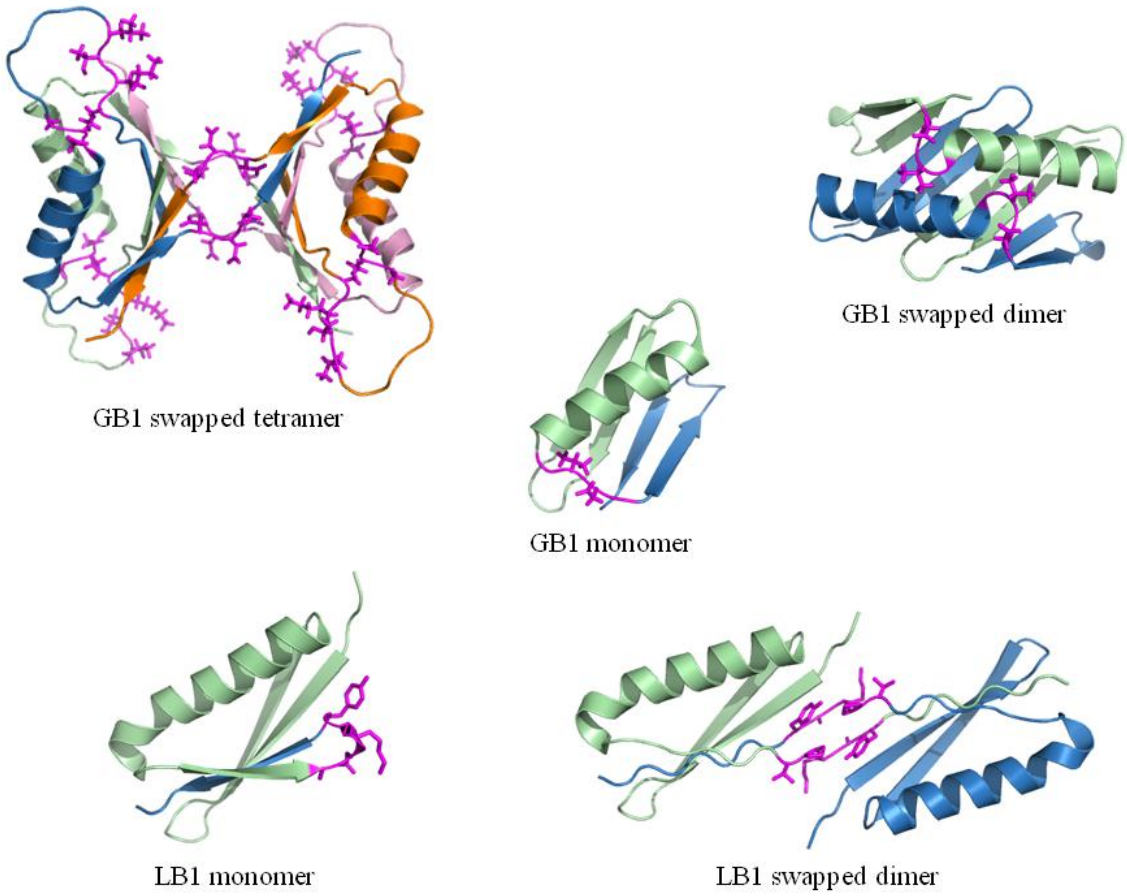

Figure 4.3 Structures of B1 domains.

In monomers, exchanged elements are colored in blue. In the dimers, individual polypeptide chains are colored in green and blue, respectively. Hinge residues are shown with their side chains in stick representation and colored in magenta.

An additional amino acid change in the domain-swapped dimer core caused a further dramatic change in structure: a symmetric tetramer ensued with inter-molecular strand-exchange involving all four units.[222] Three β-strands and the α-helix were retained in the tetramer, although their intra- and intermolecular interactions were radically different, with strand β2 of the first hairpin missing. The β3-β4 hairpin was changed to a side by side arrangement of strands β3 and β4 from one subunit, running antiparallel to β3 and β4 of another one. This topological change was accompanied by a shift in register. In addition to strand-exchange of the domain swapping kind, a new interface between surface elements of the individual chains was formed.

LB1 exhibits the same fold as the GB1 monomer,[223] however, a quite different domain-swapped structure was found for its mutants (Figure 4.3). Substitution of a glycine by alanine in the turn of the second β-hairpin caused exchange of the C-terminal β-strand between the subunits, with the wild type hairpin straightening and creating the inter-molecular β-sheet interface. These long β-strands are kinked, causing both B1 units to be rotated around the hinge region. Exchange of valine to alanine in the hydrophobic core also resulted in this type of domain-swapped structure.[135] Interestingly, in the X-ray structure, the asymmetric unit contains two wild type-like monomers and a domain-swapped dimer. Novel inter-molecular hydrophobic contacts as well as inter-molecular hydrogen bonds between the exchanged β-strands contribute to the stability of the domain swap.[135]

The above described different oligomeric B1 structures are illuminating examples for structural evolutionary paths from monomers to multimers.

4.3.3 Lectins

Several lectin structures were found to exhibit domain-swapped multimers. The first example was CV-N, which has been introduced in Chapter 3 and will be studied further in the next

chapter. The second antiviral lectin that exhibited domain swapping was Griffithsin (GRFT).[224] GRFT is a 121 amino acid protein of the red alga *Griffithsia* sp.. It exhibits antiviral activity against HIV-1 and severe acute respiratory syndrome (SARS) virus, by binding to various viral glycoproteins (gp) such as gp120, gp41, and gp160 in a monosaccharide-dependent manner.[224, 225] The structure of GRFT closely resembles jacalin lectins and comprises three repeats of a four-stranded antiparallel β-sheet. In the swapped dimer, the first two β-strands of one chain complete the β-prism of the other chain (Figure 4.4). Thus far, GRFT is the only example of a jacalin-fold protein for which a domain-swapped structure has been observed. GRFT is also the only member in its fold family that contains three carbohydrate binding sites. Other jacalins usually have a single one. The prism structure of GRFT is encoded by its triple sequence repeat. The three sugar binding sites reside in the loops of the β-hairpins formed by the second and third strand of each β-sheet.[224] Another lectin, Microcystis viridis lectin (MVL) was also suggested to show a domain-swapped structure. However, since no monomeric structure is available, it is difficult to ascertain that indeed a domain swapping has occured.[226]

Although CV-N and GRFT undergo domain swapping, the extent of the exchanged sequence is quite different. In CV-N, half of the molecule is involved in the swap, while in GRFT only the first two β-strands out of twelve are swapped. In addition, for CV-N, both monomeric and dimeric structures have been extensively characterized, while for GRFT only the dimeric structure is available.

As to their anti-HIV activities, the above lectins interact with oligosaccharides on viral envelope glycoproteins. The GRFT dimer contains six sugar binding sites, while CV-N exhibits two (monomer) or four (dimer). Both proteins are highly potent and inhibit HIV-1 at nanomolar concentrations.[224, 227] The binding sites on CV-N interact with the terminal epitopes (D1 and D3

arms) of the large, branched oligosaccharides. For GRFT, a similar binding mode has been proposed.[224, 227]

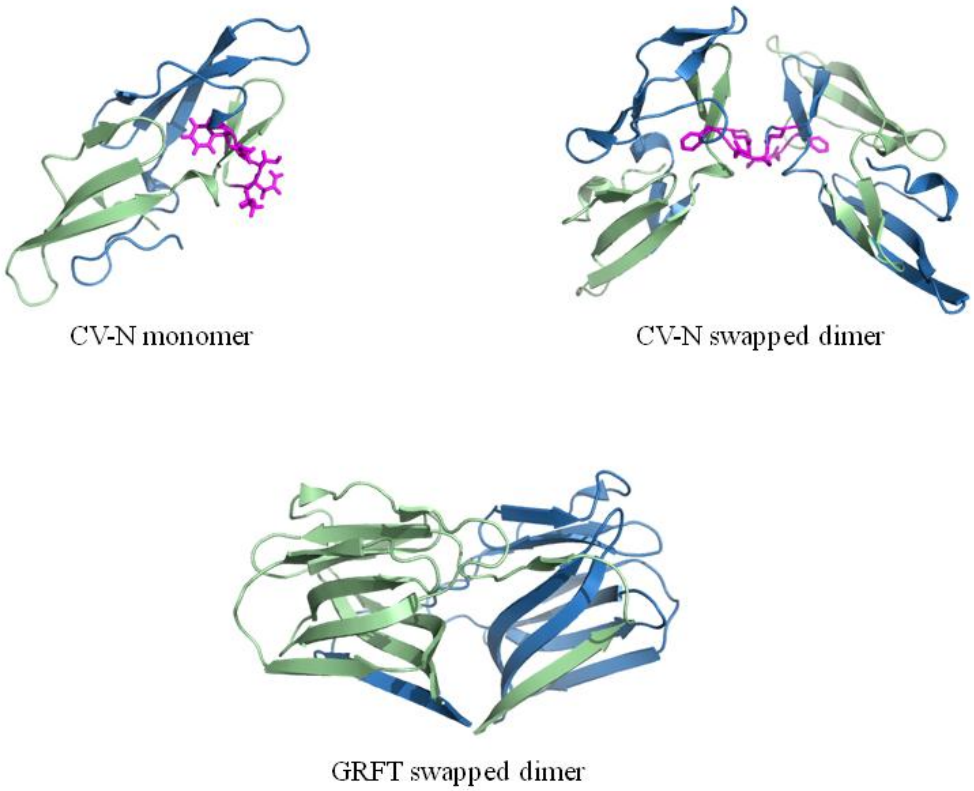

Figure 4.4 Structures of Lectins.

In the monomer, exchanged elements are colored in blue. In the dimers, individual polypeptide chains are colored in green and blue, respectively. Hinge residues are shown with their side chains in stick representation and colored in magenta.

4.4 CONCLUSIONS

Over the last decades, more and more domain-swapped protein structures have become available, and, at least for some cases, there is evidence in support of the dimer or multimer constituting biologically important species. Indeed, irrespective of whether domain swapping is a specific mechanism for regulating function *in vivo*, it is becoming clear that it is not solely an *in vitro* artifact.

Despite considerable efforts by numerous groups no unifying molecular mechanism of domain swapping has emerged: each protein seemingly behaves in a distinctive and individual fashion, and a general explanation for how proteins exchange domains still remains elusive. What seems to emerge is that domain swapping is closely associated with the unfolding/folding process of proteins. For some proteins, distinct intermediates, in which some hydrophobic part of the monomeric protein becomes exposed and, thereby, is available for interaction with a 'like' molecule may play a role, while for others, complete unfolding may occur. The fact that high protein concentration and additives (always present during crystallization) promote domain swapping suggests a switch in solute/solvent interaction. For example, exposed hydrophobic regions may no longer undergo unfavorable interactions with the aqueous solvent, but favorable ones with another polypeptide chain. In this manner an oligomeric structure can be trapped either in a crystal or an aggregate. Such behavior may also occur *in vivo* under conditions where monomer promoting factors are missing or where high local protein concentrations are induced through compartmentalization or the action of protein–protein interaction modules.

A more thorough understanding of the underlying features associated with domain swapping is certainly desirable. On one hand, domain swapping seems a means by which stable multimers can be generated under evolutionary pressure, and provides ways to improve protein stability. On the other hand, the fact that more and more proteins that exhibit disease-related aggregation also can form domain-swapped structures suggests a possible involvement in protein deposition diseases. Therefore, it may be possible to suppress aggregation by modulating domain swapping, an unexplored avenue in drug discovery.

5.0 DOMAIN SWAPPING PROCEEDS VIA COMPLETE UNFOLDING: A ^{19}F-NMR STUDY OF CYANOVIRIN-N

Work discussed in this chapter has been recently submitted for possible publication. Domain swapping creates protein oligomers by exchange of structural units between identical monomers. At present, no unifying molecular mechanism of domain swapping has emerged. Here we used the protein Cyanovirin-N and ^{19}F-NMR to investigate the process of domain swapping. CV-N is an HIV inactivating protein that can exist as a monomer or a domain-swapped dimer. We measured thermodynamic and kinetic parameters of the conversion process and determined the size of the energy barrier between the two species. The barrier is very large and of similar magnitude to that for complete unfolding of the protein. Therefore, for CV-N, overall unfolding of the polypeptide is required for domain swapping.

5.1 INTRODUCTION

Under physiological conditions most proteins exhibit a unique, narrowly distributed ensemble of conformations, broadly termed the native state. Within this native state ensemble, relatively low kinetic barriers separate the individual, very similar conformational sub-states.[36] Under specific circumstances, proteins may sample multiple sub-states, and such structural plasticity is exploited in molecular switches. For example, proteins that bind different substrates often employ alternative binding modes that optimize the intermolecular interactions, which are

facilitated by their conformational adaptability. Likewise, oligomerization may occur in different geometries, depending on the environmental conditions. Among thousands of homo-oligomers, a special type of oligomerization involves 'domain swapping'.[190] In domain-swapped structures one monomeric subunit exchanges one or more identical structural elements (domains, sub-domains or secondary structure elements) with another monomer. The three-dimensional structure of the pseudo-monomer within the domain-swapped multimer is identical to its corresponding monomer structure, except for the 'hinge' region that links the exchanged units.[190] Currently, more than 100 domain-swapped structures are deposited in the PDB.[228] The analysis of their chain lengths, structural class or amino acid composition does not reveal any special properties, suggesting that almost any protein may be capable of undergoing domain swapping, and that domain swapping is a specialized form of oligomer assembly.[229] Furthermore, domain swapping cannot be solely an *in vitro* artifact, given that some domain-swapped structures constitute biologically important species[230, 231] or cause disease-related aggregation.[232, 233] Therefore, understanding the mechanism of domain swapping is desirable.

