Integration of Molecular Dynamics Simulations with Experimental Data at Different Resolutions

by

Elif Nihal Korkmaz

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Biophysics)

at the

University of Wisconsin – Madison

2016

Date of final oral examination: May 12, 2016

The dissertation is approved by the following members of the Final Oral Committee:

Qiang Cui, Professor, Department of Chemistry
Ivan Rayment, Professor, Department of Biochemistry
Eric Strieter, Professor, Department of Chemistry
Thomas Record, Professor, Department of Biochemistry
John Markley, Professor, Department of Biochemistry

Integration of Molecular Dynamics Simulations with Experimental Data at Different Resolutions

Elif Nihal Korkmaz

Under the supervision of Professor Qiang Cui

At the University of Wisconsin – Madison

Abstract

Three-dimensional structures of macromolecules have been critical to understanding the function of macromolecules, however the conformational ensemble is difficult to determine experimentally. Computational techniques, when combined with experimental data, can provide key insights into the conformational dynamics of proteins. Molecular dynamics (MD) simulation is an effective technique to provide structure and function relationship through exploration of the conformational energy landscape accessible to macromolecules.

MD can function on different resolutions; it can provide details about individual atoms, or it can provide lower resolution information such as the motion and orientations of multi-megadalton domains. Although the technique has been improved tremendously, for making predictions on subtle effects, experimental validation is still needed. On an experimental point of view, the validity of the techniques and approximations used can be tested though computational methods. The procedure is most efficient when the right computational approach with the right resolution is coupled with experimental data available.

In my PhD work, I employed MD simulation techniques to provide a deeper understanding of the relationship between structure, conformational dynamics and function with the available experimental data obtained by X-ray crystallography, nuclear magnetic resonance (NMR) and small angle X-ray scattering (SAXS). To achieve this goal, I have focused on three different systems. The specific aims of my work include:

(1) To understand the effect of local interactions on the structure and stability through studying lymphotactins (hLtns); I used atomistic simulations in the light of the available NMR data.

(2) To obtain conformational ensembles for Ubiquitin (Ub) chains in order to investigate the selectivity of deubiquitinases (DUBs); I used a combination of homology modeling and implicit solvent molecular dynamics filtered via experimental SAXS data.

(3) To probe the role of skip residues in determining the structure and flexibility of a single repeat segment in the myosin rod and ultimately to construct a high-resolution model for a large portion of the rod; I used an arsenal of simulation techniques with different resolutions.

These three aims are designed to enhance the ability to use MD methods at different resolutions in coordination with experimental data to better understand protein function.

Acknowledgements

My name may be on the front cover of this thesis, but I am by no means its sole contributor. There are a number of people behind this piece of work who needs to be acknowledged and thanked here.

First and foremost, I express my deepest gratitude to my Ph.D. advisor, Prof. Qiang Cui, for his guidance, encouragement and patience. I would like to thank him for having his door always open for us and pushing us to achieve further than we think we can. His enthusiasm and passion for science have constantly inspired me. From him, I have learnt how to set high goals and strive for perfection in my work and life. I consider myself lucky to have received my training and I am a better scientist thanks to him.

I was really fortunate to have involved in different projects and collaborate with different groups. This led me to gain a wider perspective and think more critically. I would like to thank our collaborator Prof. Ivan Rayment for including me in the myosin project and the many fruitful discussions we had over the years. He is a great mentor and his positive attitude and enthusiasm always made me work harder. I would also like to thank Prof. Eric Strieter and Prof. Brian Volkman for including me in their projects and sharing their knowledge and expertise with me. I have learned a lot about how to do interesting and rigorous research through my partnerships with them. In addition, I would to thank Prof. Thomas Record and Prof. John Markley for being on my thesis committee and for their valuable comments.

I am also lucky to have worked with many talented scientists during these collaborations. I have admired Dr. Keenan Taylor's work ethic and intellect. It really showed that he loves what he does. He also has great patience to have dealt with my endless e-mails and questions. Michael Andreas, Dr. Grace Pham and Ambar Rana are a few of the other names with whom I had the chance to work. Their data and discussions helped to form the foundation of this thesis.

Graduate school is a long journey and I am happy to have met and worked along with some wonderful people. I would like to thank Dr. Michael Daily, Dr. Toshi Mori, Dr. Daniel Roston, Dr. Guanhua Hou, Dr. Michael Gauss, Dr. Leili Zhang, Chang Yun Son and Haiyun Jin for turning the office into a more fun environment.

I would like to thank my family and friends for their love and support. My mother Hilal Korkmaz and my sister Ipek Korkmaz deserve the biggest thanks for the daily, sometimes hour-long phone calls. My graduate school years wouldn't have been the same without the friendships of Arzu Uyar, Gamze Gursoy, Nazli Ozbek, Merve Ozen, Narae Lee, Chelcie Eller, Elif Sezin Seven, Berk Yavuzoglu, Nalan Erol, Burcu Kabatas, Omar Demerdash, and Ergun Tuncay and more. Some of them may be across the Atlantic but their friendship was always with me.

Last but not the least, I am lucky to have met Graham Erwin during graduate school. He is a great friend, my partner in crime, a good teacher and an intelligent scientist. He has proven himself to be the most patient person I've known, through driving lessons, editing my writing and enduring my constant chirping. He has contributed a great deal to my work in numerous ways; from brainstorming with me when I'm stuck to baking me the best brownies I have had. Without his kindness and support my work wouldn't have been complete; Madison wouldn't have been as much fun without him either.

Contents

List of Figures

List of Tables

Chapter 1

Wisconsin Initiative for Science Literacy: Introduction for a General Audience

1.1 Overview

The world around is always changing because molecules move. To understand main concepts in physics, chemistry and biology, it is essential to study the motion of molecules. A fundamental goal of research is to learn how molecules, specifically proteins, function. This implies that at the microscopic level the movement of atoms and molecules provide certain functions such as chemical reactions, molecular recognition and protein synthesis. In this work, I have used the power of computational methods to understand the structure and function of a few different proteins in the human body, in the light of some experimental findings.

1.2 What are Proteins and Amino Acids?

To start with, the amino acid is the building block of proteins. There are twenty naturally occurring amino acids and each one of them is different. Basically, multiple amino acids are linked together to form chains called proteins. Amino acids, linked together in different combinations, are used in different combinations to build thousands of proteins that the body needs to survive. Not just humans though, all organisms need proteins to maintain life. They form your muscles, they form your skin and they are even on the surface of cells.

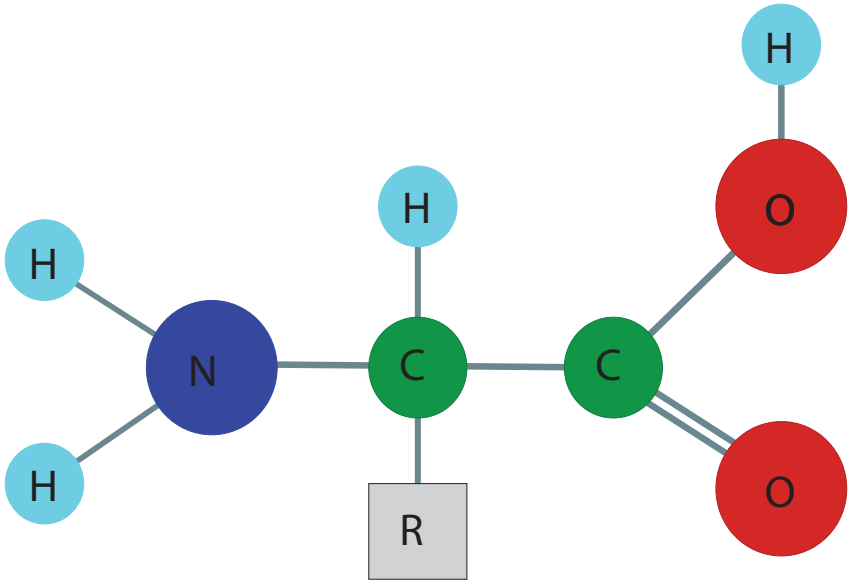


Figure 1.1: The structure of an alpha amino acid. The side chain is depicted by an R group.

Structure of all amino acids are similar. All amino acids are formed of a backbone and a side-chain. The backbone is the same for all amino acids and it is formed by a carboxyl group (COOH), an amino group (NH_2) and a second carbon that connects the those two groups with the side-chain (R) (Fig. 1.1). The side-chain is the what makes each amino acid different. There are two main categories for the side-chains: polar and non-polar. These names imply whether the electron distribution is uniform within the chain, in other words it refers to the way the side groups interact with their environment. The polar and nonpolar nature of the side-chains help the amino acids to point towards water (hydrophilic) or away from water (hydrophobic).

As proteins are synthesized, they start as a straight chain of amino acids. As the chain grows, it starts to form twists and turns; similar to a twisted piece of yarn, it starts to curl up. Some amino acids within the proteins prefer to interact with the others along the chain form this telephone cable like structure, that particular structure is called an α-helix. While others take the shape of a folded sheet and remain on the 2D plane, in this case the structure is called a β-sheet. There are also flexible regions that link those two structural elements together. When the protein is long enough, these structural elements also interact with each other.

Considering that there are twenty natural amino acids, there are endless combinations of protein sequences (primary structure), α-helices and β-sheets (secondary structure) and how those structural elements interact with other (tertiary structure) to form the ultimate structure of the protein. And the structure of the protein is crucial because it determines its function, mechanism and interactions with structures in the cell.

1.3 What are Molecular Dynamics Simulations?

Proteins form different structures, but these structures are often dynamic in nature. Each protein accommodates some degree of flexibility and movement. And it is crucial to study those conformational (structural) states to be able to understand how those proteins function and interact with others. Understanding the conformational states can help shed light on the nature of different diseases and can provide drug design and treatment options.

Molecular Dynamics (MD) simulation technique is a computational method to predict those conformational states. MD solves Newton's equation of motion over an atomistic model of a molecule to obtain the trajectory of its motion ($F = m \times a$). The positions of the all atoms in the system are recorded at specified time points through the simulation. A compilation of force constants, which is called a force field is obtained through experimental data and it is used to describe the strength of interaction between the atoms.

Cell condition are mimicked through solvating the molecule in a large water box and adding ions.

1.4 Short Summary

Determining the different structures a protein adopts is difficult experimentally, and computational techniques can be used to gather information on the conformational states of proteins. MD simulation technique is an efficient tool to obtain conformation states to better understand the structure and function relationship. The method is highly developed and has high accuracy, however experimental validation is still required. In my PhD work, I used the MD simulation technique along with the available experimental data obtained by our collaborators through X-ray crystallography, nuclear magnetic resonance (NMR) and small angle X-ray scattering (SAXS). I have focused on three main proteins to showcase how different levels of details incorporated in the simulations work with experimental data available and specific biological problem we are interested in. We have learnt that when the right computational approach is couples with the right experimental data, we can learn gain a deeper understanding on the structure and function and proteins.

Chapter 2

Introduction

2.1 Overview

Molecular dynamics (MD) simulations are powerful tools for understanding the physical basis of the structure and function of biological macromolecules. The conventional view of protein structures as relatively rigid is replaced by a dynamic model in which the functional states of a protein fluctuate around a single native state [1]. Thus the internal motions and resulting conformational changes play an essential role in their function, with major conformational changes typically taking place on timescales ranging from microseconds to seconds. MD simulations are powerful tools for modeling such motions, the characterization of which provides insight into the mechanism and function with different space and time resolutions, which are cumbersome to probe experimentally. With the recent developments in molecular dynamics algorithms, software, and computer hardware, simulation studies of biomolecular systems reached biophysically relevant time scales (microsecond to millisecond) [2–12].

Although early force field development was limited by the lack of direct comparisons between simulation and experiment, in the recent years several laboratories have demonstrated direct calculation of NMR observables from protein simulations [13–16]. Computational techniques have certainly progressed a lot in terms of speed and force field development and are becoming more powerful. However, validation of results or prediction through experimental data is crucial since the subtle effects can be overlooked due to approximation of different simulation methods. For instance, the implicit solvent method is not very efficient in predicting the ion effects since it lacks the explicit addition of water and the ions. Using experimental data helps us, not

only to improve, but also to validate predictions from computational methods.

Experimental X-ray or NMR structures are usually used as the starting point for studies of the relationship between protein structure and function. Such structures, however, do not generally describe the dynamic properties that may be as important to protein function and regulation as the atomistic positioning of backbone and side-chain atoms. MD simulations provide a unique ability to directly describe the dynamical aspects of protein structure by tracking the time-dependent positions of all atoms in the system. Moreover, from an experimental point of view, the approximations followed to improve the techniques need to be tested. Of course validation through experiments is often difficult and expensive and sometimes even not even possible due to limitations of each technique. For instance, in X-ray crystallography different conditions and solutions are used to improve crystallization or structures are fused with other proteins to enhance solubility. To test whether these modifications have an impact in the ensemble of conformations, MD simulations can be addressed.

Furthermore, without simulations experimental techniques cannot provide both spatial and time resolution. SAXS and wide angle X-ray scattering (WAXS) only provide low resolution structural information, whereas X-ray crystallography gives a static image. NMR, on the other hand, provides a small number of conformations fluctuating around the equilibrium conformation that are present at the conditions used in the experiment. Simulations are essential to accompany those experimental results to test the validity of the approximations, test hypotheses made based on the experiments and to improve limited structural knowledge obtained by experimental techniques.

Therefore, the most efficient approach to study structure and function relationship is to use a combination of experimental and computational techniques. Both approaches should be used together to improve and validate our techniques, results and hypotheses. There has been considerable effort in combining experiments with simulations. An important number of studies have focused on force field development and improvement with NMR observables. [13–17] There are several efforts to improve and filter the simulation results, but those are mostly limited to low resolution techniques via SAXS and WAXS. [18–27] However, different computational methods can be used in combination with a variety of techniques. Furthermore, some biological problems require integrating a multi-scale approach. For instance, constructing an atomistic model for the assembly myosin fibers formed by myosin rod in the thick filament of the cardiac muscle myosin essen-

tially requires atomistic level high resolution data. Although, the modeling should start with a low resolution method for construction of the initial model via homology modeling and coarse-graining. The level of detail can then be improved based on atomistic simulations and fluorescence resonance energy transfer (FRET). Different biological problems require different time and space resolutions, and the complex nature of biological processes require a multi-scale approach. In this work, I combine experimental data from our collaborators with MD simulations with varying time and space resolutions.

2.2 Innovation

The primary goal of my doctoral work is to employ MD simulation methods to better understand the relationship between protein structure, conformational ensemble, and function by incorporating the available experimental data obtained from X-ray crystallography, NMR and SAXS. Previous studies focused mainly on low-resolution principles. [21, 24, 28–31] In this work, I have used different levels of resolution while incorporating experimental data. Atomistic level studies on hLtn teach the effect of local interactions on the stability, whereas studying conformational ensembles of Ub chains teaches us how to study large conformational changes. Working on myosins, on the other hand, provides an opportunity to combine both atomistic explicit solvent and implicit solvent methods.

2.3 Overview of the Methods Used

Using the right force field along with the right solvation methods is crucial for accurate representation of systems simulated. The choice of solvation method is critical for the accuracy and the speed of the simulations, and it should be chosen carefully according to problem in hand. Explicit solvent simulations involves the presence of water molecules explicitly in the simulation box. While this approach enables addition of ions, small molecules and other solvent molecules in the simulation and provides a more reliable atomistic detail compared to implicit solvent approach, it is computationally expensive. In implicit solvent, on the other hand, effect of solvation is applied through a continuum mechanics model, and ion concentration is treated with Debye-Huckel. Lack of explicit water molecules reduces the number of atoms to be simulated dramatically, and hence allows us to simulate faster. However, implicit solvent should be used with caution since it might overlook the ionization effects.

For implicit solvent simulations, ff99SBnmr and ff14SBonlysc force fields are used along with the

Generalized Born (GB) approach. I chose to use GB along with the implicit solvent simulations owing to the fact that GB simulations provide $\sim$100 fold higher efficiency compared to the explicit solvent when GPUs are incorporated. [3, 32, 33] This means that by using implicit solvent instead of the explicit solvent, a time efficiency of 100 fold is obtained. This allows the user to larger system or longer trajectories without having to wait months or even years for explicit solvent simulations to obtain similar conformational ensembles. Longer time-scales can allow us to identify new physical interactions and structural motifs important for the stability of each fold. Production implicit solvent simulations are carried out for a minimum of 600 ns up to 3 μs. Salt concentration is regulated through Debye-Huckel theory, and the concentration is set to the physiological NaCl concentration, 0.15 M unless otherwise stated.

Explicit solvent method is better suited for studying atomistic details and electrostatic interactions; however, implicit solvent provides efficiency and thus wider sampling of the conformational space. Each method will be carefully fitted to each system studied.

2.4 Synopsis of the Biological Questions Studied

2.4.1 Conformational Duality in the Human Lymphotactin System

For the vast majority of proteins, the native conformations are determined by the sequences. [1] The functional states of a protein fluctuate around a single, minimum energy, native state. However, this empirical rule has been broken by the discovery of metamorphic proteins. The human lymphotactin (hLtn) system is a metamorphic protein that adopts two distinct conformations depending on the environmental conditions. The first conformation (hLtn10) exists at $\sim$10 °C with 0.20 M salt and exists as a monomer, while the second conformation (hLtn40) is most stable at $\sim$40 °C with no salt and exists as a dimer (Fig. 2.1). Under physiological conditions ($\sim$37 °C and 0.15 M NaCl) two folds co-exist in equilibrium. In fact, both conformations are functional and physiologically relevant. [34] hLtn10 binds to XCR1 and thereby induces intracellular calcium mobilization and chemotaxis. Yet hLtn40 binds to glycosaminoglycans (GAGs) but not the Ltn XCR1 receptor, and this interaction is necessary for leukocyte recruitment by chemokines. [34]

The overall transition between hLtn10 and hLtn40 takes place at a rate of $\sim$1 s, which is indicative of a large energy barrier. [35] Although the exact mechanism is unknown, the transition from monomer (hLtn10) to dimer (hLtn40) is not merely an isomerization process. The interconversion between the two states requires

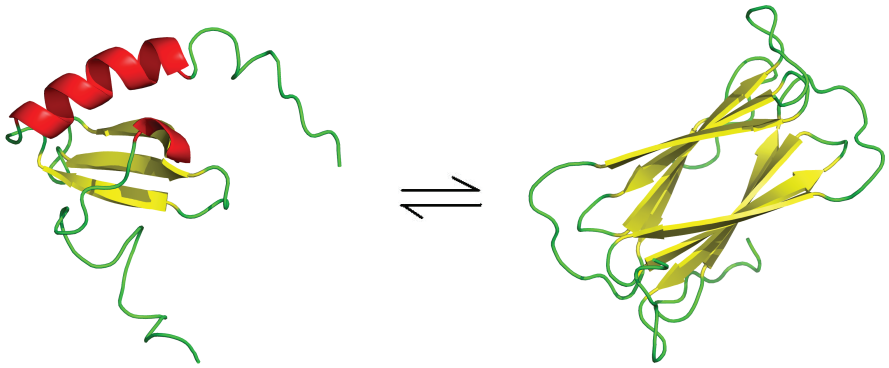


Figure 2.1: Ltn native-state interconversion. Structural rearrangement in the Ltn native state exchanges a functional GPCR agonist with the canonical monomeric chemokine fold (Ltn10) for a novel all β-sheet dimer with high affinity for extracellular matrix GAG (Ltn40). The two Ltn native state structures are equally populated at physiological conditions of temperature and ionic strength, and interconvert at a rate of $\sim$1 s^{-1}.

complete unfolding through disruption of the hydrogen bonding network defining the secondary structure and refolding. Between the two structures, all long-range interactions are replaced by a new hydrogen bonding network, tertiary and quaternary contacts. [35, 36] Residues buried in the core become solvent exposed during the transition from monomer to dimer. To accommodate the dimer interface, this structural change produces a new hydrophobic surface on hLtn40.

NMR structures from our longtime collaborator, Prof. Brian Volkman reveal that Val15, Val37, Phe39, Thr41 and Ala49 are in the hydrophobic core of hLtn10, yet they are solvent exposed in hLtn40. Leu34, Ala36, Ile38 and Ile40 are buried in hLtn40, but they reside on the surface of hLtn10. Moreover, when Trp55, normally buried in the hydrophobic core of hLtn10, is mutated to Asp, the equilibrium shifts towards hLtn40. [34, 36, 37] Comparison of surface plasmon resonance analysis of heparin binding affinities [38] with HSQC spectra of WT, R23A and R43A mutants [39] revealed that each change enhances hLtn10 stability and shifts the hLtn40 conversion to higher temperatures. This effect increases for R23A/R43A double mutant. [39] The change in unfolding free energies upon mutation of charged residues indicated that although Lys25 and Arg35 reside in the same core region with Arg23 and Arg43, their mutations do not affect the unfolding free energies. [40] Work by Formaneck et al [41] provided evidence that the Arg23–Arg43 association is stabilized with a chloride ion in solutions of higher (200 mM) salt concentration. In the absence of the stabilizing chloride, the conversion to hLtn40 prevents the electrostatic repulsion between Arg23 and Arg43 by placing them $\sim$20 Å apart. These findings imply that electrostatic interactions play an important role in hLtn10-hLtn40 conversion.

In Chapter 2, I have identified additional physical interactions and structural motifs that govern the relative stabilities of hLtn10 and hLtn40 species through extended implicit and explicit solvent simulations. Tweaking those interactions can help regulate the hLtn conversion rates, through stabilization (or destabilization) of either fold. Moreover, I have explored how different charged residues (e.g. Lys25 vs Arg23) alter the relative stability of the folds.

2.4.2 Ubiquitin Dimers and Selectivity of DUBs

One of the most common post-translational modifications in eukaryotic cells is ubiquitination, the covalent attachment of a single ubiquitin (Ub) or a chain of Ubs to the target protein. Ubiquitination alters a protein's function and its ability to bind to specific biological partners depending on the precise nature of the

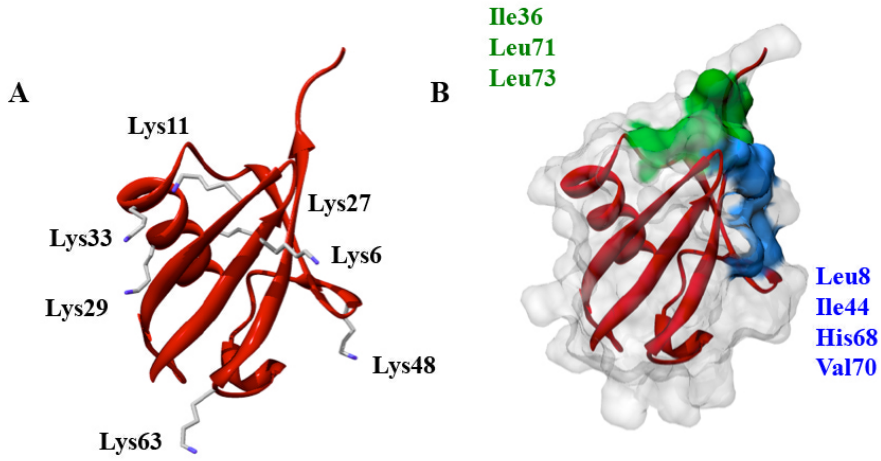


Figure 2.2: Representation of a Ub monomer (A) Lysines are marked and (B) The hydrophobic patches are shown.

ubiquitination. For instance, modification of a protein with Lys63-linked poly-Ub has a key role in signaling for many biological processes, whereas proteins bound to Lys48-linked chains are targeted for degradation by the proteasome. The overall structure of ubiquitin is compact and held together by a strong network of hydrogen bonds. Ub covalently binds to other Ubs through one of the seven lysines (Lys6, Lys11, Lys27, Lys29, Lys33, Lys48 and Lys63 on one Ub (Fig. 2.2) and the C-terminal glycine on the target Ub, forming an iso-peptide bond. [42] A hydrophobic patch on the surface formed by Leu8, Ile44 and Val70 [43, 44] enables recognition and binding to target proteins, and this patch is critical for the core processes of Ubs such as proteasomal degradation and endocytosis. However, it is not yet understood if this hydrophobic patch has a role in the formation of the poly-Ub chain itself. An additional hydrophobic patch, consisting of Ile36 [45], Leu71 and Leu73 [44], also provides binding specificity to target proteins.

Ubiquitination is reversible through deubiquitinases (DUB). DUBs directly cleave Ub modifications from the target protein or trim the attached oligomeric Ub chains. Since ubiquitination plays an important role for many cellular processes such as cell-cycle progression, viral infection and immune response [46–50], the balance of attachment and cleavage of Ub to target proteins is crucial, and defects in these processes are associated with numerous diseases, including cancers, neurological disorders and cardiac problems. [46–50] Although ubiquitination and deubiquitination processes are crucial for the homeostasis of the cell, the exact mechanism through which DUBs act is not yet known. One major challenge is the lack of structural information for most Ub chains to extract information on structure-function relationships.

The Strieter group has recently developed an elegant chemical approach that enables access to a wide array of chemically-defined Ub chains via free radical thiol-ene polimeryzation. [51, 52] With this thiolene polymerization technique, specific Ub chains can be synthesized and used as substrates in activity assays with various DUBs. It is already known that DUBs are selective in hydrolyzing Ub chains. [53] This technology led to the discovery that DUBs cleave not only Ub monomer but Ub chains, although selectively based on the size and the linkages. [52] The oligomeric Ub chains are formed via modification of a single Ub subunit with two or more Ub molecules through the seven lysine residues found in Ub structures (Lys6, Lys11, Lys27, Lys29, Lys33, Lys48 and Lys63). [54, 55]

In Chapter 3, I have constructed ensembles of conformations for defined Ub chains to investigate the selectivity of DUBs. In order to unveil molecular and mechanistic details underlying the debranching activity

of DUBs, an arsenal of techniques is used in collaboration, including small angle X-ray scattering (SAXS), homology modeling and molecular dynamics simulations.

2.4.3 Structure, Function, and Flexibility of the Conserved Skip Residues in the Cardiac Muscle Myosin and Assembly of a Larger Segment of the Myosin Rod

Muscle contraction is driven by the interaction between myosin and actin and the associated hydrolysis of ATP, but the long-range transmission of force resides in the fundamental ability of myosin and actin to self-assemble into organized thick and thin filaments respectively. This region can be divided into two parts; subfragment-2 (S2) and light meromyosin (LMM). The ability to self-assemble into thick filaments at physiological ion concentration resides in LMM. LMM is within the cardiac β-myosin and extends from approximately residues 1140 to 1935 [56]. Conversely, S2, which is soluble at low ionic strength, allows the motor domains or myosin heads to extend a variable distance away from the backbone of the thick filament to interact with actin during muscle contraction. The myosin tail is predicted to be mostly α-helical and to form a dimeric coiled-coil. The physiological importance of this region of myosin is underscored by the large number of mutations in this region of the motor that lead to cardiomyopathies.

The myosin thick filament has been the subject of many structural studies including both X-ray diffraction and electron microscopy [57–60]. It has been shown that the thick filament is bipolar, where the center of the filament is formed from an anti-parallel arrangement of myosin molecules. This leaves a central bare zone, approximately 160 nm wide that lacks myosin heads. On either side of the bare zone the myosin molecules assemble in parallel arrays to create fibers that are $\sim$1.8 μm in length.

The sequence for the myosin rod is highly conserved and displays an overall distribution of amino acids that is characteristic of an α-helical coiled-coil. Furthermore, it shows a remarkable 28 amino acid repeat that itself exhibits an alternating pattern of positively and negatively charged amino acids suggestive of a staggered interaction between adjacent rods in the thick filament [61]. Interestingly, the repeat sequence is disrupted at four locations in skeletal and cardiac myosin by the presence of additional skip residues.

The role of the skip residues has not been defined, even though the spacings of these (Fig. 2.3) are completely conserved across all class II myosins [62]. Indeed, sequence analysis of the charge distributions across a wide range of myosins suggests that the skip residues have a critical role in keeping long-range charge distributions on opposing myosin rods in phase [62]. The residues immediately surrounding the skip

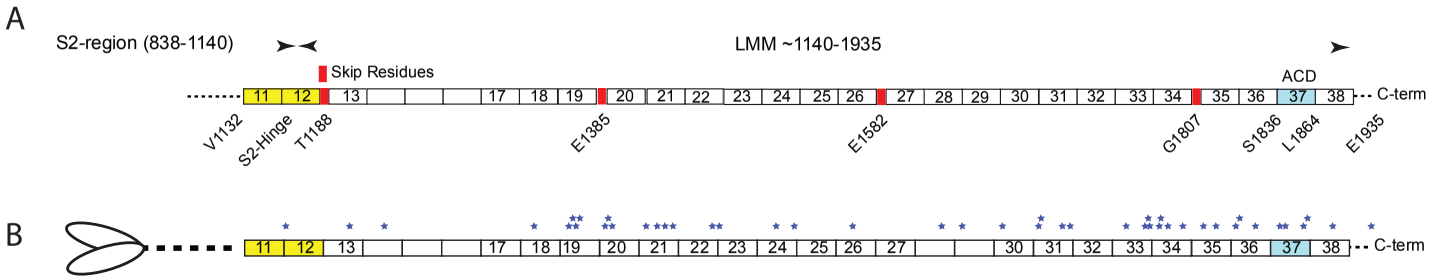


Figure 2.3: Schematic representation of the myosin repeat segments (A) Arrangement of charge repeats in skeletal/cardiac myosin. (the Skip residues are shown in red.) (B) The distribution of myopathy mutations in human β-cardiac myosin (MYH7). Forty-six mutations have been identified in the LMM region of myosin. Many of these are located in regions predicted to be involved in protein-protein interactions.

residues show a lower predicted coiled coil propensity compared to that of a canonical motif, suggesting that these regions might exhibit conformational flexibility that may be important in allowing the filaments to pack together in the thick filament.

Despite the considerable effort over the past forty years to model the arrangement of myosin rods in the thick filament [59, 60, 63], the molecular organization of the myosin rods is not well understood. The high degree of sequence conservation and sensitivity to mutations that cause cardiomyopathies strongly indicates a more ordered series of interactions within the filament. In Chapter 4, I present our work on the role of the skip residues as well as the effect of surrounding residues. Also in Chapter 5, I have shown our work on building an atomistic model for larger portion of the myosin rod by combining our knowledge and expertise on working with different resolutions.

2.5 References

[1] C. B. ANFINSEN, *Science* **181**, 223 (1973).

[2] R. O. DROR, R. M. DIRKS, J. P. GROSSMAN, H. F. XU, and D. E. SHAW, *Annual Review of Biophysics, Vol 41* **41**, 429 (2012).

[3] A. W. GOTZ, M. J. WILLIAMSON, D. XU, D. POOLE, S. LE GRAND, and R. C. WALKER, *Journal of chemical theory and computation* **8**, 1542 (2012).

[4] J. P. GROSSMAN, J. S. KUSKIN, J. A. BANK, M. THEOBALD, R. O. DROR, D. J. IERARDI, R. H. LARSON, U. BEN SCHAFER, B. TOWLES, C. YOUNG, and D. E. SHAW, *Acm Sigplan Notices* **48**, 549 (2013).

[5] J. KLEPEIS, K. LINDORFF-LARSEN, R. DROR, and D. SHAW, *Current Opinion in Structural Biology* **19**, 120 (2009).

[6] K. KOHLHOFF, D. SHUKLA, M. LAWRENZ, G. BOWMAN, D. KONERDING, D. BELOV, R. ALTMAN, and V. PANDE, *Nature Chemistry* **6**, 15 (2014).

[7] T. LANE, G. BOWMAN, K. BEAUCHAMP, V. VOELZ, and V. PANDE, *Journal of the American Chemical Society* **133**, 18413 (2011).

[8] K. LINDORFF-LARSEN, P. MARAGAKIS, S. PIANA, M. EASTWOOD, R. DROR, and D. SHAW, *Plos One* **7** (2012).

[9] J. R. MOORE, L. LEINWAND, and D. M. WARSHAW, *Circ Res* **111**, 375 (2012).

[10] V. PANDE, I. BAKER, J. CHAPMAN, S. ELMER, S. KHALIQ, S. LARSON, Y. RHEE, M. SHIRTS, C. SNOW, E. SORIN, and B. ZAGROVIC, *Biopolymers* **68**, 91 (2003).

[11] D. E. SHAW, *Abstracts of Papers of the American Chemical Society* **238** (2009).

[12] D. E. SHAW, M. M. DENEROFF, R. O. DROR, J. S. KUSKIN, R. H. LARSON, J. K. SALMON, C. YOUNG, B. BATSON, K. J. BOWERS, J. C. CHAO, M. P. EASTWOOD, J. GAGLIARDO, J. P. GROSSMAN, C. R. HO, D. J. IERARDI, I. KOLOSSVARY, J. L. KLEPEIS, T. LAYMAN, C. MCLEAVEY, M. A.

MORAES, R. MUELLER, E. C. PRIEST, Y. B. SHAN, J. SPENGLER, M. THEOBALD, B. TOWLES, and S. C. WANG, *Communications of the Acm* **51**, 91 (2008).

[13] K. BEAUCHAMP, Y. LIN, R. DAS, and V. PANDE, *Journal of Chemical Theory and Computation* **8**, 1409 (2012).

[14] V. HORNAK, R. ABEL, A. OKUR, B. STROCKBINE, A. ROITBERG, and C. SIMMERLING, *Proteins-Structure Function and Bioinformatics* **65**, 712 (2006).

[15] D.-W. LI and B. RAFAEL, *Angewandte Chemie* **122**, 6930 (2010).

[16] S. SHOWALTER and R. BRUSCHWEILER, *Journal of Chemical Theory and Computation* **3**, 961 (2007).

[17] M. BUCK, S. BOUGUET-BONNET, R. PASTOR, and A. MACKERELL, *Biophysical Journal* **90**, L36 (2006).

[18] P. BERNADO, E. MYLONAS, M. PETOUKHOV, M. BLACKLEDGE, and D. SVERGUN, *Journal of the American Chemical Society* **129**, 5656 (2007).

[19] M. A. GRAEWERT and D. I. SVERGUN, *Current Opinion in Structural Biology* **23**, 748 (2013).

[20] J. KOFINGER and G. HUMMER, *Physical Review E* **87** (2013).

[21] M. PELIKAN, G. HURA, and M. HAMMEL, *General Physiology and Biophysics* **28**, 174 (2009).

[22] M. V. PETOUKHOV, I. M. L. BILLAS, M. TAKACS, M. A. GRAEWERT, D. MORAS, and D. I. SVERGUN, *Biochemistry* **52**, 6844 (2013).

[23] K. RAVIKUMAR, W. HUANG, and S. YANG, *Journal of Chemical Physics* **138** (2013).

[24] B. ROZYCKI, Y. KIM, and G. HUMMER, *Structure* **19**, 109 (2011).

[25] N. SIBILLE and P. BERNADO, *Biochemical Society Transactions* **40**, 955 (2012).

[26] T. TENNO, K. FUJIWARA, H. TOCHIO, K. IWAI, E. MORITA, H. HAYASHI, S. MURATA, H. HIROAKI, M. SATO, K. TANAKA, and M. SHIRAKAWA, *Genes To Cells* **9**, 865 (2004).

[27] S. YANG, L. BLACHOWICZ, L. MAKOWSKI, and B. ROUX, *Proceedings of the National Academy of Sciences of the United States of America* **107**, 15757 (2010).

[28] P. C. CHEN and J. S. HUB, *Biophys J* **107**, 435 (2014).

[29] M. D. DAILY, L. MAKOWSKI, G. N. PHILLIPS, and Q. CUI, *Chemical Physics* **396**, 84 (2012).

[30] M. D. DAILY, G. N. PHILLIPS, and Q. CUI, *Plos Computational Biology* **7** (2011).

[31] A. B. DATTA, G. L. HURA, and C. WOLBERGER, *Journal of molecular biology* **392**, 1117 (2009).

[32] M. FEIG and C. BROOKS, *Current Opinion in Structural Biology* **14**, 217 (2004).

[33] J. MONGAN, C. SIMMERLING, J. MCCAMMON, D. CASE, and A. ONUFRIEV, *Journal of Chemical Theory and Computation* **3**, 156 (2007).

[34] R. TUINSTRA, F. PETERSON, S. KUTLESA, E. ELGIN, M. KRON, and B. VOLKMAN, *Proceedings of the National Academy of Sciences of the United States of America* **105**, 5057 (2008).

[35] R. TYLER, N. MURRAY, F. PETERSON, and B. VOLKMAN, *Biochemistry* **50**, 7077 (2011).

[36] E. KULOGLU, D. MCCASLIN, J. MARKLEY, and B. VOLKMAN, *Journal of Biological Chemistry* **277**, 17863 (2002).

[37] E. KULOGLU, D. MCCASLIN, M. KITABWALLA, C. PAUZA, J. MARKLEY, and B. VOLKMAN, *Biochemistry* **40**, 12486 (2001).

[38] F. C. PETERSON, E. S. ELGIN, T. J. NELSON, F. ZHANG, T. J. HOEGER, R. J. LINHARDT, and B. F. VOLKMAN, *J Biol Chem* **279**, 12598 (2004).

[39] B. F. VOLKMAN, T. Y. LIU, and F. C. PETERSON, *Methods in enzymology* **461**, 51 (2009).

[40] R. TYLER, J. WIETING, F. PETERSON, and B. VOLKMAN, *Biochemistry* **51**, 9067 (2012).

[41] M. FORMANECK, L. MA, and Q. CUI, *Journal of the American Chemical Society* **128**, 9506 (2006).

[42] W. COOK, L. JEFFREY, M. CARSON, Z. CHEN, and C. PICKART, *Journal of Biological Chemistry* **267**, 16467 (1992).

[43] R. BEAL, Q. DEVERAUX, G. XIA, M. RECHSTEINER, and C. PICKART, *Proceedings of the National Academy of Sciences of the United States of America* **93**, 861 (1996).

[44] S. RAASI, R. VARADAN, D. FUSHMAN, and C. M. PICKART, *Nature Structural & Molecular Biology* **12**, 708 (2005).

[45] J. M. WINGET and T. MAYOR, *Molecular Cell* **38**, 627 (2010).

[46] I. DIKIC, S. WAKATSUKI, and K. WALTERS, *Nature Reviews Molecular Cell Biology* **10**, 659 (2009).

[47] L. HICKE, H. SCHUBERT, and C. HILL, *Nature Reviews Molecular Cell Biology* **6**, 610 (2005).

[48] S. SCHLOSSAREK, N. FREY, and L. CARRIER, *Journal of Molecular and Cellular Cardiology* **71**, 25 (2014).

[49] H. D. ULRICH and H. WALDEN, *Nature Reviews Molecular Cell Biology* **11**, 479 (2010).

[50] N. M. WEATHINGTON and R. K. MALLAMPALLI, *Journal of Clinical Investigation* **124**, 6 (2014).

[51] V. H. TRANG, E. M. VALKEVICH, S. MINAMI, Y. C. CHEN, Y. GE, and E. R. STRIETER, *Angew Chem Int Ed Engl* **51**, 13085 (2012).

[52] E. M. VALKEVICH, R. G. GUENETTE, N. A. SANCHEZ, Y. C. CHEN, Y. GE, and E. R. STRIETER, *Journal of the American Chemical Society* **134**, 6916 (2012).

[53] J. LICCHESI, J. MIESZCZANEK, T. MEVISSEN, T. RUTHERFORD, M. AKUTSU, S. VIRDEE, F. EL OUALID, J. CHIN, H. OVAA, M. BIENZ, and D. KOMANDER, *Nature Structural & Molecular Biology* **19**, 62 (2012).

[54] F. MATTIROLI and T. SIXMA, *Nature Structural & Molecular Biology* **21**, 308 (2014).

[55] S. MISAGHI, P. GALARDY, W. MEESTER, H. OVAA, H. PLOEGH, and R. GAUDET, *Journal of Biological Chemistry* **280**, 1512 (2005).

[56] S. LOWEY, H. S. SLAYTER, A. G. WEEDS, and H. BAKER, *J Mol Biol* **42**, 1 (1969).

[57] H. A. ALKHAYAT, R. W. KENSLER, E. P. MORRIS, and J. M. SQUIRE, *J Mol Biol* **403**, 763 (2010).

[58] H. A. ALKHAYAT, E. P. MORRIS, R. W. KENSLER, and J. M. SQUIRE, *J Struct Biol* **163**, 117 (2008).

[59] R. CRAIG and J. L. WOODHEAD, *Curr Opin Struct Biol* **16**, 204 (2006).

[60] J. SQUIRE, M. CANTINO, M. CHEW, R. DENNY, J. HARFORD, L. HUDSON, and P. LUTHER, *J Struct Biol* **122**, 128 (1998).

[61] A. D. MCLACHLAN and J. KARN, *Nature* **299**, 226 (1982).

[62] R. STRAUSSMAN, J. M. SQUIRE, A. BEN-YA'ACOV, and S. RAVID, *J Mol Biol* **353**, 613 (2005).

[63] M. W. CHEW and J. M. SQUIRE, *J Struct Biol* **115**, 233 (1995).

Chapter 3

Investigation of Electrostatics and Hydrophobic Effects in the Metamorphic Protein Human Lymphotactin[1]

[1]Reprinted (adapted) with permission from "Interplay of electrostatics and hydrophobic effects in the metamorphic protein human lymphotactin", E. N. Korkmaz, B. F. Volkman and Q. Cui, *J. Phys. Chem. B* **119**, 9547 (2015). Copyright 2015 American Chemical Society.

3.1 Introduction

One classical paradigm in molecular biology, also known as Anfinsen's dogma, is that the three-dimensional native structure of a protein is entirely determined by its amino acid sequence [1]. In recent years, a growing number of proteins are known to violate this rule: their structure undergoes significant transformations upon the change of solution condition and/or binding to other molecules. Notable examples are prion proteins [2] and intrinsically disordered proteins (IDPs) [3, 4], which become more structured upon binding to their partner protein or a ligand such as DNA. Another class of relevant systems are called "metamorphic" proteins [5, 6], which undergo major topological changes upon specific perturbations. Unlike prions, the conformational transitions in metamorphic proteins are reversible, and different folds can co-exist within the cell. In contrast to IDPs, metamorphic proteins interconvert between well-structured states. The discovery of these metamorphic proteins and the study of proteins whose fold is highly sensitive to small variation in sequence [7–10] pose interesting new questions regarding the relationship between protein sequence, structure and function. Indeed, understanding how multiple native folds are encoded into a protein sequence [11] and revealing the physical mechanisms that underlie the conversion among the different fold topologies has implications in a number of important areas including structural biology, protein evolution, protein design and human disease.

A remarkable example of metamorphic proteins is human lymphotactin (hLtn) [12] in which changes in temperature and salt concentration trigger a dimerization process and all native contacts within the subunits are changed. For instance, hLtn is predominantly a monomer (hLtn10) at 10 °C with 200 mM salt while the dimeric form (hLtn40), which features significant rearrangements in the backbone hydrogen bonding contacts (see below), is more stable at 40 °C and low salt. Under physiological condition (37 °C and 150 mM salt), the two folds coexist and play different physiological functions [13, 14]. hLtn10 binds to XCR1 and thereby induces intracellular calcium mobilization and chemotaxis. hLtn40 binds to glycoaminoglycans (GAGs) but not the XCR1 receptor; GAG binding is crucial for leukocyte recruitment *in vivo*. Therefore, understanding factors that dictate the relative stability of the two folds and the conversion between them will aide the design of strategies for development of potent conformation-specific allosteric inhibitors.

The NMR structures for hLtn10 (Fig.3.1A) and hLtn40 (Fig.3.1B) reveal that the interconversion is

not a simple dimerization process. The interconversion involves disruption of all hydrogen bonds in the three-stranded antiparallel β-sheet of hLtn10 [15, 16] and formation of a new four-stranded β-sheet accompanied by the loss of the C-terminal α-helix [12]. Between the two structures, all long-range interactions are replaced by a new hydrogen bonding network, tertiary and quaternary contacts. Many residues that are buried in the core in one fold become solvent exposed after the transition, and vice versa. The structural changes lead hLtn40 to form a new hydrophobic surface in order to accommodate a dimer interface. Therefore, it is likely that the interconversion requires the complete unfolding and refolding, rather than a well-structured intermediate; this mechanistic hypothesis is consistent with the kinetic analyses of hLtn10/hLtn40 interconversion and unfolding of the two species [17].

The hLtn system represents a remarkable example for how environmental properties, such as salt and temperature modulate protein stability through the interplay of electrostatics and hydrophobic interactions [18–20]. Prior to the solution of the NMR structure for hLtn40 [12], our molecular dynamics (MD) studies [21, 22] focused on probing changes in structural stability of hLtn10 at the scale of tens of nanoseconds when temperature and salt concentration was varied separately. Although these simulations were clearly shorter than the interconversion time scale ($\sim$1 second) [17], they revealed the structural features most sensitive to the change of solution condition. For example, ion association to protein surface is highly dependent on the local sequence and enhanced with the increase in temperature [21, 22]. This is in line with the findings from Elcock [20, 23] that electrostatic (salt bridge) interactions are robust to changes in temperature and hence are better suited for promoting protein stability at high temperatures.

A specific observation from the MD simulation of hLtn10 [21] was that, at high salt (200 mM) concentration, a chloride ion was stabilized at the protein surface by simultaneously binding to the side-chains of Arg 23 and Arg 43. In the absence of salt, it is conceivable that the electrostatic repulsion between the Arg residues contributes to the conversion of hLtn10 to hLtn40, in which the pair of Arg residues are far apart. This reasoning stimulated Volkman and co-workers to more systematically explore the contributions from electrostatic interactions among side-chains to the relative stability of hLtn10/hLtn40 [24]. By inspecting the NMR structural ensembles for hLtn10 and hLtn40, they identified pairs of charged residues whose distances have changed upon interconversion. The list includes both like-charge pairs that repel in one state (e.g., Arg 23 – Arg 43 in hLtn10) and opposite-charge salt-bridges that exist in only one structure (e.g., Lys 25 – Gu 31

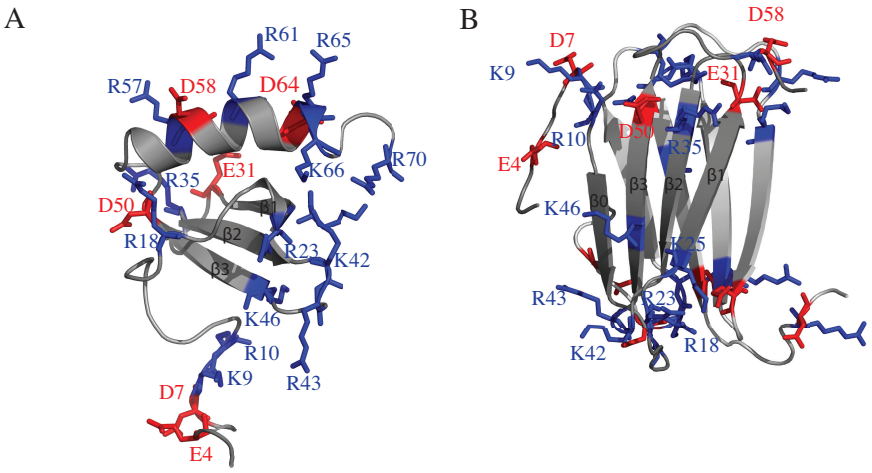


Figure 3.1: NMR structures for (A) hLtn10 (PDB code 1J9O [15]) and (B) hLtn40 (PDB code 2JP1 [12]). The charged side chains are shown with stick representation in red for negatively charged and blue for positively charged amino acids; residues are labeled only for one monomer in hLtn40 for clarity. The C-terminal tail consisting of residues 71 – 93 in Ltn10 is not included for clarity; residues after 60 in each monomer of hLtn40 are not resolved in the NMR structure.

and Arg 9 – Asp 50 in hLtn40). In some cases, the measured change in stability upon mutation was consistent with the electrostatic reasoning. For example, comparison of surface plasmon resonance analysis of heparin binding affinities [13] and HSQC spectra of WT, R23A and R43A mutants [14] revealed that both mutations enhance hLtn10 stability [24] and shift the conversion to hLtn40 at higher temperatures. The effect further increases for the R23A/R43A double mutant [14, 24]. However, mutation of many other charged residues in the list did not lead to any significant changes in stability; examples include Lys 25 and Lys 35, which reside in the same core region as Arg 23 and Arg 43 (see Fig.3.1).

One possible rationale for the lack of any significant contribution from some of the analyzed charged residues is that the conformations they adopt in solution are very different from those in the limited number of structures included in the NMR ensemble, thus they are, in fact, not involved in any major electrostatic repulsion as conjectured in Ref.24. This consideration calls for a better characterization of the structural ensembles of hLtn10 and hLtn40 under different solution conditions, and we do so with atomistic MD simulations using both explicit and implicit solvent models in this study. By comparing the interaction patterns of charged residues in the two folds under different conditions, we aim to better reveal electrostatic interactions that contribute to the relative stability of hLtn10/hLtn40. In particular, we hope to explain the trends observed in Ref.24 and identify additional interactions that can be experimentally validated.

Although the importance of salt concentration to the hLtn10/hLtn40 conversion highlights the role of electrostatics, the temperature dependence suggests that hydrophobic effects are also involved. According to the NMR structures [12, 15, 16], Val 15, Val 37, Phe 39, Thr 41 and Ala 49 are in the hydrophobic core of hLtn10 but solvent exposed in hLtn40. By contrast, Leu 34, Ala 36, Ile 38 and Ile 40 are buried in hLtn40 but reside on the surface of hLtn10. Moreover, mutation of Trp 55 to Asp, which is buried in the hydrophobic core of hLtn10, shifts the equilibrium towards hLtn40. [12, 15, 16] Therefore, another goal of our current study is to use the MD ensemble to better identify hydrophobic sites that likely dictate the relative stability of hLtn10/hLtn40. These predictions can be probed by future mutation experiments.

In the following work, we first summarize the computational models and simulation details. Next we analyze both electrostatic interaction networks and hydrophobic sites that undergo substantial changes upon the conversion from hLtn10 and hLtn40. The results are discussed in the context of available experimental data [24]. We also compare the effectiveness of explicit and implicit solvent simulations for characterizing

the solution ensembles. As a validation to the conformation ensemble sampled in the explicit solvent MD simulations, we carry out free energy perturbation calculations to explicitly probe the contribution of electrostatic repulsion between Arg 23 and Arg 43 to the stability of hLtn10 and compare to available experimental data [24]. Finally, we draw a few conclusions.

3.2 Computational Methods

3.2.1 Molecular Dynamics Simulations

All MD simulations (see Table 3.1) are carried out using the AMBER v12 program [25, 26] with the ff99SB force field [27] improved with NMR observables (ff99SBnmr1) [28–30]. Recent tests indicated that ff99SBnmr1 provides a reliable description of protein structures with both explicit and implicit solvent models [31]. All MD trajectories are analyzed mainly via AmberTools v12 [25, 26], and the Multiscale Modeling Tools for Structural Biology (MMTSB) package [32]. Most property calculations (e.g., distances between key charged residues, see Tables 3.2, 3.3) are averaged over the last half of each trajectory to ensure an adequate equilibration within the simulated time scales.

The starting structures are based on the NMR structures for hLtn10 (PDB: 1J9O [15] and 1J8I [15]) and hLtn40 (PDB: 2JP1 [12]). We simulate the two folds at two limiting experimental conditions: at 10 °C with 200 mM salt, and at 40 °C without salt. From this point forward, we refer to those conditions as s10 and n40.

In explicit solvent simulations, the TIP3P water model [33] is used along with the ion parameter set from Joung & Cheatham [34]. Production runs are obtained using constant temperature (10 °C or 40 °C) and pressure (1 bar) MD with the Anderson thermostat [35] and isotropic position scaling for pressure control. SHAKE [36] is applied to bonds involving hydrogen, allowing an integration time-step of 1 fs. The non-bonded cutoff is set to 12 Å. Particle-Mesh-Ewald is applied for electrostatic interactions using cubic spline interpolation, and 108 is set as the size of the charge grid. The simulation box is a truncated octahedron with an average box length of 98.38 Å.

In implicit solvent simulations, the gb7 Generalized Born (GB) approach [37] in AMBER 12 is adopted. Since the surface area term introduced in Amber is a rather crude approximation [38, 39], we chose to omit the surface area term (gbsa=0) as done in previous studies [40, 41]. We have selected gb7 as the GB model since

it includes a correction term to describe the solvent-excluded volume of each pair of atoms [37]. This correction results in a free energy barrier to the separation of non-bonded atoms similar to that of explicit solvents. Langevin dynamics is carried out with a collision frequency of 20 ps^{-1} at 300 K. The SHAKE algorithm [36] is applied to all bonds involving hydrogens, and an integration time step of 1 fs is used. Non-bonded interactions are computed without any cutoff, and the maximum distance between atom pairs (rgbmax) for Born radii calculations is set to 12 Å. Salt effects are described with the Debye-Hückel theory [42], and the concentration is set to 200 mM when necessary. The disordered tail of hLtn10 at the C-terminus after residue 60 is truncated to further enhance the computational speed.

The simulated time scales are summarized in Table 3.1. Explicit solvent simulations range from ∼50 to 150 ns, while the GB simulations are carried out to a much longer time scale (∼600-800 ns), which is made possible with GPU computing. [40]. As summarized in Table 3.1, in addition to the WT systems, we have also conducted simulations for several relevant mutants that implicate the pair of Arg residues (Arg 23, Arg 43) studied in Ref.24. As shown in Appendix A, pair-wise RMSD plots (Figs. A1 and A2) for the trajectories indicate that the conformational ensembles from different simulations under the same condition are similar to each other, supporting the convergence of key properties of interest here.

Analysis of trajectories are carried out by using the Amber package [25] and the Multiscale Modeling Tools for Structural Biology (MMTSB) toolbox [32]. The Cα-RMSD and Cα-RMSF are calculated using the core of the structures only (residues 25-70 for hLtn10 and residues 11-51 for hLtn40), excluding the C-terminal and N-terminal tails since they are not structured based on the lack of long range interactions in the NOE data [15]. Clustering of the trajectories is carried out using the *k*-clust command of the MMTSB package with Cα-RMSD as the similarity measure.

3.2.2 Free Energy Perturbation (FEP) and Linear Response Approximation (LRA)

As a validation to the explicit solvent conformational ensemble, we have carried out FEP calculations [43–45] to probe the impact of removing the partial charges on Arg 23 or Arg 43 on the stability of hLtn10. These calculations can be meaningfully compared to the experimental measurements on the relative stability of the WT, R23A and R43A proteins [24] since the effect of the R23A/R43A mutations on hLtn10 stability is expected to be dominated by electrostatic interactions. Indeed, the computed solvent accessible surface areas for Arg23 and Arg43 in hLtn10 are comparable to those found in reference peptides (see below) used to model

Table 3.1: Summary of MD simulations.

System	Sequence[a]	Condition	Solvent	Length (ns)
hLtn10 (s10)	WT	10 °C, 200 mM NaCl	TIP3P	90+70
hLtn10 (n40)	WT	40 °C, 0 mM NaCl	TIP3P	80
hLtn10$_{R\emptyset}$ (s10)	R23$\emptyset$, R43$\emptyset$	10 °C, 200 mM NaCl	TIP3P	46
hLtn10$_{RK}$ (s10)	R23K, R43K	10 °C, 200 mM NaCl	TIP3P	50
hLtn10 (s10)	WT	10 °C, 200 mM NaCl	gb7	800
hLtn10 (n40)	WT	40 °C, 0 mM NaCl	gb7	800
hLtn10 (s10)	FEP (R23$\rightarrow\emptyset$)	10 °C, 200 mM NaCl	TIP3P	11$\times$10
hLtn10 (s10)	LRA (R23$\rightarrow\emptyset$)	10 °C, 200 mM NaCl	TIP3P	2$\times$50
hLtn10 (s10)	FEP (R43$\rightarrow\emptyset$)	10 °C, 200 mM NaCl	TIP3P	11$\times$10
hLtn10 (s10)	LRA (R43$\rightarrow\emptyset$)	10 °C, 200 mM NaCl	TIP3P	2$\times$50
hLtn40 (s10)	WT	10 °C, 200 mM NaCl	TIP3P	150+50
hLtn40 (n40)	WT	40 °C, 0 mM NaCl	TIP3P	100
hLtn40$_{R\emptyset}$ (s10)	R23$\emptyset$, R43$\emptyset$	10 °C, 200 mM NaCl	TIP3P	50
hLtn40$_{RK}$ (s10)	R23K, R43K	10 °C, 200 mM NaCl	TIP3P	43
hLtn40 (s10)	WT	10 °C, 200 mM NaCl	gb7	800
hLtn40 (n40)	WT	40 °C, 0 mM NaCl	gb7	600

[a] R23$\emptyset$, R43$\emptyset$ indicates that the partial charges associated with Arg 23 and Arg 43 are removed. FEP indicates free energy perturbation with equally-spaced λ windows, LRA indicates linear response approximation with simulations conducted only for the end states ($\lambda = 0, 1$).

the unfolded state of hLtn10, supporting the focus on electrostatic component in the free energy calculations.

The FEP calculations follow the standard thermodynamic cycle [46, 47] to remove the partial charges on Arg 23 (or Arg 43) in two independent sets of simulations: one set involves the WT hLtn10 (excluding the C-terminal loop), while the other set involves a short peptide that mimics the unfolded state of hLtn10. For the latter, we test both a tri-peptide and a nona-peptide with four residues on each side of the Arg of interest: LPVSRIKTY and FITKRGLKY around Arg 23 and Arg 43 respectively. The FEP calculations involve eleven windows with the coupling parameter λ equally spaced between 0 and 1 with a 0.1 interval. Each window is simulated with explicit solvent MD for 10 ns, and the FEP data analysis is carried out with the Bennet Acceptance Ratio (BAR) approach [48].

The advantage of focusing on the electrostatic component here is that the free energy derivatives are expected to largely follow linear response [49–51]. Thus, we also test the LRA calculations by conducting longer ($\sim$40 ns) simulations for only the end states ($\lambda = 0, 1$). The results of the FEP and LRA calculations are summarized and compared in Table 3.4.

3.3 Results and Discussions

In the following, we first investigate the structural stability and side-chain interactions in the hLtn10 and hLtn40 folds under the two conditions that favor each fold: at 10 °C with 200 mM salt (s10), and 40 °C without salt (n40), respectively. Since the stability of each fold is dependent on the salt concentration, it is hypothesized that electrostatic interactions involving charged side-chains play important roles while the temperature dependence suggests that hydrophobic effects are also likely relevant. Thus, our analysis focuses on charged side-chains and hydrophobic residues, especially those experience different environments in the two folds, since they are expected to dictate the relative stability of the two folds. We then discuss free energy perturbation calculations that explicitly probe the contributions from Arg 23 and Arg 43 to the stability of hLtn10. Comparison of the result to experimental data serves to validate the conformational ensemble sampled in the MD simulations; moreover, the comparison helps establish the appropriate computational protocol (FEP vs. LRA) that can be potentially used as a screening tool to identify residues important to stability. Finally, we discuss implicit solvent simulations of hLtn10 and hLtn40. A comparative analysis is presented to discuss the benefits and limitations of the implicit solvent method for systems whose stability is sensitive to solution

conditions.

3.3.1 Overall Structural Stability of the WT Protein Under Different Solution Conditions

To characterize the overall structural stability, we monitor the time dependence of secondary structural elements (Fig.3.2), the time dependence of Cα-RMSD (root mean square difference relative to the starting NMR structure, Fig.3.3) and residual Cα-RMSF (root mean square fluctuations, Fig.3.4) along the explicit solvent MD trajectories.

As shown in Fig.3.2A-B, consistent with experimental observations [12, 16], the hLtn10 fold is more stable under the s10 condition; most of the secondary structures remain intact during the ∼80 ns simulation, with the exception that the N-terminal residues between Pro 20 and Arg 23 undergoes flickering fluctuations between 3_{10} helix and α-helix at the nanosecond time scale. Under the n40 condition (Fig.3.2B), the most visible differences occur in the N- and C-termini, where the loop region between Arg 9 and Leu 14 starts forming a short α-helix after 30 ns and also around Thr 79 and Thr 85 as opposed to Glu 83 to Thr 91 under the s10 condition. The Cα-RMSD (Fig.3.3A-B) and Cα-RMSF (Fig.3.4A) plots display comparable trends for s10 and n40, however RMSD of the representative member of the most populated clusters display slightly higher RMSD for n40 condition than s10 (1.3 Å vs 1.6 Å, see Fig.A3). Unexpectedly, the RMSF of hLtn10 is substantially increased in the R23K/R43K double mutant (see Fig.3.4) compared to the WT protein in particular around 30's loop which contains Glu 31 and Arg 35. When conditions and charged groups (Arg 23 and Arg 43) are altered the fluctuations of 30's loop change considerably, as can be identified on representative structures from clustering as well (see Fig.A1). Although there are no persistent side-chain contacts identified between the 30's loop and the charge rich 20's and 40's regions (see below), they are apparently coupled.

For the hLtn40 fold, the structure remains very stable in both conditions studied here; there is only minor difference in the N-terminus for one of the monomers, which forms an α-helix under the n40 condition (compare Fig.3.2C-D). The same trend is observed in the RMSD (Fig.3.3E-F) and RMSF (Fig.3.4B) plots. Clustering analysis leads to a single cluster for the hLtn40 fold under both conditions using a threshold of 2.0 Å (the value used for hLtn10). Thus, the clustering results shown in Fig.3.3E-F are obtained with a lowered threshold of 1.5 Å. The fact that hLtn40 shows a smaller number of clusters compared to hLtn10 at the same threshold is qualitatively consistent with the expectation that the dimeric structure is better stabilized than the

monomer by the fairly extensive dimer interface and it features less disordered regions [52].

3.3.2 Electrostatic Effects: Salt Ion Distribution and Network of Charged Residues

3.3.2.1 Salt ion distribution around the protein

To investigate electrostatic interactions in the two structures, we monitor the ion distribution around the protein under the s10 condition (Fig.3.5) and interaction network of charged residues under both s10 and n40 conditions (Tables 3.2 and 3.3).

Regarding the ion distribution, we do not observe any major Na^+ association (Fig.3.5C,D) for either fold, except for a weak interaction with the solvent exposed Glu 31; this association is expected due to the overall positive-charge of hLtn (+9 for hLtn10) and qualitatively consistent with our previous MD simulations [21, 22] (although we discussed an interesting enhancement of Na^+ association with the C-terminal helix at the higher temperature of 45°C). The Cl^- ions approach the protein surface frequently due to the significant number of accessible positively charged residues. As shown in the top panels of Fig.3.5, Cl^- ions have the closest encounters with Arg 23 and Arg 43 in the hLtn10 fold, in agreement with our previous observations from MD simulations [21]. The existence of such a Cl^- "binding site" is supported by the chemical shift perturbations observed in recent NMR experiments [24].

It is expected that Cl^- binding in this region helps relieve the otherwise repulsive interaction between the pair of Arg residues and other charged groups in the region, thus providing stabilization of the hLtn10 fold under the s10 condition. Removing salt therefore destabilizes hLtn10 and facilitates the conversion toward hLtn40, in which the pair of Arg residues are farther away. This idea is consistent with the experimental observations that both R23A and R43A mutations stabilized the hLtn10 fold and that the Cl^- binding site was disrupted with the R23A and R43A mutation [24]. Qualitatively consistent with these experimental observations, once we remove the partial charges of Arg 23 and Arg 43, the hLtn10 fold has substantially lower fluctuations (Fig. 3.4A) and the Cl^- ions move farther away from the protein (Fig.3.5A). Interestingly, the RMSF for the hLtn40 fold is also reduced when the partial charges of Arg 23 and Arg 43 are removed (Fig. 3.4B), although the effect appears less striking compared to hLtn10. To probe whether the multivalent character of the Arg residues is essential for the formation of the Cl^- binding site, we also carry out simulations in which both Arg 23 and Arg 43 are mutated into Lys residues. As shown in Fig.3.5, this has a small impact on

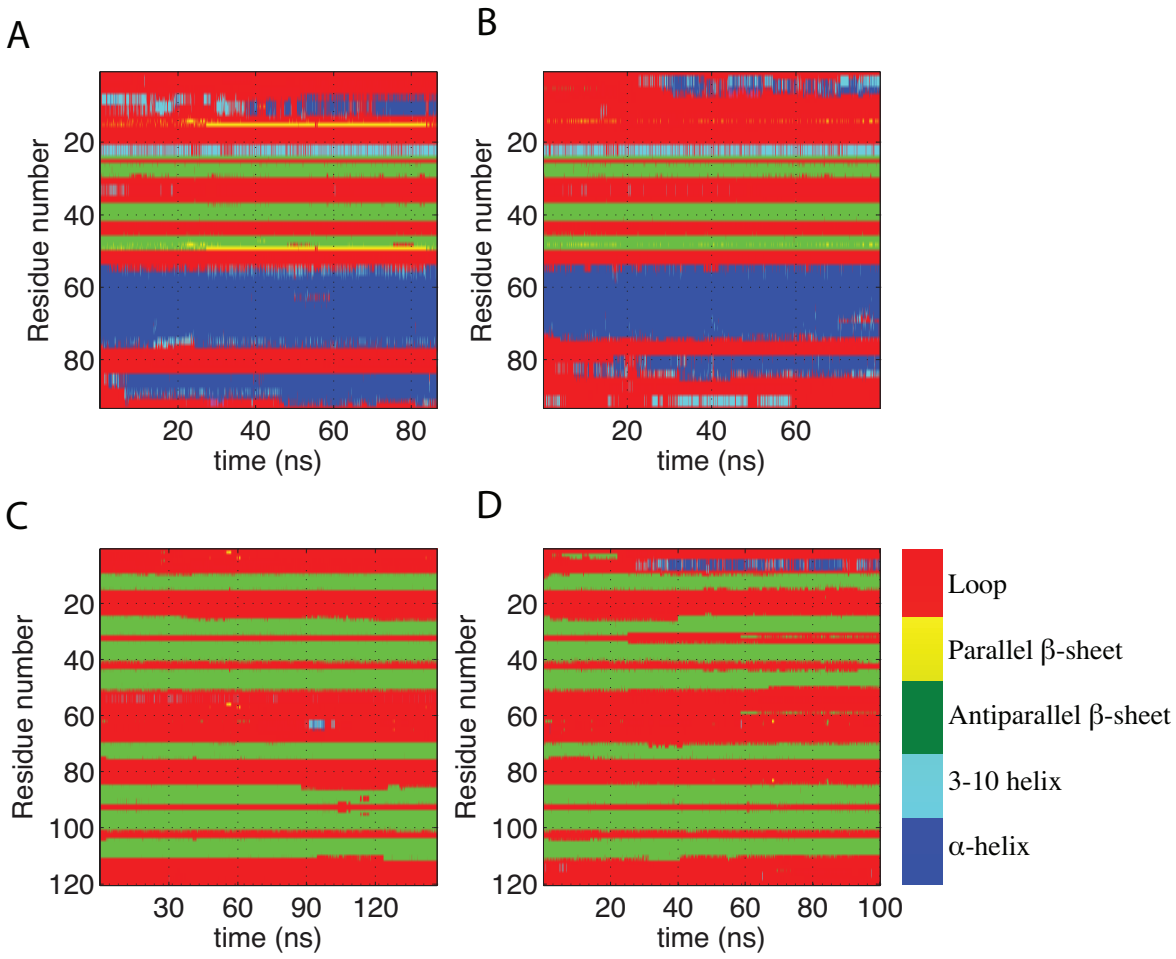


Figure 3.2: Secondary structure of protein motifs during explicit solvent simulations of WT hLtn under different solution conditions. (A) hLtn10, s10; (B) hLtn10, n40; (C) hLtn40, s10; (D) hLtn40, n40. Different colors represent different secondary structure elements, as shown on the right.

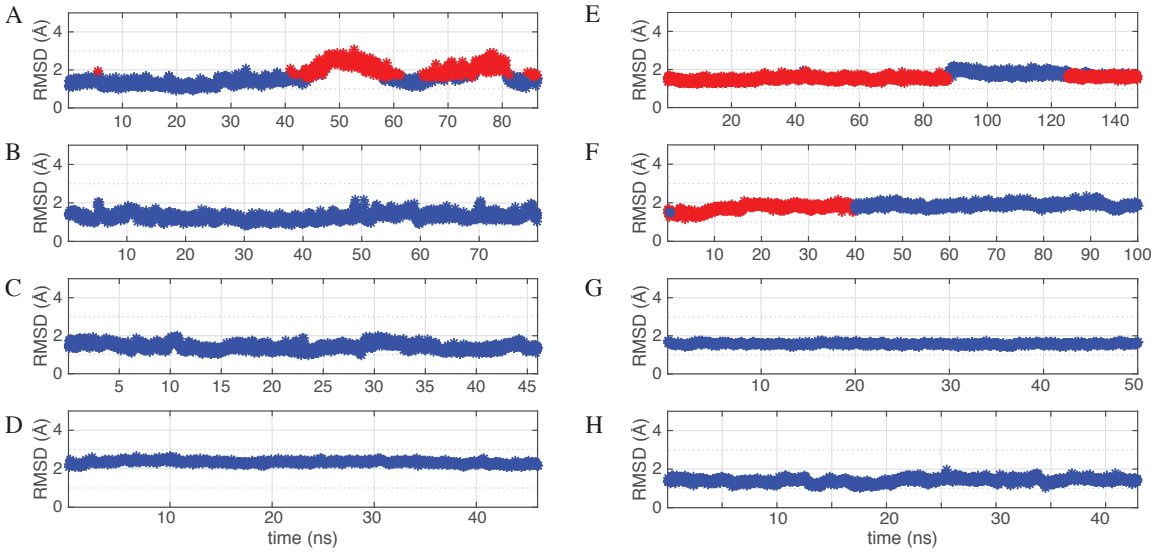


Figure 3.3: Cα Root Mean Square Deviation (RMSD, Å) vs. time during MD simulations of hLtn under different solution conditions. RMSD is calculated with respect to the initial structures (the NMR model that is closest to the average NMR conformations from 1J9O, 1J8I and 2JP1) using Cα-alignment of the core regions only (residues 25-70 for hLtn10 and residues 11-51 for hLtn40), excluding the C-terminal and N-terminal tails. Color coding is used to display different clusters obtained from clustering at 2.0 Å and 1.5 Å (using *k*-clust in MMTSB) as the threshold for hLtn10 and hLtn40, respectively. Blue displays the most populated cluster, followed by red. (A) WT hLtn10, s10; (B) WT hLtn10, n40; (C) hLtn10 RØ, s10; (D) hLtn10 RK, s10; (E) hLtn40, s10; (F) hLtn40, n40; (G) hLtn40 RØ, s10; (H) hLtn40 RK, s10. See Figs. S3-S4 for representative structures of each cluster and Figs. A1 and A2 for pairwise-RMSD values.

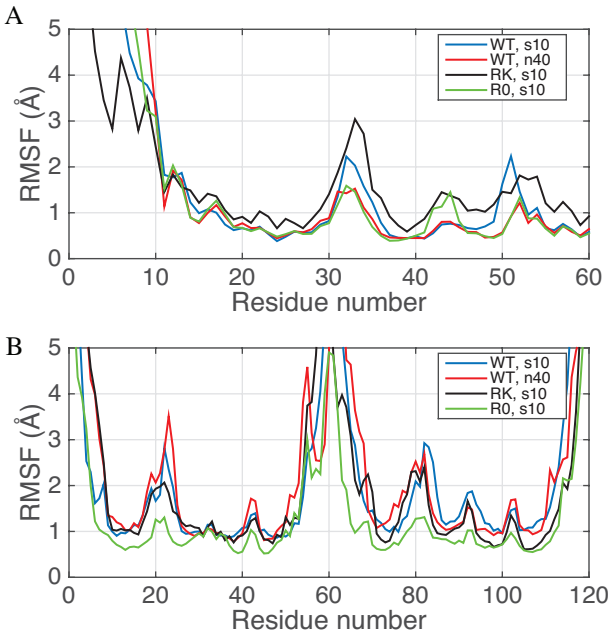


Figure 3.4: Root Mean Square Fluctuations (RMSF, Å) for explicit solvent simulations for (A) hLtn10 and (B) hLtn40. Color coding represents different models simulated: Blue: WT, s10; Red: WT, n40; Green: RØ, s10; Black: RK, s10. The RMSF is calculated based on Cα-alignment of the core regions only (residues 25-70 for hLtn10 and residues 11-51 for hLtn40), excluding the C-terminal and N-terminal tails.

the Cl^- binding, thus the existence of the Cl^- "binding site" is likely associated with the overall electrostatic properties in this region, in agreement with previous Poisson-Boltzmann calculations [21].

3.3.2.2 Network of charged residues

To more systematically identify charged residues that likely contribute to the relative stability of hLtn10/hLtn40, we have filtered the charged residue pairs that are in close vicinity of each other (within 11 Å) during MD simulations of each fold under different solution conditionsTable 3.2 and 3.3. Snapshots that illustrate key charged residues and the interaction network are shown in Fig.3.6. The distances are measured based on center of mass of nitrogen atoms (NH1, NH2) of arginine, nitrogen atom (NZ) of lysine and center of mass of oxygen atoms from glutamate (OE1, OE2) and aspartate (OD1, OD2). In addition to the WT protein, we have also examined a pair of mutants motivated by the analysis of salt ion distribution as discussed above (R23K/R43K and a mutant in which the partial charges of Arg 23/Arg 43 are removed, referred to as R∅). Since salt-bridge dynamics on protein surface can be rather stochastic in nature, it is advantageous to compare multiple simulations to seek consistent trends [53].

For hLtn10, both the WT and mutants analyzed here, only two pairs of salt bridges are highly persistent: Asp 7 – Lys 46 and Arg 18 – Asp 58, which are observed in the WT hLtn10 under the s10 condition. Recent MD simulations [52] identified salt bridge interactions Arg 18 – Asp 58 and Glu 31 – Arg 57 based on simulations at 300 K without salt; the latter was not observed in our s10 and n40 simulations. Interestingly, even the salt bridge interactions, which we have identified, break apart under the n40 condition and the two mutants cases in which these residues are not perturbed directly. Therefore, we do not anticipate that these salt-bridges make an important contribution to the stability of hLtn10 and the observation for their existence in the s10 simulation is likely fortuitous [53].