Despite considerable efforts by several experimental and computational groups, a general explanation for how proteins exchange domains still remains elusive; each protein seemingly behaves in a distinctive and individual fashion.[128, 160, 200, 211, 229, 234, 235] What seems to emerge as a common theme is that domain swapping is closely associated with the unfolding/folding process of proteins. Comparing the closed conformation of the monomeric polypeptide chain with the open conformation of the same chain in the domain-swapped structure does not immediately suggest a pathway by which all intra-molecular interactions can be replaced by inter-molecular ones. Hydrophobic contacts, hydrogen-bonding, electrostatic interactions, and even disulfide bridges can be exchanged, and only the loop region in the monomer adopts a different

conformation from the hinge in the domain-swapped dimer.[32, 229] Therefore, starting with a folded monomer structure, the expectation would be that breaking and re-establishing interactions in conjunction with backbone conformational changes in the hinge-loop may require considerable energy. We call this energy the activation energy for 3D domain swapping starting from folded monomers.[190, 229] Folding from the unfolded polypeptide chain can result in either the closed monomer or the domain-swapped dimer, with partitioning between the two products determined by their free energy difference.

Here, we experimentally investigated domain swapping by NMR using the fluorine nucleus as the NMR-active probe. Fluorine has several favorable properties: it is the smallest atom that can be substituted for a hydrogen in a molecule; it possesses a nuclear spin of 1/2, 100% natural abundance, and a high gyromagnetic ratio (0.94 of that of a proton).[236] In addition, the ^{19}F lone pair electrons can participate in non-bonded interactions with the local environment, rendering ^{19}F chemical shifts extremely sensitive to even very small changes in van der Waals contacts, electrostatic fields, and hydrogen bonding in proteins.[237] These advantages render fluorine labeling extremely attractive for NMR studies of complex systems. Although not plentiful, applications of ^{19}F-NMR have been previously used to monitor conformational changes in proteins and to evaluate kinetic parameters associated with conformational transitions.[238-242]

The system that we selected for our studies is Cyanovirin-N (CV-N),[33] a well-characterized protein with domain swapping abilities.[138, 243] Using ^{19}F-NMR, we investigated the thermodynamics and kinetics of the conversion process between monomeric form and domain-swapped dimer for the wild type (wt) CV-N and its variants (Figure 5.1). Our results permit us to assess for the first time the energy landscape for interconversion between monomer and domain-swapped dimer, including the energy barrier height between the two states. To the best of our

knowledge, our work represents the first example of directly probing and determining the activation barrier for a protein when it undergoes domain swapping.

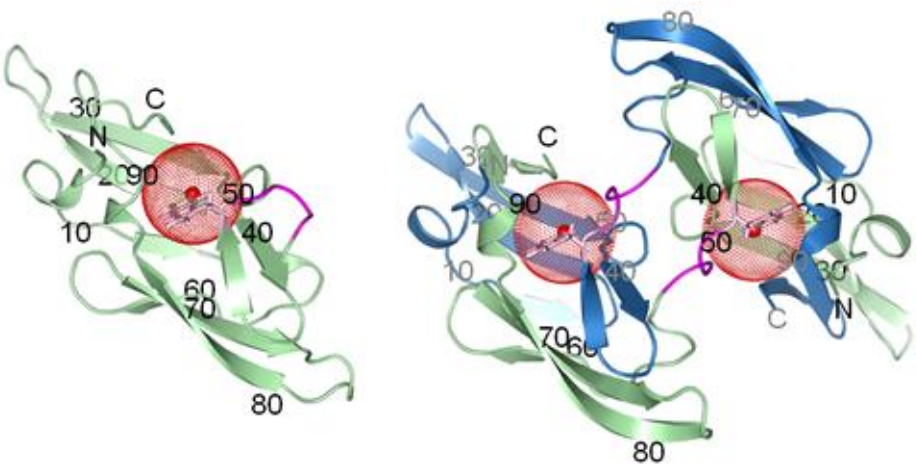

Figure 5.1 Structures of wt CV-N monomer (left, PDB ID: 2EZM) and domain-swapped dimer (right, PDB ID: 3EZM).

Ribbon diagrams are shown with chains A and B colored in green and blue, respectively, and the hinge-loop in magenta. The side chain of W49 is shown in stick representation (pink) with a red sphere of radius 5 Å drawn around the fluorine atom at position 5 of the tryptophan ring. Amino acid sequence positions are labeled for every 10^{th} residue, in black for chain A and in gray for chain B.

5.2 EXPERIMENTS AND METHODS

5.2.1 Sample preparation

The genes for mutant variants (CV-N^{P51G}, CV-$N^{\Delta Q50}$) of wt CV-N were prepared using the QuikChange Site-directed Mutagenesis kit (Stratagene Corp., La Jolla, CA). The presence of the desired mutations was confirmed by sequencing. All proteins were expressed using the pET26b(+) (Novagen Inc., Madison, WI) vector in *Escherichia coli* BL-21 (DE3). Cultures were grown at 37 ℃ in modified minimal medium, and 5-^{19}F-DL-tryptophan (Sigma-Aldrich Corp.,

St. Louis, MO) was added to the medium at a final concentration of 500 mg/L 15 minutes prior to induction with 0.5 mM IPTG. Cells were harvested 3 hours after induction by centrifugation and suspended in ice-cold PBS buffer (40 ml/1 L culture) for opening by sonication. Insoluble material was removed by centrifugation. The soluble protein present in the supernatant was fractionated by anion-exchange chromatography on a Q HP column (GE Healthcare, Piscataway, NJ) using a linear gradient of NaCl (0-1000 mM) for elution. Additional purification was achieved by gel filtration on Superdex 75 (HiLoad 2.6 × 60 cm, GE Healthcare, Piscataway, NJ), equilibrated in 20 mM sodium phosphate buffer (pH 6.0). Fractions with different quaternary states were collected: monomeric wt CV-N, monomeric CV-N^{P51G}, and dimeric CV-$N^{\Delta Q50}$. A sample of domain-swapped dimeric wt CV-N was obtained by incubating an ~ 10 mM monomeric sample at 39 ℃ for a week.[195] Dimeric domain-swapped CV-N^{P51G} was obtained by unfolding ~ 4 mM monomer in 8 M GdnHCl overnight, followed by extensive dialysis against 20 mM sodium phosphate buffer (pH 6.0) at 4 ℃ overnight for refolding. The domain-swapped dimer species was separated from the monomer species on a Superdex 75 gel filtration column equilibrated in 20 mM sodium phosphate, pH 6.0, containing 0.02% sodium azide, 2 mM DTT at 4 ℃. The extent of fluorine labeling (> 95%), purity and identity of all proteins were assessed and verified by mass spectrometry and SDS-PAGE. All samples were prepared in 20 mM sodium phosphate buffer, pH 6.0, and kept at 4 ℃ until used. D_2O was added to a final concentration of 8% to all NMR samples.

5.2.2 Differential Scanning Calorimetry (DSC)

20 mM sodium phosphate buffer (pH 6.0) was degassed overnight, and samples at a protein concentration of 1 mg/mL were dialyzed against the degassed buffer for at least 12 hours. DSC measurements were carried out using a VP-DSC instrument (MicroCal Inc., Northampton, MA)

at a heating scan rate of 1 ℃ per minute from 20 ℃ to 100 ℃. Data were analyzed using the Microcal Origin 7.0 software (MicroCal Inc., Northampton, MA).

5.2.3 NMR spectroscopy

Experiments were performed on Bruker Avance 600 or 900 MHz NMR spectrometers equipped with TCI triple-resonance, z-axis gradient cryoprobes (Bruker, Billerica, MA). External 2,2-dimethyl-2-silapentene-5-sulfonate (DSS) solution (1mM) was used for ^{1}H chemical shift referencing.[244] ^{19}F-NMR spectra were obtained on a Bruker Avance 600 spectrometer equipped with a Bruker CP TXO triple-resonance, X-nuclei observe, z-axis gradient cryoprobe (Bruker, Billerica, MA). External trifluoroacetic acid (TFA) solution (10 mM) was used for ^{19}F chemical shift referencing.[241, 245] The temperature was calibrated using 100% ethylene glycol.[246]

5.2.4 Data analysis

Conversion between CV-N monomer and CV-N domain-swapped dimer on an accessible timescale occurs only at elevated temperatures.[195] The conversion was followed by NMR. The fractions of polypeptide chains in the monomeric and dimeric states, f_M and f_D, were determined from the relative intensities of their associated resonances, using either ^{19}F- or ^{1}H-spectra. Integration of the peak areas (volumes) was carried out in Topspin (Bruker, Billerica, MA). The absolute concentrations of CV-N monomer [M] and CV-N dimer [D] were calculated based on their respective initial concentrations, C_M and C_D, before incubation at elevated temperatures as:

$$\begin{cases} [\mathrm{M}] = C_M \cdot \frac{f_M}{f_M + f_D} = 2C_D \cdot \frac{f_M}{f_M + f_D} \\ [\mathrm{D}] = \frac{1}{2}\left(C_M \cdot \frac{f_D}{f_M + f_D}\right) = \frac{1}{2}\left(2C_D \cdot \frac{f_D}{f_M + f_D}\right) \end{cases} \quad (5.1)$$

These equations are derived using the following properties: (i) each dimer contains two polypeptide chains, while each monomer contains only one; (ii) the total number of polypeptide chains (participating in either monomers or dimers) is conserved, i.e., [M] + 2[D] = constant.

For domain swapping, both conversions $D \xrightarrow{k_1} 2M$ and $2M \xrightarrow{k_{-1}} D$ occur simultaneously. According to classical chemical kinetics theory,[247] the order of a reaction and the rate constant *k* for a reaction can be obtained by monitoring the change in the concentration of the reactant during the time course of the reaction and fitting the data by appropriate models. The reaction isobserved in our case to obey a first-order reaction kinetics such that the integrated rate law reads:

$$[A] = [A]_0 \exp(-k_a t) \tag{5.2}$$

where [A] is the instantaneous concentration of the reactant (monomer or dimer) and k_a is the effective rate constant ($k_a = k_1 + k_{-1}$). Additionally, the relative resonance intensity ratio f_M/f_D at equilibrium is governed by the ratio of k_1/k_{-1}, allowing for the extraction of k_1 and k_{-1} values.

The temperature dependence of the reaction rate constant *k* permits us to calculate the Gibbs free energy of activation $\Delta G^{\ddagger}$ at any given temperature using the Eyring equation:

$$k = \frac{k_B T}{h} \cdot e^{-\frac{\Delta G^{\ddagger}}{RT}} \tag{5.3}$$

which leads to:

$$\ln \frac{k}{T} = -\frac{\Delta H^{\ddagger}}{R} \cdot \frac{1}{T} + \frac{\Delta S^{\ddagger}}{R} + \ln \frac{k_B}{h} \tag{5.4}$$

using $\Delta G^{\ddagger} = \Delta H^{\ddagger} - T\,\Delta S^{\ddagger}$, with the gas constant R = 1.986 cal K^{-1}mol^{-1}, the Boltzmann factor k_B = 1.38 x 10^{-23} J/K, and the Planck's constant h = 6.63 x 10^{-34} J s. Plotting ln(*k*/T) *vs.* 1/T yields a straight line with slope equal to $-\Delta H^{\ddagger}/R$.