Other than the brief discussion of Arg 23/Arg 43 in Ref.24, previous MD studies have not elaborated on the contributions from like-charge electrostatic interactions to the stabilities of hLtn10 and hLtn40 [21, 22, 52]. Motivated by the discussion of Volkman et al. [24], in this work we also examine other like-charge pairs observed in the MD simulations under conditions s10 and n40 as well as for two mutant cases involving Arg 23 and Arg 43. The Arg 23 – Arg 43 pair remains fairly close, reaching less than 8 Å in the s10 simulation due partly to the interaction with Cl^- ions. A few other pairs of cationic residues are also in proximity to each other (see Table 3.2), including Lys 8 – Lys 46, Lys 25 – Lys 42, Lys 42 – Lys 66, and Lys 42 – Arg 70, Arg

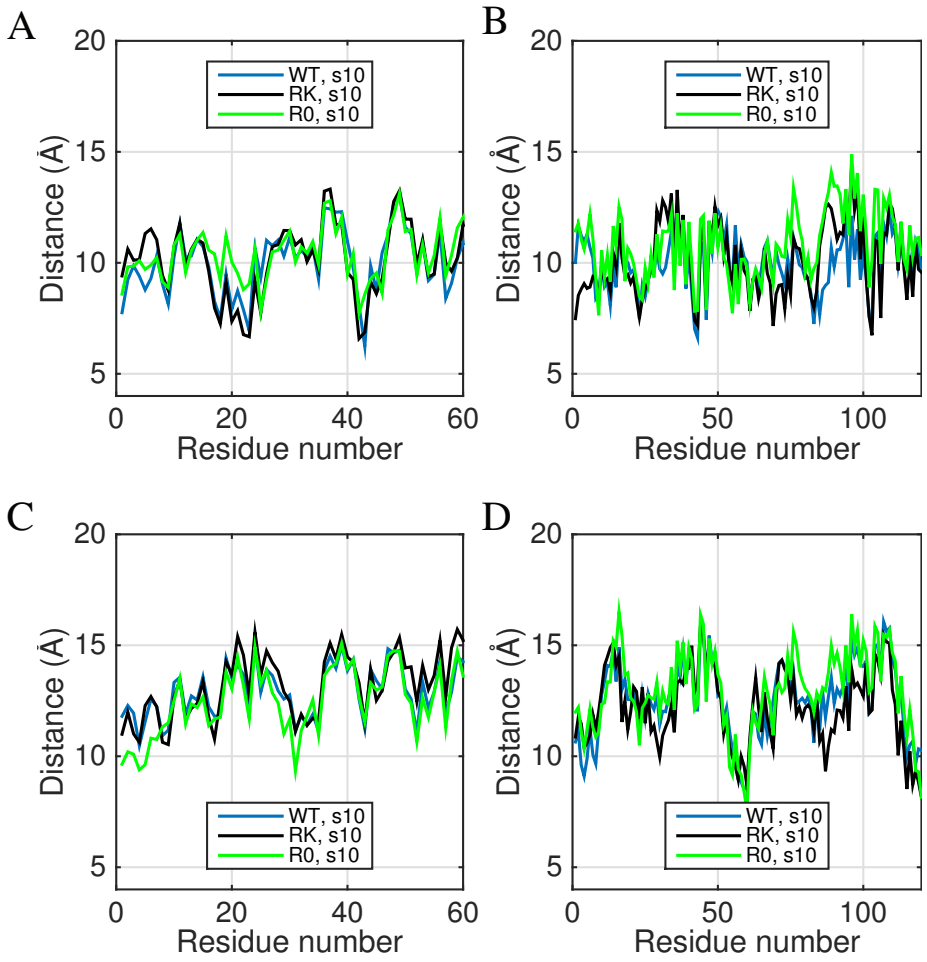


Figure 3.5: Minimal distance between the center of mass of the charged amino acid side-chains and salt ions during the MD simulations for (A, C) hLtn10 and (B, D) hLtn40. Distances to Cl^- are plotted in the top two panels and the distances to Na^+ are shown in the bottom two figures. Color coding represents different models simulated: Blue: WT, s10; Red: WT, n40; Green: RØ, s10; Black: RK, s10.

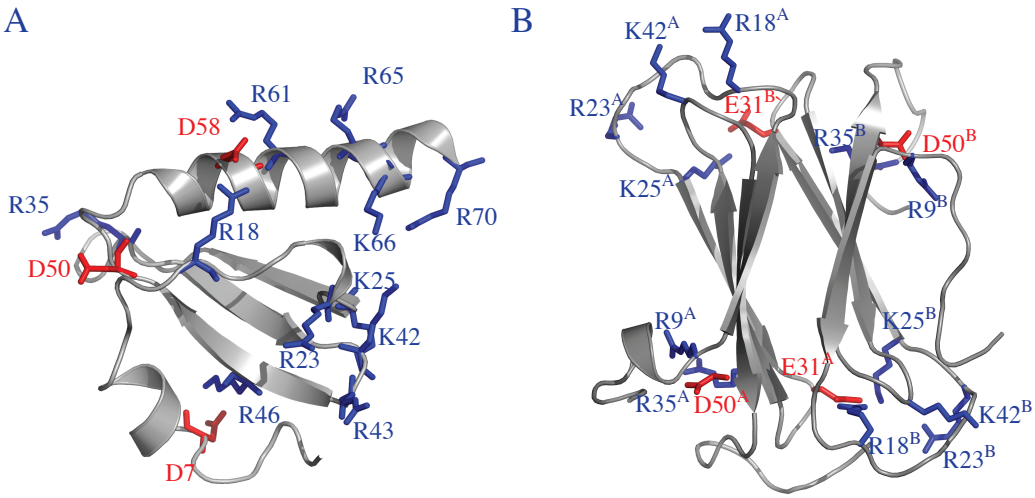


Figure 3.6: Identified charged interactions from explicit solvent MD simulations shown on three dimensional structures for (A) hLtn10 and (B) hLtn40. For distances, see Tables 3.2 and 3.3.

61 – Arg 65 and Lys 66 – Arg 70; Lys 8 – Arg 35 are also observed to be fairly close to each other in the n40 simulation. By comparing to the mutant simulations, however, it seems that only Lys 25 – Lys 42, Lys 42 – Lys 66, Lys 66 – Arg 70 represent consistently close "repulsive pairs" in addition to Arg 23 – Arg 43, although the lower limit of the Lys 25 – Lys42 and Lys 42 – Lys 66 distance is larger than that for Arg 23 – Arg 43. These observations are qualitatively consistent with the experimental observation that only R23A and R43A mutants exhibited substantially higher, $\sim$1 kcal/mol, stability than the WT hLtn10. Mutations R9A, K25A and K66A, by contrast, were found to have minimal impact on the stability. As we suspected (see Introduction), although Arg 9 and Lys 66 are observed to form "repulsive pairs" with Lys 46 and Lys 42, respectively, in the NMR ensemble (the Lys 42 – Lys 66 separation is as low as 5.4$\pm$1.4 Å, see Table 3.2), these repulsive configurations are not sampled with high population in the MD solution ensemble. Moreover, Lys 42 resides in close vicinity to Lys 25, Lys 66 and Arg 70; thus removing only Lys 66 is likely not sufficient to alleviate the repulsive interactions. By contrast, Lys 42 is worth further studies since it is engaged in multiple repulsive interactions with Lys 25, Lys 66 and Arg 70.

With hLtn40 (Table 3.3), we observe a larger number of salt-bridge interactions, consistent with recent MD simulations [52]. However, we only focus on the pairs that are consistently observed in multiple sets of independent simulations under different conditions. The only pair that emerges (with N-O distance $\sim$4 - 5 Å) under all explicit solvent simulations, including the s10 set, is an intramolecular protein salt bridge between Arg 35 and Asp 50. It is interesting that the same pairwise interaction is observed in hLtn10 (Table 3.2), but features longer distances (typically $>$ 6 Å), presumably due to the different local structural features of the two hLtn folds. Therefore, it appears that mutations involving this salt bridge might be tested for altering the relative stability of hLtn10 and hLtn40. We predict that removing this salt bridge has a larger impact on hLtn40 than on hLtn10, thus shifting the population towards the latter state.

Among the number of inter-subunit salt-bridge interactions, Glu 31 – Lys 25* (* is used to indicate residue from a different subunit in the dimer) is the only one that appears to be persistent in n40 (the condition that stabilizes hLtn40) and also most s10 simulations. However, it undergoes very large fluctuations in WT s10 condition. Therefore, we anticipate that removing this inter-protein salt-bridge, especially under conditions similar to n40, shifts the population toward hLtn10. This shift is precisely what was observed experimentally. Although the R25A mutation had little impact on the stability of hLtn10. The mutation shifts

Table 3.2: Distances (Å) for charged residue pairs in hLtn10 that remain spatially close in the MD ensembles.

	WT s10[c] TIP3P	WT n40 TIP3P	RØ [d] s10 TIP3P	RK [d] s10 TIP3P	WT s10 gb7	WT n40 gb7
Salt-Bridges[b]						
Glu 4 – Arg 35 (25.6±3.8)	27.0±2.6/15.6±6.0	10.4±6.5	15.2±2.5	15.4±3.7	6.6±3.0	20.2±2.8
Asp 7 – Lys 46 (13.8±3.9)	3.1±0.3/12.2±2.3	17.7±2.9	12.5±3.7	10.9±2.1	18.1±1.6	19.8±2.1
Arg 18 – Asp 58 (8.0±0.6)	4.8±1.8/8.4±1.6	9.4±2.2	8.2±1.3	13.5±20.1	6.4±3.2	8.3±2.9
Arg 35 – Asp 50 (10.4±1.1)	12.5±3.1/6.6±2.7	8.0±2.8	9.0±1.4	6.3±2.8	12.0±3.4	10.8±3.3
Repulsive pairs[b]						
Lys 8 – Lys 46 (11.9±2.9)	9.1±1.4/14.4±3.5	20.2±3.0	9.8±2.8	27.4±29.2	19.7±1.9	20.1±3.1
Lys 8 – Arg 35 (22.1±1.7)	21.1±2.6/19.7±3.0	11.3±5.0	15.2±1.3	28.3±27.7	21.1±3.3	14.8±3.5
Arg 23 – Arg 43 (9.7±1.9)	9.6±2.1/9.7±1.9	10.3±2.2	5.0±1.0	8.9±1.6	8.8±2.2	9.5±2.5
Lys 25 – Lys 42 (6.0±1.4)	10.0±1.7/9.3±1.2	9.4±1.2	9.2±1.2	9.5±1.2	8.9±1.6	8.8±1.6
Lys 42 – Lys 66 (5.4±1.4)	9.9±2.0/7.9±2.6	7.7±2.8	6.9±2.5	6.7±2.4	—[e]	—
Lys 42 – Arg 70 (9.2±2.4)	10.9±2.9/9.2±2.0	9.7±3.4	12.4±2.4	9.4±2.6	—	—
Arg 61 – Arg 65 (8.2±1.6)	10.7±1.9/6.3±1.5	8.8±2.7	9.7±1.9	10.3±1.5	—	—
Lys 66 – Arg 70 (9.4±2.5)	7.0±1.0/6.5±1.4	8.5±3.6	9.2±2.0	7.3±1.0	—	—

[a] Distances are measured based on center of mass of nitrogen atoms (NH1, NH2) of Arginine, nitrogen atom (NZ) of Lysine and center of mass of oxygen atoms from Glutamate (OE1, OE2) and Aspartate (OD1, OD2). [b] Values in parentheses are values calculated based on the NMR ensembles and standard deviations (PDB: 1J9O [15] and 1J8I [15]). [c] Values before and after the slash are from two independent simulations. [d] RØ indicates that the partial charges associated with Arg 23 and Arg 43 are removed. RK indicates the R23K, R43K mutant. [e] Not present because the C-terminal tail beyond residue 60 is not included in the implicit solvent simulations.

the hLtn40/hLtn10 equilibrium towards the latter [24].

Concerning like-charge pairs, the only pair that shows consistent proximity among independent explicit solvent simulations is the Arg 18 – Lys 42 pair; Asp 58 – Glu 31 also appears to be close in n40 simulations but moves apart in WT s10 simulation. Since Arg 18 also appears to be close to Glu 31 in multiple simulations of hLtn40, removing charge at this site (Arg 18) probably has compensating effects and thus leads to minimal perturbation to the hLtn40 stability; similar arguments can be made for Glu 31 and Asp 58. Lys 42, by contrast, does not appear to be implicated in other major charged interactions. Therefore, K42A is an interesting mutation that may stabilize hLtn40, although, as discussed above, K42A is likely to relieve several repulsive electrostatic interactions in hLtn10. In other words, K42A is not expected to have a large impact on the relative stability of hLtn10 and hLtn40.

3.3.2.3 Electrostatic contributions from Arg 23 and Arg 43 to hLtn10 stability

The analysis of charged residues discussed above is qualitative in nature. To probe the importance of electrostatic interactions at a quantitative level, we carry out FEP and LRA calculations. We focus here on the electrostatic contributions from the Arg 23/Arg 43 pair because they have been probed experimentally [24] and thus a comparison will help validate the conformational ensemble sampled in the explicit solvent simulations. The van der Waals contributions are not expected to be important with the sampled distance separation ($\sim$7-11 Å) between these Arg residues.

The FEP calculation results using the tri-peptide as the model for the unfolded state (Table 3.4) lead to estimates for the mutation effects on the hLtn10 stability ($\Delta\Delta G_{R23A}$, $\Delta\Delta G_{R43A}$). The results are on the order of +1.0 to +0.4 kcal/mol, which are in agreement with the experimental data. The statistical error for the computed free energy difference using BAR is on the order of 0.13 kcal/mol, suggesting that the FEP results for the mutation effects are meaningful. Using the nona-peptides as the model for the unfolded state leads to very similar results for the R43A mutation, but not for the R23A mutation (-0.3 kcal/mol). The difference in tri-peptide and the nona-peptide calculations highlight the challenging nature of the calculation since the effect of interest is relatively small in magnitude ($\sim$ 1 kcal/mol).

The LRA results also show the correct trends based on the average free energy values (e.g., +1.9 and +1.3 kcal/mol with the tri-peptide reference), independent of the model used for the unfolded state. The uncertainty in computed free energy values, however, is on the same order of magnitude. Therefore, it is

Table 3.3: Distances (Å) for charged residue pairs in hLtn40 that remain spatially close in the MD ensembles.

	WT s10[c] TIP3P	WT n40 TIP3P	RØ [d] s10 TIP3P	RK [d] s10 TIP3P	WT s10 gb7	WT n40 gb7
Salt-Bridges[b]						
Arg 9 – Asp 50 (5.0±1.4)	7.2±0.4/11.4±1.3	4.3±0.7	6.0±2.2	6.9±2.4	16.6±3.9	12.2±3.4
Glu 31 – Arg 18* (10.3±3.5)	12.9±19.4/6.5±1.6	2.8±0.1	4.9±0.8	6.6±2.2	15.1±2.3	14.6±5.1
Glu 31 – Arg 23* (11.1±1.6)	17.1±17.5/4.6±2.1	4.5±1.3	7.5±1.4	10.1±2.3	6.3±2.3	12.3±3.0
Glu 31 – Lys 25* (3.4±0.9)	11.9±19.5/5.6±1.5	7.2±1.6	3.4±1.0	3.1±0.6	12.5±2.2	14.5±2.7
Asp 58 – Arg 18* (14.5±4.8)	24.4±14.4/15.6±1.7	7.2±1.1	14.2±3.4	10.0±2.7	24.1±5.6	22.8±3.1
Asp 58 – Arg 23* (15.8±3.2)	32.9±12.3/13.1±2.5	4.7±3.2	16.1±5.0	8.0±4.2	19.7±3.7	19.1±3.3
Arg 35 – Asp 50 (6.3±2.3)	4.7±0.6/7.4±1.3	5.5±1.6	4.5±0.4	5.1±1.5	4.7±0.4	4.9±2.7
Arg 35 – Asp 58 (11.9±3.0)	8.2±1.2/21.5±3.2	16.1±1.3	17.4±1.7	19.1±1.4	10.0±2.6	12.5±1.6
Repulsive pairs[b]						
Arg 9 – Arg 35 (9.9±2.9)	10.0±0.5/17.8±2.0	7.1±1.6	7.3±2.1	7.7±2.3	20.3±3.7	15.6±3.7
Arg 18 – Lys 42 (10.7±2.0)	9.6±1.8/9.9±0.7	7.4±2.0	6.6±1.3	9.9±1.5	19.0±2.6	9.2±2.3
Glu 31– Asp 58 (8.3±3.0)	19.8±2.2/15.1±2.4	6.5±1.2	15.6±3.2	5.8±1.5	17.1±4.3	9.8±0.6
Arg 35 – Arg 57 (11.1±3.4)	8.8±1.5/12.1±3.9	17.0±2.1	13.7±2.1	19.2±2.0	8.0±3.4	13.0±2.0

[a] Distances are measured based on center of mass of nitrogen atoms (NH1, NH2) of Arginine, nitrogen atom (NZ) of Lysine and center of mass of oxygen atoms from Glutamate (OE1, OE2) and Aspartate (OD1, OD2). [b] Values in parentheses are values calculated based on the NMR ensembles (PDBid 2JP1 [12]). "*" indicates residue from the other monomer in the dimer. [c] Values before and after the slash are from two independent simulations. [d] RØ indicates that the partial charges associated with Arg 23 and Arg 43 are removed. RK indicates the R23K, R43K mutant.

difficult to draw firm conclusions based on LRA calculations with $\sim$ 50 ns of sampling. This is expected considering that the Arg residues are involved in a complex network of compensating electrostatic interactions and they are close to structurally mobile regions (30's and 40's loops). Moving forward, it seems that FEP combined with BAR is the more robust protocol for screening for sites that make important contributions to hLtn10/hLtn40 stability. According to the analysis of charged residue network discussed above (see Table 3.5), sites of interest are Arg 35, Lys 42 and Asp 50 in the context of hLtn40 stability. However, FEP calculations for hLtn40 with explicit solvent are computationally demanding, thus we leave those to a future study.

3.3.3 Important Hydrophobic Residues

Apart from charged interactions, hydrophobic residues that switch environment (e.g., the level of solvent accessibility) are expected to be important to the interconversion of the two folds. Identifying those hydrophobic residues will lead to new strategies for modulating the relative stability of hLtn10/hLtn40. With this aim in mind, we analyze the solvent accessible surface area (SASA) values for the two folds in different simulations. Fig. 3.7 summarizes the difference in SASA (ΔSASA) between hLtn10 in s10 and hLtn40 in n40. The results reveal that Val 12, Ile 29, Leu 34 and Leu 45 are more solvent exposed whereas Leu 19, Ile 24, Phe 39 and Trp 55 tend to stay more buried in hLtn10. Thus mutating the first batch to polar/charged residues favors hLtn10 over hLtn40, while the opposite is expected to be true for mutating the second batch to polar/charged residues.

When calculated over the NMR ensemble, ΔSASA shows similar trends for Ile 29, Leu 34 and Trp 55. However, additional sites are observed to change SASA considerably between the two native states: Val 12, Leu 19, Ile 24, Phe 39 and Leu 45. This finding is in line with the fact that the position of side-chains may adopt different conformations in solution than from those included in the limited number of conformers in the NMR ensemble. So far, the only hydrophobic site that has been experimentally tested is Trp 55. The mutation of Trp 55 to Asp shifts the population towards hLtn40 [12, 15, 16]. It would be of interest to explore additional sites identified through the MD ensemble here (summarized in Table 3.5).

3.3.4 Comparison Between Explicit and Implicit Solvent Simulations

Considering the sensitivity of hLtn structure to solution conditions, the system is particularly interesting from the perspective of testing the applicability of implicit solvent simulations.

Table 3.4: Computed mutation effect (in kcal/mol) on the stability of hLtn10[a]

	tri-peptide unfolded state	nano-peptide unfolded state	hLtn10 folded state
ΔG_{R23A}	177.3±0.1 (179.7±1.6)	176.0±0.1 (180.0±2.2)	176.3±0.1 (177.8±2.3)
$\Delta\Delta G_{R23A}$[b]	1.0±0.1 (1.9±2.8)	-0.3±0.1 (2.2±3.2)	–
ΔG_{R43A}	177.2±0.1 (179.6±0.9)	177.3±0.1 (179.0±1.1)	176.8±0.1 (178.3±2.9)
$\Delta\Delta G_{R43A}$[b]	0.4±0.1 (1.3±3.0)	0.5±0.9 (0.7±3.1)	–

[a] Values without parentheses are based on free energy perturbation (FEP) and data processing using BAR. Values with parentheses are based on linear response approximation (LRA) calculations. Only electrostatic components have been included in both FEP and LRA calculations. [b] The corresponding experimental values from Ref.24 are: 0.8±0.1 kcal/mol for R23 and 0.7±0.1 kcal/mol for R43.

Table 3.5: Predicted residues that perturb the stability of hLtn10 and hLtn40 based on the current MD simulations

Residues	Suggested mutation	Predicted Effect
Arg 35 – Asp50	Remove salt-bridge	Decreased hLtn40 stability
Lys 42	to negatively charged residue/Ala	Increased hLtn10 stability
Lys 42	Ala	Increases hLtn40 stability
Trp 55, Ile 24 Phe 39, Leu 19	to charged/polar residues	Favors hLtn40 over hLtn10
Ile 29, Leu 45, Leu 34, Val 12	to charged/polar residues	Favors hLtn10 over hLtn40

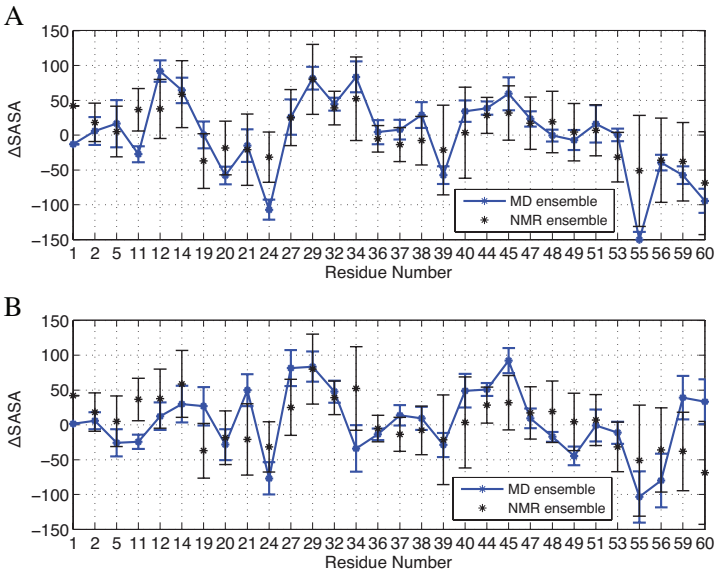


Figure 3.7: The difference of Solvent Accessible Surface Area (ΔSASA, in $Å^2$) between hLtn10 under the s10 and hLtn40 under the n40 condition from (A) explicit and (B) implicit solvent simulations (blue) are plotted against ΔSASA calculated over the NMR ensembles. Only results for hydrophobic residues are shown.

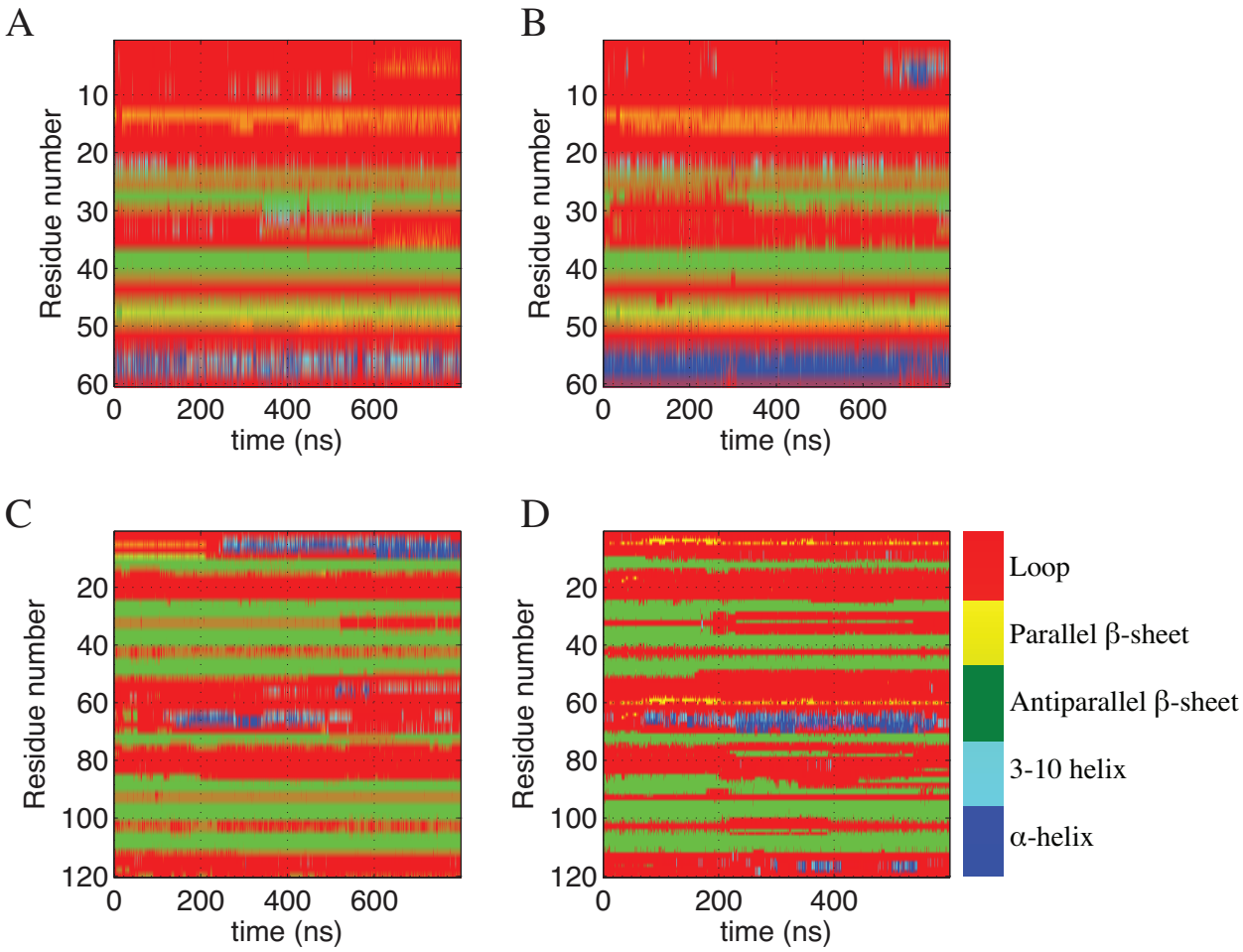


Figure 3.8: Secondary structure of protein motifs during implicit solvent (gb7) simulations of WT hLtn under different solution conditions. (A) hLtn10, s10; (B) hLtn10, n40; (C) hLtn40, s10; (D) hLtn40, n40. The C-terminal tail of hLtn10 following residue 60 has been truncated. Different colors represent different secondary structure elements as shown on the right.

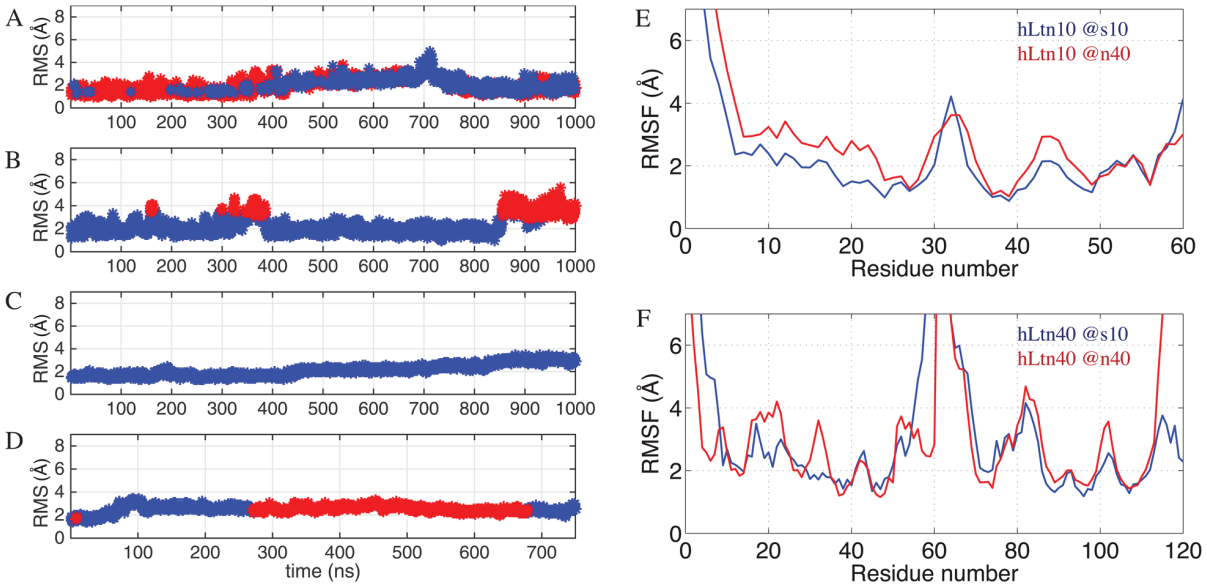


Figure 3.9: RMSD and RMSF calculated for hLtn10 and hLtn40 from implicit solvent simulations. Similar to Fig.3.3, RMSD is calculated with respect to the initial NMR structures used via Cα-aligment and RMSD values are color-coded via different clusters obtain at 4 Å threshold for (A) hLtn10, s10; (B) hLtn10, n40; (C) hLtn40, s10; (D) hLtn40, n40. A larger threshold than for explicit solvent simulations is used since the implicit solvent simulations are substantially longer and sample more diverse structures. The RMSF values are for (E) hLtn10 and (F) hLtn40 simulated under different conditions (compare to Fig.3.4); Blue: s10, Red: n40.

At the secondary structure level, the gb7 results are similar to the explicit solvent simulations (compare Figs.3.2 and 3.8): hLtn10 is somewhat more stable under the s10 condition than n40, especially for the first β-sheet. Under high temperature, β1 is less stable. For hLtn40, a similar trend is observed with β-sheet 1 and 2 become looser at higher temperature (compare Fig.3.8C-D). Looking at the RMSD and RMSF plots (Fig.3.9), it appears that the gb7 captures the expected behavior, i.e., hLtn10 is structurally less stable (with RMSD $>$ 5 Å) for some time periods due to motions of the N-terminal helix and fluctuations of the 30's and 40's loops under the n40 condition (see representative structures from clustering analysis in Fig.A3). hLtn40 is somewhat closer to the experimental structure under the n40 condition despite the instability around the β-sheets mentioned above. Overall, however, the RMSFs from the implicit solvent simulations are higher than those from explicit solvent simulations, only by $\sim$ 1 Å for most residues.

Regarding the interactions of charged residues, for hLtn10, the gb7 and explicit simulations do not differ significantly (Table 3.2). For example, the Arg 23 – Arg 43 pair remains close; without the explicit ions, the pair actually is closer by about 1 Å compared to explicit solvent simulations. For salt-bridge interactions, WT explicit solvent simulation finds stable pairs between Asp 7 – Lys 46 and Arg 18 – Asp 58, while those are not formed in the gb7 simulations. On the other hand, as we commented on above, these persistent salt-bridges are likely fortuitously observed in the particular explicit solvent s10 simulations because they are not observed in other explicit solvent simulations either. In this regard, the longer time scale available to the gb7 simulations likely leads to more extensive and realistic samplings.

For hLtn40, the persistent pairs observed in the explicit solvent simulations are Arg 35 – Asp 50, Glu 31 – Lys 25* and Arg 18 – Lys 42. Among them, Arg 35 – Asp 50 remains short and stable in the gb7 simulations, and Glu 31 – Lys 25* remain somewhat close but quite a bit longer than what explicit solvent simulations predict, especially under the n40 condition.

Regarding the change of SASA values associated with the hydrophobic residues in the two folds, comparison of the explicit solvent (Fig.3.7 top panel) and implicit solvent (Fig.3.7 bottom panels) observes parallel behaviors. Largest peaks (i.e., change of environment between the two folds) such as Ile 24, Ile 29, Leu 45 and Trp 55 are still prominent in the implicit solvent simulations. Nevertheless, there are quantitative differences. Val 21, Ile 24, Ile 29, Phe 39, Leu 45, Trp 55 and Val 59 show notably higher SASA in hLtn10 in implicit solvent simulation compared to explicit solvent.

Collectively, these observations indicate that the gb7 model provides an adequate description of the structural ensembles of hLtn10 and hLtn40 under different solution conditions, especially concerning the overall structural features and interactions among charged residues. This is rather remarkable considering the simple model used to describe salt effect (Debye-Hückel) and that the surface area term is not included in the GPU based gb7 computations. Nevertheless, for higher temperature simulations (40 °C), it appears that the stability of certain structural elements is not as well described by gb7 since in most GB simulations the surface tension coefficient in the SASA term is not taken to be temperature dependent.

3.4 Conclusions

Lymphotactin is a metamorphic protein which exists as a monomer (hLtn10) at 10 °C with 200 mM salt and as a dimer (hLtn40) at 40 °C and low salt. It is a remarkable example for how environmental conditions can alter protein stability via electrostatic and hydrophobic interactions. Since the stability is partially driven by salt, electrostatic interactions have an important role in maintaining stability; however, it requires extensive sampling to identify charge pairs that are potentially important for stability. To tackle this challenge, we have designed simulations with explicit and implicit solvent models, under different conditions, including wild type and mutant models. We have found supporting evidence in line with the hypothesis that Arg 23 and Arg 43 is a repulsive pair; either removal of the charges in these locations or Cl- binding can help stabilize the hLtn10 structure. While focusing on the network of charged interactions, we have identified Asp 7 – Lys 46 and Arg 18 – Asp 58 as salt-bridges and Lys 25 – Lys 42, Lys 42 – Lys 66 and Lys 66 – Arg 70 as like-charge pairs that are persistent among simulations and are potentially involved in the stability of hLtn10. We predict that as the common residue in all the like-charge pairs, Lys 42 deserves further attention and that the mutation of Lys 42 to a negatively charged amino acid is predicted to increase the hLtn10 stability. Consistent with previous studies [52] for the dimer, we have observed a larger number of salt bridges than hLtn10. Arg 35 – Asp 50, Glu31 – Lys 25* are the most consistent salt-bridges among all simulations and disrupting those interactions is predicted to cause destabilization of the hLtn40 system. Additionally, analysis of hydrophobic interactions reveals that mutating Trp 55, Ile 24, Phe 39 and Leu 10 to charged/polar residues would favor hLtn40 over hLtn10, whereas mutating Ile 29, Leu 45, Leu 34 and Val 12 to charged/polar residues would favor hLtn10 over hLtn40.

In addition to studying the network of electrostatic interactions, we have carried out FEP and LRA

calculations on Arg 23 and Arg 43 to quantify their contribution to the stability of hLtn10. We have shown that removal of charges from Arg 23 and Arg 43 leads to a stabilizing effect on hLtn10, in agreement with experiment, and that FEP in combination with the BAR protocol is a more robust method than LRA for screening sites that contribute to the stability of hLtn.

Finally, our results on the GB simulations of hLtn reveal that at the room temperature, GB provides a satisfactory description of the structural ensembles for hLtn10 and hLtn40, especially for the overall structural stability and interactions among charged residues. At higher temperatures, however, the GB model adopted here does not provide an accurate description for the stability of some of the structural elements.

3.5 References

[1] C. B. ANFINSEN, *Science* **181**, 223 (1973).

[2] S. B. PRUSINER, *Proc. Nat. Acad. Sci. USA* **95**, 13363 (1998).

[3] H. J. DYSON and P. E. WRIGHT, *Nat. Rev. Mol. Cell Biol.* **6**, 197 (2005).

[4] V. N. UVERSKY and A. K. DUNKER, *Biochim. Biophys. Acta - Prot. & Proteom.* **1804**, 1231 (2010).

[5] A. MURZIN, *Science* **320**, 1725 (2008).

[6] P. BRYAN and J. ORBAN, *Curr. Opin. Struct. Biol.* **20**, 482 (2010).

[7] P. ALEXANDER, Y. HE, Y. CHEN, J. ORBAN, and P. BRYAN, *Proc. Natl. Acad. Sci. USA* **106**, 21149 (2009).

[8] Y. HE, Y. CHEN, P. A. ALEXANDER, P. N. BRYAN, and J. ORBAN, *Structure* **20**, 283 (2012).

[9] B. CAO and R. ELBER, *Proteins: Struct., Funct., & Bioinform.* **78**, 985 (2010).

[10] C. G. ROESSLER, B. M. HALL, W. J. ANDERSON, W. M. INGRAM, S. A. ROBERTS, W. R. MONTFORT, and M. H. CORDES, *Proc. Natl. Acad. Sci. USA* **105**, 2343 (2008).

[11] C. CAMILLONI and L. SUTTO, *J. Chem. Phys.* **131**, 245105 (2009).

[12] R. TUINSTRA, F. PETERSON, S. KUTLESA, E. ELGIN, M. KRON, and B. VOLKMAN, *Proc. Natl. Acad. Sci. USA* **105**, 5057 (2008).

[13] F. C. PETERSON, E. S. ELGIN, T. J. NELSON, F. ZHANG, T. J. HOEGER, R. J. LINHARDT, and B. F. VOLKMAN, *J. Biol. Chem.* **279**, 12598 (2004).

[14] B. F. VOLKMAN, T. Y. LIU, and F. C. PETERSON, *Methods Enzymol.* **461**, 51 (2009).

[15] E. KULOGLU, D. MCCASLIN, M. KITABWALLA, C. PAUZA, J. MARKLEY, and B. VOLKMAN, *Biochem.* **40**, 12486 (2001).

[16] E. KULOGLU, D. MCCASLIN, J. MARKLEY, and B. VOLKMAN, *J. Biol. Chem.* **277**, 17863 (2002).

[17] R. TYLER, N. MURRAY, F. PETERSON, and B. VOLKMAN, *Biochem.* **50**, 7077 (2011).

[18] K. A. DILL, *Biochem.* **29**, 7133 (1990).

[19] S. D'AMICO, J. C. MARX, C. GERDAY, and G. FELLER, *J. Biol. Chem* **278**, 7891 (2003).

[20] A. H. ELCOCK, *J. Mol. Biol.* **284**, 489 (1998).

[21] M. FORMANECK, L. MA, and Q. CUI, *J. Am. Chem. Soc.* **128**, 9506 (2006).

[22] L. MA and Q. CUI, *Biochem.* **45**, 14466 (2006).

[23] A. S. THOMAS and A. H. ELCOCK, *J. Am. Chem. Soc.* **126**, 2208 (2004).

[24] R. TYLER, J. WIETING, F. PETERSON, and B. VOLKMAN, *Biochem.* **51**, 9067 (2012).

[25] D. CASE, T. DARDEN, T. CHEATHAM III, C. SIMMERLING, J. WANG, R. DUKE, R. LUO, R. WALKER, W. ZHANG, and K. MERZ, *University of California, San Francisco* **1**, 3 (2012).

[26] R. SALOMON-FERRER, D. CASE, and R. WALKER, *WIREs Comput Mol Sci.* **3**, 198 (2013).

[27] V. HORNAK, R. ABEL, A. OKUR, B. STROCKBINE, A. ROITBERG, and C. SIMMERLING, *Proteins: Struct., Funct., & Bioinf.* **65**, 712 (2006).

[28] S. SHOWALTER and R. BRUSCHWEILER, *J. Chem. Theory Comput.* **3**, 961 (2007).

[29] D. LI and R. BRUSCHWEILER, *J. Phys. Chem. Lett.* **1**, 246 (2010).

[30] D. LI and R. BRÜSCHWEILER, *Angew. Chem. Int. Ed.* **122**, 6930 (2010).

[31] K. BEAUCHAMP, Y. LIN, R. DAS, and V. PANDE, *J. Chem. Theory Comput.* **8**, 1409 (2012).

[32] M. FEIG, J. KARANICOLAS, and C. L. BROOKS, *J. Mol. Graph. & Model.* **22**, 377 (2004).

[33] W. L. JORGENSEN, J. CHANDRASEKHAR, R. W. MADURA, JEFFRY D IMPEY, and M. L. KLEIN, *J. Chem. Phys.* **79**, 926 (1983).

[34] I. S. JOUNG and T. E. CHEATHAM, *J. Phys. Chem. B* **112**, 9020 (2008).

[35] H. C. ANDERSEN, *J. Chem. Phys.* **72**, 2384 (1980).

[36] J.-P. RYCKAERT, G. CICCOTTI, and H. J. BERENDSEN, *J. Comp. Phys.* **23**, 327 (1977).

[37] J. MONGAN, C. SIMMERLING, J. MCCAMMON, D. CASE, and A. ONUFRIEV, *J. Chem. Theory Comput.* **3**, 156 (2007).

[38] J. WEISER, P. S. SHENKIN, and W. C. STILL, *J. Comput. Chem.* **20**, 217 (1999).

[39] V. TSUI and D. A. CASE, *Biopolymers* **56**, 275 (2000).

[40] A. W. GOTZ, M. J. WILLIAMSON, D. XU, D. POOLE, S. LE GRAND, and R. C. WALKER, *J. Chem. Theory Comput.* **8**, 1542 (2012).

[41] H. NGUYEN, D. R. ROE, and C. SIMMERLING, *J. Chem. Theory Comput.* **9**, 2020 (2013).

[42] J. SRINIVASAN, M. W. TREVATHAN, P. BEROZA, and D. A. CASE, *Theo. Chem. Acc.* **101**, 426 (1999).

[43] T. P. STRAATSMA and J. A. MCCAMMON, *Annu. Rev. Phys. Chem.* **43**, 407 (1992).

[44] P. A. KOLLMAN, *Chem. Rev.* **93**, 2395 (1993).

[45] T. SIMONSON, G. ARCHONTIS, and M. KARPLUS, *Acc. Chem. Res.* **35**, 430 (2002).

[46] J. L. GAO, K. KUCZERA, B. TIDOR, and M. KARPLUS, *Science* **244**, 1069 (1989).

[47] M. R. SHIRTS and D. L. MOBLEY, *Methods in Mol. Biol.* **924**, 271 (2013).

[48] C. H. BENNETT, *J. Comp. Phys.* **22**, 245 (1976).

[49] R. M. LEVY, M. BELHADJ, and D. B. KITCHEN, *J. Chem. Phys.* **95**, 3627 (1991).

[50] T. SIMONSON, *Proc. Acad. Natl. Sci. U.S.A.* **99**, 6544 (2002).

[51] A. WARSHEL and A. DRYGA, *Proteins: Struct., Funct., Bioinf.* **79**, 3469 (2011).

[52] S. R. HARVEY, M. PORRINI, A. KONIJNENBERG, D. J. CLARKE, R. C. TYLER, P. R. LANGRIDGE-SMITH, C. E. MACPHEE, B. F. VOLKMAN, and P. E. BARRAN, *J. Phys. Chem. B* **118**, 12348 (2014).

[53] L. MA, N. SUNDLASS, R. T. RAINES, and Q. CUI, *Biochem.* **50**, 266 (2011).

Chapter 4

Comparison of Native and Non-native Ubiquitin Dimers through Molecular Dynamics and Small Angle X-Ray Scattering Reveals Analogous Structures [1]

PREFACE This chapter contains portion of my co-first-author manuscript published in Prot. Sci with Grace Pham and Ambar Rana. My contributions to this work include designing and performing all of the molecular simulations and their analysis. G. H. Pham designed and synthesized the ubiquitin dimers with isopeptide and thiolene linkages and collected the small angle X-ray scattering data. A. S. J. B. Rana designed, performed and analyzed the kinetic assays.

[1] Reprinted (adapted) with permission from "Comparison of Native and Non-native Ubiquitin Oligomers Reveals Analogous Structures and Reactivities", G. H. Pham‡ , A. S. J. B. Rana‡ , E. N. Korkmaz‡ , V. H. Trang, Q. Cui, L. A. Leinwand and E. R. Strieter *Prot. Sci.* **25**, 456-471 (2016) ‡Contributed equally.

4.1 Introduction

Post-translational modifications (PTMs) dramatically expand the functions of proteins. Understanding the functional consequences of PTMs on a molecular level requires access to modified proteins for structural and biochemical studies – a challenge that is driving research in numerous labs. A PTM of particular interest, due to its involvement in virtually every biochemical pathway, is protein ubiquitination [1]. Ubiquitin (Ub) is a small 76 amino acid protein that is covalently attached to substrates through an isopeptide bond between the C-terminal glycine of Ub and the ε-amino group of lysine residues. A set of dedicated enzymes (E1, E2 and E3) site-selectively conjugates Ub to substrate proteins and builds Ub chains by connecting subunits via the seven internal lysine residues (K6, K11, K27, K29, K33, K48, K63) or the N-terminus [2]. Enzymatic synthesis can be applied in vitro to generate select Ub conjugates with varying topologies [3–5]. However, the scope of enzymes identified for site-selective ubiquitination currently does not approach the vast array of Ub conjugates suggested to regulate cellular pathways [6].

To address this limitation, semi-synthetic and total-synthetic approaches have been developed to access native and non-native Ub conjugates. Strategies to generate non-native linkages benefit from their high-yields and tunability for regiospecific modifications. There have been a number of success stories involving application of Ub conjugates with non-native linkages to probe biological systems [7–10]. Notably, non-native Ub chains were used to determine the minimal Ub signal required for efficient proteasomal targeting of substrate proteins [11]. However, there is still some skepticism toward the use of non-native mimics of isopeptide bonds for biochemical and biophysical studies to facilitate new discoveries. Such concerns are especially germane to the study of deubiquitinases (DUBs), a class of enzymes that directly target and cleave isopeptide bonds [12]. In order to be recognized and processed by DUBs, Ub conjugates bearing non-native linkages should recapitulate native structural features. The conformations and flexibility of select non-native Ub linkages have previously been evaluated with computational methods [13]. However, to effectively address whether non-native linkages can adopt native chain conformations and functions, a comprehensive solution-phase characterization is needed.

Our lab recently demonstrated that thiol-ene coupling (TEC) could be used to rapidly build a diverse array of Ub chains with minimal synthetic manipulations [14, 15]. TEC chemistry affords a thioether Gly-

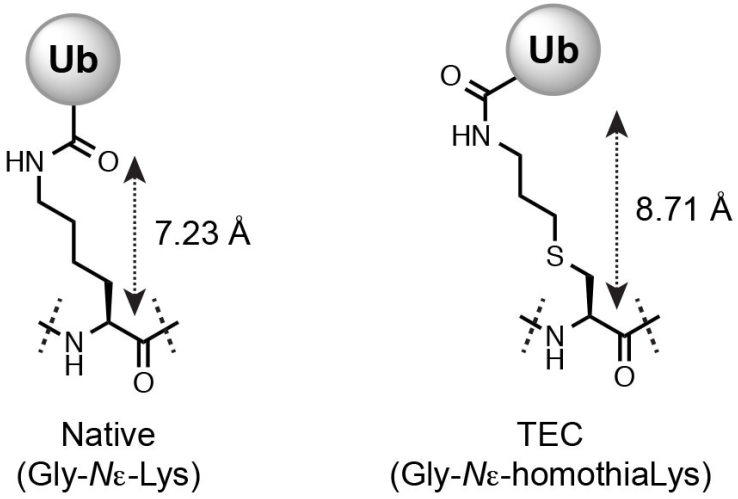


Figure 4.1: Structures of native and non-native isopeptide linkages. The distance from the Cα-atom of lysine to the carbonyl carbon of Ub Gly76 is shown for each linkage. This is based on density functional theory geometry optimized structures.

Nε-homothiaLys linkage that is 1.5 Å longer than the native isopeptide bond (Fig. 4.1). We previously demonstrated that DUBs could still selectively cleave the artificial linkage in end-point hydrolysis assays. To evaluate the extent to which the Gly-Nε-homothiaLys linkage can mimic the native isopeptide bond, we first sought to compare the Ub chain conformations generated by different linkage types. For this task, we selected small angle X-ray scattering (SAXS), a popular technique for evaluating the conformations of flexible multi-domain proteins in solution [16, 17]. Among the different Ub chains, the K48-linked Ub dimer (Ub2) is the best characterized in terms of structure, with more than 50 reported conformations derived from X-ray and NMR analyses [18–23]. We took advantage of this wealth of existing structural information to compare the SAXS-derived structural properties of native and TEC-derived K48-linked Ub2. SAXS analysis of both dimers reveals comparable scattering profiles that are matched to similar structures from a library of reported structures. Despite differences in the linkage itself, the conformations of native and non-native K48-Ub2 are markedly similar. As an alternative structural characterization protocol, multiple microsecond atomistic simulations are carried out for Ub2 and representative structures from the molecular dynamics trajectories are used to fit the SAXS profiles. Since the low-resolution nature of SAXS data prevents a precise determination of relative orientation of the two monomers, the MD derived ensemble of conformations differ somewhat from those fitted with the experimental structural library; however, the results again point to similar conformational ensembles for the native and TEC-derived K48-Ub2.

In the second tack, we compared steady-state kinetics of DUB cleavage reactions with TEC-derived and native Ub dimers. There are five families of DUBs: the Ub C-terminal hydrolases (UCHs), Ub specific proteases (USPs), ovarian tumor DUBs (OTUs), Machado-Joseph disease proteases (MJDs), and the zinc-dependent Jab1/Mpn/Mov34 proteases (JAMMs). The physiological substrates for many of these enzymes are unknown, but many DUBs exhibit selective hydrolysis activity toward Ub chains [24]. For example, members of the JAMM family (e.g., AMSH and BRCC36) cleave K63-linked chains exclusively [25, 26], and ten out of the 16 human OTU DUBs preferentially to cleave only one or two linkage types [27]. For TEC-derived chains to be of use for systematic interrogation of DUB activity, hydrolytic cleavage of the non-native linkage must occur with native-like efficiency and selectivity. We demonstrate that these criteria are satisfied with several different DUBs and Ub dimers. Comparing DUB activity with native and non-native substrates also provided a unique opportunity to gain new mechanistic insights.

4.2 Methods

4.2.1 Synthesis of Non-Native and Native Ub Dimers

To generate non-native Ub2 via thiol-ene coupling (TEC), Ub-KxC harboring a C-terminal allylamine adduct (2 mM) and lithium acyl phosphinate (LAP) (0.5 mM) were combined in 250 mM NaOAc buffer pH 5. The reaction mixture (1.8 mL total) was cooled to 4°C and irradiated with 365 nM light for 30 minutes using an OmniCure series 1500 light source placed 15 cm above the sample. Native Ub2 derivatives were generated using human E1 Ub-activating enzyme (0.1 μM) and linkage specific E2 conjugating enzymes as previously reported. For both the TEC and enzymatic reactions, the products were purified using a Superdex 75 HiLoad size exclusion column at a flow rate of 0.2 mL/min, collecting 2.5 mL fractions over 0.7 column volumes.

4.2.2 SEC-SAXS Measurements and Data Processing

SEC-SAXS (Size-Exclusion Chromatography – Small Angle X-ray Scattering) experiments were performed at BioCAT (beamline 18-ID, Advanced Photon Source at Argonne National Labs) [28]. The camera included a focused 12 KeV (1.03 Å) X-ray beam, a 1.5 mm quartz capillary sample cell, a sample to detector distance of ∼2.5 m, and a Mar165 CCD detector. The q-range sampled was $\sim$ 0.0065 – 0.3 Å^{-1}. To ensure sample monodispersity, we used an in-line SEC setup, which included an AKTA-pure FPLC unit and a Superdex-200 10/300 GL column (GE Healthcare Life Sciences). The column outlet was directly connected to the SAXS sample cell. Two second exposures were collected every six seconds during the gel-filtration chromatography run. Samples were analyzed at room temperature in 10mM Tris buffer pH 7.5, 50 mM NaCl, 1 mM DTT and 1 mM EDTA. Exposures before and after the elution of the sample were averaged and used as the buffer curve, and the exposures during elution (co-incident with the UV peak on the chromatogram) were treated as protein plus buffer curves. Data were corrected for background scattering by subtracting the buffer curve from protein plus buffer curves. $P(r)$functions and R_g values were determined from the scattering data using PRIMUS and GNOM from the ATSAS software package [29].

4.2.3 Molecular Dynamics

Atomistic simulations were carried out for Ub2. To facilitate sampling, independent MD simulations for Ub2 with a native K48-linker started with four experimental structures that differ in the degree of com-

pactness: 2O6V ($R_g \sim$ 16.4 Å) [30], 2PE9 ($R_g \sim$ 15.8 Å) [21], 3AUL ($R_g \sim$16.7 Å) [19] and 2KDF (R_g: 21.4 - 23.9 Å) [23] for K48-linkage as well as 2JF5 (R_g: 22.8 Å) [31], 2RR9 ($R_g \sim$ 23.5 Å), 3A1Q ($R_g \sim$ 22.5 Å) [32] and 3H7P (R_g: 23.1 Å) [33] for K63-linkage. Moreover, dimers bearing non-native linkages were also simulated in which Gly-Nε-homothiaLys (thiol-ene) or Gly-Nε-thiaLys (short thiol-ene) was introduced at position-48 or position-63 of one Ub subunit to connect with the C-terminus (Gly76) of another. With these different models, MD simulations were performed on the μs time-scale to ensure sampling a diverse range of conformations. The Amber v12 Molecular Dynamics package [34–38] was chosen for the simulations since the implicit solvent model implementation on a GPU affords a 100-fold speed enhancement compared to a single CPU. Specifically, the ff14SBonlysc Amber force field was chosen [39–42] along with gb8 as the implicit solvent (Generalized Born) model. [41] Production gb8 simulations were carried out for a minimum of 800 ns using a 1 fs time step (see Table 4.1 for a summary). Langevin dynamics were used with a collision frequency of 20 ps^{-1} at 300 K. The SHAKE algorithm was applied to bonds with hydrogen atoms with a tolerance of 10^{-5} Å. [43] The non-bonded cutoff was set at 9999 Å, and the maximum distance between atom pairs (rgbmax) for Born radii calculations was maintained at 12 Å. Debye-Huckel theory was applied to treat salt (implicitly) and the concentration was set to a physiologically relevant concentration (0.15 M).

To quantify the degree of conformational sampling, we calculated the torsion angle, γ, over a pool of conformations formed by merging all trajectories for each linkage. To calculate γ, we used the center of mass (COM) of each hydrophobic patch (L8, I44, H68 and V70) as a pivot point (Fig. 4.2). The torsion angle γ was concentrated around $\pm$ 150° but all angles are sampled between $\pm$ 180° (Fig. 4.2A and B), verifying that our simulations were capable of visiting a wide range of different conformations.

4.2.4 Analysis of MD Trajectories and Determination of Conformational Ensembles Using Experimental SAXS Profiles as Reference

The MD trajectories for different Ub2 systems were combined and then analyzed with a clustering protocol using the k-clust command in MMTSB package [44] with Cα-RMSD as the similarity measure. SAXS profiles for each cluster centroid were calculated using Fast SAXS [45–47] and scored with respect to the experimental SAXS profile. SAXS fitting was carried out with three scoring functions that have been used in previous computational analyses [46–51]; with both, a lower score means a better match.

For the first scoring function,The FoXS (Fast open-source X-ray Scattering) server was used to cal-

Table 4.1: Lengths of atomistic molecular dynamics simulations for K48- and K63-linked Ub2.

Starting structure	Linkage	Lengths (ns)
2O6V	K48/Isopeptide	1000
2PE9	K48/Isopeptide	1000
3AUL	K48/Isopeptide	1000
2KDF	K48/Isopeptide	1000
2KDF	K48/Thiolene	1000
2KDF	K48/ShorterThiolene	800
2JF5	K63/Isopeptide	1000
2RR9	K63/Isopeptide	1000×2
2RR9	K63/Thiolene	1000×2
3A1Q	K63/Isopeptide	1000×2
3A1Q	K63/Thiolene	1000×2
I	K63/Isopeptide	1000
3H7P	K63/Thiolene	1000×2
3H7P	K63/Shorter Thiolene	1000×2

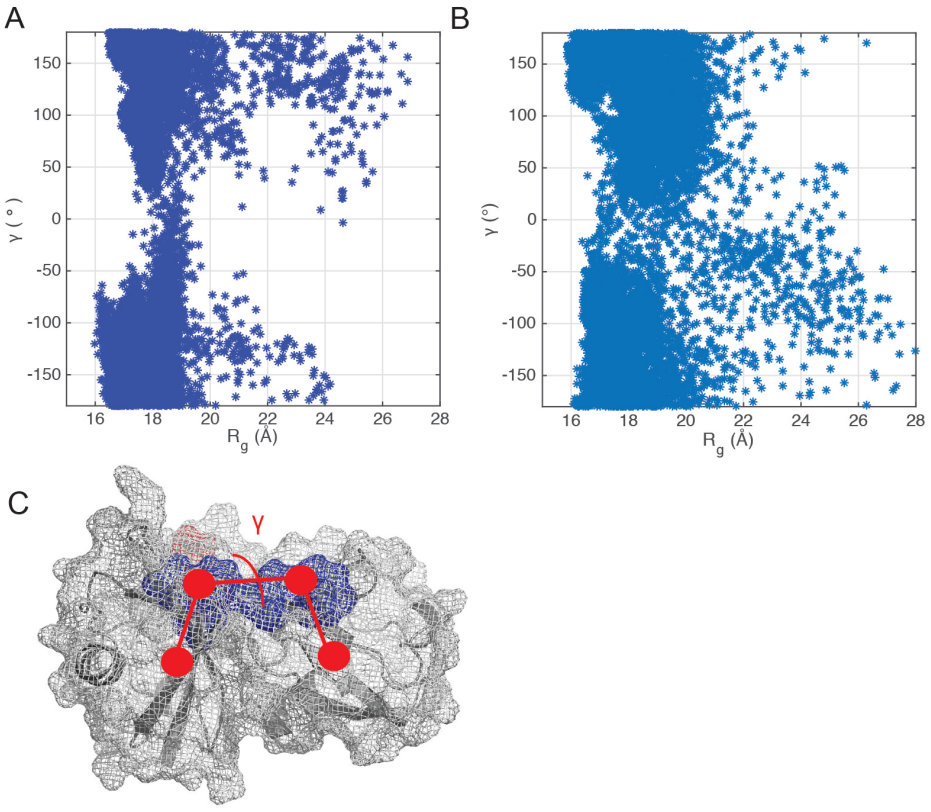


Figure 4.2: Distribution of R_g along γ for the (A) K48-linked and (B) K63-linked Ubiquitin dimers. The distributions of R_g and γ indicate that the MD simulations access a broad range of conformations. (C) γ is defined as the torsional angle created by the center of mass and the hydrophobic patches of the two subunits (residues L8, I44, H68 and V70) to quantify the degree of rotation of the subunits with respect to each other.

culate and compare SAXS scattering profiles for the library of reported structures [52, 53]. SAXS scattering profiles were computed based on the Debye Formula:

$$I(q) = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(q) f_j(q) \frac{sin(qd_{ij})}{qd_{ij}} \tag{4.1}$$

The intensity, $I(q)$, is a function of the momentum transfer, q; f_i(q) and f_j(q) represent the isotropic atomic form factors of the atoms i and j; d_{ij} is the distance between atoms i and j, and N is the total number of atoms in the molecule. SAXS profiles were computed with default settings assigned for the form factors – hydration layer, implicit hydrogens and excluded volume adjustments. Maximum q value was adjusted to match experimental data, q_{max} = 0.25 $Å^{-1}$, while profile size was left at a default value of 500. To assess the extent to which SAXS can be used to qualitatively distinguish different structures, we compared normalized $P(r)$ curves generated from the library of the-linked Ub2 structures. One hundred percent overlap in the $P(r)$ curves of two structures corresponded to identical shapes.

Experimental SAXS profiles were compared to FoXS computed profiles of reported structures. The quality of fit was evaluated by the χ_1 function:

$$\chi_1 = \sqrt{\frac{1}{M} \sum_{i=1}^{M} \left(\frac{I_{exp}(q_i) - cI(q_i)}{\sigma(q_i)} \right)^2} \tag{4.2}$$

M is the number of points in the profile, c is the scaling parameter and σ is the experimental error. The linear least squares minimization was performed to find the value of c that led to a low χ value, corresponding to a good fit of computed profile to experimental data.

The second scoring function [46, 47] is given as,

$$\chi_2^2 = \sum_{j=1}^{M} \frac{1}{\delta log I^2(q_j)} \left(log I_{calc}(q) - log I_{exp}(q) - \Delta \right)^2 \tag{4.3}$$

where q_{min} and q_{max} are the lower and upper limits of the q-range from the experimental scattering profile I_{exp}(q) and δlog $I(q)$ are the experimental uncertainties of log I_{exp}(q). The value of Δ corresponds to the offset between the theoretical and experimental SAXS profiles (log $I_{ca;}$ and log I_{exp}) at $q = q_{min}$. A good fit should have a score lower than 0.001 [46, 47]. The third scoring function is defined as the following [51],

$$\chi_3^2 = \frac{1}{K-1} \sum_{j=1}^{K} \left[\frac{\mu I(q_j) - I_{exp}(q_j) - \Delta}{\sigma(q_j)} \right]^2 \tag{4.4}$$

$$\mu = \sum_{j=1}^{K} \frac{I(q_j) I_{exp}(q_j)}{\sigma^2 I(q_j)} \Bigg/ \sum_{j=1}^{K} \frac{I(q_j)^2}{\sigma^2 I(q_j)} \tag{4.5}$$

where K is the number of experimental points, σ(q) are the standard deviations, and μ is a scaling factor. A good fit should have a score lower than 2.000 [51].

With the computed χ values, optimal weights for different cluster centroids are computed based on the best fit of the ensemble-averaged SAXS profile to the experimental data for q up to 0.20 $Å^{-1}$; the results are summarized in Table 4.3.

4.2.5 Steady-State Kinetic Analyses of DUB-Catalyzed Cleavage Reactions

Stock solutions of DUBs and Ub dimers were prepared in a 50 mM Tris-Cl buffer at pH 7.5 containing 150 mM NaCl and 5 mM DTT. Kinetic assays were performed by varying the concentration of Ub dimers while maintaining a constant concentration of the indicated DUB (50 nM of USP7, 200nM and 2μM of OTUD7B, and 500 nM of AMSH) at 37 °C. Reactions were quenched by addition of 6X Laemmli sample loading buffer. To visualize monoUb formation, the quenched reactions were run on a 15 % denaturing SDS-PAGE gel along with monoUb standard, followed by SYPRO(R) Ruby staining. ImageJ was used to quantify the appearance of monoUb over time. Experimental data were processed using Prism 5.02 (GraphPad Software).

4.3 Results and Discussions

4.3.1 SAXS Analysis of Native and TEC-Derived K48-Linked Ub2

Among the eight possible linkages between Ub subunits in a dimer, K48 has been extensively characterized using both X-ray scattering and NMR analyses. The reported structures of K48-linked Ub2 exhibit a range of conformations from compact (1AAR [18], 2PE9 [21] and 3AUL [19]) to extended (2KDF [23]). Consistent with this structural diversity, the conformational behavior of K48-linked Ub2 is complex with multiple states coexisting in solution [54–56]. To assess whether TEC-derived K48-linked Ub2 adopts a structure similar to that of the native dimers, we chose to use SAXS – an experimental technique for mapping the three-dimensional conformations of macromolecules in solution [16, 17]. The applicability of SAXS was recently demonstrated by work with native K63-linked Ub dimers, which showed that the open and closed

states of these oligomers could be distinguished [55]. Based on these results, we surmised that SAXS would be sufficiently sensitive to detect any major structural differences between native and non-native dimers.

The experimental SAXS profiles of native and TEC-derived 48-linked Ub2 reveal similar structural features for both dimers (Fig. 4.3A). The radii of gyration, R_g, based on the Guinier approximation, are 19.35 $\pm$ 0.08 Å and 19.25 $\pm$ 0.12 Å for native and non-native Ub2, respectively (Fig. 4.3B and Table 4.2). The distance distribution, $P(r)$, curves for both samples show a major peak at 20 Å and a subtle shoulder after 40 Å with maximum distance (D_{max}) of 70 Å for native Ub2 and 65 Å for non-native Ub2 (Fig. 4.3C and Table 4.2). Comparison of the normalized $P(r)$ curves shows approximately 95 % overlap. Lastly, Kratky curves for both native and non-native dimer converge at high q, suggesting well-folded structures (Fig. 4.3D).

The theoretical scattering profiles for the reported open and compact structures of K48-linked Ub2 are distinct from the experimental data. With the open structure (2KDF [23]), derived from Ub2 bound to S5a, the scattering profile curves downward in the low q region, whereas compact structures (1AAR [18], 2PE9 [21], 3AUL [19]) give rise to positive curvature (Fig. 4.4A). Within the same region, the scattering intensities for both the native and non-native dimers occupy the space between the open and compact states, suggesting the average structure in solution is one that is a mixture of open and compact states. The differences between the experimental and theoretical scattering profiles are even more apparent when examining the $P(r)$ curves (Fig. 4.4B). The compact states have smaller dimensions ($D_{max} \sim 60$ Å) compared to the experimental curves with minimal features around 40 Å. By contrast, the open state has a Dmax of $\sim$ 75 Å and two distinct peaks, indicative of an extended structure. In addition to the aforementioned compact and extended structures, there is a range of other intermediate K48-linked Ub2 structures in the presence and absence of binding partners [20, 22, 57–60].

4.3.2 SAXS Analysis of Native and TEC-Derived K63-Linked Ub2

K63-linked Ub2 has also been extensively characterized using X-ray scattering and NMR analyses. Similar to K48-linked Ub2, the conformational behavior of K63-linked Ub2 is reported to be complex with multiple states coexisting in solution. The experimental SAXS profiles of native and TEC-derived 63- linked Ub2 reveal similar structural features for both dimers (Fig. 4.3A). R_g values for native and non-native Ub2 are 20.47 $\pm$ 0.07 Å and 21.58 $\pm$ 0.10 Å, respectively (Fig. 4.3B). $P(r)$ curves for both samples show a major peak at 20 Å and a relatively distinct shoulder at 40 Å (Fig. 4.3C). Kratky curves for both native and non-native

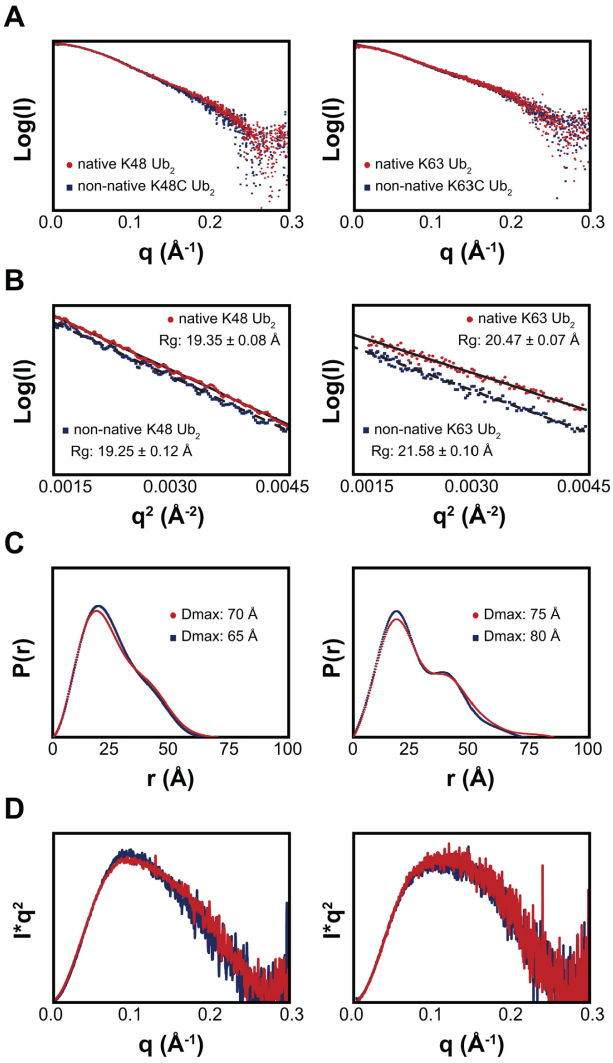


Figure 4.3: SAXS analysis of native and non-native TEC-derived K48-linked (left) and K63-linked (right) Ub2 reveals similar shapes. (A) SAXS intensity profiles of native (red) and TEC-derived (blue) Ub dimers. (B) The Kratky, (C) Guinier and (D) the distance distribution plots suggest well-folded structures with comparable dimensions. Guinier plots are staggered to clearly display both datasets.

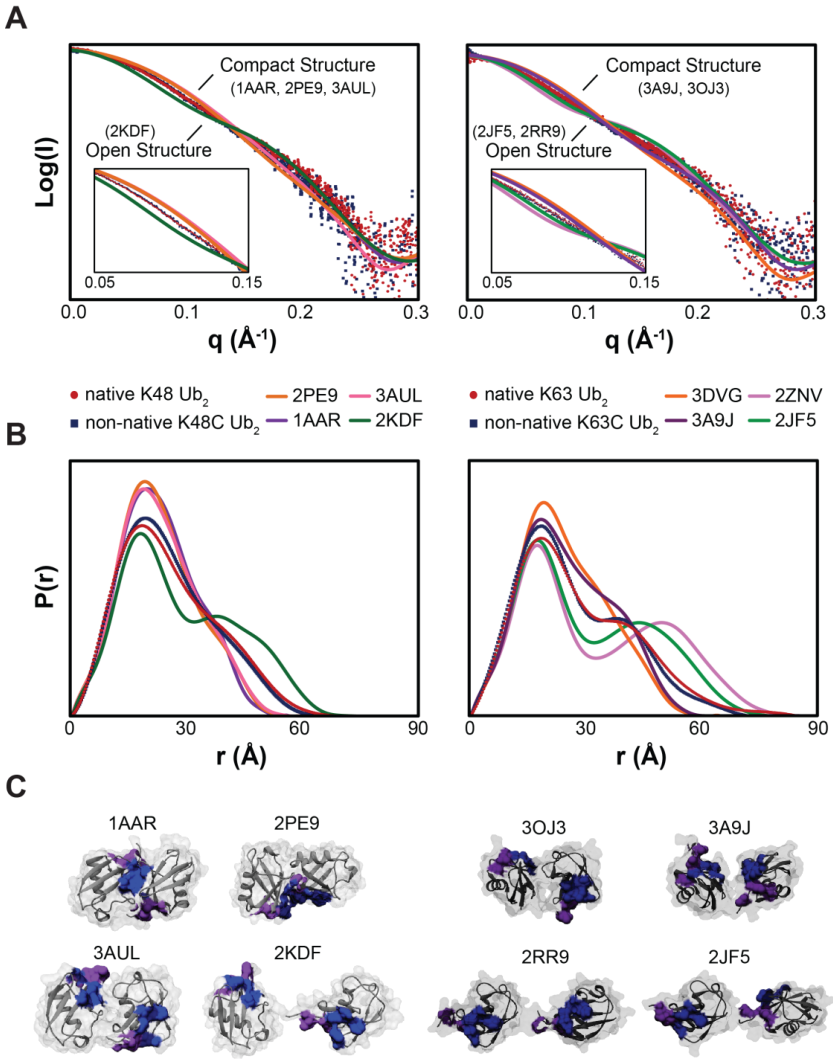


Figure 4.4: Comparison between experimental and theoretical scattering curves for open and extended structures of K48-linked (left) and K63-linked (right) Ub2. (A) Experimental (dots) and theoretical (lines) scattering intensity profiles. The theoretical scattering data were generated using CRYSOL. Insert shows expansion of $q = 0.05$ to $0.15\,\AA^{-1}$ (B) Experimental (dots) and theoretical (lines) distance distribution plots. (C) Compact and extended structures are shown, including PDB: 1AAR, 2PE9, 3AUL, 2KDF, 3DVG, 2JF5, 3A9J, 2ZNV. Hydrophobic patches centered on Ile44 and Ile36 are shown in blue and purple, respectively. Structures were visualized in Chimera.

Table 4.2: Summary of SAXS parameters.

SAXS Parameters	Native K48-linked Ub2	TEC K48C-derived Ub2
R_g (Å)	19.35 ± 0.08	19.25 ± 0.12
D_{max} (Å)	$\sim$70	$\sim$65
V_{porod} ($Å^3$)	24,353	23,653

SAXS Parameters	Native K63-linked Ub2	TEC K63C-derived Ub2
R_g (Å)	20.47 ± 0.07	21.58 ± 0.10
D_{max} (Å)	$\sim$75	$\sim$80
V_{porod} ($Å^3$)	25,515	26,131

dimer converge at high q, suggesting well-folded structures (Fig. 4.3D)

4.3.3 Determination of Conformational Ensembles from MD simulations and Experimental SAXS Data

The relative orientation of the two Ub monomers in the open conformer(s) is potentially difficult to determine using SAXS data alone due to the low-resolution nature of the technique. We have shown that using of an experimental structural library points to collect populations of open conformers for both native and TEC-derived Ub2. (Fig. B3) Therefore, to achieve structural diversity, we have carried out atomistic simulations for the K48-linked and K63-linked Ub2 starting from many independent structures. By comparing calculated SAXS profiles for representative structures from the MD trajectories to measured scattering data allows us to characterize the conformational diversity of the native and TEC-derived Ub2.

For simulations with K48-linked Ub2, all starting structures, even the most extended with R_g 25 Å, converged to modestly compact conformations $R_g \sim$ 18-19 Å. Compared to Ub2 simulations with a native linker, simulations with a TEC-derived linker generated relatively more compact conformations with the average R_g slightly below 19 Å. Clustering analysis was done over all collected MD trajectories, regardless of the chemical nature of the linker. For the ten clusters obtained, SAXS profiles were computed for the representative members from each cluster (Supporting Information Fig. B1 and Table B1) and reweighted to best fit the experimentally measured SAXS profiles of the native and TEC-derived dimers (Fig. 4.5). This procedure led to clusters 8, 9, and 10 as representatives for the native Ub2 ensemble with a population of 25 %, 45 % and 30 %, respectively. For the TEC-derived K48-linked Ub2, clusters 8 and 9 were selected with a population of 28 % and 72 %, respectively. As shown in Table 4.3, clusters 8 and 9 have a similar level of compactness with R_g values of $\sim$18.5 Å, while cluster 10 is more open in nature and has an R_g of 21.1 Å. Accordingly, the ensemble-averaged R_g values for the native and the TEC-derived dimers are estimated to be 19.31 $\pm$ 1.31 Å and 18.51 $\pm$ 1.00 Å, respectively.