The equilibrium constant K_{eq} and the Gibbs free energy change $\Delta G_{D\text{-}M}$ for the conversion reaction are given by:

$$K_{eq} = [M]_{eq}^2 / [D]_{eq} \tag{5.5}$$

$$\Delta G_{D\text{-}M} = -RT \ln K_{eq} \tag{5.6}$$

5.3 RESULTS AND DISCUSSION

5.3.1 CV-N system

CV-N is a 101 amino acid cyanobacterial lectin that was originally isolated from an aqueous extract of *Nostoc ellipsosporum*.[33] CV-N exhibits potent anti-HIV activity and is being developed as a general virucidal agent against HIV and other enveloped viruses.[33] The original solution structure found the protein to be monomeric[33] while in the subsequently solved X-ray structures domain-swapped dimers were observed[138, 243] (Figure 5.1). Manipulating experimental conditions, both quaternary states can be generated for CV-N, and the CV-N system has been used extensively for biophysical, structural, and functional studies.[33, 90, 97, 138, 195, 243, 248-251] The monomer structure exhibits a compact, bilobal fold with C2 pseudo-symmetry. Each domain comprises a triple-stranded β-sheet with a β-hairpin packed on top. A helical linker is located in the middle of the sequence. In the domain-swapped dimer structure, this linker acts as a hinge to open the monomers which pair up to form a dimer exhibiting essentially the same interactions as present in the monomer, but now inter-molecular. Residues in the hinge region (Q50-N53) provide important determinants for domain swapping. For instance, changing the single proline at position 51 to glycine results in substantial stabilization of the mutant, compared to the wild type, for both the monomer and the domain-swapped dimer.[195] The S52P mutant yields predominantly dimeric protein,[195] and the deletion mutant, ΔQ50, exists solely as a domain-swapped dimer.[97]

CV-N contains only one tryptophan (W49) in its sequence, and the side chain sits at the junction between the pseudo-symmetric halves, close to the pseudo two-fold axis, occupying a pivotal region during domain swapping. We therefore introduced 5-^{19}F-tryptophan into CV-N (Figure 5.1), for exploring the mechanism of domain swapping by ^{19}F-NMR. Incorporation of a

single or a few 5-^{19}F-tryptophan residues into proteins has been shown previously to cause no discernible effects on global and local structure or thermodynamic stability of ^{19}F labeled proteins.[237, 241, 242]

5.3.2 ^{19}F spectroscopy

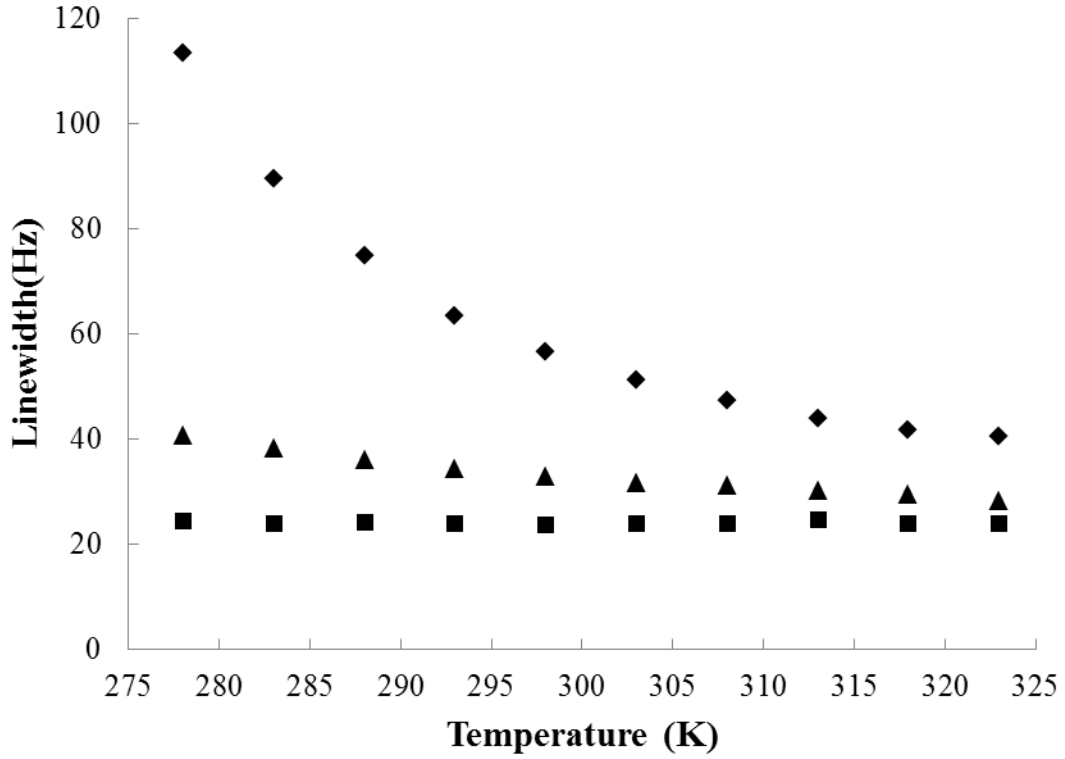

Figure 5.2 Linewidths of 5-^{19}F-tryptophan resonances as a function of temperature. Results are presented for free tryptophan (squares), that in CV-N^{P51G} monomer (triangles), and in the CV-N^{P51G} domain-swapped dimer (diamonds).

Since there is only one tryptophan in CV-N sequence, a single ^{19}F resonance is expected in the 1D ^{19}F spectrum. If, on the other hand, more than one species of the same protein exists, multiple resonances corresponding to the number of the species will be observed. Given the extreme sensitivity of the ^{19}F chemical shift to conformational and electronic influences, combined with

its large chemical shift range, little overlap in the ^{19}F spectra of F-labeled proteins ensues.[242] In addition, the temperature dependence of the ^{19}F chemical shift is small in the present case, with chemical shift differences of 0.12 ppm and 0.28 ppm observed for free 5-^{19}F-tryptophan and monomeric CV-N^{P51G}, respectively, between 278 and 323 K. In addition, essentially identical linewidths were observed for free 5-^{19}F-tryptophan over the temperature range 278-323 K, indicating that the rotational correlation time does not appreciably vary within this temperature range (Figure 5.2). For the CV-N monomer and the domain-swapped dimer, however, increases in linewidths were noted in the ^{19}F resonance when the temperature was reduced, reflecting the slower overall tumbling of the protein at lower temperature. This effect was more pronounced for dimer, due its larger size (Figure 5.2).

Figure 5.3 displays the ^{19}F spectra of 5-^{19}F-tryptophan labeled CV-N at 298 K. and pertinent spectral parameters are listed in Table 5.1. Interestingly, the single amino acid change from proline to glycine at position 51 did not significantly affect the chemical shift and linewidth of the ^{19}F resonance of the 5-^{19}F-tryptophan labeled CV-N monomer species. However, a significant difference was observed for the CV-N^{P51G} dimer, with the ^{19}F resonance substantially upfield shifted, compared to wt CV-N monomer, wt CV-N dimer, and CV-N^{P51G} monomer. In addition, the linewidth for the wt CV-N dimer (71.83 Hz) was noticeably larger than that of the CV-N^{P51G} dimer (56.42 Hz). Since W49 is adjacent to the hinge-loop region, these observations suggest that the influence of msec motions imparted by slow *cis-trans* isomerization of the proline containing wt CV-N hinge is removed in the CV-N^{P51G} variant. This is consistent with the fact that the wild type sequence contains a proline residue, and prolines are known for slow *cis-trans* isomerization and imparting rigidity to polypeptide backbones.[252] Since the ^{19}F resonance of 5-^{19}F-tryptophan labeled wt CV-N monomer and domain-swapped dimer species

are partially overlapping, we used the well separated Nε1 proton resonances of the tryptophan side chain of the monomer and the domain-swapped dimer[195] for monitoring the conversion time course for wt CV-N.

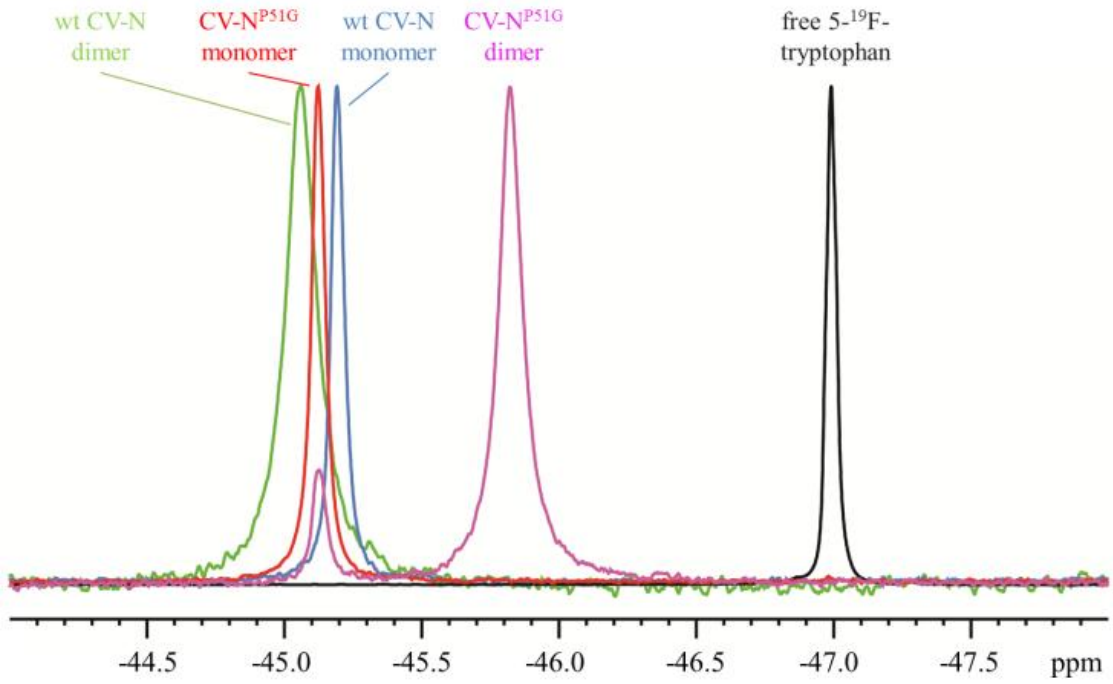

Figure 5.3 ^{19}F-NMR spectra of 5-^{19}F-tryptophan labeled CV-N samples and free 5-^{19}F-tryptophan at 298 K.

Table 5.1 ^{19}F-NMR parameters of 5-^{19}F-Tryptophan labeled CV-N samples at 298 K

	free 5-^{19}F-tryptophan	wt CV-N		CV-N^{P51G}	
		M	D	M	D
resonance frequency (ppm)	-46.99	-45.19	-45.06	-45.12	-45.82
linewidth at half-height (Hz)	23.69	31.60	71.83	32.79	56.42

5.3.3 Kinetics of the conversion between domain-swapped dimer and monomer

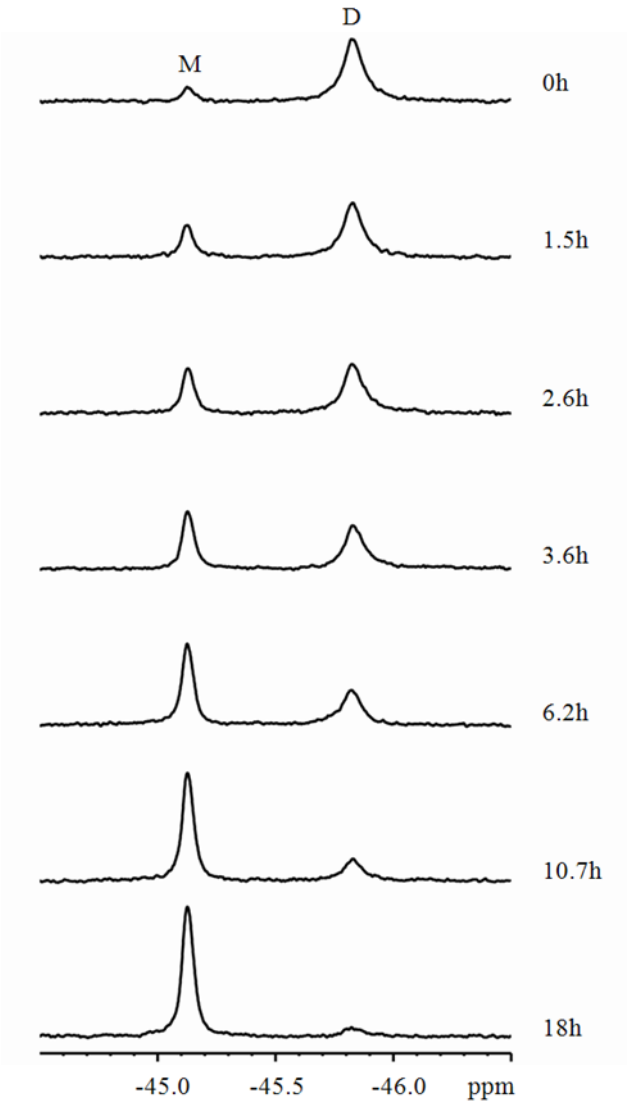

Figure 5.4 ^{19}F-NMR spectra recorded at 298 K following the conversion process from domain-swapped dimer to monomer of 5-^{19}F-tryptophan labeled CV-N^{P51G} at 330.5 K.

The length of incubation at 330.5 K is indicated at the right side of each spectrum. NMR spectra were recorded at 298 K to prevent any conversion during the time of the NMR measurement.

For CV-N^{P51G}, the monomer and domain-swapped dimer ^{19}F resonances are well separated and conversion between the two species can be followed readily using 1D spectra (Figure 5.4). The predominantly dimeric sample was incubated at 330.5 K for increasing amounts of time, and ^{19}F

spectra were recorded at 298 K, where the conversion process is slowed sufficiently to not interfere with accurate determination of the relative intensities/amounts. The data provided in Figure 5.4 clearly show that after ~ 4 hours of incubation at 330.5 K, ~ 50% of the swapped-dimer species had converted into monomer. Spectra were also recorded for the CV-N^{P51G} dimer conversion at other temperatures, as well as for the wt CV-N conversion process. The excellent spectral quality allowed to fit the data using eq 5.2 and permitted us to extract rate constants, for example: k_1 of $3.3*10^{-5}$ s^{-1} for the reaction D $\rightarrow$ 2M at 330.5K.