MD simulations for K63-linked Ub2 also show a wide range of conformational sampling. The simulations were initiated from extended structures 2JF5 and 3H7P with R_g values of $\sim$23 Å, which converged to more compact structures. From the ten clusters for K63-linked Ub2 (Supporting Information Fig. B2 and Table B2), clusters 3, 7, 9 and 10 comprise the best conformational ensemble for native K63- Ub2 with populations of 30 %, 24 %, 12 % and 34 %, respectively. Clusters 3, 7, and 9 are relatively compact while cluster

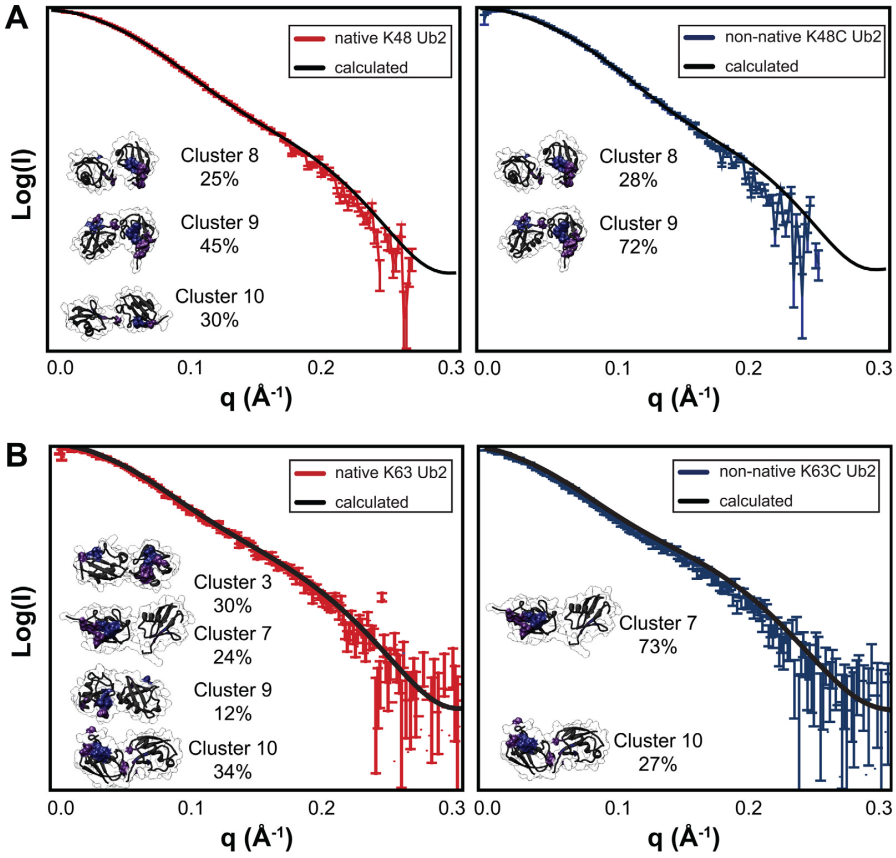


Figure 4.5: Comparison of MD ensembles for native and TEC-derived K48- and K63-linked Ub2. SAXS profiles are calculated via reweighting the clusters. The experimental SAXS profiles (red, blue) and calculated fits (black) are shown for (A) the native and TEC-derived K48-linked Ub2 and (B) native and TEC-derived K63-linked Ub2. Each representative member is shown with its respective percentages in the population. Hydrophobic patches centered on Ile36 and Ile44 are shown in purple and blue, respectively.

Table 4.3: Fitted Populations of Native and TEC-derived Ub2 based on Conformations Obtained from Atomistic Simulations Matched to Experimental SAXS Profiles.

R_g (Å)	Cluster	Fitted Population (%)	Score 1[a] (χ_1)	Score 2[b] ($\chi_2^2 <$ 1E-03)	Score 3[c] ($\chi_3^2 < 20$)
			Non-native K48C Ub2		
18.45 ± 1.24	8[d]	25	1.3	0.8E-04	1.2
18.53 ± 0.90	9[d]	45	2.0	0.5E-04 2.8	
21.17 ± 1.79	10[d]	30	1.8	2.2E-04	6.8
			Non-native K48C Ub2		
18.45 ± 1.24	8[d]	28	1.0	3.6E-04	1.3
18.53 ± 0.90	9[d]	72	0.9	2.2E-04	0.7
21.17 ± 1.79	10[d]	–	–	–	–
			Native K63 Ub2		
17.56 ± 0.78	3[e]	30	3.4	1.5E-02	11.7
18.77 ± 0.98	7[e]	24	3.1	1.2E-02	2.9
19.87 ± 2.03	9[e]	12	3.3	1.2E-02	2.3
23.69 ± 1.65	10[e]	34	3.5	1.1E-02	14.7
			Non-native K63C Ub2		
17.56 ± 0.78	3[e]	–	–	–	–
18.77 ± 0.98	7[e]	73	1.7	7.3E-02	3.3
19.87 ± 2.03	9[e]	–	–	–	–
23.69 ± 1.65	10[e]	27	2.7	4.4E-02	9.3

[a] Equation (2).

[b] Equation (3).

[c] Equation (4).

[d] MD clusters from K48-Ub2 simulations.

[e] MD clusters from K63-Ub2 simulations.

10 is more extended (Table I). R_g values for the selected clusters range from 17.6 Å to 23.7 Å. Clusters 7 and 10 are sufficient to represent the best fit conformational ensemble for TEC-derived K63-linked Ub2. The calculated ensemble-averaged R_g values for native and non-native Ub2 are 20.21 $\pm$ 0.70 Å and 20.10 $\pm$ 0.45 Å, respectively.

As shown in Table 4.3 (and SI Tables B1, reftab:simu63) the general quality of fitting to the experimental SAXS profiles is similar regardless of the scoring function. Scoring functions 2 and 3 favor the MD ensembles, while scoring function 1 slightly favors the ensembles determined using the experimental structural library. For both K48- and K63-linked Ub2 conformational ensemble analyses, the MD-derived ensembles and ensembles obtained from extant structures (See Figures B3, B4, B5 and B6) vary in terms of the precise R_g values and relative orientations of Ub monomers. However, the overall shapes and composition of the ensembles are similar, reflective of similar SAXS scattering profiles for native and TEC-derived Ub2.

4.3.4 Steady-State Kinetic Analyses of Ub Dimer Cleavage by Deubiquitinases

The SAXS patterns for native and non-native 48-linked Ub2 reflect similar conformations, however, it is possible that the unnatural linkage can affect recognition and hydrolysis by DUBs. To address this point, we compared hydrolysis of native and non-native Ub2 by representative members from three different DUB families. Of the five DUB families, three are known to selectively dismantle Ub chains: USPs, JAMMs, and OTUs [12]. We sought to investigate whether the selectivity and efficiency of these DUBs are preserved with non-native substrates. We selected the DUBs USP7, AMSH, and Cezanne as representative members of each family. To expand the scope of our analysis beyond K48-linked Ub2, we also assessed the hydrolysis of K63-linked and K11-linked Ub2. Collectively, these three Ub chain types are the most abundant Ub signals in asynchronous eukaryotic cells [61–64].

USPs are known to remove mono-Ub from protein substrates, but a handful of DUBs from this family can also dismantle Ub chains *invitro* with low selectivity [65]. Using USP15, we assessed whether this type of DUB can differentiate between the native and TEC-derived isopeptide linkage (Fig. 4.6A). For hydrolysis of native and TEC-derived K6-linked Ub2, K_m and k_{cat} values were two-fold higher for the native Ub2, translating to the same specificity constant (kcat/Km) on the order of 10^3 M^{-1} s^{-1} for both native and TEC-derived substrates (Table 4.4). With K48- and K63-linked Ub2, a two-fold higher K_m was observed with the native substrates while there was no significant difference in k_{cat} values. Overall, native and TEC-derived K48- and

K63-linked Ub2 were similarly hydrolyzed by USP15, with specificity constants on the order of 10^4 M^{-1} s^{-1} (Table 4.4). Hydrolysis of K6-linked Ub2 is an order of magnitude lower compared to the other two linkages tested; both native and TEC-derived substrates followed the same trend.

In contrast to USP7, AMSH is a JAMM DUB known for specifically cleaving K63-linked Ub chains. To direct linkage specificity, AMSH interacts with the surface surrounding K63 in the proximal Ub, which is linked to the C-terminus of the distal Ub. Upon measuring the kinetics of AMSH-catalyzed reactions, we find that the parameters are essentially the same for both dimer types, with k_{cat}/K_m on the order of 10^3 M^{-1} s^{-1} (Fig. 4.6B). This relatively low catalytic efficiency is consistent with reported kinetic parameters for the AMSH catalytic domain [66]. That the kinetic parameters, K_m and k_{cat}, are similar for the hydrolysis of native and non-native K63-linked Ub2 suggests the interactions important for both binding and catalysis are preserved with non-native Ub2.

OTUD7B is an OTU DUB that preferentially hydrolyzes K11-linked Ub chains [67]. We set out to examine (i) whether linkage selectivity is governed by the efficiency with which a Michaelis complex is formed (K_m) or steps following complex formation (k_{cat}), and (ii) whether a non-native linkage affects this preference. We observe a five-fold difference in k_{cat} and a marginal difference in K_m, which manifests in a 3-fold higher specificity constant (k_{cat}/K_m) for the native K11-linked Ub2 compared to the TEC-derived dimer (Fig. 4.7C). To evaluate whether OTUD7B maintains a preference for the K11-linkage with TEC-derived substrates, we also assessed hydrolysis of K48- and K63-linked Ub2 (Table 4.4). A 10-fold increase in enzyme concentration was necessary to induce appreciable hydrolysis of K48- and K63-linked Ub2. For native and TEC-derived Ub2, the specificity constant k_{cat}/K_m drops from 10^4 M^{-1} s^{-1} with K11-linked Ub2 to 10^3 M^{-1} s^{-1} with K48- or K63-linked Ub2. The disparity in hydrolysis can mainly be attributed to differences in k_{cat} values (40-fold for native Ub2, 10-fold for non-native Ub2). The kinetic results suggest OTUD7B achieves selectivity through a k_{cat} effect for both the native and TEC-derived K11-linkage.

OTUB1 is an OTU DUB known for specifically cleaving K48-linked Ub chains. [68] For this particular DUB, the rate of hydrolysis of TEC-derived K48 linkage is slower compared to its native counterpart. To assess whether this was the case for all OTU DUBs, we also investigated the activity of OTUB2 and OTUD7B. OTUB2 is known to preferentially hydrolyze K63-linked Ub chains, [68] and our kinetic analysis also reports the same trend (Fig. 4.7A) When OTUB2 was tested with native and TEC-derived K48-linked and K63-

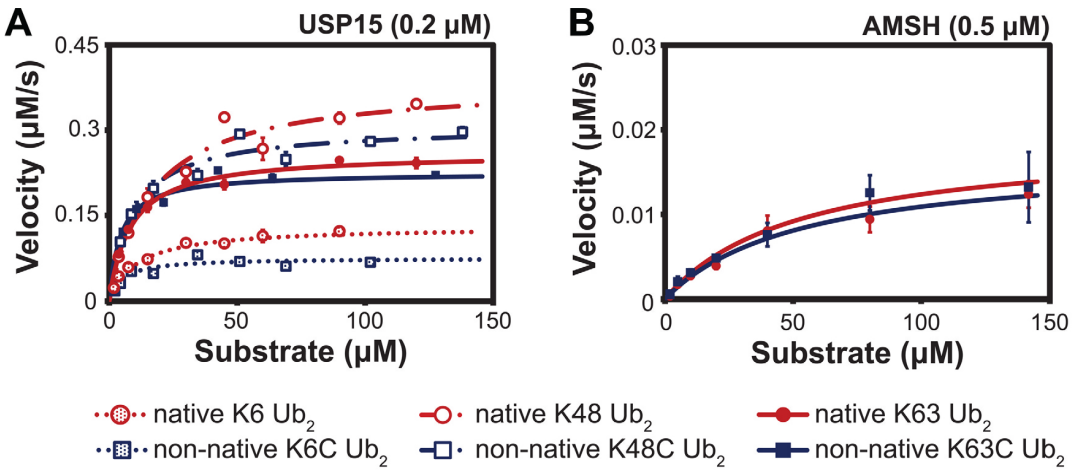


Figure 4.6: Hydrolysis of native and TEC-derived Ub2 by USP15 and AMSH. Michaelis–Menten analysis for hydrolysis of native (circles) and TEC-derived (squares) Ub2 by (A) USP15 (See Fig.B7, B8 and B9), and (B) AMSH (See Fig.B10).

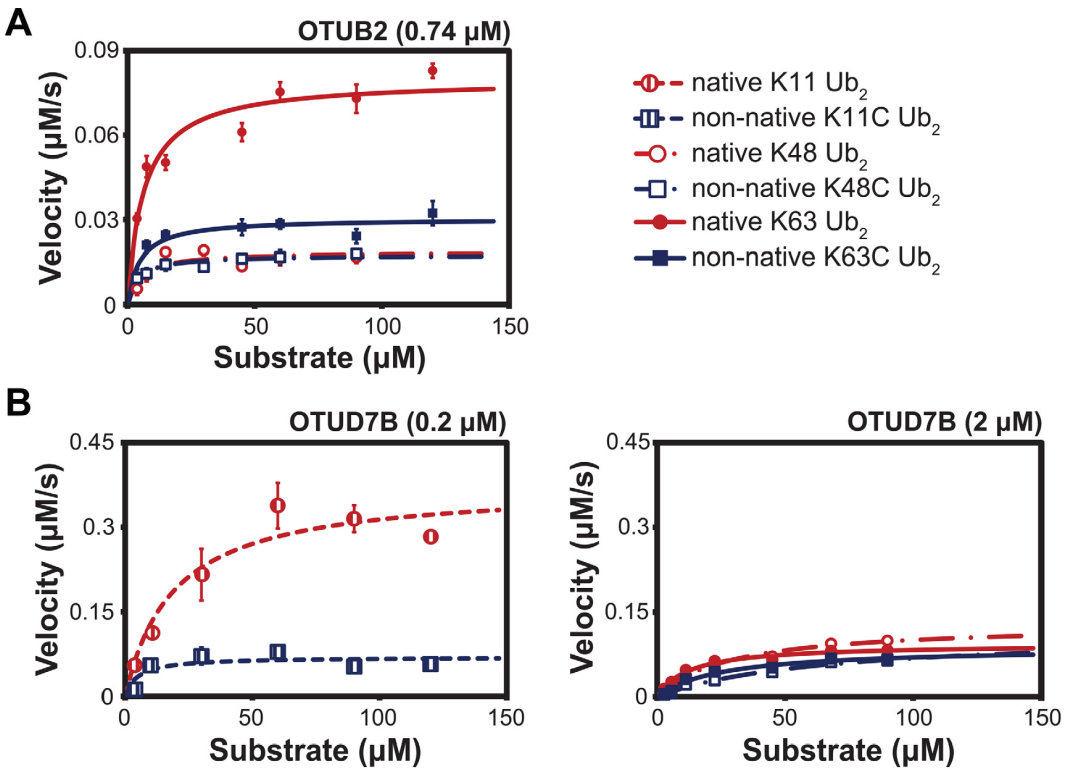


Figure 4.7: Hydrolysis of native and TEC-derived Ub2 by OTUB2 and OTUD7B. Michaelis–Menten analysis for hydrolysis of native (circles) and TEC-derived (squares) Ub2 by (A) OTUB2 (See Fig B11, B12 and B13), and (B) OTUD7B (See Fig.B14 and B15).

Table 4.4: List of kinetic parameters measured for all experiments for hydrolysis of native and TEC-derived Ub2 by USP15, AMSH, OTUB2 and OTUD7B.

Enzyme	Substrate	k_{cat} (min^{-1})	K_m (μM)	k_{cat}/K_m ($M^{-1}s^{-1}$)
USP15	K6 Ub2	39 ± 1	11 ± 1	60×10^3
	K6C Ub2	22 ± 1	5 ± 1	80×10^3
	K48Ub2	115 ± 4	17 ± 2	120×10^3
	K48C Ub2	92 ± 2	9 ± 1	170×10^3
	K63 Ub2	78 ± 1	8 ± 1	160×10^3
	K63C Ub2	68 ± 2	4 ± 1	260×10^3
AMSH	K64 Ub2	2 ± 0.3	54 ± 18	0.6×10^3
	K64C Ub2	2 ± 0.5	53 ± 27	0.7×10^3
OTUB2	K48Ub2	1.5 ± 0.1	5 ± 2	5×10^3
	K48C Ub2	1.4 ± 0.1	4 ± 1	6×10^3
	K63 Ub2	6.5 ± 0.2	6 ± 1	17×10^3
	K63C Ub2	2.5 ± 0.1	5 ± 1	8×10^3
OTUD7B	K11 Ub2	112 ± 9	19 ± 5	98×10^3
	K11C Ub2	21 ± 2	5 ± 2	69×10^3
	K48Ub2	3.9 ± 0.1	29 ± 3	2×10^3
	K48C Ub2	3.4 ± 0.3	65 ± 10	1×10^3
	K63 Ub2	2.8 ± 0.1	13 ± 1	4×10^3
	K63C Ub2	2.6 ± 0.1	25 ± 4	2×10^3

linked Ub2, similar Km values were observed for all four substrates and higher k_{cat} values were observed for K63-linked over K48-linked substrates (Fig. 4.7C). The difference in k_{cat} values is four-fold among native substrates and two-fold with non-native substrates. With both native and TEC-derived substrates, the enzyme is able to preferentially hydrolyze K63-linked Ub2 with higher efficiency than K48-linked Ub2.

4.4 Conclusions

Chemical strategies to generate Ub chains allow excellent control over chain length and regioselectivity, both of which are required for rigorous investigation of Ub chain structure and function. Whenever conjugation chemistry generates a non-native linkage in place of the native isopeptide linkage, it is important to assess whether the chemical surrogate can direct Ub chain behavior in a native-like manner. To evaluate the extent to which the TEC-derived non-native Gly-Nε-homothiaLys linkage mimics the native isopeptide bond, we compared both linkage types in the context of Ub dimers.

SAXS is a relatively low-resolution technique, so it is not possible to derive absolute structural information from SAXS scattering data alone. However, SAXS profiles can be compared to evaluate structural variability. Indeed, SAXS profiles calculated from atomistic molecular dynamics simulations (along with a library of extant K48-linked Ub2 structures) show a range of overall shapes, reflecting the known structural diversity. Upon inspecting the experimental SAXS profiles for native and TEC-derived K48-linked Ub2, it was evident the average conformation for both samples do not match the most compact or extended structures. This finding is consistent with recent in-solution NMR and FRET measurements [54, 56]. Computational studies with K48-linked Ub2 and a coarse-grained potential function also reported a multi-state energy landscape with a number of intermediate conformers in addition to the compact and extended structures [55].

Integrating SAXS data with various sets of conformers enables further comparison of native and TEC-derived Ub2. Using an the conformational pool created by simulations, we find that the the most representative structures quite similar. Fitting representative structures from the MD trajectories to the experimental SAXS profile leads to slightly different structural ensembles compared to the MES/experimental structural library combination, especially in terms of the relative orientations of the monomers. This is not entirely surprising since SAXS is a low-resolution technique and the structure library has limited conformational diversity. Importantly, however, the MD ensembles reweighted by the experimental SAXS profiles support that the native and TEC-derived Ub2 have similar conformational ensembles, with the native Ub2 being slightly more ex-

tended. This subtle difference in size could reflect structural variation at the molecular level and can be further characterized with high-resolution structural analyses.

Kinetic investigations of DUB activity have uncovered similar reactivity profiles between native and non-native dimers. AMSH is a K63 linkage-selective DUB that engages oligomers through multiple contacts within the proximal and distal units of Ub2. Upon comparing experimental SAXS profiles for native and TEC-derived K63-linked Ub2 with the calculated SAXS profile for K63-linked Ub2 cocrystalized with AMSH (PDB: 2ZNV), both native and TEC-derived Ub2 showed a moderate 70 % $P(r)$ overlap with the reported structure. 2ZNV captures an extended conformation of K63-linked Ub2, with R_g values 24–25 Å, whereas the R_g values for free K63-linked Ub2 in solution is smaller, 20.5 Å for native and 21.8 Å for TEC-derived. As native and TEC-derived K63-linked Ub2 produced equivalent kinetic parameters, it suggests that they are both able to make the necessary contacts with AMSH and adopt the conformation needed for hydrolysis.

The OTU subfamily of DUBs is also known to be linkage selective. For example, OTUB1 prefers to hydrolyze K48-linked Ub chains, OTUB2 prefers K63-linked chains, and OTUD7B prefers K11-linked chains. When OTUB1 was tested with native and TEC-derived K48-linked Ub2, the native Ub2 was hydrolyzed faster than the TEC-derived substrate (data not shown). Because OTUB1 could not accommodate the non-native linkage, either due to different recognition of the dimer structure or just of the linkage itself, we expanded our analysis to other OTU DUBs, including OTUB2 and OTUD7B. The catalytic domain of OTUB2 shares 58 % sequence identity with that of OTUB1. OTUB2 is known to preferentially hydrolyze K63-linked Ub chains, and our kinetic analysis confirms this trend for both native and TEC-derived substrates. OTUB2 interacts with K48- and K63-linked Ub2 similarly, and achieves selectivity for K63-linked Ub2 in a k_{cat} dependent manner. Unlike OTUB1, OTUB2 processes native and non-native Ub2 substrates with similar efficiency and selectivity. We also evaluated cleavage activity of OTUD7B with native and TEC-derived K11-, K48- and K63- linked Ub2. Similar to observations with OTUB2, k_{cat} is the major determinant for substrate selectivity. It is possible that selectivity arises from a necessary conformational change, in substrate and/or enzyme, which positions the isopeptide linkage into the proper orientation for catalysis. Because hydrolysis is best achieved with K11-linked Ub2, both native and TEC-derived, the positioning of this Ub2 is ideal for OTUD7B recognition and catalysis.

A similar k_{cat} effect was also observed with USP15, which interacts with K6-, K48- and K63- linked

Ub2 with similar efficiency but shows relatively slower hydrolysis of K6-linked Ub2. The recognition of K6-linked Ub2 is unfavorable enough to lower k_{cat}/K_m, but does not significantly ablate hydrolysis. Consequently, unlike OTUB2 and OTUD7B, USP15 is promiscuous toward different Ub2 substrates and demonstrates a relatively subtle preference for K48- and K63-linked Ub2. The same substrate preference was observed for both native and TEC-derived substrates, suggesting that the non-native substrates are efficacious mimics of their native counterparts.

Our comparative analysis of native and TEC-derived Ub2 demonstrate that the non-native Ub2 can be used to investigate DUB activity. Similar analyses can be applied to assess the structure and recognition of other non-native Ub conjugates. As analogs of other post-translational modifications, such as phosphorylation and glycosylation, have enabled leaps in mechanistic understanding in their respective fields, non-native Ub chains have a promising role in Ub research. When possible, new discoveries acquired with non-native Ub chains should be confirmed with the relevant native chains to better capture the molecular details influencing Ub functions.

4.5 References

[1] C. GRABBE, K. HUSNJAK, and I. DIKIC, *Nat Rev Mol Cell Biol* **12**, 295 (2011).

[2] C. M. PICKART, *Annu Rev Biochem* **70**, 503 (2001).

[3] K. C. DONG, E. HELGASON, C. YU, L. PHU, D. P. ARNOTT, I. BOSANAC, D. M. COMPAAN, O. W. HUANG, A. V. FEDOROVA, D. S. KIRKPATRICK, S. G. HYMOWITZ, and E. C. DUEBER, *Structure* **19**, 1053 (2011).

[4] M. K. HOSPENTHAL, S. M. FREUND, and D. KOMANDER, *Nat Struct Mol Biol* **20**, 555 (2013).

[5] C. M. PICKART and S. RAASI, *Methods Enzymol* **399**, 21 (2005).

[6] D. KOMANDER and M. RAPE, *Annu Rev Biochem* **81**, 203 (2012).

[7] T. ABEYWARDANA and M. R. PRATT, *Chembiochem* **15**, 1547 (2014).

[8] G. H. PHAM and E. R. STRIETER, *Curr Opin Chem Biol* **28**, 57 (2015).

[9] L. SPASSER and A. BRIK, *Angew Chem Int Ed Engl* **51**, 6840 (2012).

[10] E. R. STRIETER and D. A. KORASICK, *ACS Chem Biol* **7**, 52 (2012).

[11] J. S. THROWER, L. HOFFMAN, M. RECHSTEINER, and C. M. PICKART, *EMBO J* **19**, 94 (2000).

[12] J. HEIDEKER and I. E. WERTZ, *Biochemical Journal* **467**, 191 (2015).

[13] T. DRESSELHAUS, N. D. WEIKART, H. D. MOOTZ, and M. P. WALLER, *Rsc Advances* **3**, 16122 (2013).

[14] V. H. TRANG, E. M. VALKEVICH, S. MINAMI, Y. C. CHEN, Y. GE, and E. R. STRIETER, *Angew Chem Int Ed Engl* **51**, 13085 (2012).

[15] E. M. VALKEVICH, R. G. GUENETTE, N. A. SANCHEZ, Y. C. CHEN, Y. GE, and E. R. STRIETER, *J Am Chem Soc* **134**, 6916 (2012).

[16] C. E. BLANCHET and D. I. SVERGUN, *Annual Review of Physical Chemistry, Vol 64* **64**, 37 (2013).

[17] R. P. RAMBO and J. A. TAINER, *Annual Review of Biophysics, Vol 42* **42**, 415 (2013).

[18] W. J. COOK, L. C. JEFFREY, M. CARSON, Z. J. CHEN, and C. M. PICKART, *Journal of Biological Chemistry* **267**, 16467 (1992).

[19] T. HIRANO, O. SERVE, M. YAGI-UTSUMI, E. TAKEMOTO, T. HIROMOTO, T. SATOH, T. MIZUSHIMA, and K. KATO, *The Journal of biological chemistry* **286**, 37496 (2011).

[20] S. LIU, Y. H. CHEN, J. LI, T. HUANG, S. TARASOV, A. KING, A. M. WEISSMAN, R. A. BYRD, and R. DAS, *Structure* **20**, 2138 (2012).

[21] Y. RYABOV and D. FUSHMAN, *Journal of the American Chemical Society* **129**, 7894 (2007).

[22] R. VARADAN, M. ASSFALG, and D. FUSHMAN, *Ubiquitin and Protein Degradation, Pt B* **399**, 177 (2005).

[23] N. ZHANG, Q. WANG, A. EHLINGER, L. RANDLES, J. W. LARY, Y. KANG, A. HARIRINIA, A. J. STORASKA, J. L. COLE, and D. FUSHMAN, *Molecular cell* **35**, 280 (2009).

[24] M. S. RITORTO, R. EWAN, A. B. PEREZ-OLIVA, A. KNEBEL, S. J. BUHRLAGE, M. WIGHTMAN, S. M. KELLY, N. T. WOOD, S. VIRDEE, N. S. GRAY, N. A. MORRICE, D. R. ALESSI, and M. TROST, *Nat Commun* **5**, 4763 (2014).

[25] E. M. COOPER, C. CUTCLIFFE, T. Z. KRISTIANSEN, A. PANDEY, C. M. PICKART, and R. E. COHEN, *Embo Journal* **28**, 621 (2009).

[26] Y. SATO, A. YOSHIKAWA, A. YAMAGATA, H. MIMURA, M. YAMASHITA, K. OOKATA, O. NUREKI, K. IWAI, M. KOMADA, and S. FUKAI, *Nature* **456**, 274 (2008).

[27] T. E. T. MEVISSEN, M. K. HOSPENTHAL, P. P. GEURINK, P. R. ELLIOTT, M. AKUTSU, N. ARNAUDO, R. EKKEBUS, Y. KULATHU, T. WAUER, F. EL OUALID, S. M. V. FREUND, H. OVAA, and D. KOMANDER, *Cell* **154**, 169 (2013).

[28] E. MATHEW, A. MIRZA, and N. MENHART, *J Synchrotron Radiat* **11**, 314 (2004).

[29] M. V. PETOUKHOV, D. FRANKE, A. V. SHKUMATOV, G. TRIA, A. G. KIKHNEY, M. GAJDA, C. GORBA, H. D. MERTENS, P. V. KONAREV, and D. I. SVERGUN, *J Appl Crystallogr* **45**, 342 (2012).

[30] M. EDDINS, R. VARADAN, D. FLUSHMAN, C. PICKART, and C. WOLBERGER, *Journal of Molecular Biology* **367**, 204 (2007).

[31] D. KOMANDER, F. REYES-TURCU, J. D. LICCHESI, P. ODENWAELDER, K. D. WILKINSON, and D. BARFORD, *EMBO reports* **10**, 466 (2009).

[32] Y. SATO, A. YOSHIKAWA, M. YAMASHITA, A. YAMAGATA, and S. FUKAI, *The EMBO journal* **28**, 3903 (2009).

[33] S. D. WEEKS, K. C. GRASTY, L. HERNANDEZ-CUEBAS, and P. J. LOLL, *Proteins: Structure, Function, and Bioinformatics* **77**, 753 (2009).

[34] D. A. CASE, T. E. CHEATHAM, T. DARDEN, H. GOHLKE, R. LUO, K. M. MERZ, A. ONUFRIEV, C. SIMMERLING, B. WANG, and R. J. WOODS, *J Comput Chem* **26**, 1668 (2005).

[35] D. CASE, T. DARDEN, T. CHEATHAM III, C. SIMMERLING, J. WANG, R. DUKE, R. LUO, R. WALKER, W. ZHANG, and K. MERZ, *There is no corresponding record for this reference* .

[36] D. CASE, T. DARDEN, T. CHEATHAM III, C. SIMMERLING, J. WANG, R. DUKE, R. LUO, R. WALKER, W. ZHANG, and K. MERZ, *University of California, San Francisco* **1**, 3 (2012).

[37] D. A. PEARLMAN, D. A. CASE, J. W. CALDWELL, W. S. ROSS, T. E. CHEATHAM III, S. DEBOLT, D. FERGUSON, G. SEIBEL, and P. KOLLMAN, *Computer Physics Communications* **91**, 1 (1995).

[38] R. SALOMON-FERRER, D. CASE, and R. WALKER, *Wiley Interdisciplinary Reviews-Computational Molecular Science* **3**, 198 (2013).

[39] J. GRAF, P. H. NGUYEN, G. STOCK, and H. SCHWALBE, *Journal of the American Chemical Society* **129**, 1179 (2007).

[40] V. HORNAK, R. ABEL, A. OKUR, B. STROCKBINE, A. ROITBERG, and C. SIMMERLING, *Proteins-Structure Function and Bioinformatics* **65**, 712 (2006).

[41] H. NGUYEN, D. R. ROE, and C. SIMMERLING, *Journal of Chemical Theory and Computation* **9**, 2020 (2013).

[42] L. WICKSTROM, A. OKUR, and C. SIMMERLING, *Biophysical Journal* **97**, 853 (2009).

[43] J.-P. RYCKAERT, G. CICCOTTI, and H. J. BERENDSEN, *Journal of Computational Physics* **23**, 327 (1977).

[44] M. FEIG, J. KARANICOLAS, and C. L. BROOKS, *J Mol Graph Model* **22**, 377 (2004).

[45] K. RAVIKUMAR, W. HUANG, and S. YANG, *Journal of Chemical Physics* **138** (2013).

[46] S. YANG, L. BLACHOWICZ, L. MAKOWSKI, and B. ROUX, *Proceedings of the National Academy of Sciences of the United States of America* **107**, 15757 (2010).

[47] S. YANG, M. PARISIEN, F. MAJOR, and B. ROUX, *Journal of Physical Chemistry B* **114**, 10039 (2010).

[48] M. D. DAILY, L. MAKOWSKI, G. N. PHILLIPS, and Q. CUI, *Chemical Physics* **396**, 84 (2012).

[49] M. A. JAMROS, L. C. OLIVEIRA, P. C. WHITFORD, J. N. ONUCHIC, J. A. ADAMS, D. K. BLUMENTHAL, and P. A. JENNINGS, *Journal of Biological Chemistry* **285**, 36121 (2010).

[50] B. ROZYCKI, Y. KIM, and G. HUMMER, *Structure* **19**, 109 (2011).

[51] M. PELIKAN, G. L. HURA, and M. HAMMEL, *Gen Physiol Biophys* **28**, 174 (2009).

[52] D. SCHNEIDMAN-DUHOVNY, M. HAMMEL, and A. SALI, *Nucleic Acids Res* **38**, W540 (2010).

[53] D. SCHNEIDMAN-DUHOVNY, M. HAMMEL, J. A. TAINER, and A. SALI, *Biophys J* **105**, 962 (2013).

[54] K. BERLIN, C. A. CASTANEDA, D. SCHNEIDMAN-DUHOVNY, A. SALI, A. NAVA-TUDELA, and D. FUSHMAN, *J Am Chem Soc* **135**, 16595 (2013).

[55] Y. WANG, C. TANG, E. WANG, and J. WANG, *PLoS Comput Biol* **10**, e1003691 (2014).

[56] Y. YE, G. BLASER, M. H. HORROCKS, M. J. RUEDAS-RAMA, S. IBRAHIM, A. A. ZHUKOV, A. ORTE, D. KLENERMAN, S. E. JACKSON, and D. KOMANDER, *Nature* **492**, 266 (2012).

[57] Y. C. JUANG, M. C. LANDRY, M. SANCHES, V. VITTAL, C. LEUNG, D. F. CECCARELLI, A. R. F. MATEO, J. N. PRUNEDA, D. MAO, R. K. SZILARD, S. ORLICKY, M. MUNRO, P. S. BRZOVIC, R. E. KLEVIT, F. SICHERI, and D. DUROCHER, *Molecular Cell* **46**, 549 (2012).

[58] M. Y. LAI, D. N. ZHANG, N. LARONDE-LEBLANC, and D. FUSHMAN, *Biochimica Et Biophysica Acta-Molecular Cell Research* **1823**, 2046 (2012).

[59] J. F. TREMPE, N. R. BROWN, M. E. M. NOBLE, and J. A. ENDICOTT, *Acta Crystallographica Section F-Structural Biology and Crystallization Communications* **66**, 994 (2010).

[60] A. D. J. VAN DIJK, D. FUSHMAN, and A. M. J. J. BONVIN, *Proteins-Structure Function and Bioinformatics* **60**, 367 (2005).

[61] E. B. DAMMER, C. H. NA, P. XU, N. T. SEYFRIED, D. M. DUONG, D. CHENG, M. GEARING, H. REES, J. J. LAH, A. I. LEVEY, J. RUSH, and J. PENG, *J Biol Chem* **286**, 10457 (2011).

[62] W. KIM, E. J. BENNETT, E. L. HUTTLIN, A. GUO, J. LI, A. POSSEMATO, M. E. SOWA, R. RAD, J. RUSH, M. J. COMB, J. W. HARPER, and S. P. GYGI, *Mol Cell* **44**, 325 (2011).

[63] S. A. WAGNER, P. BELI, B. T. WEINERT, M. L. NIELSEN, J. COX, M. MANN, and C. CHOUDHARY, *Mol Cell Proteomics* **10**, M111 013284 (2011).

[64] I. ZIV, Y. MATIUHIN, D. S. KIRKPATRICK, Z. ERPAPAZOGLOU, S. LEON, M. PANTAZOPOULOU, W. KIM, S. P. GYGI, R. HAGUENAUER-TSAPIS, N. REIS, M. H. GLICKMAN, and O. KLEIFELD, *Mol Cell Proteomics* **10**, M111 009753 (2011).

[65] D. KOMANDER, M. J. CLAGUE, and S. URBE, *Nat Rev Mol Cell Biol* **10**, 550 (2009).

[66] C. W. DAVIES, L. N. PAUL, M. I. KIM, and C. DAS, *J Mol Biol* **413**, 416 (2011).

[67] A. BREMM, S. M. FREUND, and D. KOMANDER, *Nat Struct Mol Biol* **17**, 939 (2010).

[68] M. J. EDELMANN, A. IPHÖFER, M. AKUTSU, M. ALTUN, K. DI GLERIA, H. B. KRAMER, E. FIEBIGER, S. DHE-PAGANON, and B. M. KESSLER, *Biochemical Journal* **418**, 379 (2009).

Chapter 5

Skip Residues Modulate the Structural Properties of the Myosin Rod and Guide Thick Filament Assembly [1]

PREFACE This chapter contains portion of my co-first-author manuscript published in PNAS with Keenan Taylor and Massimo Buvoli. My contributions to this work include designing and performing all of the molecular simulations and their analysis. K.C.T. designed and prepared myosin fusion proteins, performed the crystallographic studies and analyzed data. M.B. and A.B. designed, performed and analyzed the myosin assembly experiments in vivo, analyzed data. N.T.H prepared and crystallized proteins.

[1] Reprinted (adapted) with permission from "Skip residues modulate the structural properties of the myosin rod and guide thick filament assembly", K. C. Taylor‡ , M. Buvoli‡ , E. N. Korkmaz‡ , A. Buvoli, Y. Zheng, N. T. Heinze, Q. Cui, L. A. Leinwand and I. Rayment *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3806 (2015) ‡Contributed equally. Copyright 2015 National Academy of Sciences.

5.1 Introduction

Muscle contraction is primarily driven by the interactions between actin and myosin and the associated ATP hydrolysis, but the long-range transmission of force is based on the intrinsic ability of both proteins to self-assemble into organized filaments. The myosin thick filament is a well-characterized bipolar structure. The central area or bare zone, is $\sim$160 nm wide and is structurally defined by the packing of anti-parallel myosin molecules crosslinked at the sarcomeric M band by scaffold proteins [1, 2]. On either side of the bare zone, parallel arrays of staggered myosin molecules assemble into the characteristic A-band that is $\sim$1.6 nm in length and is centered between two Z-lines, where actin filaments are crosslinked at the Z-disk [3, 4].