The same analysis was repeated for a series of temperatures. The time-courses for the conversion of the wt CV-N swapped dimer at different temperatures are displayed, and the resonance intensities exhibited an exponential decrease at each temperature, as shown in Figure 5.5A. Not surprisingly, faster rates were observed at higher temperatures. Using the experimentally determined temperature dependence of the rate constant *k*, the activation enthalpy $\Delta H^{\ddagger}_{D\text{-}M}$, entropy $\Delta S^{\ddagger}_{D\text{-}M}$ and Gibbs energy $\Delta G^{\ddagger}_{D\text{-}M}$ for the conversion from domain-swapped dimer to monomer was calculated using eqs 5.3 and 5.4 (Figure 5.5A inset).

The series of gray data points in Figure 5.5A represents the conversion at 325.5 K, the fastest reaction for wt CV-N domain-swapped dimer (k_1 = $8.2*10^{-5}$ s^{-1}). At a very similar temperature, 327.8 K, conversion for the CV-N^{P51G} domain-swapped dimer was the slowest reaction in the series (k_1 = 4.3 x 10^{-6} s^{-1}, black data points in Figure 5.5B), and required more than six days to reach the equilibrium. Therefore, the accessible temperature windows for the conversion reaction for wt CV-N and CV-N^{P51G} are distinctly different and non-overlapping: at 327.8 K, the conversion for wt CV-N is too fast, while the conversion for CV-N^{P51G} at 325.5 K is too slow. As a consequence, temperature dependent $\Delta G^{\ddagger}_{D\text{-}M}$ values could only be extracted for different sets of temperatures (Table 5.2). Given that smaller activation energies are seen with

increasing temperatures, it is safe to assume that the $\Delta G^{\ddagger}_{D\text{-}M}$ for the wt CV-N domain-swapped dimer conversion at 327.8 K should be lower than 25.2 kcal/mol, the measured $\Delta G^{\ddagger}_{D\text{-}M}$ for the wt CV-N domain-swapped dimer conversion at 325.5 K. Comparison of this value with the $\Delta G^{\ddagger}_{D\text{-}M}$ for CV-N^{P51G} (27.3 kcal/mol at 327.8 K) reveals that less energy is required for the wt CV-N conversion than for the CV-N^{P51G} dimer at the same temperature. This is consistent with the experimentally observed faster equilibration during the conversion of wt CV-N dimer into monomer.

Since equivalent experiments were carried out for wt CV-N and CV-N^{P51G}, we can directly compare the activation barriers for conversion. The $\Delta H^{\ddagger}$ values are listed in Table 5.2. Interestingly, these $\Delta H^{\ddagger}$ values are very similar in magnitude to the unfolding enthalpy changes, ΔH, observed by DSC. Since both wt CV-N and CV-N^{P51G} comprise monomeric and dimeric species that can undergo interconversions, we used a unique mutant, CV-N$^{\Delta Q50}$, that exists only as an unfolded monomer or a folded domain-swapped dimer for the control DSC experiment. The $\Delta H_{D\text{-}U}$ value for CV-N$^{\Delta Q50}$ unfolding was 141.9 kcal/mol; this value is of the same order of magnitude as the activation enthalpy $\Delta H^{\ddagger}_{D\text{-}M}$ for the conversion from domain-swapped dimer to monomer for wt CV-N (152.6 kcal/mol) and CV-N^{P51G} (161.7 kcal/mol) extracted for the NMR kinetic study. This surprising result implies that the monomer/swapped dimer conversion proceeds via complete unfolding of the protein, rather than partially un/folded states.

We also followed the reverse reaction for wt CV-N, namely conversion from monomer to domain-swapped dimer (Figure 5.5C). At 325.5 K, the reaction was carried out twice to evaluate and confirm the reliability of the experimental data. Both datasets agree extremely well (magenta and blue symbols) and can be fit to the same curve. In addition, the extracted $\Delta H^{\ddagger}_{M\text{-}D}$ value for the conversion of the wt CV-N monomer to the domain-swapped dimer (144.8 kcal/mol) agrees

well with the DSC result (129.9 kcal/mol) and the derived value (125.3 kcal/mol) for the CV-N^{P51G} monomer to domain-swapped dimer conversion. This is very gratifying and again implies that complete unfolding is involved in the conversion process.

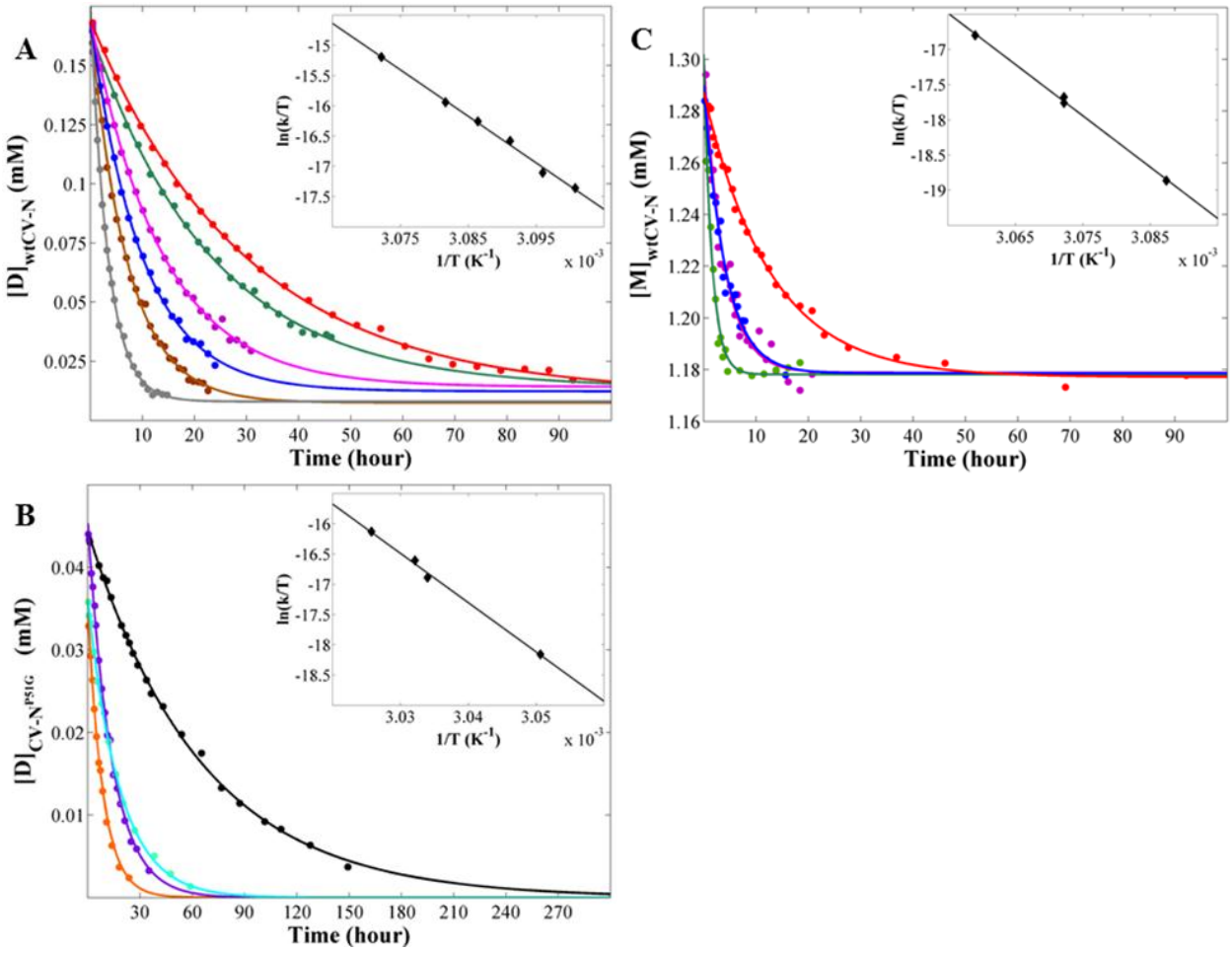

Figure 5.5 Time dependence of the conversion reactions for wt CV-N and CV-N^{P51G} at different temperatures.

Each point represents the concentration of the domain-swapped dimer (or monomer) species at a particular point in time as measured by the relative intensities of the dimer and monomer resonances. The inset shows the temperature dependence of reaction rate constant. The data fits a straight line whose slope ($-\frac{\Delta H^{\ddagger}}{R}$) and intercept ($\frac{\Delta S^{\ddagger}}{R} + \ln\frac{k_B}{h}$) yield the activation enthalpy $\Delta H^{\ddagger}$ and entropy $\Delta S^{\ddagger}$, respectively, using eq 5.4. (A) The conversion from wt CV-N domain-swapped

dimer to monomer. The incubation temperatures are: 322.5 K, red; 323 K, green; 323.5 K, magenta; 324 K, blue; 324.5 K, brown; and 325.5, gray. (B) The conversion from CV-N^{P51G} domain-swapped dimer to monomer. The incubation temperatures are: 327.8 K, black; 329.6 K, cyan; 329.8 K, purple; and 330.5 K, orange. (C) The conversion for wt CV-N monomer to domain-swapped dimer. The incubation temperatures are: 323.9 K, red; 325.5 K (1), blue; 325.5 K (2), magenta; and 326.9 K, green.

Both conversion reactions (monomer to dimer and dimer to monomer) exhibit exponential time dependence, suggesting that both are first order reactions. This observation appears to be at odds with the assumption that a molecular reaction of the type M + M → D might be a second order reaction. Although puzzling at first, the observed first order kinetics is in perfect agreement with the fact that complete unfolding occurs in the conversion reaction. The observations are indeed consistent with the presence of the rate-limiting steps of M → U and D → 2U for conversion of monomer to domain-swapped dimer and conversion from domain-swapped dimer to monomer, respectively. Each conversion process consists of two steps, with the unfolded state (U) as the intermediate.

Our current system is particularly suitable to investigate the kinetics given our excellent fluorine labeling efficiency. However, even if incomplete labeling were the case, resulting in sample heterogeneity,[241] it should be possible to follow the first order reaction and determine the reaction rate constant. Kinetic parameters (but not thermodynamic ones) are extracted from the temperature dependence of the reaction rate, and thus do not depend on the concentration. Therefore, only the labeled fraction of the protein is contributing to the data and correct kinetic information is obtained.

In addition to the Gibbs free energy barrier $\Delta G^{\ddagger}$ and the activation enthalpy $\Delta H^{\ddagger}$ discussed above, the average entropy change $\Delta S^{\ddagger}$ can also be extracted using eq 5.4. The entropy

change was 391.3 cal/(mol K) for the wt CV-N domain-swapped dimer to monomer conversion, ~ 30 cal/(mol.K) larger than the value extracted for the wt CV-N monomer to dimer conversion of 362.5 cal/(mol K). Given that in the conversion reaction one dimer molecule converts into two unfolded single-chain molecules, the total number of molecules in the system increases while the number of polypeptide chains remains the same. Therefore, the system becomes more disordered and its entropy change is larger than for unfolding of a single folded to and unfolded chain, for which no increase in the number of molecules occurs. The slight increase in entropy for the CV-N^{P51G} domain-swapped dimer conversion compared to the wt CV-N dimer (410.0 cal/mol K) can be explained by the increased flexibility in the linker introduced by the P51G mutation.

Table 5.2 Energetics of domain swapping and protein unfolding of wt CV-N and its variants

	Kinetic parameters for domain swapping measured by NMR			**Thermodynamic properties for unfolding measured by DSC**		
	wt CV-N (M-D)	**wt CV-N (D-M)**	**CV-N^{P51G} (D-M)**	**CV-N^{P51G} (M-U)**	**CV-N^{P51G} (D-U)**	**CV-$N^{\Delta Q50}$ (D-U)**
$\Delta H^{\ddagger}$ **or** ΔH **(kcal mol^{-1})**	144.8 ±21.5	152.6 ±14.5	161.7 ±31.8	129.9 ±1.1	171.0 ±3.5	141.9 ±0.5
$\Delta S^{\ddagger}$ **(cal/mol K)**	362.5 ±65.9	391.3 ±44.9	410.0 ±96.5	-	-	-
$\Delta G^{\ddagger}$ **(kcal mol^{-1})**	26.8 ±0.1 (325.5K)	25.2 ±0.1 (325.5K)	27.3 ±0.1 (327.8K)	-	-	-
k_1 **or** k_{-1} **× 10^6 (s^{-1})**	6.6 ±0.3 (325.5K)	82.0 ±2.6 (325.5K)	4.3 ±0.5 (327.8K)	-	-	-

5.3.4 Equilibrium properties

The data presented in Figure 5.5 also allows for the extraction of the monomer-dimer equilibrium constant, K_{eq}, since the final flat part of each curve at long conversion times yields the equilibrium concentration. For the conversion starting from the wt CV-N domain-swapped dimer all reactions reached a similar equilibrium concentration of 11.2 ± 2.8 μM. Taking the

reaction D → 2M into account, we then extracted an average equilibrium constant K_{eq} of 15.3 mM, which leads to a Gibbs free energy $\Delta G_{D\text{-}M}$ of 2.4 ± 0.3 kcal/mol at 293 K based on eq 5.6. Neglecting a possible, small temperature dependence in K_{eq}, for the temperature interval from 322.5 K to 325.5K, this reaction Gibbs energy $\Delta G_{D\text{-}M}$ can be equated with the difference for thermal unfolding of the wt CV-N domain-swapped dimer, compared to twice the value for the unfolding of the wt CV-N monomer.

Although the mechanism(s) for unfolding by chaotrops, such as urea and guanidine hydrochloride (GdnHCl) may be different from thermal unfolding, it is expected that the energy difference between monomer and dimer for the two unfolding reactions is similar. In particular, it is reasonable to assume that the energy difference between reactants and products of the unfolding reaction is mainly determined by their intrinsic interaction difference. Previously reported unfolding free energies for wt CV-N monomer and the obligate domain-swapped dimer form are $\Delta G^{wt}_{M\text{-}U}$ = 4.2 ± 0.2 kcal/mol and $\Delta G^{\Delta Q50}_{D\text{-}U}$ = 10.6 ± 0.5 kcal/mol, respectively,[94, 195] yielding a chemical reaction energy of about 2.2 (10.6 – 2 × 4.2) kcal/mol. Since the previous chemical unfolding and the current thermal conversion/unfolding were performed for identical buffer conditions and temperature (293 K), it is satisfying to observe the excellent agreement between these values.

The conversion of the CV-N^{P51G} domain-swapped dimer into monomer (Figure 5.5B) yields a final equilibrium concentration of dimer around zero, given the experimental precision. (A very small amount of dimer (< 5%) cannot reliably be distinguished from the noise in the spectra.) In order to derive a lower limit K_{eq} value we used the last/smallest available concentration as the approximate equilibrium concentration and obtained a value of K_{eq} = 2.9 ± 0.9 mM.

For both, wt CV-N and CV-N^{P51G}, the interconversion ΔG is very small, in excellent agreement with the fact all interactions within the monomeric and swapped-dimeric structures are extremely similar; only the hinge-loop conformation is different. Therefore, any measurable free energy difference has to be associated with the hinge-loop that can either introduce or relieve strain in the monomer-dimer interconversion.

5.3.5 The energy landscape of domain swapping

The available thermodynamic and kinetic parameters (Table 5.2) permit a reconstruction of the overall energy landscape for domain swapping of CV-N^{P51G} (black profile). This is depicted in Figure 5.6, with the unfolding enthalpies for the monomer $\Delta H_{M\text{-}U}$ and domain-swapped dimer $\Delta H_{D\text{-}U}$ of CV-N^{P51G} obtained from DSC measurements and the activation enthalpy $\Delta H^{\ddagger}_{D\text{-}M}$ for the CV-N^{P51G} dimer to monomer conversion extracted from the ^{19}F-NMR study. The activation enthalpy $\Delta H^{\ddagger}_{M\text{-}D}$ for the CV-N^{P51G} monomer to dimer conversion can also be estimated since the activation enthalpy difference between the monomer → dimer and the dimer → monomer reaction ($\Delta H^{\ddagger}_{M\text{-}D} - \Delta H^{\ddagger *}_{D\text{-}M}$) should be equal to their unfolding enthalpy difference ($\Delta H_{M\text{-}U} - \Delta H^{*}_{D\text{-}U}$). The asterisks indicate that half the dimer values from Table 5.2 have to be used for the normalization, to ascertain that an identical number of polypeptide chains is taken into account. A similar treatment yields the energy landscape for wt CV-N (gray profile). The wt CV-N $\Delta H^{\ddagger}_{M\text{-}D}$ and $\Delta H^{\ddagger}_{D\text{-}M}$ values were extracted from the NMR study and $\Delta H_{D\text{-}U}$ for unfolding of the CV-N$^{\Delta Q50}$ domain-swapped dimer was determined by DSC. As can be easily appreciated, the activation barrier for domain swapping is comparable in magnitude to the unfolding barrier for both wt CV-N and CV-N^{P51G}. In addition, as observed previously,[195] the single amino acid change in P51G mutant stabilizes both monomer and domain-swapped dimer of this variant.

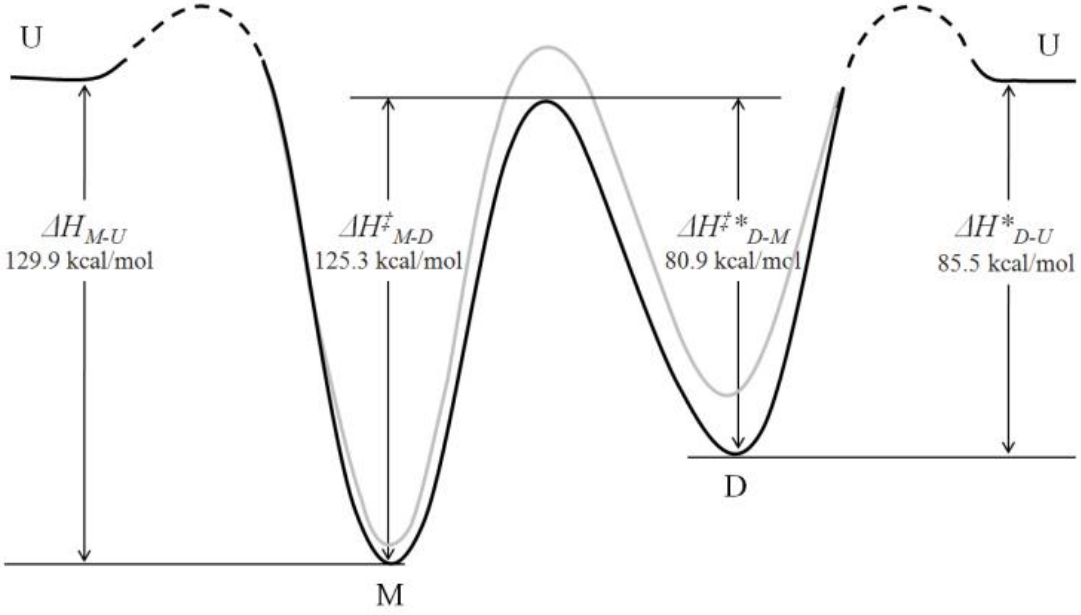

Figure 5.6 Energy diagram for domain swapping of CV-N^{P51G} (black) and wt CV-N (gray).

5.4 CONCLUSION

We carried out an extensive investigation of the thermodynamic and kinetic behavior for domain swapping of wt CV-N and CV-N^{P51G}, primarily using ^{19}F-NMR. Both proteins can exist at room temperature either as monomers or domain-swapped dimers in solution, indicating that the equilibrium free energies of both quaternary states are comparable. However, interconversion between these quaternary states is slow at room temperature or below. Therefore, the kinetic barrier between the monomer and domain-swapped dimer for CV-N is to be significant (of the order of ~100 ± 20 kcal/mol). Indeed, we determined here that this barrier is of similar magnitude to that for complete unfolding, suggesting that, at least for CV-N, complete unfolding is required for domain swapping to occur.

6.0 CONCLUSION AND FUTURE WORK

6.1 METHODS FOR INVESTIGATEING CONFORMATIONAL DYNAMICS

A previous study in our group collected 64 non homologous proteins, each containing a pair of structures solved by NMR and X-ray crystallography.[52] When comparing the residue fluctuations predicted by the GNM with the RMSDs among NMR ensembles, a correlation coefficient of 0.76 was obtained; however, the correlation between the GNM predictions and the X-ray crystallographic B-factors was found to be 0.59, only. To find an answer to this difference, we performed a further study, as described in Chapter 2. In this latter study, we found that intermolecular contacts between neighboring proteins occupying adjacent crystal lattice sites suppress the mobility of particular residues. As a result, these portions exhibit lower B-factors, and therefore appear to have lower RMSFs in their residue coordinates compared to GNM predictions. Therefore, the B-factors do not necessarily provide an accurate description of equilibrium dynamics. Instead, we generated X-ray ensembles based on the distance constraints extracted from X-ray structures. These ensembles were shown to correlate well with both the GNM predictions and the NMR ensemble data, suggesting that multiple conformations deduced from X-ray diffraction could satisfactorily describe the accessible conformational space under native state conditions.

Different experimental approaches usually investigate molecular-to-systems dynamics at different time scales.[3] However, in the above study, we also found that the dynamics explored by

different methods correlated with each other, despite their different time scales. To explain such observations, we performed multiple MD simulations with different simulation lengths, ranging from 1 ns to 400 ns, as described in Chapter 3. We found that the distribution of residue fluctuations is practically insensitive to the simulation length, while the amplitudes and correlation times of molecular motions appeared to increase with simulation time, within the limits permitted by the constraints exerted by the native contact topology in addition to covalent bonds. Additionally, the PCA of the generated trajectories revealed that the global mode deduced from of 1 ns long simulations and that from 400 ns long simulations exhibit a correlation coefficient of 0.77. Our results suggest that the protein tends to sample the same essential modes (reconfiguration directions) in long and short simulations, albeit at different sizes. This concordance supports the view that global motions are robustly defined by the shape of the native energy minimum, and the preferred mechanisms of reconfiguration may be detected even in short simulations, provided that the multiple runs are performed and dominant features are extracted by a PCA.

These two studies highlight the importance of using ensembles of structures for a given protein, so as to visualize the conformational space accessible to a given protein. Likewise, it is important to perform ensembles of simulations in order to gain an accurate understanding of the conformational dynamics accessible to the protein. Furthermore, dynamic methods can be advantageously used to investigate a broad range of the time scales for proteins, especially their low frequency motions, which are often related to biological activities.

6.2 DOMAIN SWAPPING

The 38 real/quasi-domain-swapped proteins we compiled from the literature show extremely diverse properties from their primary structures to quaternary structures, indicating almost any protein may be capable of undergoing domain swapping. According to our study, the first question we tried to answer was how to locate the hinge residues of domain-swapped proteins in their monomeric conformations. Although only the native contact topology is not sufficient based on our GNM analysis, the identification of conserved residues and co-evolving residue pairs may be a feasible way to provide criteria information in future studies.[253] The second concern is the molecular mechanism that underlie domain swapping, which appears closely associated with the unfolding/folding process of proteins.

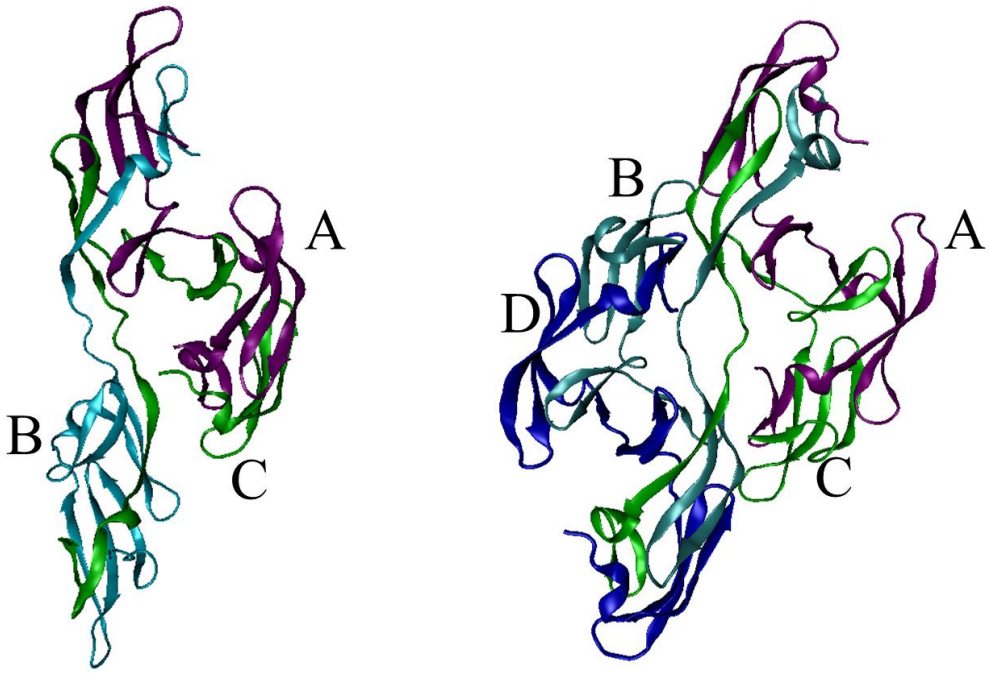

Figure 6.1 Structures of CV-N^{P51G} domain-swapped trimer (left) and tetramer (right).

As a specific case for experimental investigation, the conversion process between monomer and domain-swapped dimer for CV-N was studied by ^{19}F-NMR. This novel method allowed us to determine the thermodynamic and kinetic determinants of CV-N domain swapping, and we found that complete unfolding is required for CV-N domain swapping. This method may be further utilized to examine other domain-swapped systems to examine whether complete unfolding is a common mechanism of domain swapping or not. Moreover, the recently solved crystal structures of CV-N domain-swapped trimer and tetramer show new conformations as intermediates during refolding, providing further support for complete unfolding as a mechanism for CV-N domain swapping. Therefore, identifying intermediate conformations emerges here as the next step, toward directlyof investigating domain swapping.

BIBLIOGRAPHY

1. Falke JJ. Enzymology. A moving story. Science 2002;295:1480-1481.

2. Bahar I, Lezon TR, Bakan A, Shrivastava IH. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. Chem Rev 2010;110:1463-1497.

3. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature 2007;450:964-972.

4. 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.

5. Bahar I, Chennubhotla C, Tobi D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. Curr Opin Struct Biol 2007;17:633-640.

6. Frauenfelder H, McMahon BH, Austin RH, Chu K, Groves JT. The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. Proc Natl Acad Sci U S A 2001;98:2370-2374.