The motor activity of myosin resides in the globular N-terminal region or subfragment 1 whereas the remainder of the molecule forms an extended dimeric α-helical coiled-coil. This rod-like section can be divided into two parts: subfragment-2 which allows the motors to extend away from the thick filament and light meromyosin (LMM) which promotes both parallel and anti-parallel myosin filament formation [5–7].

The sequence of the myosin rod shows the classical seven residue heptad repeat that is considered the hallmark of coiled-coil structures. However, it is also characterized by a remarkable dipolar charge profile, repeated every 28 amino acid residues, that is predicted to assist the staggered assembly of adjacent rods in the thick filament [5]. In sarcomeric myosins, the cyclic pattern of 38 dipolar charge repeats is interrupted by four widely spaced extra amino acids called skip residues (Fig.5.1A). These are by convention located at the end of different 28 amino acid repeats following position c of the heptad motif [5]. Insertion of a single residue (or deletion of six residues) introduces a discontinuity in the phasing of the heptad repeats that results in deformation of the α-helical coiled-coil. Such skip residues and stutters (deletions of three residues) are predicted to introduce regions of flexibility in the coiled-coil by causing local unwinding of the two α-helices; in contrast, stammers (deletions of four residues) are predicted to cause local over-winding of the supercoil [8]. Although the number and spacing of the skip residues are conserved across all sarcomeric myosins, both smooth muscle and non-muscle isoforms, which assemble differently, lack the second skip residue [9].

The role of myosin skip residues has not yet been defined. Early studies have associated their positions with the four rod bends observed by EM on purified molecules [9], and linear modeling of the charge

distribution of the rod has suggested that skip residues could play a role in properly staggering adjacent rod molecules [10]. Nevertheless, deletion of the two skip residues contained in the LMM does not alter myosin solubility or paracrystal formation in vitro [11].

We report here the first structural data, molecular dynamical properties and role in myosin assembly of the regions encompassing the four skip residues of a sarcomeric cardiac myosin. Our data reveal that the first three skip residues are structurally comparable and induce a unique local relaxation of the coiled-coil super-helical pitch. However, we find that the functional importance of each of the first three skip residues in promoting myosin assembly in vivo is different. Surprisingly, we discovered that the fourth skip residue lies within a highly flexible molecular hinge that is necessary for myosin incorporation in the bare zone of sarcomeres.

5.2 Methods

5.2.1 Molecular Dynamics

All simulations were carried out using the AMBER v12 Molecular Dynamics package with the f99SB force field improved with NMR observables (ff99SBnmr) [12–14]. The Generalized Born implicit solvent approach was adopted using specifically the gb7 model as this is $\sim$100 fold faster compared to explicit solvent simulations when graphical processor units are used [15–17]. Production gb7 simulations were carried out for 1000 ns using 1 fs as the time step. Langevin dynamics was applied with a collision frequency of 20 ps-1 at 300 K. The SHAKE [18] algorithm was applied to bonds involving hydrogen with a tolerance of 10-5. Nonbonded cutoff was set up as 9999 Å, and the maximum distance between atom pairs for Born radii calculations was set up to be 12 Å. Salt concentration was set to the physiological molarity of 0.15 M.

5.2.2 Explicit Solvent Control Simulations

The validity of implicit solvent simulations was tested through explicit solvent simulations (Fig.C4). In explicit solvent simulations, due to speed limitations, the simulations were carried out up to 160 ns. TIP3P water model was used along with ion parameters from Joung and Cheatham [19]. SHAKE was applied to bonds involving hydrogen atoms. Andersen thermostat was selected to maintain the temperature at 300 K [20]. Pressure was held constant at 1 bar with a relaxation time of 1 ps. The electrostatic interactions were treated with particle mesh Ewald with a grid size of about 1 Å, and Lennard-Jones interactions were treated with a

cutoff of 12 Å.

Simulated models have been summarized in Table 5.1 below.

5.3 Results

Three distinct lines of investigation were followed to gain insight into the role of the four skip residues in myosin filament formation. The three-dimensional structures of the regions surrounding the four myosin skip residues were determined by X-ray crystallography. Molecular dynamics simulations were carried out to reveal the effect of skip residues on the structure and the dynamics of the region. Finally, myosin assembly into sarcomeres in cardiac cells examined the function of the skip residues in a biological context.

5.3.1 The Coiled-coil Surrounding the First-Three Skip Residues is Locally Unwound

Sections of the human β-cardiac myosin (MYH7 gene; GenBank: M58018.1) ranging from 50 to 70 amino acids roughly centered on each of the four skip residues, were fused at their N-termini to globular folding domains, Xrcc4 and Gp7, that have been shown to enhance crystallization propensity and solubility of isolated sections of extended coiled-coils [21, 22]. Crystallization and structural determination of three of the structures, was obtained only after the addition of the helical bundle domain from microtubule binding protein Eb1 at the C-terminus of the myosin fragments to create double-sided fusions (Tables C1 and C2, and Fig.C1) [21]. In general, the stabilization domains might influence the folding, structure, or functional properties of coiled-coil fragments. However, Xrcc4 and Gp7 have been successfully employed to determine the structures of overlapping segments of tropomyosin and a component of the yeast spindle pole body [21–23], and GCN4, the stereotypical leucine zipper, has been used extensively for structural and functional studies of segments of coiled-coil proteins [24].

Comparison of the crystal structures reveals that the regions surrounding the first three-skip residues share a remarkable degree of structural similarity (Fig.5.1). The Cα root mean square deviations (RMSD) between Skip 1 and 2 and 3 are 0.92 Åand 1.1 Årespectively. The close similarity of the first three structures and the presence of multiple copies in the asymmetric unit for Skip 1, which show an RMSD difference of 0.51 Å, suggest that the structural features described below are a fundamental property of the myosin rod and not an artifact of crystal packing.

The polypeptide chain surrounding T1188 (Skip 1), E1385 (Skip 2) and E1582 (Skip 3) adopts a non-

Table 5.1: Simulated models and lengths of simulations.

Simulated Model	Length (ns)
Skip 1	1000
Skip 1 – ΔS1	950
Skip 2	1000
Skip 2 – ΔS2	1000
Skip 3*	230
Skip 3	1000
Skip 3 – ΔS3	1000
Skip 3 – Recoiled	1000
Skip 3 – Recoiled – ΔS3*	290
Skip 3 – Recoiled – ΔS3	1000
Xcrr4 – Skip 3 fused	700
Skip 4	1000
Skip 4 – ΔS4	1550

All simulations are carried out with implicit solvent model (gb7), unless they are labelled with *. In those cases the explicit solvent model with TIP3P was employed. Skip 3 - recoiling includes a series of mutations that change the distribution of hydrophobic amino acids towards a canonical coiled coil. The resulting sequence is LEHEEGKILRAQLEFNQIKAEIERLAAEVDEELEQAVRNHLRVVDSLQTSLD (The bold represent the mutated residues and the skip residue is underlined).

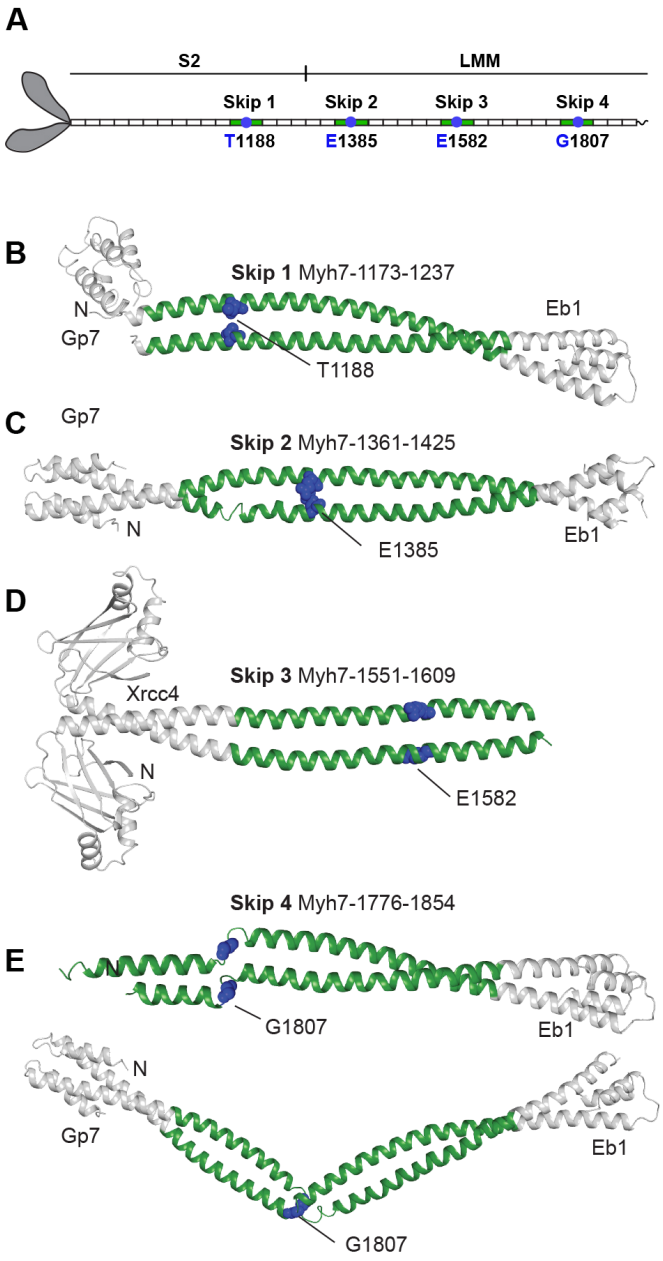

A
S2
LMM
Skip 1
Skip 2
Skip 3
Skip 4
T1188
E1385
E1582
G1807
B
Skip 1 Myh7-1173-1237
N
Gp7
Eb1
T1188
C
Gp7
Skip 2 Myh7-1361-1425
N
Eb1
E1385
D
Skip 3 Myh7-1551-1609
Xrcc4
N
E1582
Skip 4 Myh7-1776-1854
E
N
G1807
Eb1
N
Gp7
Eb1
G1807

Figure 5.1 *(previous page)*: Cartoon of their location in the myosin rod and structures of the four human β-cardiac myosin (MYH7) skip residues. (A) shows the relative location of the skip residues in the myosin rod which depicts the 38 dipolar charge repeats, each of which is formed by 28 amino acid residues [5]. Each fusion protein consists of an N-terminal globular element, either Gp7 or Xrcc4, (white) prior to a section of MYH7 (green). A C-terminal fusion, Eb1 (white), is also present in all constructs except for Skip 3: Xrcc4-L1551-N1609. Each skip residue is colored in blue and depicted in sphere representation. The N-terminus of each construct is indicated. (B) Gp7-K1173-I1238-Eb1 (Skip 1). (C) Gp7-L1361-I1425-Eb1 (Skip 2). (D) Xrcc4-L1551-N1609 (Skip 3). (E) Two crystallographically independent dimers within the asymmetric unit are shown for Gp7-A1777-T1854-Eb1 (Skip 4). Gp7 is disordered in the crystal lattice for the first of the two dimers shown in panel E.

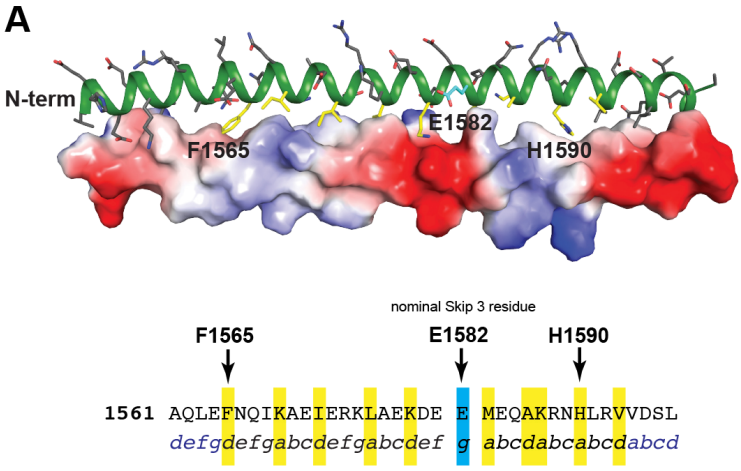
canonical coiled-coil characterized by local unwinding. These regions, which extend ∼17 and 11 residues on the N- and C-terminal sides of the skip residues, respectively, show a substantial increase in the super-helical pitch such that the helices run approximately parallel to each other [25, 26]. This effectively changes the regular a - g heptad designation of the residues surrounding each skip residue as previously proposed [5]. As a result, side chains that would be predicted to be solvent exposed by assignment to a regular heptad repeat are now buried along the interface. Furthermore, because the distortion extends over four heptads, it is difficult to attribute the changes in the coiled-coil to the insertion of one particular residue in this region. This unwinding was predicted from the original sequence analysis, though the structural similarity between the first three skip regions was not [5].

5.3.2 The Coiled-coil Surrounding Skips 1, 2, and 3 Accommodates the Break in the Heptad Repeat Similarly

Since the structures of Skips 1, 2, and 3 accommodate the insertion of the skip residue in a similar manner, the description of the structures will focus on Skip 3, which was determined to the highest resolution (Table C2).

Based on the position of residues around the conventional coiled-coil helical wheel, Skip 3 residue, E1582, is solvent exposed and lies in the g position which differs from the earlier prediction that it would lie between c and d (Fig.5.2A) [5, 27]. The approximate heptad designation of the 28 residue repeats either side of Skip 3 based on their helical disposition is shown in Fig.5.2A, where the residues that contribute to the hydrophobic core of the helical assembly are designated either a or d according to their observed position in the helical interface.

The transition from a canonical coiled-coil starts at ∼F1565 where the phenyl side chains on opposing polypeptide chains stack against each other in the hydrophobic interface which necessitates a distortion of the coiled-coil that places the α-carbon in a nominal d position rather than a (Fig.5.2A). Thereafter, the helices run approximately parallel to each other with the overall positions of the side chains following a distorted abcdefg repeat until the skip residue, E1582. The eleven residues on the C-terminal side of Skip 3 exhibit greater deviations from the canonical repeat showing an unconventional abcdabcabcd distribution; subsequently the residues in the coiled-coil return to a conventional repeat. Given the overall structural similarity between Skip 1, 2, and 3, it is not surprising that the distribution of hydrophobic residues that constitute the interface is also


A
N-term
F1565
E1582
H1590
nominal Skip 3 residue
F1565
E1582
H1590
1561 AQLEFNQIKAEIERKLAEKDE E MEQAKRNHLRVVDSL
defgdefgabcdefgabcdef g abcdabcabcdabcd

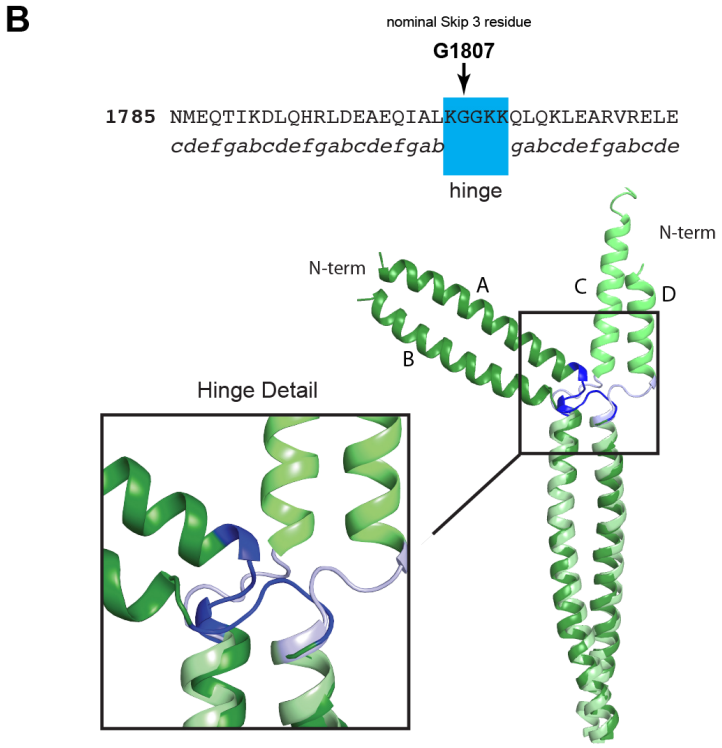

B
nominal Skip 3 residue
G1807
1785 NMEQTIKDLQHRLDEAEQIALKGGKKQLQKLEARVRELE
cdefgabcdefgabcdefgab gabcdefgabcde
hinge
N-term
N-term
A
B
C
D
Hinge Detail

Figure 5.2 *(previous page)*: Structural analysis of Skip 3 and Skip 4. (A) A cartoon representation of the coiled-coil surrounding E1582 (Skip 3) and surface electrostatic representation is shown on the lower-helix. Packing residues in the non-conical coiled-coil region of the upper-helix are colored in yellow on the upper-helix. The skip residue is colored in cyan. The protein sequence surrounding the nominal skip residue, E1582, with the observed coiled-coil position registry shown below. Residues that are in a standard packing arrangement are in blue, while the atypical region is colored black. Residues packing along the distorted interface are highlighted in yellow and the skip residue is in cyan. (B) The sequence and structure of Skip 4. The protein sequence surrounding the nominal Skip 4 residue, G1807, is shown with the coiled-coil position registry below. Residues involved in the hinge region are indicated. A C-terminal structural alignment of residues Q1811 – T1854 in chains A and B superimposed on C and D for the two independent molecules in the asymmetric unit for Skip 4 is also shown. This reveals the conformational variability in the Skip 4 hinge. The stabilization and folding domains were omitted from all structural figures.

conserved among the skip regions.

Interestingly, there are numerous solvent exposed hydrophobic residues surrounding the Skip 3 residue. These include L1559, L1563, I1568, and L1591. Their biological significance is unclear; however this pattern of hydrophobic exposed residues is highly conserved across species (Fig.C2). Likewise, as judged by sequence alignment, there is also remarkable similarity in the electrostatic surfaces of Skip 1, 2 and 3, (Figs.C2, C3). The existence of exposed hydrophobic side chains makes a priori assignment of residues to an accurate heptad pattern difficult.

5.3.3 The Fourth Skip Residue Forms a Highly Flexible Hinge

Remarkably, the fourth skip residue, G1807, is contained within a highly flexible hinge that disrupts the heptad repeat in a completely different manner from that observed for the first three skip residues (Fig. 1E). There are two crystallographically independent dimers present within the unit cell where both dimers show ordered electron density for the hinge. Importantly each dimer adopts a dramatically different conformation for its hinge (Fig. 1E). The key implication from these structures is that the hinge could adopt a set of conformations in vivo that are likely to be different from the ones seen in vitro because the latter reflect the necessities of crystallographic packing. Although the role of glycine residues as sites of flexibility had not been previously recognized, the hinge encompassing Skip 4 includes two of them, which probably contributes to its function. Glycine residues are uncommon in canonical coiled-coils and are often viewed as helix breakers, however the myosin rod contains twelve of them, in the LMM portion [28].

The hinge, extending from residues 1806 to 1810 (Fig. 5.2B), introduces a discontinuity in the coiled-coil, however the α-helices on either side of the break follow a uniform and normal super-helical pitch. The helical regions on either side of the discontinuity are highly similar in both crystallographically independent dimers where they align with a RMSD of 0.57 Å and 1.09 Å for the N- and C-terminal regions respectively (Fig. 2B). While the residues surrounding Skip 4 adopt a canonical coiled-coil there are some noteworthy features. First, the d and a positions of the heptad repeat preceding Skip 4 are occupied by alanine residues that reduce the hydrophobicity of the dimer interface. Second, the hydrophobic moieties (I1803 and L1805) in the heptad preceding the hinge are solvent exposed. Finally, two glutamine residues (Q1811 and Q1813) immediately C-terminal to the hinge are located such that they may hydrogen bond to exposed carbonyl oxygen atoms resulting from the break in the helicity. The unique nature of this hinge and its importance in bipolar

thick filaments is observed in both molecular dynamics simulations and in-vivo experiments as discussed below.

5.3.4 Dynamics of the Coiled-coil Regions Flanking the Skip Residues

Since the crystal structure provides only a static view of the myosin rod, the structural dynamics of the coiled-coil surrounding all four skip residues were assessed by molecular dynamics. To achieve simulations on the micros time scale, graphical processing units along with implicit solvent models were utilized (Fig. C4). Comparative analysis demonstrated that the conformational dynamics are not altered by the addition of the globular folding domains or the use of explicit solvent (Figs. C2 and C5). Consequently, all simulations were performed solely on sections of myosin coiled-coil containing the skip residues.

5.3.5 The Skip Residue is an Integral Component of the Coiled-coil

The crystal structures show that the designated skip residues are only one residue in substantial regions of distorted coiled-coil. Thus, the question arises whether the skip residues lead to coiled-coil instability or if the surrounding regions have adopted a stable state that accommodates their insertion. To address this question, simulations of the four regions encompassed by the crystallographic studies were performed with and without (ΔS constructs) the presence of each skip residue to probe the role of the skip residue within each structural environment (Table C1).

The simulations were analyzed in three ways. At a static level, the effect of removing the skip residues was assessed by examining the average distance between the center of mass of the two helices (DCOM) for the final 500 ns of each simulation, whereas the conformational variability was determined from both the root mean square fluctuations (RMSF) of individual residues and clustering analysis of the models. DCOM was determined as a moving window covering seven consecutive α-carbons to measure the degree of coiling: smaller values of DCOM imply tighter packing of the helices and a lowered super helical pitch, whereas a larger DCOM value indicates a non-close-packed state and an increased super helical pitch. Analysis of the simulations for Skip 3 residue shows that its removal allows the region to move as a whole towards a canonical coiled-coil, with a decreased super helical pitch (Fig. 3A). This effect is especially evident for the C-terminal side of the skip residue. Skips 1 and 2 follow the same pattern though to a lesser extent Fig. C6). Thus, the residues identified with Skips 1, 2, and 3 are associated with an increase in the helical pitch over a stretch of

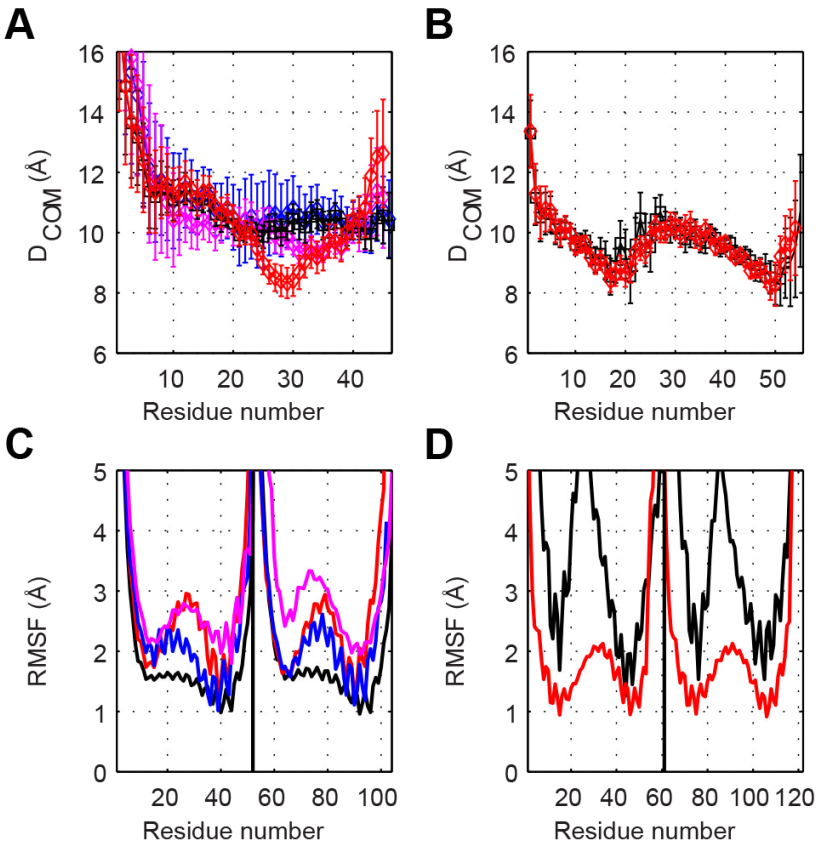

A
B
C
D
D_{COM} (Å)
RMSF (Å)
Residue number
16
14
12
10
8
6
5
4
3
2
1
0
10 20 30 40
10 20 30 40 50
20 40 60 80 100
20 40 60 80 100 120

Figure 5.3 *(previous page)*: Analysis of the molecular dynamics simulations for Skip 3 and Skip 4. DCOM and RMSF (Root Mean Square Fluctuation) for the regions surrounding Skip 3, panels A and C respectively, and Skip 4, panels B, and D respectively. The analyses of the wild-type simulations are shown in black, the skip residue deletion simulations (ΔS) are depicted in red. The simulation of the recoiled Skip 3 wild-type sequence (S3-R) in which the distribution of hydrophobic residues matches that expected for a canonical coiled coil, but still includes Skip 3 is shown in pink whereas the recoiled Skip 3 deletion (ΔS3-R) is depicted in blue. DCOM is the distance between two α-helices calculated from the center of masses of Cα atoms for seven consecutive amino acids and is inversely correlated with the degree of coiling in the simulations of models. Cα-RMSF values are an indication of the degree if flexibility, as well as stability. Regions with higher RMSF values have larger degree of flexibility. The measurements were averaged over the final 500 ns of each simulation to allow sufficient sampling for relaxation and ensure convergence of ensembles. Residue numbers correspond to L1551–D1602 and K1783–S1843 for Skip 3 and Skip 4 respectively. The sequence of the recoiled wild-type sequence is: (LEHEEGKILRAQLEFNQIKAEIERLAAEVDEELEQAVRNHLRVVDSLQTSLD) where Skip 3 is underlined and the mutated residues are shown in bold.

$\sim$29 residues. In contrast, deletion of Skip residue 4 has a different effect. The flanking sections of coiled-coil are not significantly disturbed by the absence of G1807, likely owing to the fact that in the WT structure this residue resides in a highly flexible loop flanked by canonical coiled-coils (Fig. 5.3B). The differences between Skip 4 and the other skip residues are also seen in the conformational dynamics.

5.3.6 Deletions of Skip 3 and Skip 4 Have Different Effects on the Conformational Dynamics and Structural Stability

Analysis of the Cα-based root mean square fluctuations in the simulations with wild-type (WT) and without the Skip residues (ΔS1, ΔS2, ΔS3 and ΔS4) shows that removal of Skip 1, 2, or 3 results in greater fluctuations than the WT (Fig. 5.3C, and Fig C6). This implies that deletion of these residues would lead to enhanced flexibility in these skip regions. Conversely, removal of Skip 4 leads to less variation in the positions of individual residues (Fig. 5.3D).

The clustering analysis is consistent with the local fluctuations. In the presence of Skip 3 there is one structural cluster that persists throughout the simulation while in its absence (ΔS3) there is a large increase in the number of well-populated structural clusters (Fig. 5.4A,B). The same trend is seen for Skip 1 and 2 (Fig. C6). In contrast, the WT structure for the Skip 4 region shows a very large distribution of structural states throughout the simulation, which are reduced dramatically on deletion of that residue (Fig. 5.4C,D). Importantly, both WT and ΔS4 simulations maintain a structural ensemble dominated by canonical coiled-coils but removal of the Skip 4 residue reduced the conformational diversity within the ensemble.

5.3.7 Computational Recoiling of Skip 3

Deletion of Skip 1, 2, or 3 is predicted to lead to greater conformational variability, which is consistent with the concept that these regions have evolved to accommodate the presence of an additional residue through local relaxation of the super-helical pitch. The simulations also predict that deletion (ΔS1, ΔS2 and ΔS3) moves the structure towards a more canonical coiled-coil as indicated by the smaller values of DCOM, though the resulting pattern of hydrophobic amino acid residues is inconsistent with a normal heptad repeat pattern. Thus, any potential deleterious effects caused by deletion of the skip residues could be attributed to either the absence of the residue itself or destabilization of the surrounding regions. To resolve this issue, the sequence surrounding Skip 3 was reconfigured to retain the same charge distribution but exhibit the normal pattern

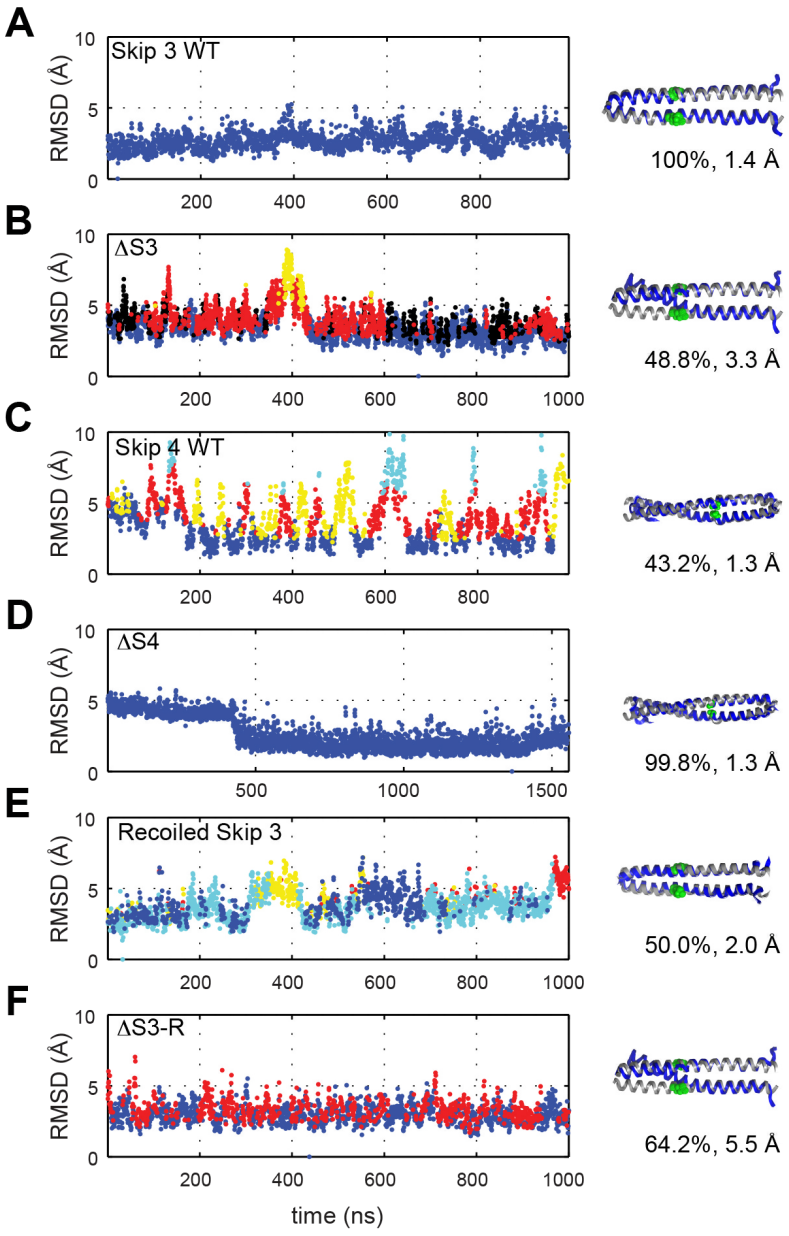

A
Skip 3 WT
RMSD (Å)
100%, 1.4 Å
B
ΔS3
48.8%, 3.3 Å
C
Skip 4 WT
43.2%, 1.3 Å
D
ΔS4
99.8%, 1.3 Å
E
Recoiled Skip 3
50.0%, 2.0 Å
F
ΔS3-R
64.2%, 5.5 Å
time (ns)

Figure 5.4 *(previous page)*: Diversity of ensembles formed by the regions surround Skip 3 and 4 in the presence and absence of the skip residues. (A) Skip 3 WT, (B) Skip 3 deletion ΔS3, (C) Skip 4 WT, (D) Skip 4 deletion ΔS4, (E) recoiled Skip 3 wild-type sequence, and (F) recoiled Skip 3 deletion (ΔS3-R). In this figure, Cα-RMSD with respect to the representative member of the most populated cluster was plotted against simulation time. The color-coding represents the different clusters formed where blue depicts the dominating clusters. The Cα-alignments of representative members onto the crystal structure and model structure, for Skip 3 and Skip 4 respectively, are shown on the right where the representative and initial structures are shown in blue and grey respectively. Skip residues are depicted in green. The percentage of each cluster and the Cα-RMSD to initial structure is given under each structure.

of hydrophobic residues found in a canonical coiled-coil interface (Fig. 5.3). This was simulated with and without the nominal Skip 3 residue where the recoiled wild-type sequence and recoiled Skip 3 deletion are designated S3-R and ΔS3-R respectively. These models show that introduction of a canonical coiled-coil pattern leads to greater conformational flexibility in the presence of Skip 3; but in the absence leads to less conformational variability in the clustering analysis and lower root mean square fluctuations (Fig. 5.3A, C and Fig. 5.4E,F).

5.3.8 The Role of the Skip Residues in the Assembly of Bipolar Thick Filaments

To test the biological role of the skip residues, rat α-cardiac myosin rod constructs (WT and deleted for individual skip residues) tagged with GFP at their amino termini were transfected into neonatal rat ventricular cardiac myocytes (NRVMs) and examined for their ability to be incorporated into thick filaments by live cell imaging [29]. When single cardiomyocytes were surveyed by confocal microscopy, we found that no skip residue deletion fundamentally prevented myosin incorporation into the endogenous sarcomeres (Fig. 5.5A). However, deletion of Skip 3 (ΔS3) caused myosin cytoplasmic aggregates that show a sponge-like structure in about 40 – 50% of the transfected cells (Fig. 5.5A; Fig. C7A). When the expression levels of the mutants were boosted, no negative effects on sarcomere structure were observed, but a higher number of large aggregates were detected in ΔS3 transfected cells (Fig. C7B). Cardiomyocytes were cotransfected with both mutant GFP- and mCherry-tagged WT myosin rods to determine if the skip residue deletions alter the distribution of myosin within the thick filament. As previously reported, GFP- and mCherry-tagged labeled WT myosin rods are uniformly incorporated and fully colocalized along the $\sim$ 1.6 microm long A band [29] (Fig. 5.5B, WT). Furthermore, the fluorescence signals are absent from the I band corresponding to the thin actin filaments, as well as from the H band corresponding to the central part of the sarcomere where myosin anti-parallel interactions take place in the bare zone. Lastly, the lack of fluorescence in the bare zone indicates that both the GFP and mCherry molecules are correctly positioned at the beginning of the parallel interactions; these occur on each side of the sarcomere at an estimated relative distance of about $\sim$1600 Å, that also corresponds to the theoretical myosin rod length [30]. Co-transfection of the mCherry-tagged WT and GFP-ΔS3 myosin rods (Fig. 5.5B; Fig. C7C) confirmed the phenotype of the ΔS3 construct but also revealed its dominant negative effects, as judged by the presence of WT mCherry-tagged myosin molecules in the cytoplasmic aggregates (Fig. 5.5B, ΔS3). However, this kind of analysis exposed an unexpected and

critical role of the Skip 4 residue in promoting anti-parallel assembly of myosin. Surprisingly, the emission signals of the ΔS4 GFP- and WT mCherry-tagged myosin molecules do not visibly overlap close to the center of the bipolar myosin filaments. Indeed, the red mCherry fluorescence is predominantly detectable in this region (Fig. 5.5B, ΔS4). In contrast, the WT and ΔS4 tagged molecules are homogenously distributed along the rest of the thick filaments where myosin rods pack in a parallel manner. To better characterize the ΔS4 phenotype, Linescan analysis, which measures the intensity value of each fluorophore over the length of the sarcomere, was then performed. The plot derived from an evaluation of 425 single sarcomeres imaged from three independent transfections revealed that the ΔS4 green fluorescence signal has both an offset of the peaks toward the Z lines, as well as a substantially deeper valley in proximity of the bare zone (Fig. 5.5C). Thus lack of GFP signal strongly suggests that the ΔS4 mutant is incapable of forming the anti-parallel interactions occurring in the core of the bare zone where full overlap of the rods appear to be demanded [30]. To confirm the ΔS4 phenotype in different experimental conditions, cardiomyocytes were then stained with an antibody recognizing only the endogenous full length myosin molecules or imaged in time-course experiments. The same peculiar distribution and spatial relationship with the WT myosin molecules was also observed in fixed cells (Fig. 5.5D, Ab-59) as well as shortly after transfection, when the transcription rate is high (Fig. 5.5D, 36 h, 48 h).

5.3.9 Skip Residues are Coupled with Proximal Non-canonical Residues

Mutations predicted to convert or recoil the skip regions into more stable structures were then introduced to determine whether unwinding of the regions surrounding Skip 1, 2, or 3 observed in the WT structures is functionally important. As seen with the skip residue deletions, the recoiled myosins were also incorporated into the endogenous sarcomeres (Fig. C7D). The number of cells showing cytoplasmic aggregates did not increase when transfected with the Skip 3 deletion recoiling construct (ΔS3-R) ($\sim$40 – 50%), but higher magnification inspection showed that many of the thick filaments containing the ΔS3-R protein have a less defined structure (ΔS3-R HMV). Moreover, higher expression levels caused, as previously observed with the ΔS3 construct, a greater number of large mutant aggregates (Fig. C7D ΔS3-R). Myosin aggregates were also observed in a small percentage of cells ($\sim$2 – 4%) transfected with the Skip 1 recoiled construct (Fig. C7D ΔS1-R). Interestingly, recoiling/mutagenesis converting the Skip 2 region into the corresponding sequence of non-muscle/smooth muscle myosins, which lack Skip 2 residue (Fig. C4, did not cause any noticeable myosin

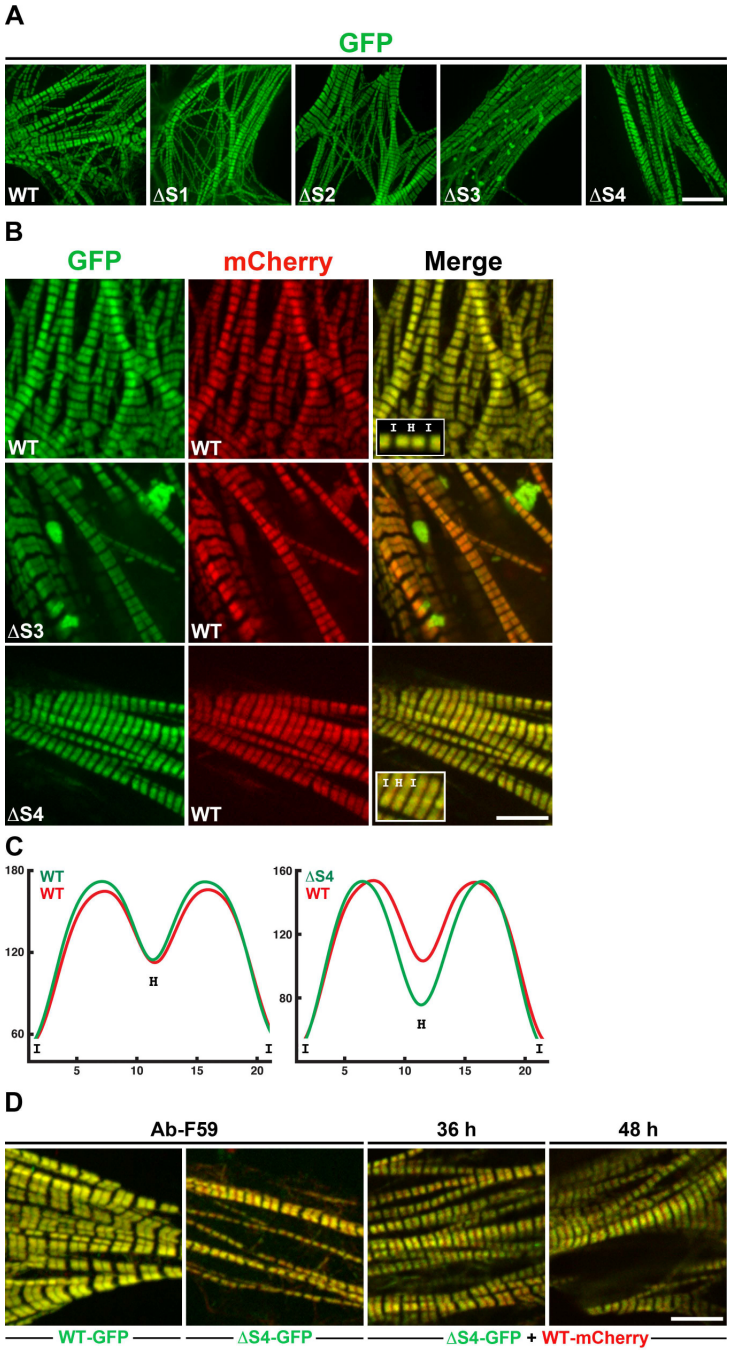

A
GFP
WT
ΔS1
ΔS2
ΔS3
ΔS4
B
GFP
mCherry
Merge
WT
WT
I H I
ΔS3
WT
ΔS4
WT
I H I
C
180
120
60
WT
WT
H
I
I
5
10
15
20
160
120
80
ΔS4
WT
H
I
I
5
10
15
20
D
Ab-F59
36 h
48 h
WT-GFP
ΔS4-GFP
ΔS4-GFP + WT-mCherry

Figure 5.5 *(previous page)*: (A) Cardiomyocytes electroporated with WT and mutant GFP-tagged skip residue deletion constructs (ΔS) were imaged by confocal microscopy 96 h later (Bar: 10 μm). (B) Cardiomyocytes were co-transfected with mutant GFP- and WT mCherry-tagged constructs as indicated and were imaged by confocal microscopy 96 h later. The two boxes in the WT and ΔS4 merge panels show the high magnification view of the sarcomeric I band and H zone (bare zone) and the lack of colocalization between the mutant GFP- and the WT mCherry-tagged myosins (Bar: 5 μm). (C) Linescan analysis showing the relative intensity across the sarcomere of WT GFP and mCherry (left) and ΔS4 GFP– and WT mCherry– tagged myosins (right). Cells from 3 independent transfections were imaged and a total of 420 sarcomeres for each graph analyzed. (D) Colocalization of ΔS4 GFP construct with the endogenous myosin, and time course incorporation into the sarcomeres. (Ab-F59): cardiomyocytes were transfected with WT or ΔS4 GFP-tagged myosin constructs; 96 h later cells were fixed and stained with F59 anti-myosin primary antibody that recognizes only the myosin head domain, and the Alexa Fluor 568 secondary antibody with orange-red emission color. All panels are overlays of GFP and mCherry fluorescence signals. (36 h; 48 h): cardiomyocytes co-tranfected with ΔS4 GFP- and WT mCherry-tagged myosin rod constructs were imaged by confocal microscopy 36 and 48 h later (Bar: 5 μm).

phenotype (Fig. C7D ΔS2-R). Cotransfection with mCherry tagged WT and ΔS-R constructs validated the importance of Skip 3 for proper myosin assembly (Fig. C8, ΔS3-R) but failed to detect any phenotype for either Skip 1 or 2 recoiling mutants (Fig. C8, ΔS1-R, ΔS2-R). The lack of aberrant formation of parallel myosin arrangement and the more obvious exclusion from the bare zone for ΔS4-R, as observed for ΔS4, confirmed the specialized role of the Skip 4 residue/hinge region in bipolar myosin assembly (Fig. C8,, ΔS4-R).

Finally, to address whether the distortion or flexibility is important for the Skip 3 region the skip residue was reintroduced into the recoiled mutant (S3-R) and its effect examined in vivo. In this case the skip residue introduces instability into an otherwise stable coiled-coil as demonstrated by molecular dynamics simulation (Figs. 3A,C and 4E,F). Cells expressing this mutant show the same cytoplasmic aggregates observed with ΔS3-R mutant (Fig. C7, S3-R). This implies that a structurally stable but unwound structure surrounding Skip 3 is functionally important.

5.3.10 Inter-changeability of the Skip Regions

The finding that deletion of Skip 3 and 4 causes myosin aggregates or exclusion from the bare zone raises the question whether their biological function is defined by the location of the distortion in the coiled coil or the specific sequence at that location. To address this question, 14 residues on either side of Skip 3 and Skip 4 were substituted with the corresponding region of Skip 2 that we have shown does not itself play an important role in myosin assembly, and shares only structural similarity, but limited sequence identity with the Skip 3 region (Fig. 5.6A).

Examination of GFP replacement mutants transfected into NRVMs alone, revealed no anomalous distribution of the tagged myosins in the cytoplasm, or defects in the incorporation into the sarcomere (Fig. 5.6B, S2-S3 Repl, S2-S4 Repl). However, cotransfection with WT mCherry myosin shows an assembly defect associated only with the Skip 4 replacement which is exactly the opposite of what was observed when Skip 4 was deleted: imaging and linescan analysis revealed accumulation but not exclusion of the S2-S4 construct in the bare zone that becomes completely filled by the GFP fluorescence signal (Fig. 5.6C, S2-S4 Repl; 5.6D).

5.4 Discussions

Considerable effort has been devoted at the ultrastructural level to determine the organization of the myosin rods in the sarcomeric thick filament [3, 30]. Likewise, there is a large database of sequences that

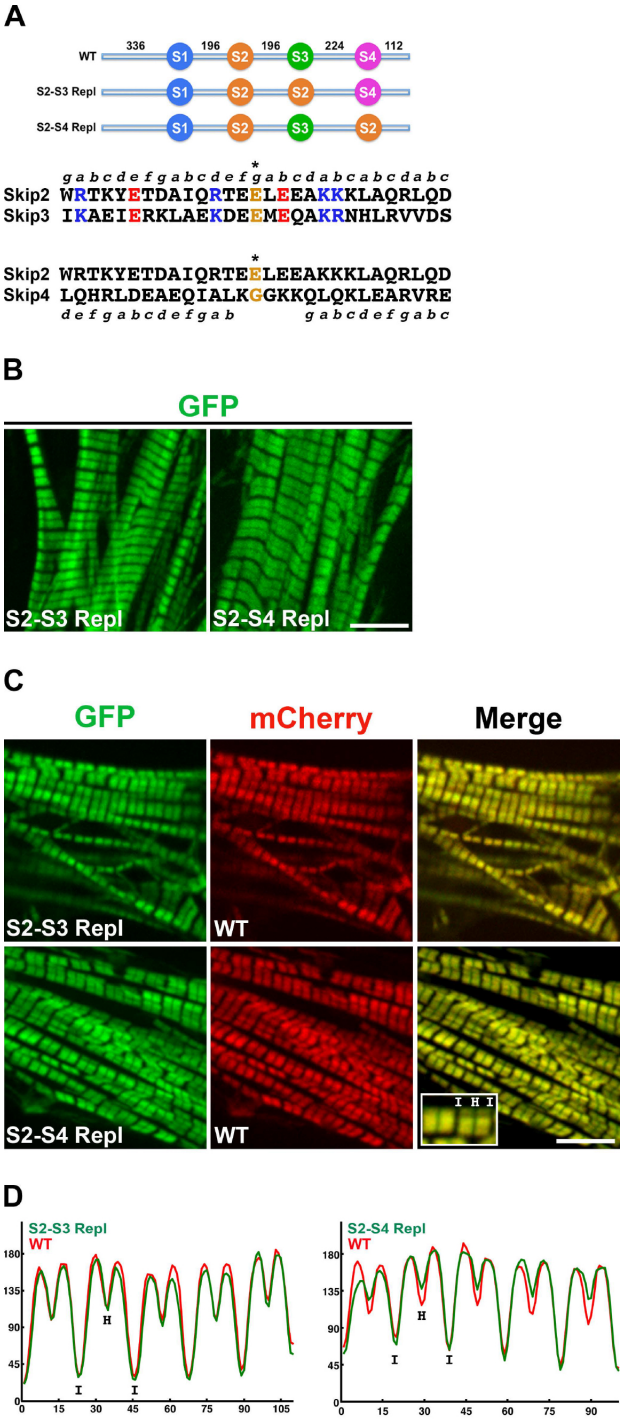

A
WT
S2-S3 Repl
S2-S4 Repl
336
196
196
224
112
S1
S2
S3
S4
g a b c d e f g a b c d e f g a b c d a b c a b c d a b c
Skip2 WRTKYETDAIQRTEELEEAKKKLAQRLQD
Skip3 IKAEIERKLAEKDEEMEQAKRNHLRVVDS
Skip2 WRTKYETDAIQRTEELEEAKKKLAQRLQD
Skip4 LQHRLDEAEQIALKGGKKQLQKLEARVRE
d e f g a b c d e f g a b g a b c d e f g a b c
B
GFP
S2-S3 Repl
S2-S4 Repl
C
GFP
mCherry
Merge
S2-S3 Repl
WT
S2-S4 Repl
WT
I H I
D
S2-S3 Repl
WT
S2-S4 Repl
WT
180
135
90
45
0
H
I
I
0 15 30 45 60 75 90 105
0 15 30 45 60 75 90

Figure 5.6 *(previous page)*: Functional activity of mutants carrying duplications of 28 amino acids encompassing the Skip 2 residue. (A, top) Topology of the duplications showing the number of amino acids separating the skip residues from each other or from the beginning/end of the myosin rod. (A, bottom) alignment of 28 amino acids surrounding the Skip 2 residue with the corresponding Skip 3 and 4 regions (replaced by the duplications). The observed coiled-coil position registry are shown above and below the sequences together with the conserved charge distribution for Skip 2 and 3 (B) Cardiomyocytes were transfected with GFP-tagged myosin constructs as indicated. (C) Cardiomyocytes were co-transfected with mutant GFP- and WT mCherry-tagged myosin constructs as indicated. The box in the S2-S4 Repl merge images shows the high magnification view of the of the sarcomeric I band and H zone. Cells in both B and C panels were imaged by confocal microscopy 96 h after transfection. Bars, 5 μm. (D) Linescan analysis showing the relative intensity across five sarcomeres of S2-S3 Repl GFP- and WT mCherry-tagged myosins (left) and S2-S4 Repl GFP- and WT mCherry-tagged myosins (right). x-axis: pixel distance (0.086 μm/pixel); y-axis: fluorescence intensity. The location of the I band and H zone are reported.

indicate that the myosin rod is highly conserved [31]. What is lacking is an understanding of the connection between sequence and macromolecular assembly. The structural and functional data presented in this study shed new light on the function of the four skip residues that interrupt the cyclic pattern of the myosin rod. It has been previously argued that by introducing conformational instability in the coiled-coil, skip residues allow myosin rods to wrap along the cylindrical thick filament [9, 32]. Our structural studies show that the first three skip residues cause the coiled-coil to unwind for a region of 29 amino acid residues over which the α-helices run approximately parallel to each other. As such the designation of a single residue as a "skip" residue is very difficult because the distortion extends over $\sim$ 4 heptad repeats. Rather, the discussion of the role of the skip residues in the assembly of the myosin rod must focus on the structural properties of the entire region more than a single residue. The structures determined here suggest that the regions surrounding the first three skip residues exhibit a stable but distorted conformational state. Moreover, molecular dynamics simulations show that the skip residues maintain the relaxation of the super-helical pitch and their removal increases the flexibility of the surrounding regions. Although not assessed by our calculations, the structures are consistent with enhanced long range flexibility introduced by unwinding of the coiled-coil.

The myosin distortions introduced by the first three skip residues are both structurally and functionally different from those observed in tropomyosin and reflect the different roles that these molecules play in the sarcomere. The structure of the first three skip regions has evolved to ensure regional and adequate flexibility of the rod required for assembly of the long myosin coiled-coil into the thick filament. On the contrary, tropomyosin conformational flexibility has been achieved by the inclusion of alanine residues in the a and d positions at seven locations [33, 34]. This molecular arrangement, which results in looser packing of the α-helices, but maintains the overall coiled-coil structure, allows the molecule to wrap continuously around the actin filament, and to adopt different local positions depending upon its regulatory state [35, 36]. Distortions from a canonical coiled-coil are common and are often implicated in mediating protein-protein interactions as observed in intermediate filaments [37]. While several structural examples of stammer/stutters structures have been reported, skip residues have been observed mainly in antiparallel single chain coiled coils but there are very few structures for parallel dimeric coiled-coils that include skip residues [38]. Thus, the comparison between the first three and the last myosin skip residues reveals how the insertion of a single residue can be differently accommodated within a long parallel coiled coil to diversify structural function, and highlights the

importance of experimental characterization of regions that do not show a canonical repeat.

It is accepted that the myosin rod arose from gene duplication of an ancestral rod unit [39]. Our data show that during myosin evolution the same unwound structure has been conserved, even though the amino acid sequence has diverged considerably over the passage of time. The question arises whether the first three skip regions, which are by themselves highly conserved across the myosin II superfamily, play an equal role in thick filament assembly. The cellular data presented here show that the first three skip residues are not functionally equivalent in promoting myosin assembly and their hierarchical importance appears to be set by their position along the rod. The modification of the region surrounding Skip 3 has a greater affect on assembly than modification of either Skip 1 or 2. Only deletion of the Skip 3 residue causes both aggregation in the cytoplasm and reduced incorporation of the mutant protein into the sarcomere. Indeed, any modifications of Skip 3 that change the pitch or introduce flexibility lead to the same cell phenotype. This speaks in favor of the importance of a defined structure. Myosin assembly is not affected by the replacement of the Skip 3 region with the corresponding Skip 2 region that introduces 23 amino acid changes out of 28, but maintains the same structure, and the overall charge and distribution of surface exposed hydrophobic residues (Fig. C3) [10].

Interestingly, recoiling of Skip 1, which is located within the hinge between subfragment-2 and LMM and presumably allows the myosin heads to commute between the thick and thin filaments, shows only a mild phenotype (Fig. C7, Panel D). Although NRVMs contract spontaneously, our assay only measures myosin incorporation into the thick filaments and not the effects on sarcomere contractility. Thus, a role for the Skip 1 region in controlling specific steps that take place during actomyosin sliding cannot be ruled out. The lack of phenotype observed by converting the Skip 2 region into non-muscle/smooth myosin isoforms, which lack the Skip 2 residue, is also puzzling. Clearly, the formation of side-polar versus bipolar filaments is driven by more than the absence or presence of the Skip 2 region and implies that this region could play a sarcomeric role not shared with non muscle myosins [9].

Our data clearly show structural and functional differences between the first three skip residues and the last one. The Skip 4 residue is located in a highly flexible loop encompassing 3-4 amino acids separating conventional coiled-coil regions that constitute a true hinge. Thus, defining of one of these residues as the Skip 4 residue is somewhat arbitrary. Removal of the skip residue from the hinge does not prevent incorporation in the parallel arms of the sarcomere but leads to exclusion of the mutant myosin from the bare zone. This

implies that the C-terminal portion of the myosin rod requires a high degree of movement to interact with other components of the sarcomere or to assemble in a different fashion in the bare zone. This is the first evidence that myosin requires conformational flexibility to form antiparallel assemblies that constitute the bare zone.

Interestingly, the myosin-titin interaction site has been mapped to a 17-residue sequence located 8 amino acids downstream of the Skip 4 residue. Since the two proteins run approximately parallel to each other along the sarcomere, it has been proposed that the bending of the rod induced by the skip 4 residue could allow the two molecules to interact [40]. Thus, one explanation for the phenotype observed with the Skip 4 mutant constructs, might be that the reduction in flexibility introduced by the mutations in the rod could affect the interaction between the two proteins. However, mutagenesis of the predicted titin binding site, which contains a unique rod arrangement of negative residues (RELENELE), does not compromise myosin assembly nor alter its ability to be correctly incorporated into the sarcomere (Fig. C9. However, reduced hinge flexibility (as probably occurs in the S2-S4 mutant) could affect accurate myosin antiparallel assembly by causing loss of interaction with other M-resident proteins [7].

In contrast to parallel assembly, which shows a characteristic polar repeat consistent with a 98-residue stagger that underlies the 143 Å distance between crowns in the thick filament, there is no obvious sequence signature that would indicate how the myosin rods align antiparallel to each other in the bare zone [5, 41, 42]. This implies that the assembly cannot be described by a simple binary association of only two antiparallel coiled-coils. The bare zone exhibits D32 dihedral symmetry which demands a regular assembly of myosin molecules [43] and image reconstruction shows a thickening in the bare zone consistent with the overlap of myosin rods. Therefore a complex arrangement is expected because both parallel and antiparallel interactions take place in the bare zone. Based on packing considerations each myosin rod is expected to interact with at least five other molecules in this densely packed region [32]. Indeed, modeling of the thick filament predicts that the C-terminal portion of the rod assembles differently in the bare zone than in the arms of the sarcomere [30, 44]. The observation that the hinge at Skip 4 is necessary for incorporation into the bare zone is suggestive of a discontinuous interaction between myosin rods or interaction with multiple molecules. Altogether these findings provide evidence that only the Skip 3 and 4 residue regions are essential for parallel and antiparallel myosin interactions. While the similar structure of the first three skip residues has been functionally diversified by their location along the rod, the more complex structure of the bare zone has required the evolution of a

specialized hinge structure.

More than 130 mutations causing hypertrophic and dilated cardiomyopathy as well as different distal myopathies have been mapped to the β-myosin rod; of interest, some of them are located near the four skip residues. In particular three of them, E1573K R1588P and L1591P which lie in the solvent exposed c, b, and b positions of the heptad repeat respectively, are located within the skip 3 region resolved in this study. The first of these mutations has been implicated in Ebstein anomaly, which is a rare congenital heart malformation affecting the tricuspid valve as well as the right ventricle whereas the others result in Laing distal myopathy, characterized by slowly progressive skeletal muscle weakness with variable degrees of cardiac impairment [45–47]. To determine the potential impact of these mutations on the structure of the skip 3 region, molecular dynamics were then carried out (Fig. C10). While the E1573K mutation has little effect, which implies that this charge reversal influences packing interactions between adjacent myosin rods, the proline substitutions have a much greater effect. In fact, they cause greater local conformational variability, and separation of the α-helices. Moreover, they introduce a bend into the coiled-coil and a change in pitch. Thus the proline substitutions analyzed are predicted to have both a local and long-range effects on the conformation of myosin, which are expected to alter the rod interactions within the thick filament.

In addition to providing the first structural and functional characterization of four coiled-coil regions of a sarcomeric myosin rod, the multidisciplinary approach adopted here lays the groundwork for defining the functional impact of disease causing mutations on the coiled-coil structure of the myosin rod.

5.5 References

[1] R. CRAIG, *J Mol Biol* **109**, 69 (1977).

[2] M. GAUTEL, *Curr Opin Cell Biol* **23**, 39 (2011).

[3] H. A. AL-KHAYAT, R. W. KENSLER, J. M. SQUIRE, S. B. MARSTON, and E. P. MORRIS, *Proc Natl Acad Sci U S A* **110**, 318 (2013).

[4] R. CRAIG and J. L. WOODHEAD, *Curr Opin Struct Biol* **16**, 204 (2006).

[5] A. D. MCLACHLAN and J. KARN, *Nature* **299**, 226 (1982).

[6] R. L. SOHN, K. L. VIKSTROM, M. STRAUSS, C. COHEN, A. G. SZENT-GYORGYI, and L. A. LEINWAND, *J Mol Biol* **266**, 317 (1997).

[7] R. C. THOMPSON, M. BUVOLI, A. BUVOLI, and L. A. LEINWAND, *FEBS Lett* **586**, 3008 (2012).

[8] D. A. PARRY, R. D. FRASER, and J. M. SQUIRE, *J Struct Biol* **163**, 258 (2008).

[9] G. OFFER, *J Mol Biol* **216**, 213 (1990).

[10] R. STRAUSSMAN, J. M. SQUIRE, A. BEN-YA'ACOV, and S. RAVID, *J Mol Biol* **353**, 613 (2005).

[11] S. J. ATKINSON and M. STEWART, *J Cell Sci* **99 (Pt 4)**, 823 (1991).

[12] K. A. BEAUCHAMP, Y. S. LIN, R. DAS, and V. S. PANDE, *J Chem Theory Comput* **8**, 1409 (2012).

[13] D. CASE, T. DARDEN, I. T.E. CHEATHAM, C. SIMMERLING, J. WANG, R. DUKE, R. LUO, R. WALKER, W. ZHANG, K. MERZ, B. ROBERTS, S. HAYIK, A. ROITBERG, G. SEABRA, J. SWAILS, A. GOETZ, I. KOLOSSVRY, K. WONG, F. PAESANI, J. VANICEK, R. WOLF, X. W. J. LIU, S. BROZELL, T. STEINBRECHER, H. GOHLKE, Q. CAI, X. YE, J. WANG, M.-J. HSIEH, G. CUI, D. ROE, D. MATHEWS, M. SEETIN, R. SALOMON-FERRER, C. SAGUI, V. BABIN, T. LUCHKO, S. GUSAROV, A. KOVALENKO, and P. KOLLMAN, 2012.

[14] D. W. LI and R. BRUSCHWEILER, *Angew Chem Int Ed Engl* **49**, 6778 (2010).

[15] A. W. GOTZ, M. J. WILLIAMSON, D. XU, D. POOLE, S. LE GRAND, and R. C. WALKER, *J Chem Theory Comput* **8**, 1542 (2012).

[16] M. S. LEE, F. R. SALSBURY, and C. L. BROOKS, *Journal of Chemical Physics* **116**, 10606 (2002).

[17] J. MONGAN, C. SIMMERLING, J. A. MCCAMMON, D. A. CASE, and A. ONUFRIEV, *Journal of Chemical Theory and Computation* **3**, 156 (2007).

[18] J. P. RYCKAERT, G. CICCOTTI, and H. J. C. BERENDSEN, *Journal of Computational Physics* **23**, 327 (1977).

[19] I. S. JOUNG and R. CHEATHAM, T. E., *J Phys Chem B* **112**, 9020 (2008).

[20] T. A. ANDREA, W. C. SWOPE, and H. C. ANDERSEN, *Journal of Chemical Physics* **79**, 4576 (1983).

[21] J. FRYE, V. A. KLENCHIN, and I. RAYMENT, *Biochemistry* **49**, 4908 (2010).

[22] V. A. KLENCHIN, J. J. FRYE, M. H. JONES, M. WINEY, and I. RAYMENT, *The Journal of biological chemistry* **286**, 18240 (2011).

[23] Y. NITANAI, S. MINAKATA, K. MAEDA, N. ODA, and Y. MAEDA, *Adv Exp Med Biol* **592**, 137 (2007).

[24] S. DEISS, B. HERNANDEZ ALVAREZ, K. BAR, C. P. EWERS, M. COLES, R. ALBRECHT, and M. D. HARTMANN, *J Struct Biol* **186**, 380 (2014).

[25] F. H. C. CRICK, *Acta Cryst.* **6**, 685 (1953).

[26] G. GRIGORYAN and W. F. DEGRADO, *Journal of Molecular Biology* **405**, 1079 (2011).

[27] E. K. O'SHEA, J. D. KLEMM, P. S. KIM, and T. ALBER, *Science* **254**, 539 (1991).

[28] M. M. GROMIHA and D. A. PARRY, *Biophys Chem* **111**, 95 (2004).

[29] M. BUVOLI, A. BUVOLI, and L. A. LEINWAND, *J Mol Biol* **415**, 807 (2012).

[30] H. A. AL-KHAYAT, R. W. KENSLER, E. P. MORRIS, and J. M. SQUIRE, *J Mol Biol* **403**, 763 (2010).

[31] G. E. WHITE and H. P. ERICKSON, *J Struct Biol* **154**, 111 (2006).

[32] M. W. CHEW and J. M. SQUIRE, *J Struct Biol* **115**, 233 (1995).

[33] J. H. BROWN, *Protein Sci* **19**, 1366 (2010).

[34] J. H. BROWN, K. H. KIM, G. JUN, N. J. GREENFIELD, R. DOMINGUEZ, N. VOLKMANN, S. E. HITCHCOCK-DEGREGORI, and C. COHEN, *Proc Natl Acad Sci U S A* **98**, 8496 (2001).

[35] X. E. LI, K. C. HOLMES, W. LEHMAN, H. JUNG, and S. FISCHER, *J Mol Biol* **395**, 327 (2010).

[36] W. J. ZHENG, B. BARUA, and S. E. HITCHCOCK-DEGREGORI, *Biophysical Journal* **105**, 1882 (2013).

[37] A. A. CHERNYATINA, D. GUZENKO, and S. V. STRELKOV, *Curr Opin Cell Biol* **32**, 65 (2015).

[38] S. V. STRELKOV and P. BURKHARD, *J Struct Biol* **137**, 54 (2002).

[39] A. D. MCLACHLAN, *J Mol Biol* **169**, 15 (1983).

[40] A. HOUMEIDA, J. HOLT, L. TSKHOVREBOVA, and J. TRINICK, *J Cell Biol* **131**, 1471 (1995).

[41] A. D. MCLACHLAN and J. KARN, *J Mol Biol* **164**, 605 (1983).

[42] D. RICKETSON, C. A. JOHNSTON, and K. E. PREHODA, *Proceedings of the National Academy of Sciences of the United States of America* **107**, 20964 (2010).

[43] P. K. LUTHER, P. M. MUNRO, and J. M. SQUIRE, *J Mol Biol* **151**, 703 (1981).

[44] J. M. SQUIRE, *J Mol Biol* **77**, 291 (1973).

[45] T. CULLUP, P. J. LAMONT, S. CIRAK, M. S. DAMIAN, W. WALLEFELD, R. GOODING, S. V. TAN, J. SHEEHAN, F. MUNTONI, S. ABBS, C. A. SEWRY, V. DUBOWITZ, N. G. LAING, and H. JUNGBLUTH, *Neuromuscul Disord* **22**, 1096 (2012).

[46] A. V. POSTMA, K. VAN ENGELEN, J. VAN DE MEERAKKER, T. RAHMAN, S. PROBST, M. J. BAARS, U. BAUER, T. PICKARDT, S. R. SPERLING, F. BERGER, A. F. MOORMAN, B. J. MULDER, L. THIERFELDER, B. KEAVNEY, J. GOODSHIP, and S. KLAASSEN, *Circ Cardiovasc Genet* **4**, 43 (2011).

[47] G. TASCA, E. RICCI, S. PENTTILA, M. MONFORTE, V. GIGLIO, P. OTTAVIANI, G. CAMASTRA, G. SILVESTRI, and B. UDD, *Neuromuscul Disord* **22**, 640 (2012).

Chapter 6

A Composite Approach Towards a Complete Model of the Myosin Rod[1]

PREFACE This chapter contains my co-first-author manuscript published in Proteins: Structure, Function, and Bioinformatics with Keenan Taylor and Micheal Andreas. My contributions to this work include designing and performing all of the molecular simulations and their analysis. K.C.T. and M.P.A. designed and prepared myosin fusion proteins, performed the crystallographic studies and analyzed data. M.B. and A.B. designed, performed and analyzed the myosin assembly experiments in vivo, analyzed data. N.T.H prepared and crystallized proteins.

[1]Reprinted (adapted) with permission from "A Composite Approach Towards a Complete Model of the Myosin Rod", E. N. Korkmaz‡ , K. C. Taylor‡ , M. P. Andreas‡ , G. Ajay, N. T. Heinze, Q. Cui and I. Rayment *Proteins* **184**, 1097-0134 (2016) ‡Contributed equally.

6.1 Introduction

Cardiac and skeletal muscles contain sets of interdigitated thick and thin filaments. Muscle contraction occurs when the thick filaments, which are built primarily from myosin II, actively slide past the actin-containing thin filaments. Myosin plays both a structural and enzymatic role in this fundamental biological process. The globular N-terminal domains (heads) hydrolyze ATP and interact with actin to generate force, whereas the C-terminal region forms a long α-helix that dimerizes to form a coiled-coil. The coiled-coil is known as the myosin rod. The extended region of myosin can be split in two sections: Subfragment-2 (S2) and light meromyosin (LMM), where these are connected by a flexible section of the coiled-coil that functions as a hinge. In cardiac β-myosin LMM extends from approximately amino acid residues 1140 to 1935. [1] LMM self-assembles to form the backbone of the thick filament at physiological ionic strength. At this time, there is considerable knowledge of the structure and function of the globular motor regions of myosin, but many aspects of the molecular organization of the myosin rods in the thick filament are still unresolved. This has important implications since a significant number of mutations that lead to skeletal and cardiac myopathies are located in the myosin rod. [2–4]

The reason that the structure of the myosin thick filament is not well understood is due to the lack of high resolution structural information for LMM and the complexity of the assembly. Although the myosin rod is predicted to consist of a coiled-coil, which is in principle a simple motif, there are distinct locations where the sequence deviates from a canonical structure. [5] Indeed, there are four conserved locations that contain an additional "skip" residue that disrupts the coiled-coil heptad repeat. These skip residues have been proposed to introduce flexibility into the myosin rod to allow assembly into the thick filament. [6, 7] Surrounding these skip residues, the sequence exhibits a 28 amino acid repeat and contains an alternating pattern of positively and negatively charged amino acids, a feature that has been implicated in the staggered interaction between adjacent myosin molecules in the thick filament. [6] In addition, the coiled-coil propensity in the proposed hinge region between S2 and LMM is considerably lower than the sequence signature for a stable coiled-coil. This region is adjacent to the first skip residue and has been proposed to provide an area of flexibility that allows the motor domains swing away from the myosin thick filament to interact with the actin thin filaments. [6, 8]

Previous studies on tropomyosin and intermediate filaments have shown that the local pitch of a coiled-coil is profoundly influenced by the amino acid sequence, especially in those regions that deviate from a canonical heptad repeat. [9, 10] As such, it is not possible to accurately predict the structure of the myosin rod with the accuracy needed to build a realistic model for the rod and to understand the interactions between adjacent myosin molecules in the thick filament. Clearly there is a need for high resolution structural data.

The sheer size of the myosin rod at approximately 1600 Å in linear dimension, prohibits a direct crystallographic study. Thus, the only feasible approach is to divide the rod into smaller sections that might be more amenable to structural study. Indeed, this is the standard approach for investigating large intractable proteins, but it is more complicated for fragments of the myosin because the rod exhibits a strong salt-dependent aggregation. A direct result of this property is that fragments of myosin readily form paracrystals at low ionic strength. [11–15] This is due in part to periodic clusters of positively and negatively charged residues that repeat every 28 amino acid residues for the entire length of the myosin rod noted above. [5] In sarcomeric myosins the rod contains 38 of these repeats. [6] The latter property has doubtlessly contributed to the difficulties of crystallizing short sections of the myosin rod that might otherwise appear amenable to structural study. Indeed, until recently there were no structures available for the myosin rod except for the N-terminal 126 amino acid residues of human cardiac β-myosin II S2, which exhibits for most of its length a canonical coiled-coil and has no tendency to aggregate at low ionic strength. [16]

Apart from the problems of self-assembly into paracrystals, a second issue is whether fragments of myosin contain all of the information necessary to fold correctly and yield a soluble protein. The folding of coiled-coils is generally highly cooperative where initiation of folding is often controlled by trigger sequences that have a higher coiled-coil propensity than the bulk of the protein. These sequence signatures usually lie towards the N-terminus of the coiled-coil. [17–22] Removal of these motifs frequently prevents proper folding or loss of dimerization of the resultant fragment. Indeed, a recent paper by Wolny et al., shows that there appears to be a folding signal located in the myosin tail between 1301 and 1330 that leads to high helicity in even short peptides (7 heptads) that contain this sequence. Conversely, five other down-stream peptides that contained only seven heptads of the myosin rod showed limited α-helical content. [23] This problem of stabilizing coiled coils that lack their native trigger sequence has been circumvented for some coiled-coils by incorporation of a short stable coiled-coil, such as that seen in the well-studied transcription

factor GCN4. [19, 24, 25] This strategy could be applied to sections of LMM but is not predicted to eliminate self-assembly into paracrystals at low ionic strength.

The problem of myosin self-assembly has now been solved by using globular fusion proteins in place of GCN4. [26] This strategy was developed for investigating the overlap complex for tropomyosin and components of the yeast spindle pole body and has been recently applied to myosin. [26–28] In those structural studies, the globular domains from human DNA ligase binding protein Xrcc4 and the bacteriophage ϕ-29 scaffolding protein Gp7 were fused to the N-terminus of the targeted coiled-coil domains. [29–31] The resultant fusion proteins were expressed with good solubility in E. coli and crystallized readily. Following a similar strategy, the structure of the regions surrounding the four skip residues in human cardiac β-myosin II have been reported which now opens up the possibility of determining the structure of the entire myosin rod. [26] The use of globular domains also solves the problem of the limited stability of short sections of the myosin rod. [23]

In principle, a model for the entire LMM (∼800 amino acid residues) could be assembled from the structures of 15-20 overlapping fragments that each span 50-70 amino acid residues. The initial stages of this project are reported here and establishes proof of concept for the approach. This study reports a composite model for residues L1526-R1689 of human cardiac β-myosin generated from five overlapping structures, covering ∼24 % of LMM. This study also describes the degree of stabilization that can be obtained from a suite of folding domains and their effect on protein expression and crystallization. Molecular dynamic (MD) simulations of myosin coiled-coil with and without a folding domain provided a means to assess changes in dynamic properties brought about by the fusion. Additionally the fusion structures presented here have a large degree of sequence overlap that provides a basis for evaluating the effects of the fusion protein and crystallographic contacts on the targeted coiled-coil. Lastly, surface analysis shows that the composite model exhibits a substantial hydrophobic surface area which plays a role in the higher order assembly of bipolar thick filaments.

6.2 Materials and Methods

All cloning was performed as described previously. [28] Briefly, QuikChange cloning was used to amplify and insert sections of the MYH7 gene into vectors containing the desired solubilization domain. DNA encoding human cardiac myosin rod was purchased as an image clone from Open Biosystems. The use

of QuikChange cloning allowed all constructs to be made without introduction of cloning artifacts, which is critical for maintaining the correct coiled-coil registration between the folding domains and the target gene fragment. The sequences of all constructs were verified over their entire open reading frames.

6.2.1 Construct Design

All fusion protein constructs were cloned into a modified pET24D (EMD) plasmid or pKLD37, a modified pET31b plasmid (EMD) by QuikChange cloning. [32–35] A His-tag and an rTEV cleavage recognition site were introduced N-terminally to constructs that contained Gp7, Xrcc4, and GCN4. [36] The Paircoil algorithm was used to predict regions of coiled-coil and their helical registration. [37] A complete description of all constructs is given in (Table 6.1).