7. 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.

8. 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.

9. Tobi D, Bahar I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc Natl Acad Sci U S A 2005;102:18908-18913.

10. Xu C, Tobi D, Bahar I. Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T<-->R2 transition. J Mol Biol 2003;333:153-168.

11. Yang LW, Bahar I. Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure 2005;13:893-904.

12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235-242.

13. Blundell TL, Hendrickson WA. What is 'current opinion' in structural biology? Curr Opin Struct Biol 2011;21:447-449.

14. Marion D, Driscoll PC, Kay LE, Wingfield PT, Bax A, Gronenborn AM, Clore GM. Overcoming the overlap problem in the assignment of 1H NMR spectra of larger proteins by use of three-dimensional heteronuclear 1H-15N Hartmann-Hahn-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy: application to interleukin 1 beta. Biochemistry 1989;28:6150-6156.

15. Clore GM, Gronenborn AM. New methods of structure refinement for macromolecular structure determination by NMR. Proc Natl Acad Sci U S A 1998;95:5891-5898.

16. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D Biol Crystallogr 1999;55:849-861.

17. Schlichting I, Berendzen J, Chu K, Stock AM, Maves SA, Benson DE, Sweet RM, Ringe D, Petsko GA, Sligar SG. The catalytic pathway of cytochrome p450cam at atomic resolution. Science 2000;287:1615-1622.

18. Akke M, Palmer AG. Monitoring Macromolecular Motions on Microsecond to Millisecond Time Scales by R1 ñ-R1 Constant Relaxation Time NMR Spectroscopy. Journal of the American Chemical Society 1996;118:911-912.

19. Lipari G, Szabo A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. Journal of the American Chemical Society 1982;104:4559-4570.

20. 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.

21. 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.

22. 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.

23. Michalet X, Weiss S, Jager M. Single-molecule fluorescence studies of protein folding and conformational dynamics. Chem Rev 2006;106:1785-1813.

24. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nat Struct Biol 2002;9:646-652.

25. 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.

26. Haliloglu T, Bahar I, Erman B. Gaussian Dynamics of Folded Proteins. Physical Review Letters 1997;79:3090.

27. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins 1998;33:417-429.

28. Tirion MM. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys Rev Lett 1996;77:1905-1908.

29. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D. Intrinsic dynamics of an enzyme underlies catalysis. Nature 2005;438:117-121.

30. Jolliffe I. Principal Component Analysis. In. Encyclopedia of Statistics in Behavioral Science: John Wiley & Sons, Ltd; 2005.

31. Bennett MJ, Choe S, Eisenberg D. Domain swapping: entangling alliances between proteins. Proc Natl Acad Sci U S A 1994;91:3127-3131.

32. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11:1285-1299.

33. 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.

34. Hartmann H, Parak F, Steigemann W, Petsko GA, Ponzi DR, Frauenfelder H. Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K. Proc Natl Acad Sci U S A 1982;79:4967-4971.

35. Frauenfelder H, Parak F, Young RD. Conformational substates in proteins. Annu Rev Biophys Biophys Chem 1988;17:451-479.

36. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science 1991;254:1598-1603.

37. Bonvin AM, Rullmann JA, Lamerichs RM, Boelens R, Kaptein R. "Ensemble" iterative relaxation matrix approach: a new NMR refinement protocol applied to the solution structure of crambin. Proteins 1993;15:385-400.

38. Levin EJ, Kondrashov DA, Wesenberg GE, Phillips GN, Jr. Ensemble refinement of protein crystal structures: validation and application. Structure 2007;15:1040-1052.

39. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature 2005;433:128-132.

40. Cruickshank DWJ. The determination of the anisotropic thermal motion of atoms in crystals. Acta Crystallographica 1956;9:747-753.

41. Lumry R. Protein substructures and folded stability. Biophys Chem 2002;101-102:81-92.

42. Duan Y, Wang L, Kollman PA. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. Proc Natl Acad Sci U S A 1998;95:9897-9902.

43. 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.

44. Ringe D, Petsko GA. Study of protein dynamics by X-ray diffraction. Methods Enzymol 1986;131:389-433.

45. 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.

46. Cui Q, Bahar I. Normal Mode Analysis: Theory and applications to biological and chemical systems. Chapman & Hall/CRC; 2006.

47. Yang LW, Eyal E, Bahar I, Kitao A. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. Bioinformatics 2009;25:606-614.

48. Bahar I, Rader AJ. Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 2005;15:586-592.

49. Ma J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure 2005;13:373-380.

50. Tama F, Brooks CL. Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. Annu Rev Biophys Biomol Struct 2006;35:115-133.

51. Nicolay S, Sanejouand YH. Functional modes of proteins are among the most robust. Phys Rev Lett 2006;96:078104.

52. 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.

53. Kundu S, Melton JS, Sorensen DC, Phillips GN, Jr. Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J 2002;83:723-732.

54. Schomaker V, Trueblood KN. On the rigid-body motion of molecules in crystals. Acta Cryst B 1968;24:63-76.

55. Kondrashov DA, Van Wynsberghe AW, Bannen RM, Cui Q, Phillips GN, Jr. Protein structural variation in computational models and crystallographic data. Structure 2007;15:169-177.

56. 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.

57. Poon BK, Chen X, Lu M, Vyas NK, Quiocho FA, Wang Q, Ma J. Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-A crystallographic resolution. Proc Natl Acad Sci U S A 2007;104:7869-7874.

58. Song G, Jernigan RL. vGNM: a better model for understanding the dynamics of proteins in crystals. J Mol Biol 2007;369:880-893.

59. Kondrashov DA, Zhang W, Aranda R, Stec B, Phillips GN, Jr. Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. Proteins 2008;70:353-362.

60. Hinsen K. Structural flexibility in proteins: impact of the crystal environment. Bioinformatics 2008;24:521-528.

61. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779-815.

62. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 1999;285:1711-1733.

63. Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC. PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr 2002;58:1948-1954.

64. Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 1996;14:51-32.

65. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235-242.

66. 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.

67. 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.

68. Amadei A, Linssen AB, Berendsen HJ. Essential dynamics of proteins. Proteins 1993;17:412-425.

69. Best RB, Lindorff-Larsen K, DePristo MA, Vendruscolo M. Relation between native ensembles and experimental structures of proteins. Proc Natl Acad Sci U S A 2006;103:10901-10906.

70. Apaydin MS, Conitzer V, Donald BR. Structure-based protein NMR assignments using native structural ensembles. J Biomol NMR 2008;40:263-276.

71. Bahar I, Lezon TR, Yang LW, Eyal E. Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys 2010;39:23-42.

72. Cavanagh J, Venters RA. Protein dynamic studies move to a new time slot. Nat Struct Biol 2001;8:912-914.

73. 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.

74. Hall JB, Fushman D. Variability of the 15N chemical shielding tensors in the B3 domain of protein G from 15N relaxation measurements at several fields. Implications for backbone order parameters. J Am Chem Soc 2006;128:7855-7870.

75. 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.

76. 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.

77. 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.

78. 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.

79. Bui JM, Gsponer J, Vendruscolo M, Dobson CM. Analysis of sub-tauc and supra-tauc motions in protein Gbeta1 using molecular dynamics simulations. Biophys J 2009;97:2513-2520.

80. 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.

81. 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.

82. Caves LS, Evanseck JD, Karplus M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. Protein Sci 1998;7:649-666.

83. Romo TD, Grossfield A. Validating and improving elastic network models with molecular dynamics simulations. Proteins 2011;79:23-34.

84. Smith LJ, Daura X, van Gunsteren WF. Assessing equilibration and convergence in biomolecular simulations. Proteins 2002;48:487-496.

85. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal Component Analysis and Long Time Protein Dynamics. The Journal of Physical Chemistry 1996;100:2567-2572.

86. 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.

87. 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.

88. 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.

89. 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.

90. 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.

91. 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.

92. 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.

93. 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.

94. 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.

95. 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.

96. 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.

97. 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.

98. Li DW, Showalter SA, Bruschweiler R. Entropy localization in proteins. J Phys Chem B 2010;114:16036-16044.

99. 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.

100. 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.

101. 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.

102. Kitao A, Go N. Investigating protein dynamics in collective coordinate space. Curr Opin Struct Biol 1999;9:164-169.

103. 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.

104. 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.

105. 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.

106. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223-230.

107. Tuinstra RL, Peterson FC, Kutlesa S, Elgin ES, Kron MA, Volkman BF. Interconversion between two unrelated protein folds in the lymphotactin native state. Proc Natl Acad Sci U S A 2008;105:5057-5062.

108. Luo X, Tang Z, Xia G, Wassmann K, Matsumoto T, Rizo J, Yu H. The Mad2 spindle checkpoint protein has two distinct natively folded states. Nat Struct Mol Biol 2004;11:338-345.

109. Marianayagam NJ, Sunde M, Matthews JM. The power of two: protein dimerization in biology. Trends Biochem Sci 2004;29:618-625.

110. Lawrence SH, Ramirez UD, Tang L, Fazliyez F, Kundrat L, Markham GD, Jaffe EK. Shape shifting leads to small-molecule allosteric drug discovery. Chem Biol 2008;15:586-596.

111. Huang DB, Ainsworth CF, Stevens FJ, Schiffer M. Three quaternary structures for a single protein. Proc Natl Acad Sci U S A 1996;93:7017-7021.

112. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235-242.

113. Bennett MJ, Choe S, Eisenberg D. Refined structure of dimeric diphtheria toxin at 2.0 A resolution. Protein Sci 1994;3:1444-1463.

114. CRESTFIELD AM, STEIN WH, MOORE S. On the aggregation of bovine pancreatic ribonuclease. Arch Biochem Biophys 1962;Suppl 1:217-222.

115. Remington S, Wiegand G, Huber R. Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 A resolution. J Mol Biol 1982;158:111-152.

116. Anderson WF, Ohlendorf DH, Takeda Y, Matthews BW. Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. Nature 1981;290:754-758.

117. Fita I, Rossmann MG. The NADPH binding site on beef liver catalase. Proc Natl Acad Sci U S A 1985;82:1604-1608.

118. Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 1991;253:657-661.

119. Bennett MJ, Eisenberg D. Refined structure of monomeric diphtheria toxin at 2.3 A resolution. Protein Sci 1994;3:1464-1475.

120. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536-540.

121. Schreuder HA, de BB, Dijkema R, Mulders J, Theunissen HJ, Grootenhuis PD, Hol WG. The intact and cleaved human antithrombin III complex as a model for serpin-proteinase interactions. Nat Struct Biol 1994;1:48-54.

122. Fridmann-Sirkis Y, Kent HM, Lewis MJ, Evans PR, Pelham HR. Structural analysis of the interaction between the SNARE Tlg1 and Vps51. Traffic 2006;7:182-190.

123. Pryor PR, Jackson L, Gray SR, Edeling MA, Thompson A, Sanderson CM, Evans PR, Owen DJ, Luzio JP. Molecular basis for the sorting of the SNARE VAMP7 into endocytic clathrin-coated vesicles by the ArfGAP Hrb. Cell 2008;134:817-827.

124. Lewis RJ, Brannigan JA, Muchova K, Barak I, Wilkinson AJ. Phosphorylated aspartate in the structure of a response regulator protein. J Mol Biol 1999;294:9-15.

125. Lewis RJ, Muchova K, Brannigan JA, Barak I, Leonard G, Wilkinson AJ. Domain swapping in the sporulation response regulator Spo0A. J Mol Biol 2000;297:757-770.

126. Buckle AM, Fersht AR. Subsite binding in an RNase: structure of a barnase-tetranucleotide complex at 1.76-A resolution. Biochemistry 1994;33:1644-1653.

127. Zegers I, Deswarte J, Wyns L. Trimeric domain-swapped barnase. Proc Natl Acad Sci U S A 1999;96:818-822.

128. Stroud JC, Wu Y, Bates DL, Han A, Nowick K, Paabo S, Tong H, Chen L. Structure of the forkhead domain of FOXP2 bound to DNA. Structure 2006;14:159-166.

129. Manion MK, O'Neill JW, Giedt CD, Kim KM, Zhang KY, Hockenbery DM. Bcl-XL mutations suppress cellular sensitivity to antimycin A. J Biol Chem 2004;279:2159-2165.

130. O'Neill JW, Manion MK, Maguire B, Hockenbery DM. BCL-XL dimerization by three-dimensional domain swapping. J Mol Biol 2006;356:367-381.

131. Benoff B, Lawson C, Berman H, Carey J. Long-range effects on structure in a temperature-sensitive mutant of trp repressor. Unknown 2012.

132. Lawson CL, Benoff B, Berger T, Berman HM, Carey J. E. coli trp repressor forms a domain-swapped array in aqueous alcohol. Structure 2004;12:1099-1108.

133. Hatherley D, Graham SC, Turner J, Harlos K, Stuart DI, Barclay AN. Paired receptor specificity explained by structures of signal regulatory proteins alone and complexed with CD47. Mol Cell 2008;31:266-277.

134. Honnappa S, Okhrimenko O, Jaussi R, Jawhari H, Jelesarov I, Winkler FK, Steinmetz MO. Key interaction modes of dynamic +TIP networks. Mol Cell 2006;23:663-671.

135. O'Neill JW, Kim DE, Johnsen K, Baker D, Zhang KY. Single-site mutations induce 3D domain swapping in the B1 domain of protein L from Peptostreptococcus magnus. Structure 2001;9:1017-1027.

136. Max KE, Zeeb M, Bienert R, Balbach J, Heinemann U. Common mode of DNA binding to cold shock domains. Crystal structure of hexathymidine bound to the domain-swapped form of a major cold shock protein from Bacillus caldolyticus. FEBS J 2007;274:1265-1279.

137. Mueller U, Perl D, Schmid FX, Heinemann U. Thermal stability and atomic-resolution crystal structure of the Bacillus caldolyticus cold shock protein. J Mol Biol 2000;297:975-988.

138. Yang F, Bewley CA, Louis JM, Gustafson KR, Boyd MR, Gronenborn AM, Clore GM, Wlodawer A. Crystal structure of cyanovirin-N, a potent HIV-inactivating protein, shows unexpected domain swapping. J Mol Biol 1999;288:403-412.

139. Yamasaki M, Li W, Johnson DJ, Huntington JA. Crystal structure of a stable dimer reveals the molecular basis of serpin polymerization. Nature 2008;455:1255-1258.

140. Liu Y, Hart PJ, Schlunegger MP, Eisenberg D. The crystal structure of a 3D domain-swapped dimer of RNase A at a 2.1-A resolution. Proc Natl Acad Sci U S A 1998;95:3437-3442.

141. Liu Y, Gotte G, Libonati M, Eisenberg D. A domain-swapped RNase A dimer with implications for amyloid formation. Nat Struct Biol 2001;8:211-214.

142. Wlodawer A, Borkakoti N, Moss DS, Howlin B. Comparison of two independently refined models of ribonuclease-A. Acta Cryst B 1986;42:379-387.

143. Pesenti ME, Spinelli S, Bezirard V, Briand L, Pernollet JC, Tegoni M, Cambillau C. Structural basis of the honey bee PBP pheromone and pH-induced conformational change. J Mol Biol 2008;380:158-169.

144. Pesenti ME, Spinelli S, Bezirard V, Briand L, Pernollet JC, Campanacci V, Tegoni M, Cambillau C. Queen bee pheromone binding protein pH-induced domain swapping favors pheromone release. J Mol Biol 2009;390:981-990.

145. Khandelwal P, Keliikuli K, Smith CL, Saper MA, Zuiderweg ER. Solution structure and phosphopeptide binding to the N-terminal domain of Yersinia YopH: comparison with a crystal structure. Biochemistry 2002;41:11425-11437.

146. Smith CL, Khandelwal P, Keliikuli K, Zuiderweg ER, Saper MA. Structure of the type III secretion and substrate-binding domain of Yersinia YopH phosphatase. Mol Microbiol 2001;42:967-979.

147. Martin JR, Craven CJ, Jerala R, Kroon-Zitko L, Zerovnik E, Turk V, Waltho JP. The three-dimensional solution structure of human stefin A. J Mol Biol 1995;246:331-343.

148. Staniforth RA, Giannini S, Higgins LD, Conroy MJ, Hounslow AM, Jerala R, Craven CJ, Waltho JP. Three-dimensional domain swapping in the folded and molten-globule states of cystatins, an amyloid-forming structural superfamily. EMBO J 2001;20:4774-4781.

149. Sridharan S, Razvi A, Scholtz JM, Sacchettini JC. The HPr proteins from the thermophile Bacillus stearothermophilus can form domain-swapped dimers. J Mol Biol 2005;346:919-931.

150. Garcia-Pino A, Martinez-Rodriguez S, Wahni K, Wyns L, Loris R, Messens J. Coupling of domain swapping to kinetic stability in a thioredoxin mutant. J Mol Biol 2009;385:1590-1599.

151. Roos G, Garcia-Pino A, Van BK, Brosens E, Wahni K, Vandenbussche G, Wyns L, Loris R, Messens J. The conserved active site proline determines the reducing power of Staphylococcus aureus thioredoxin. J Mol Biol 2007;368:800-811.

152. Hakansson M, Svensson A, Fast J, Linse S. An extended hydrophobic core induces EF-hand swapping. Protein Sci 2001;10:927-933.

153. Jimenez B, Poggi L, Piccioli M. Monitoring the early steps of unfolding of dicalcium and mono-Ce3+-substituted forms of P43M calbindin D9k. Biochemistry 2003;42:13066-13073.

154. Yang W, Wilkins AL, Ye Y, Liu ZR, Li SY, Urbauer JL, Hellinga HW, Kearney A, van der Merwe PA, Yang JJ. Design of a calcium-binding protein with desired structure in a cell adhesion molecule. J Am Chem Soc 2005;127:2085-2093.

155. Murray AJ, Lewis SJ, Barclay AN, Brady RL. One sequence, two folds: a metastable structure of CD2. Proc Natl Acad Sci U S A 1995;92:7337-7341.

156. Kukimoto-Niino M, Sakamoto A, Kanno E, Hanawa-Suetsugu K, Terada T, Shirouzu M, Fukuda M, Yokoyama S. Structural basis for the exclusive specificity of Slac2-a/melanophilin for the Rab27 GTPases. Structure 2008;16:1478-1490.

157. Chavas LM, Torii S, Kamikubo H, Kawasaki M, Ihara K, Kato R, Kataoka M, Izumi T, Wakatsuki S. Structure of the small GTPase Rab27b shows an unexpected swapped dimer. Acta Crystallogr D Biol Crystallogr 2007;63:769-779.

158. Schiering N, Casale E, Caccia P, Giordano P, Battistini C. Dimer formation through domain swapping in the crystal structure of the Grb2-SH2-Ac-pYVNV complex. Biochemistry 2000;39:13376-13382.

159. Ettmayer P, France D, Gounarides J, Jarosinski M, Martin MS, Rondeau JM, Sabio M, Topiol S, Weidmann B, Zurini M, Bair KW. Structural and conformational requirements for high-affinity binding to the SH2 domain of Grb2(1). J Med Chem 1999;42:971-980.

160. Byeon IJ, Louis JM, Gronenborn AM. A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. J Mol Biol 2003;333:141-152.

161. Ramoni R, Spinelli S, Grolli S, Conti V, Merli E, Cambillau C, Tegoni M. Deswapping bovine odorant binding protein. Biochim Biophys Acta 2008;1784:651-657.

162. Tegoni M, Ramoni R, Bignetti E, Spinelli S, Cambillau C. Domain swapping creates a third putative combining site in bovine odorant binding protein dimer. Nat Struct Biol 1996;3:863-867.

163. Knaus KJ, Morillas M, Swietnicki W, Malone M, Surewicz WK, Yee VC. Crystal structure of the human prion protein reveals a mechanism for oligomerization. Nat Struct Biol 2001;8:770-774.

164. Antonyuk SV, Trevitt CR, Strange RW, Jackson GS, Sangar D, Batchelor M, Cooper S, Fraser C, Jones S, Georgiou T, Khalili-Shirazi A, Clarke AR, Hasnain SS, Collinge J. Crystal structure of human prion protein bound to a therapeutic antibody. Proc Natl Acad Sci U S A 2009;106:2554-2558.

165. Wolff N, Izadi-Pruneyre N, Couprie J, Habeck M, Linge J, Rieping W, Wandersman C, Nilges M, Delepierre M, Lecroisey A. Comparative analysis of structural and dynamic properties of the loaded and unloaded hemophore HasA: functional implications. J Mol Biol 2008;376:517-525.

166. Czjzek M, Letoffe S, Wandersman C, Delepierre M, Lecroisey A, Izadi-Pruneyre N. The crystal structure of the secreted dimeric form of the hemophore HasA reveals a domain swapping with an exchanged heme ligand. J Mol Biol 2007;365:1176-1186.

167. Rosenfeld RJ, Garcin ED, Panda K, Andersson G, Aberg A, Wallace AV, Morris GM, Olson AJ, Stuehr DJ, Tainer JA, Getzoff ED. Conformational changes in nitric oxide synthases induced by chlorzoxazone and nitroindazoles: crystallographic and computational analyses of inhibitor potency. Biochemistry 2002;41:13915-13925.

168. Crane BR, Rosenfeld RJ, Arvai AS, Ghosh DK, Ghosh S, Tainer JA, Stuehr DJ, Getzoff ED. N-terminal domain swapping and metal ion binding in nitric oxide synthase dimerization. EMBO J 1999;18:6271-6281.

169. Green SM, Gittis AG, Meeker AK, Lattman EE. One-step evolution of a dimer from a monomeric protein. Nat Struct Biol 1995;2:746-751.

170. Loll PJ, Lattman EE. The crystal structure of the ternary complex of staphylococcal nuclease, Ca2+, and the inhibitor pdTp, refined at 1.65 A. Proteins 1989;5:183-201.

171. Suino-Powell K, Xu Y, Zhang C, Tao YG, Tolbert WD, Simons SS, Jr., Xu HE. Doubling the size of the glucocorticoid receptor ligand binding pocket by deacylcortivazol. Mol Cell Biol 2008;28:1915-1923.

172. Schoch GA, D'Arcy B, Stihle M, Burger D, Bar D, Benz J, Thoma R, Ruf A. Molecular switch in the glucocorticoid receptor: active and passive antagonist conformations. J Mol Biol 2010;395:568-577.

173. Wiesmann C, Ultsch MH, Bass SH, de Vos AM. Crystal structure of nerve growth factor in complex with the ligand-binding domain of the TrkA receptor. Nature 1999;401:184-188.

174. Ultsch MH, Wiesmann C, Simmons LC, Henrich J, Yang M, Reilly D, Bass SH, de Vos AM. Crystal structures of the neurotrophin-binding domain of TrkA, TrkB and TrkC. J Mol Biol 1999;290:149-159.

175. Zdanov A, Schalk-Hihi C, Gustchina A, Tsang M, Weatherbee J, Wlodawer A. Crystal structure of interleukin-10 reveals the functional dimer with an unexpected topological similarity to interferon gamma. Structure 1995;3:591-601.

176. Josephson K, Jones BC, Walter LJ, DiGiacomo R, Indelicato SR, Walter MR. Noncompetitive antibody neutralization of IL-10 revealed by protein engineering and x-ray crystallography. Structure 2002;10:981-987.

177. Sue SC, Lee WT, Tien SC, Lee SC, Yu JG, Wu WJ, Wu WG, Huang TH. PWWP module of human hepatoma-derived growth factor forms a domain-swapped dimer with much higher affinity for heparin. J Mol Biol 2007;367:456-472.

178. Sue SC, Chen JY, Lee SC, Wu WG, Huang TH. Solution structure and heparin interaction of human hepatoma-derived growth factor. J Mol Biol 2004;343:1365-1377.

179. Ivanov D, Tsodikov OV, Kasanov J, Ellenberger T, Wagner G, Collins T. Domain-swapped dimerization of the HIV-1 capsid C-terminal domain. Proc Natl Acad Sci U S A 2007;104:4353-4358.

180. Byeon IJ, Meng X, Jung J, Zhao G, Yang R, Ahn J, Shi J, Concel J, Aiken C, Zhang P, Gronenborn AM. Structural convergence between Cryo-EM and NMR reveals intersubunit interactions critical for HIV-1 capsid function. Cell 2009;139:780-790.

181. Yu XL, Hu T, Du JM, Ding JP, Yang XM, Zhang J, Yang B, Shen X, Zhang Z, Zhong WD, Wen N, Jiang H, Zhu P, Chen ZN. Crystal structure of HAb18G/CD147: implications for immunoglobulin superfamily homophilic adhesion. J Biol Chem 2008;283:18056-18065.

182. Luo J, Teplyakov A, Obmolova G, Malia T, Wu SJ, Beil E, Baker A, Swencki-Underwood B, Zhao Y, Sprenkle J, Dixon K, Sweet R, Gilliland GL. Structure of the EMMPRIN N-terminal domain 1: dimerization via beta-strand swapping. Proteins 2009;77:1009-1014.

183. Zhang HP, Nagashima T, Hayashi F, Yokoyama S. Solution structure of the RGS domain of Regulator of G-protein signaling 7. to be published 2005.

184. Soundararajan M, Willard FS, Kimple AJ, Turnbull AP, Ball LJ, Schoch GA, Gileadi C, Fedorov OY, Dowler EF, Higman VA, Hutsell SQ, Sundstrom M, Doyle DA, Siderovski DP. Structural diversity in the RGS domain and its interaction with heterotrimeric G protein alpha-subunits. Proc Natl Acad Sci U S A 2008;105:6457-6462.

185. Jedrzejczak R, Dauter Z, Dauter M, Piatek R, Zalewska B, Mroz M, Bury K, Nowicki B, Kur J. Structure of DraD invasin from uropathogenic Escherichia coli: a dimer with swapped beta-tails. Acta Crystallogr D Biol Crystallogr 2006;62:157-164.

186. Cota E, Jones C, Simpson P, Altroff H, Anderson KL, du ML, Guignot J, Servin A, Le BC, Mardon H, Matthews S. The solution structure of the invasive tip complex from Afa/Dr fibrils. Mol Microbiol 2006;62:356-366.

187. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 2010;38:D142-D148.

188. Diederichs K, Jacques S, Boone T, Karplus PA. Low-resolution structure of recombinant human granulocyte-macrophage colony stimulating factor. J Mol Biol 1991;221:55-60.

189. Milburn MV, Hassell AM, Lambert MH, Jordan SR, Proudfoot AE, Graber P, Wells TN. A novel dimer configuration revealed by the crystal structure at 2.4 A resolution of human interleukin-5. Nature 1993;363:172-176.

190. Bennett MJ, Schlunegger MP, Eisenberg D. 3D domain swapping: a mechanism for oligomer assembly. Protein Sci 1995;4:2455-2468.

191. Ogihara NL, Ghirlanda G, Bryson JW, Gingery M, DeGrado WF, Eisenberg D. Design of three-dimensional domain-swapped dimers and fibrous oligomers. Proc Natl Acad Sci U S A 2001;98:1404-1409.

192. Chen YW, Stott K, Perutz MF. Crystal structure of a dimeric chymotrypsin inhibitor 2 mutant containing an inserted glutamine repeat. Proc Natl Acad Sci U S A 1999;96:1257-1261.

193. Bergdoll M, Remy MH, Cagnon C, Masson JM, Dumas P. Proline-dependent oligomerization with arm exchange. Structure 1997;5:391-401.

194. Rousseau F, Schymkowitz JW, Wilkinson HR, Itzhaki LS. Three-dimensional domain swapping in p13suc1 occurs in the unfolded state and is controlled by conserved proline residues. Proc Natl Acad Sci U S A 2001;98:5596-5601.

195. Barrientos LG, Louis JM, Botos I, Mori T, Han Z, O'Keefe BR, Boyd MR, Wlodawer A, Gronenborn AM. The domain-swapped dimer of cyanovirin-N is in a metastable folded state: reconciliation of X-ray and NMR structures. Structure 2002;10:673-686.

196. Mazzarella L, Capasso S, Demasi D, Dilorenzo G, Mattia CA, Zagari A. Bovine Seminal Ribonuclease - Structure at 1.9-Angstrom Resolution. Acta Crystallographica Section D-Biological Crystallography 1993;49:389-402.

197. Canals A, Pous J, Guasch A, Benito A, Ribo M, Vilanova M, Coll M. The structure of an engineered domain-swapped ribonuclease dimer and its implications for the evolution of proteins toward oligomerization. Structure 2001;9:967-976.

198. Liu Y, Gotte G, Libonati M, Eisenberg D. Structures of the two 3D domain-swapped RNase A trimers. Protein Sci 2002;11:371-380.

199. Sirota FL, Hery-Huynh S, Maurer-Stroh S, Wodak SJ. Role of the amino acid sequence in domain swapping of the B1 domain of protein G. Proteins 2008;72:88-104.

200. Murray AJ, Head JG, Barker JJ, Brady RL. Engineering an intertwined form of CD2 for stability and assembly. Nat Struct Biol 1998;5:778-782.

201. Newcomer ME. Protein folding and three-dimensional domain swapping: a strained relationship? Curr Opin Struct Biol 2002;12:48-53.

202. Gronenborn AM. Protein acrobatics in pairs--dimerization via domain swapping. Curr Opin Struct Biol 2009;19:39-49.

203. Rousseau F., Schymkowitz J., Itzhaki L.S. Implications of 3D domain swapping for protein folding, misfolding and function. In: 2010.

204. Bennett MJ, Sawaya MR, Eisenberg D. Deposition diseases and 3D domain swapping. Structure 2006;14:811-824.

205. Parge HE, Arvai AS, Murtari DJ, Reed SI, Tainer JA. Human CksHs2 atomic structure: a role for its hexameric assembly in cell cycle control. Science 1993;262:387-395.

206. Seeliger MA, Spichty M, Kelly SE, Bycroft M, Freund SM, Karplus M, Itzhaki LS. Role of conformational heterogeneity in domain swapping and adapter function of the Cks proteins. J Biol Chem 2005;280:30448-30459.

207. Wolynes P, Luthey-Schulten Z, Onuchic J. Fast-folding experiments and the topography of protein folding energy landscapes. Chem Biol 1996;3:425-432.

208. Yang S, Cho SS, Levy Y, Cheung MS, Levine H, Wolynes PG, Onuchic JN. Domain swapping is a consequence of minimal frustration. Proc Natl Acad Sci U S A 2004;101:13786-13791.

209. Cho SS, Levy Y, Onuchic JN, Wolynes PG. Overcoming residual frustration in domain-swapping: the roles of disulfide bonds in dimerization and aggregation. Phys Biol 2005;2:S44-S55.

210. Koharudin LM, Viscomi AR, Jee JG, Ottonello S, Gronenborn AM. The evolutionarily conserved family of cyanovirin-N homologs: structures and carbohydrate specificity. Structure 2008;16:570-584.

211. Ding F, Prutzman KC, Campbell SL, Dokholyan NV. Topological determinants of protein domain swapping. Structure 2006;14:5-14.

212. Kundu S, Jernigan RL. Molecular mechanism of domain swapping in proteins: an analysis of slower motions. Biophys J 2004;86:3846-3854.

213. Malevanets A, Sirota FL, Wodak SJ. Mechanism and energy landscape of domain swapping in the B1 domain of protein G. J Mol Biol 2008;382:223-235.

214. Fink AL. Protein aggregation: folding aggregates, inclusion bodies and amyloid. Fold Des 1998;3:R9-23.

215. Gotte G, Bertoldi M, Libonati M. Structural versatility of bovine ribonuclease A. Distinct conformers of trimeric and tetrameric aggregates of the enzyme. Eur J Biochem 1999;265:680-687.

216. Piccoli R, Di DA, D'Alessio G. Co-operativity in seminal ribonuclease function. Kinetic studies. Biochem J 1988;253:329-336.

217. Nenci A, Gotte G, Bertoldi M, Libonati M. Structural properties of trimers and tetramers of ribonuclease A. Protein Sci 2001;10:2017-2027.

218. Sambashivan S, Liu Y, Sawaya MR, Gingery M, Eisenberg D. Amyloid-like fibrils of ribonuclease A with three-dimensional domain-swapped and native-like structure. Nature 2005;437:266-269.

219. Byeon IJ, Louis JM, Gronenborn AM. A captured folding intermediate involved in dimerization and domain-swapping of GB1. J Mol Biol 2004;340:615-625.

220. Jee J, Byeon IJ, Louis JM, Gronenborn AM. The point mutation A34F causes dimerization of GB1. Proteins 2008;71:1420-1431.

221. Louis JM, Byeon IJ, Baxa U, Gronenborn AM. The GB1 amyloid fibril: recruitment of the peripheral beta-strands of the domain swapped dimer into the polymeric interface. J Mol Biol 2005;348:687-698.

222. Kirsten FM, Dyda F, Dobrodumov A, Gronenborn AM. Core mutations switch monomeric protein GB1 into an intertwined tetramer. Nat Struct Biol 2002;9:877-885.

223. Wikstrom M, Drakenberg T, Forsen S, Sjobring U, Bjorck L. Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein L. Comparison with the IgG-binding domains of protein G. Biochemistry 1994;33:14011-14017.

224. Ziolkowska NE, O'Keefe BR, Mori T, Zhu C, Giomarelli B, Vojdani F, Palmer KE, McMahon JB, Wlodawer A. Domain-swapped structure of the potent antiviral protein griffithsin and its mode of carbohydrate binding. Structure 2006;14:1127-1135.

225. Mori T, O'Keefe BR, Sowder RC, Bringans S, Gardella R, Berg S, Cochran P, Turpin JA, Buckheit RW, Jr., McMahon JB, Boyd MR. Isolation and characterization of griffithsin, a novel HIV-inactivating protein, from the red alga Griffithsia sp. J Biol Chem 2005;280:9345-9353.

226. Williams DC, Jr., Lee JY, Cai M, Bewley CA, Clore GM. Crystal structures of the HIV-1 inhibitory cyanobacterial protein MVL free and bound to Man3GlcNAc2: structural basis for specificity and high-affinity binding to the core pentasaccharide from n-linked oligomannoside. J Biol Chem 2005;280:29269-29276.

227. Ziolkowska NE, Wlodawer A. Structural studies of algal lectins with anti-HIV activity. Acta Biochim Pol 2006;53:617-626.

228. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235-242.

229. Liu L, Gronenborn AM. Domain swapping in Proteins. In: 2011.

230. Cafaro V, De LC, Piccoli R, Bracale A, Mastronicola MR, Di DA, D'Alessio G. The antitumor action of seminal ribonuclease and its quaternary conformations. FEBS Lett 1995;359:31-34.

231. Czjzek M, Letoffe S, Wandersman C, Delepierre M, Lecroisey A, Izadi-Pruneyre N. The crystal structure of the secreted dimeric form of the hemophore HasA reveals a domain swapping with an exchanged heme ligand. J Mol Biol 2007;365:1176-1186.

232. Sanders A, Jeremy CC, Higgins LD, Giannini S, Conroy MJ, Hounslow AM, Waltho JP, Staniforth RA. Cystatin forms a tetramer through structural rearrangement of domain-swapped dimers prior to amyloidogenesis. J Mol Biol 2004;336:165-178.

233. Yamasaki M, Li W, Johnson DJ, Huntington JA. Crystal structure of a stable dimer reveals the molecular basis of serpin polymerization. Nature 2008;455:1255-1258.

234. Chen YW, Stott K, Perutz MF. Crystal structure of a dimeric chymotrypsin inhibitor 2 mutant containing an inserted glutamine repeat. Proc Natl Acad Sci U S A 1999;96:1257-1261.

235. Ogihara NL, Ghirlanda G, Bryson JW, Gingery M, DeGrado WF, Eisenberg D. Design of three-dimensional domain-swapped dimers and fibrous oligomers. Proc Natl Acad Sci U S A 2001;98:1404-1409.

236. Dolbier WR. Guide to fluorine nmr for organic chemists. In: 2009.

237. Campos-Olivas R, Aziz R, Helms GL, Evans JN, Gronenborn AM. Placement of 19F into the center of GB1: effects on structure and stability. FEBS Lett 2002;517:55-60.

238. Abbott GL, Blouse GE, Perron MJ, Shore JD, Luck LA, Szabo AG. 19F NMR studies of plasminogen activator inhibitor-1. Biochemistry 2004;43:1507-1519.

239. Ahmed AH, Loh AP, Jane DE, Oswald RE. Dynamics of the S1S2 glutamate binding domain of GluR2 measured using 19F NMR spectroscopy. J Biol Chem 2007;282:12773-12784.

240. Toptygin D, Gronenborn AM, Brand L. Nanosecond relaxation dynamics of protein GB1 identified by the time-dependent red shift in the fluorescence of tryptophan and 5-fluorotryptophan. J Phys Chem B 2006;110:26292-26302.

241. Schuler B, Kremer W, Kalbitzer HR, Jaenicke R. Role of entropy in protein thermostability: folding kinetics of a hyperthermophilic cold shock protein at high temperatures using 19F NMR. Biochemistry 2002;41:11670-11680.

242. Danielson MA, Falke JJ. Use of 19F NMR to probe protein structure and conformational changes. Annu Rev Biophys Biomol Struct 1996;25:163-195.

243. Botos I, O'Keefe BR, Shenoy SR, Cartner LK, Ratner DM, Seeberger PH, Boyd MR, Wlodawer A. Structures of the complexes of a potent anti-HIV protein cyanovirin-N and high mannose oligosaccharides. J Biol Chem 2002;277:34336-34342.

244. Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD. 1H, 13C and 15N chemical shift referencing in biomolecular NMR. J Biomol NMR 1995;6:135-140.

245. Maurer T, Kalbitzer HR. Indirect Referencing of 31P and 19F NMR Spectra. J Magn Reson B 1996;113:177-178.

246. Van Geet AL. Calibration of the methanol and glycol nuclear magnetic resonance thermometers with a static thermistor probe. Analytical Chemistry 1968;40:2227-2229.

247. Atkins P, De Paula J. Atkins' Physical Chemistry. Oxford University Press; 2006.

248. Fromme R, Katiliene Z, Giomarelli B, Bogani F, Mc MJ, Mori T, Fromme P, Ghirlanda G. A monovalent mutant of cyanovirin-N provides insight into the role of multiple interactions with gp120 for antiviral activity. Biochemistry 2007;46:9199-9207.

249. Rohl CA, Baker D. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. J Am Chem Soc 2002;124:2723-2729.

250. Tanaka H, Chiba H, Inokoshi J, Kuno A, Sugai T, Takahashi A, Ito Y, Tsunoda M, Suzuki K, Takenaka A, Sekiguchi T, Umeyama H, Hirabayashi J, Omura S. Mechanism by which the lectin actinohivin blocks HIV infection of target cells. Proc Natl Acad Sci U S A 2009;106:15633-15638.

251. Zweckstetter M, Bax A. Prediction of Sterically Induced Alignment in a Dilute Liquid Crystalline Phase: Aid to Protein Structure Determination by NMR. Journal of the American Chemical Society 2000;122:3791-3792.

252. Brandts JF, Halvorson HR, Brennan M. Consideration of the Possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues. Biochemistry 1975;14:4953-4963.

253. Liu Y, Gierasch LM, Bahar I. Role of Hsp70 ATPase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with NEFs. PLoS Comput Biol 2010;6.