6.2.2 Protein Expression and Purification

Myosin fusion proteins were expressed in an E. coli BL21-CodonPlus (DE3)-RIL cell line (Stratagene). Cells were grown in lysogeny broth (LB) medium under an appropriate antibiotic selection at 37 °C with shaking to an A600 of ∼1.0, then cooled on ice for 15 min; at this point, 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) was added and the cells were grown for an additional 16 h at 16 °C before harvesting by centrifugation. Cells were washed with 50 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) pH 7.6, 50 mM NaCl, 1 mM ethylenediaminetetraacetic acid (EDTA) at a ratio of 1L of buffer per 6ThL of culture prior to flash freezing in liquid nitrogen and storage at -80 °C.

All protein purification steps were carried out at 4 °C. 10 g of cell were lysed in 100 mL of lysis buffer (50 mM HEPES pH 7.6, 50 mM NaCl, 20 mM imidazole, 0.5 mg/mL lysozyme) with 1 mM PMSF, 50 nM Leupeptin (Peptide International), 70 nM E-65 (Peptide International), 2 nM Aprotinin (ProSpec), and 2 M AEBSF (Gold BioTechnology) by sonication. Lysate was clarified by centrifugation at 125,000 g for 30 min at 4 °C in a Ti-45 rotor. The supernatant was loaded onto a 5 mL Ni-NTA (nickel-nitrilotriacetic acid) column (QIAGEN) by gravity and washed with 25 column volumes of buffer A (25 mM HEPES pH 7.6, 300 mM NaCl, 20 mM imidazole). The column was then washed with an additional 5 column volumes buffer A with 40 mM imidazole. Protein was eluted in four column volumes of buffer A with 200 mM imidazole. 1 mM β-mercaptoethanol was added to all buffers for purification of constructs containing cysteine residues. The His-tag was removed by incubation with a 1:40 molar ratio of rTEV protease to myosin fusion protein at

Table 6.1: Fusion Constructs[a]

MyH7-1361–1406
MSYYHHHHHHDYDIPTSENLYFQGLSKANSEVAQWRTKYETDAIQRTEELEEAKKKLA
QRLQEAEEAVEA

GCN4-MyH7-1361–1406
MGSSHHHHHHHHDYDIPTSENLYFQGASMSVKELEDKVEELLSKNYHLENEVARLKKL
LSKANSEVAQWRTKYETDAIQRTEELEEAKKKLAQRLQEAEEAVEAVN

Gp7-MyH7-1361–1406
MSYYHHHHHHDYDIPTSENLYFQGGSGPLKPEEHEDILNKLLDPELAQSERTEALQQLR
VNYGSFVSEYNDLTKSLSKANSEVAQWRTKYETDAIQRTEELEEAKKKLAQRLQEAEE
AVEAVNA

Xrcc4-MyH7-1361–1406
MSYYHHHHHHDYDIPTSENLYFQGGSGERKISRIHLVSEPSITHFLQVSWEKTLESGFVI
TLTDGHSAWTGTVSESEISQEADDMAMEKGKYVGELRKALLSGAGPADVYTFNFSKE
SCYFFFEKNLKDVSFRLGSFNLEKVENPAEVIRELICYCLDTTAENQAKNEHLQKENER
LQRVLSKANSEVAQWRTKYETDAIQRTEELEEAKKKLAQRLQEAEEAVEAVNA

Gp7-MyH7-MyH7-1526–1571
MSHHHHHHHHDYDIPTSENLYFQGGSGPLKPEEHEDILNKLLDPELAQSERTEALQQLR
VNYGSFVSEYNDLTKSHEKLEKVRKQLEAEKMELQSALEEAEASLEHEEGKILRAQLEF
NQIKAE

Xrcc4-MyH7-1562–1622
MSYYHHHHHHDYDIPTSENLYFQGGSGERKISRIHLVSEPSITHFLQVSWEKTLESGFVIT
LTDGHSAWTGTVSESEISQEADDMAMEKGKYVGELRKALLSGAGPADVYTFNFSKESC
YFFFEKNLKDVSFRLGSFNLEKVENPAEVIRELICYCLDTTAENQAKNEHLQLEFNQIKA
EIERKLAEKDEEMEQAKRNHLRVVDSLQTSLDAETRSRNEALRVKKKME**GDL**[b]

Xrcc4-MyH7-1590–1657
MSYYHHHHHHDYDIPTSENLYFQGGSGERKISRIHLVSEPSITHFLQVSWEKTLESGFVIT
LTDGHSAWTGTVSESEISQEADDMAMEKGKYVGELRKALLSGAGPADVYTFNFSKESC
YFFFEKNLKDVSFRLGSFNLEKVENPAEVIRELICYCLDTTAENQAKNEHHLRVVDSLQT
SLDAETRSRNEALRVKKKMEGDLNEMEIQLSHANRMAAEAQKQVKSLQSLLKDTQIQL

Xrcc4-MyH7-1631–1692
MSYYHHHHHHDYDIPTSENLYFQGGSGERKISRIHLVSEPSITHFLQVSWEKTLESGFVIT
LTDGHSAWTGTVSESEISQEADDMAMEKGKYVGELRKALLSGAGPADVYTFNFSKESC
YFFFEKNLKDVSFRLGSFNLEKVENPAEVIRELICYCLDTTAENQAKNEHANRMAAEAQ
KQVKSLQSLLKDTQIQLDDAVRANDDLKENIAIVERRNNLLQAELEELR**AVV**[b]

[a] The myosin segment is underlined.

[b] The residues highlighted in bold were not observed in the crystal lattice though were present in the protein.

4 °C in 25 mM HEPES pH 7.6, 100 mM NaCl, 0.1 mM EDTA, 0.5 mM tris(2-carboxyethyl)phosphine. The NaCl concentration was increased to 300 mM and the cleaved protein was then loaded onto a 2 mL Ni-NTA column equilibrated in buffer A without imidazole. Myosin constructs were eluted in four column volumes of buffer A, and rTEV protease was eluted with buffer A containing 200 mM imidazole. Fusion proteins were concentrated in an Amicon Ultra-15 30 KDa cutoff (Millipore) to between 10 and 25 mg/mL prior to overnight dialysis into storage buffer (10 mM HEPES pH 7.6, 100 mM NaCl). The protein was then flash-frozen in 30 μL droplets in liquid nitrogen and stored at -80 °C.

6.2.3 Crystallization

Crystals of Gp7-1526-1571 were grown at 4 °C by vapor diffusion from a 1:1 mixture of 15 mg/mL protein solution and a polyethylene glycol (PEG) solution consisting of 16 % (w/v) PEG 8000, 400 mM malonate pH 7.2, and 100 mM triethanolamine pH 7.5. Rod-shaped crystals grew over the course of several days to final dimensions of 400 μm x 50 μm x 50 μm. All manipulations of crystals prior to freezing were carried out at 4 °C. Crystals were cryo-protected by first being transferred to synthetic mother liquor solution consisting of 18 % (w/v) PEG 8000, 400 mM malonate pH 7.2, 100 mM triethanolamine pH 7.5, and 100 mM NaCl followed by stepwise transfer to a final solution of 18 % (w/v) PEG 8000, 400 mM malonate pH 7.2, 100 mM triethanolamine pH 7.5, 100 mM NaCl, and 10 % (w/v) ethylene glycol. Crystals were flash-frozen by plunging them into liquid nitrogen.

Crystals of reductively methylated Xrcc4-1562-1622 were grown at 4 °C by vapor diffusion from a 1:1 mixture of 13 mg/ mL protein solution and polyethylene glycol (PEG) solution consisting of 23 % (w/v) PEG 4000, 500 mM NaCl, 100 mM triethanolamine pH 8.0. Shard-shaped crystals grew over the course of 7 days to an average dimension of 250 μm x 100 μm x 75 μm. All manipulations of the crystals were performed at 4 °C. Crystals were cryo-protected by first being transferred to synthetic mother liquor solution consisting of 23 % (w/v) PEG 4000, 500 mM NaCl, 100 mM triethanolamine pH 8.0 followed by stepwise transfer to a final solution of 23 % (w/v) PEG 4000, 500 mM NaCl, 100 mM triethanolamine pH 8.0, 12.5 % (w/v) ethylene glycol, 250 mM CaCl2. Crystals were flash frozen by being rapidly plunged into liquid nitrogen. [38]

Crystals of reductively methylated [38] Xrcc4-1590-1657 were grown at 4 °C by vapor diffusion from a 1:1 mixture of 15 mg/ mL protein solution and methyl ether polyethylene glycol (MEPEG) solution consisting of 16 % (w/v) MEPEG 2000, 250 mM KNO3, 100 mM MOPS pH 7.0. Hexagonal shaped crystals formed

over the course of 7 days to an average dimension of 600 μm x 600 μm x 200 μm. Crystals were cryo-protected by transferring to an initial synthetic motherliqour solution consisting of 16 % (w/v) MEPEG 2000, 200 mM KNO3, 100 mM MOPS pH 7.0 followed by a stepwise transfer into a final solution consisting of 20 % (w/v) MEPEG 2000, 200 mM KNO3 and 12 % (w/v) ethylene glycol. Crystals were flash frozen by being rapidly plunged into liquid nitrogen.

Crystals of Xrcc-1631-1692 were grown at 4 °C by vapor diffusion from a 1:1 mixture of 14 mg/ mL protein solution and MEPEG solution consisting of 14 % (w/v) MEPEG 5000, 200 mM glycine, 100 mM bistrispropane pH 7.0. Crystals were cryo-protected by transferring to a cryoprotection solution of 20 % (w/v) MEPEG 5000, 100 mM NaCl, 200 mM glycine, 100 mM bis-tris propane pH 7.0, followed by a stepwise transfer into a final solution consisting of 20 % (w/v) MEPEG 5000, 200 mM glycine, 100 mM bis-tris propane pH 7.0, 12.5 % (w/v) ethylene glycol, 200 mM CaCl2. Crystals were then flash frozen by plunging into liquid nitrogen.

6.2.4 Data Collection and Structure Determination

X-ray diffraction data were collected at beam line SBC 19-ID (Advanced Photon Source). The datasets were integrated and scaled using HKL3000. [39, 40] X-ray data collection statistics are given in Table 6.2. The structures of Gp7-1526-1571, Xrcc4-1562-1622, Xrcc4-1590-1657, and Xrcc4-1631-1692 were solved by molecular replacement with Phaser [41, 42] using either residues 2 – 52 of Gp7 (1NO4) or residues 1 – 142 of Xrcc4 (1IK9) as search models. [30, 31] Following density modification by Parrot, initial models of Xrcc4-1562-1622, Xrcc4-1590-1657, and Xrcc4-1631-1692 were built in Buccaneer. [43, 44] Subsequent iterative cycles between manual model building in Coot followed by restrained refinement in Refmac 5.6 were used to generate the penultimate structural coordinates. [45, 46] Final refinements for Xrcc4-1590-1657 were performed by TLS and restrained refinement in Refmac 5.6. 42 Two chains were present in the asymmetric unit, and each chain was assigned as an individual TLS group for the entire chain. Final refinements on Xrcc4-1562-1622 and Xrcc4-1631-1692 structures were performed using Phenix Refine. [47] For Xrcc4-1562-1622, four chains were present in the asymmetric unit, with chain A having 3 TLS groups, chain B having 2 TLS groups, chain C having 3 TLS groups, and chain D having 2 TLS groups. Xrcc4-1631-1692 utilized 2 TLS groups for each chain at the junction between the fusion protein and myosin, totaling 4 TLS groups. [46] An initial model of Gp7-1526-1571 was constructed with Phenix AutoBuild and was refined with alternating

Table 6.2: Crystallographic data collection and refinement statistics

Data Collection	Gp7-1526-1571	Gp7-1526-1571	Xrcc4-1590-1657 Xrcc4-1631-1692
Space group	P1	P21	C2 C2
Cell dimensions			
a, b, c (Å)	56.1, 64.5, 70.5	83.1, 57.3, 112.1	223.6, 84.0, 39.4 200, 44.4, 74.7
α, β, γ (°)	70.5, 77.5, 74.2	β	β=92.8 β =108.2
Wavelength (Å)	0.9791	0.9791	0.97924 0.9791
Resolution (Å)[a])	50-2.1 (2.14-2.10)	50-3.1 (3.15 -3.1)	50-2.1 (2.14-2.10) 50-2.3 (2.31-2.27)
$R^{a,b}_{merge}$	0.065 (0.43)	0.070 (0.37)	0.090 (0.37) 0.068 (0.33)
$< I > / < \sigma I >^a$	23.1 (2.67)	27.3 (4.9)	34.1 (2.7) 40.2 (3.45)
Completeness (%)[a]	97.6 (96.3)	99.1 (99.4)	99.4 (98.1) 99.9 (99.6)
Redundancy[a]	3.8 (2.3)	6.0 (5.9)	4.8 (3.1) 6.3(4.9)
Beamline	19-ID	19-ID	19-ID 19-ID
Refinement			
Resolution (Å)[a]	50-2.1 (2.13-2.09)	50-3.1(3.18-3.10)	111.7-2.3 (2.36-2.3) 50-2.3 (2.36-2.30)
No. reflections[c]	67072 (3282)	18959 (948)	30554 (1645) 28089 (1404)
R_{work}/R_{free}	0.21/0.26	0.24/0.28	0.21/0.26 0.22/0.24
No. atoms			
Protein	6360	6341	3392 3191
Water	164	2	54 108
Ligand	0	0	0 0
Average β-factors (Å^2)	43	82	59.1 63.4
R.m.s. deviations			
Bond lengths (Å)	0.008	0.004	0.012 0.004
Bond angles (°)	0.935	0.741	1.414 0.78
Ramachandran (%)			
Most favored	99.7	98.5	98.8 98.7
Allowed	0.3	1.5	1.2 1.3
Disallowed	0	0	0 0
TLS Groups	16	10	2 4
PDB Accession Number	5CJ1	5CJ4	5CHX 5CJ0

[a] Data in parentheses represent the highest resolution shell

[b] $R_{merge} = \sum |I_{hkl} - I_{hkl}| / \sum |I_{hkl}|$

[c] Data in parentheses represent the number of reflections used for the calculation of R_{free}.

[d] $R_{factor} = \sum |F_{obs} - F_{calc}| / \sum |F_{obs}|$

where R_{work} refers to the R_{factor} for the data used in the refinement and R_{free} refers to the R_{factor} for 5% of the data that were excluded from the refinement.

Table 6.3: Simulated Models and Lengths of Simulations.

Model	Trajectory Length (ns)
MyH7-1551-1602[a]	1000
Gp7-MyH7-1551-1602	1000
Xrcc4-MyH7-1551-1603	1400
MyH7-1526-1689	1100
MyH7-1526-1689	520
MyH7-1526-1689	540

[a] Residues 1551–1602 or 1551–1603 were used in the simulation since these constituted the structurally ordered residues in the crystal structure for Xrcc4-1551–1609 (accession number 4XA4).

rounds of manual model building in Coot followed by restrained refinement in Phenix. [47] In the refinement two TLS groups per chain were defined by the junction between the fusion protein and myosin. [48] The eight chains in the asymmetric unit were divided into a total of 16 TLS groups. Refinement statistics are presented in Table 6.2.

6.2.5 Reconstruction of Myosin from Structural Fragments

The initial structure for the simulations was created in the MODELLER Homology Modeling Package [49–53] starting from the aligned structures for Gp7-1526-1571, Xrcc4-1562-1622, Xrcc4-1590-1657, Xrcc4-1631-1692 and Xrcc4-1551-1609. After alignment, overlapping segments of the structures were trimmed and the orientation of the side chains were energy minimized and the conformation of the trimmed chain ends were optimized to obtain a complete segment consisting of 164 amino acids (MyH7-1526-1689) of the myosin rod surrounding the Skip 3 residue.

All molecular dynamics simulations were carried out using the AMBER v14 MD program. [54–57] The Generalized Born (GB), an implicit solvent approach was chosen because it provides $\sim$100 fold efficiency compared to the explicit solvent when graphical processing units are used, [58–60] and long simulations are required to allow adequate structural relaxations of models constructed based on crystal structures. Specifically, the ff99SB force field improved with NMR observables (ff99SBnmr) was chosen [61–66] along with the gb7 model. [60] Production gb7 simulations were carried out for a minimum of 500 ns using a 1 fs time step (2 fs was used for the independent runs). Details of the simulations are summarized in Table 6.3. Langevin dynamics was followed with a collision frequency of 20 ps-1 at 300 K. The SHAKE algorithm was applied to bonds with hydrogen atoms with a tolerance of 10-5 Å. [67] The non-bonded cutoff was set as 9999 Å and the maximum distance between atom pairs (rgbmax) for Born radii calculations was kept at 12 Å. Salt was treated implicitly via Debye-Huckel theory and the concentration was set to the physiological concentration of 0.15 M. The collected trajectories were analyzed with AmberTools v15 and the Multiscale Modeling Tools for Structural Biology (MMTSB) package. [55, 57, 58]

To investigate the effect of the head groups, the isolated Skip 3 region, residues 1551-1603 from the MyH7-1551-1609 structure, was simulated along with fusions with Gp7 and Xrcc4 (Table 6.3).

In order to examine the convergence of the simulation of the contiguous model, three independent simulations were carried out for a minimum of 520 ns each. The analysis was performed over the combined

trajectory of the three simulations where all averages were calculated for the entire pool of conformations. To determine if the simulations had converged the distribution of the clusters among the three individual trajectories was plotted after performing clustering using the k-clust algorithm in the MMTSB package [58] (Fig. D1. It was observed that each cluster was sampled in every simulation. Additionally, the averaged DCOM and super helical pitch for the individual simulations Figures D2 and D3 were in qualitative and quantitative agreement with the averages for the aggregate simulations.

6.2.6 CD and fluorescence Spectroscopy

Circular dichroism spectra were recorded from 190 to 300 nm in an AVIV model 420 CD spectrophotometer. Protein samples were diluted to an A280 of 0.35 in buffer containing 50 mM sodium phosphate pH 7.0, 0.5 mM tris(2-carboxyethyl)phosphine. Thermal melting curves were recorded by monitoring the change in tryptophan emission intensity in a QuantaMaster Model C-60/2000 spectrofluorometer. Fluorescence emissions were recorded from 300 to 500 nm. The excitation wavelength was 290 nm. The temperature was increased by 4 degree increments starting at 4 °C. The sample was equilibrated at each temperature point for 10 minutes before measurement.

6.3 Results and Discussions

6.3.1 Fusion Proteins do not Significantly Influence the Backbone of the Target Coiled-coil

Four crystal structures of human cardiac β-myosin LMM are reported here that in combination with the previously determined structure for Skip 3 (Xrcc4-1551-1609 containing ∼8 heptads of myosin) [26] encompass residues 1526-1688 (Fig. 6.1). In total this section extends over repeat 25 to the final heptad of repeat 30. These new fragments are defined as Gp7-1526-1571, Xrcc4-1562-1622, Xrcc4-1590-1657, and Xrcc4-1631-1692 that contain ∼7, 9, 10, and 8 heptads of coiled-coil respectively. The four new structures overlap with Xrcc4-1551-1609 on the N- or C-terminal side for a total coverage of LMM of 164 residues (∼23 heptads). These structures are N-terminal fusions with either Gp7 or Xrcc4 and were designed such that the overlapping regions of LMM coiled-coil could be used to assemble a larger composite structure. The junctions between the folding domain and myosin were designed to preserve the coiled-coil registry and minimize disruption of inter-chain salt bridges as has been previously described. [27]

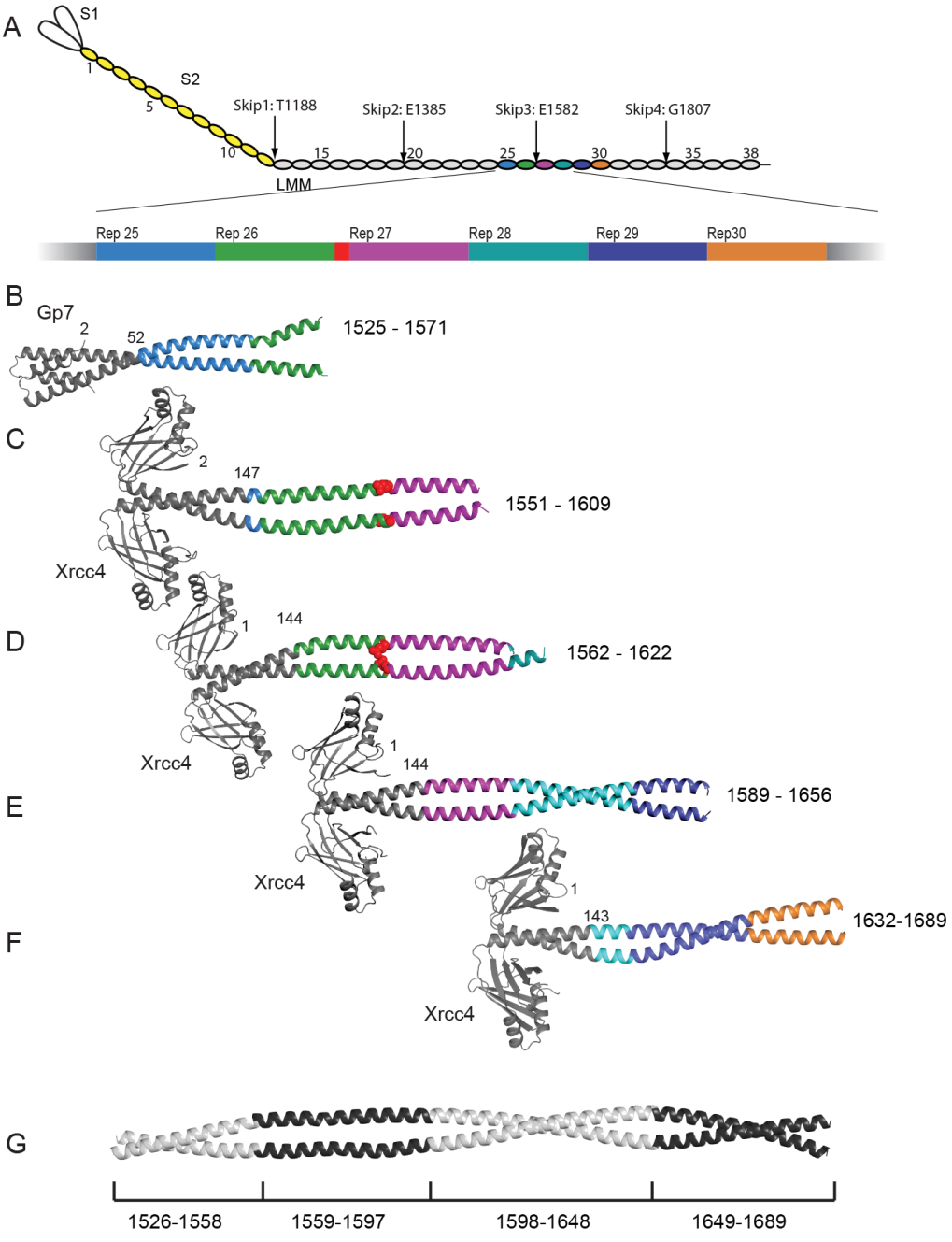

A
S1
S2
Skip1: T1188
Skip2: E1385
Skip3: E1582
Skip4: G1807
LMM
Rep 25
Rep 26
Rep 27
Rep 28
Rep 29
Rep30
B
Gp7
1525 - 1571
C
Xrcc4
1551 - 1609
D
Xrcc4
1562 - 1622
E
Xrcc4
1589 - 1656
F
Xrcc4
1632-1689
G
1526-1558
1559-1597
1598-1648
1649-1689

Figure 6.1 *(previous page)*: Structures of myosin fusion proteins and preliminary composite model. (A) A representation of myosin in which each 28 amino acid repeat of the C-terminal coiled-coil is an oval. S1, S2, and LMM are colored white, yellow and grey respectively. The numbering is shown for every fifth repeat and the positions of the skip residues are indicated. Repeats 25 through 30 are colored differently and the third skip residue, E1582, is shown in red. The fusion proteins are colored in grey while the myosin repeats are colored as in panel A for (B) Gp7-L1526-E1571, (C) Xrcc4-L1551-N1609, (D) Xrcc4-Q1562-L1622. (E) Xrcc4-H1590-L1657, (F) Xrcc4-A1632-R1689 and (G) A simple composite model for L1526-R1689 of human cardiac β-myosin. This was assembled from four of the five overlapping structures taken from Gp7-L1526-E1571, Xrcc4-L1551-N1609, Xrcc4-H1590-L1657, and Xrcc4-A1632-R1689. The residues incorporated from each structure are listed below. The coordinates for Xrcc4-L1551-N1609 were taken from the RCSB with accession 4XA4. Figures 6.1, 6.2, 6.3, 6.4, 6.5, 6.8, 6.8, 6.9, and 6.10 were prepared in part with Pymol (http://www.pymol.org/).

As noted earlier, isolated fragments of LMM exhibit a strong dependence salt dependence on solubility and tendency to form paracrystals which confound the ability to crystallize these proteins for high resolution structural studies. This problem was solved by including a globular domain excised from either Xrcc4 or Gp7. These proteins were selected from the protein data bank as small but structurally distinct domains that lead into well ordered dimeric coiled-coils. The intact Xrcc4 is involved in DNA repair and binds to both DNA ligase IV (RCSB accession number 1IK9) [31] and DNA. Here, only the globular domain of the protein that consists of $\sim$140 residues and is composed mostly of β-strands was used, but is referred to as Xrcc4 for simplicity in this context. Likewise, the bacteriophage ϕ29 scaffolding protein Gp7 consists of a small helical bundle domain that leads into an extended coiled-coil (RCSB accession number 1NO4) [30] where only the first $\sim$50 residues that contain the globular domain were incorporated in these fusions. These are denoted as Gp7-fusions for simplicity also.

Experimentally it was determined that the fusion proteins that exhibited the best characteristics for structural and biophysical studies contained $\sim$8-10 heptads of LMM. Longer fusions did not crystallize well and demonstrated increasing salt-dependent solubility. On average, six constructs were prepared for every successful structural determination.

A major question in assembling structures from fragments derived from fusion proteins is whether the conformation of the resultant pieces is influenced by crystal packing forces or by the interface between the folding domain and the fragment of myosin. In the case of Xrcc4-1562-1622 there are two dimers in the asymmetric unit that have a Cα root mean square deviation of 0.63 Å^2 over 89 α-carbon atoms of the target myosin coiled-coil. This indicates that crystallographic packing does not have a major influence on the resultant structure in this instance. In the same vein, the two fusions to Xrcc4 (Fig 6.1C and 6.1D) both overlap between residues 1562-1609 of myosin. This significant overlap is more than the minimum required for assembly into a larger model, but serves as a means to probe the influences of the fusion protein on the target section of coiled-coil. Each of the two Xrcc4-1562-1622 dimers aligns to Xrcc4-1551-1609 with a Cα root mean square deviation of 1.0 Å^2 and 1.2 Å^2 over 76 and 80 α-carbon atoms respectively (Fig 6.3). This strongly supports the hypothesis that the backbone conformations of the selected fragment of the coiled-coil are not notably perturbed by crystallographic contacts or the fusion to a folding domain.

To further test this hypothesis, MD simulations were performed on the isolated section of MyH7-

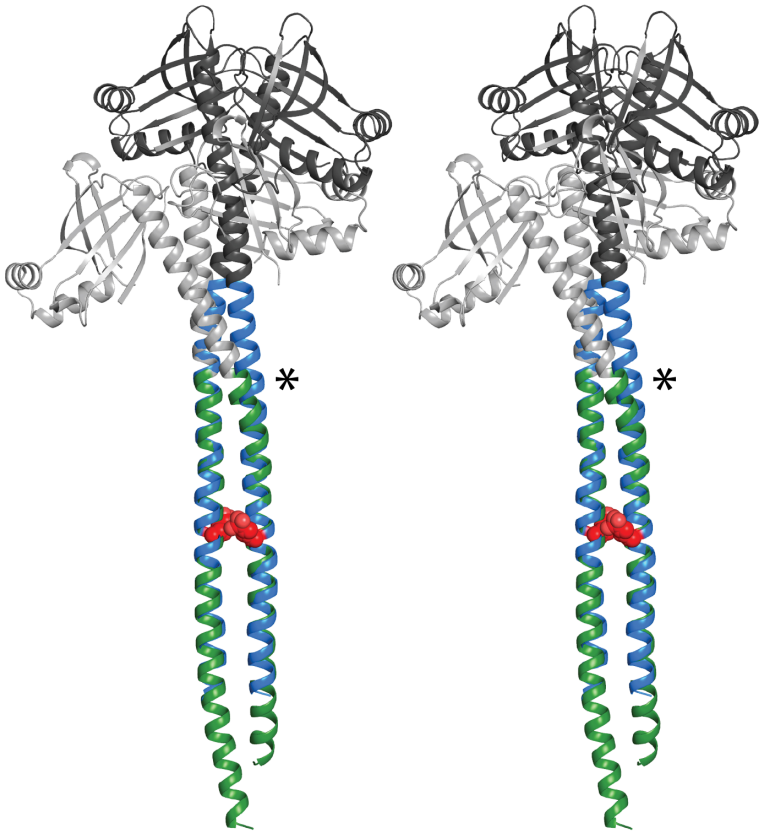


Figure 6.2: Stereo view of a structural alignment between Xrcc4-1551-1609 and Xrcc4-1562-1622. For Xrcc4-1551-1609 the Xrcc4 portion is colored in black and 1551-1609 of myosin is in blue. For Xrcc4-1562-1622 the Xrcc4 portion is colored in grey and 1562-1622 of myosin is in green. The structures are represented in cartoon and the skip residue is shown in red spheres. The primary point of divergence of the structures in the target coiled-coil is indicated with an asterisk.

1551-1602 and fused to either Gp7 or Xrcc4 (Table 6.3). Structures for the MyH7-1551-1602 segment were extracted from all three simulations to form a pool of conformations. Hierarchical clustering was performed over the ensemble of conformations for MyH7-1551-1603 using the k-clust algorithm in the MMTSB package. [58] Using a Cα root mean square distance (RMSD) of 4.75 Å as the clustering threshold, four clusters were observed. Figure 6.3A displays distribution of clusters amongst the simulations along with Cα-RMSD with respect to the Xrcc4-1551-1609 crystal structure. The folding domains have a slight stabilizing effect on the conformational ensembles. This is manifested in the average Cα-RMSD. The individual MyH7-1551-1602 simulation has an average Cα-RMSD of 3.8 Å as opposed to 2.4 Å and 3.3 Å from Gp7- and Xrcc4-bound forms. Additionally, the individual MyH7-1551-1602 simulation samples a slightly larger range of conformations; this simulation contains two extra clusters that are not found in the simulation of the Gp7- and Xrcc4-bound constructs. However, those clusters are within 2.7 Å and 2.0 Å Cα-RMSD with respect to the crystal structure (possibly due to the fluctuations at the N-terminal) and have low populations. Cα root mean square fluctuations (Cα-RMSF) exhibit similar trends (Fig. 6.3B). Thus, it can be concluded that the folding domain has a modest stabilizing effect (as supported by the biophysical measurements discussed below), and that the two most dominant conformational ensembles of the coiled-coils are not disturbed.

The conformations sampled in all three simulations are well within comparable RMSD and RMSF values, however this does not necessarily quantify the changes in the degree of coiling. To explicitly quantify the degree of coiling, the distance between center of masses of the two α-helices (DCOM) was calculated as a moving average of seven consecutive α-carbon atoms (Fig. 6.3C). All three simulations of the Skip 3 region show similar DCOM trends, except for a single heptad repeat at both the C- and N- termini. Hence, the MD simulations support the conjecture that the folding domains and C-terminal truncations do not alter the coiling patterns or the super-helical pitch except for the first and last heptad of the target section of coiled-coil.

6.3.2 Differences in the Super-Helical Pitch at the Fusion Junction do not Propagate into the Target Coiled-coil

Xrcc4-1562-1622 and Xrcc4-1551-1609 encompass a stretch of LMM that contains the third skip residue, E1582, located between repeat 26 and 27 that breaks the phase of the coiled-coil heptad repeat. The skip residue causes a significant local increase in the super-helical pitch encompassing 17 and 11 residues N- and C-terminal to the insertion. This distorted region of coiled-coil clearly starts at F1565, which places a

bulky hydrophobic residue in a core d position (Fig. 6.2 and 6.2). The two Xrcc4 constructs described herein both contain this residue, but in different positions relative to the N-terminal fusion interface. The junction between Xrcc4 and myosin in Xrcc4-1551-1609 is located two full heptads N-terminal to F1565. In contrast, the junction in Xrcc4-1562-1622 is only three residues N-terminal to F1565, or about one turn of an α-helix. As a consequence, the fusion junctions of these two constructs occur at different positions in the coiled-coil registry and include different Xrcc4 residues. The first myosin residue in Xrcc4-1551-1609 is a d position occupied by a leucine, while the first myosin residue in Xrcc4-1562-1622 is a glutamine in an e position (Fig. 6.4).

The major difference between these two structures occurs in the region surrounding F1565 near the fusion junction in Xrcc4-1562-1622, as indicated by an asterisk in Figures 6.2 and 6.4. For approximately two turns of the α-helix, just C-terminal to the folding domain interface, the coiled-coil is more tightly wound in Xrcc4-1562-1622 as compared to that region of myosin in Xrcc4-1551-1609. The coiled-coil of Xrcc4, measured from residues 118 to 153, has a super-helical pitch of about 160 Å while the skip region of myosin's super-helical pitch is measured at about 900 Å as calculated by the Crick coiled-coil parameterization (CCCP) server. [68] It is unlikely that the structural difference is the result of crystallographic packing, given that conformational variability between crystal structures is typically observed in loops and surface residues, and not in packing residues such as those in the coiled-coil. [69] Rather, this distortion appears to be the result of a mismatch of super-helical pitch between Xrcc4 and myosin. Except for the first five residues adjacent to the fusion protein the target myosin coiled-coils are highly similar. Fundamentally this implies that the structural influence of the folding domain does not propagate significantly beyond the first heptad of the target coiled-coil.

Xrcc4-1551-1609 is the highest resolution structure for that section of myosin coiled-coil and hence used in the construction of the composite model below Figures 6.1C and 6.1F. The junction between the Xrcc4 folding domain and the target myosin coiled has a smooth transition, where the coiled-coil parameters are well matched. Both Xrcc4-1551-1609 and Xrcc4-1562-1622 exhibit stable coiled-coil C-termini with no significant differences over regions of overlap.

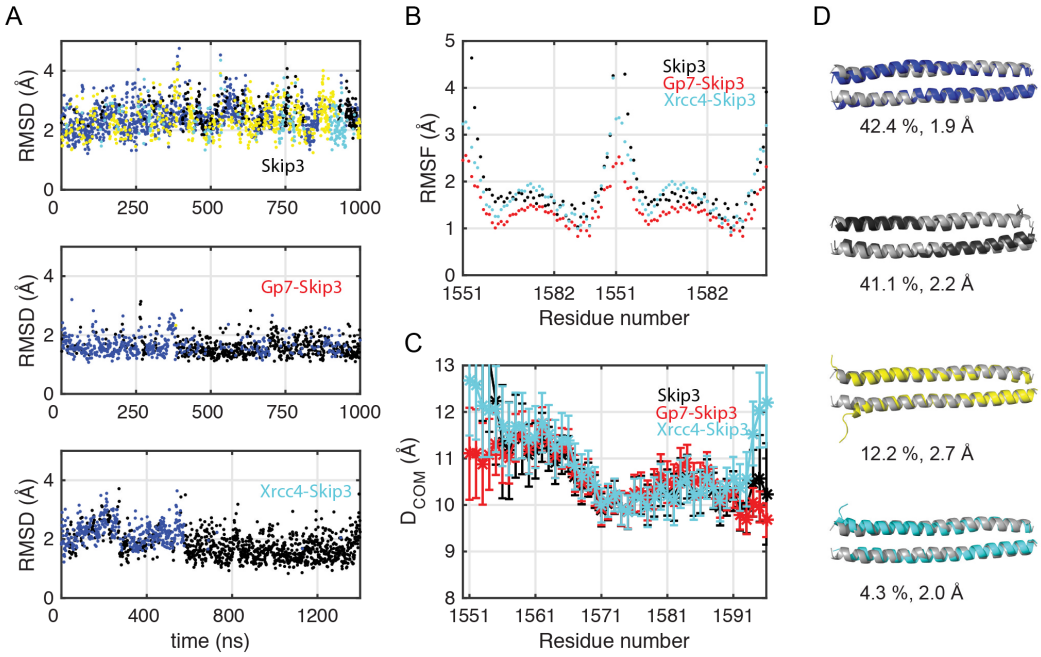


Figure 6.3: Analysis of the molecular dynamics simulations for Skip 3 in the absence and presence of a fusion partner. (A) Clustering of the Skip 3 region (MyH7-1551-1602) conformations extracted from Skip 3 alone, Gp7-Skip 3 and Xrcc4-Skip 3 simulations, is presented in three individual panels along with the Cα-RMSD with respect to the Xrcc4-1551-1609 crystal structure. Hierarchical clustering was carried out with the k-clust protocol in MMTSB using a Cα-RMSD of 4.75 Å as the similarity measure. Different colors represent different clusters. (B) Root Mean Square Fluctuations (RMSF) based on the Cα-atoms (Black- isolated Skip 3 simulation, Red: Gp7-Skip 3, Turquoise: Xrcc4-Skip 3). (C) DCOM trend for Skip 3 (Black), Gp7-Skip 3 (red) and Xrcc4-Skip 3 (turquoise) simulations. (D) Representative members from each cluster are shown along with their overall population percentages and Cα-RMSD with respect to the Xrcc4-1551-1609 crystal structure.

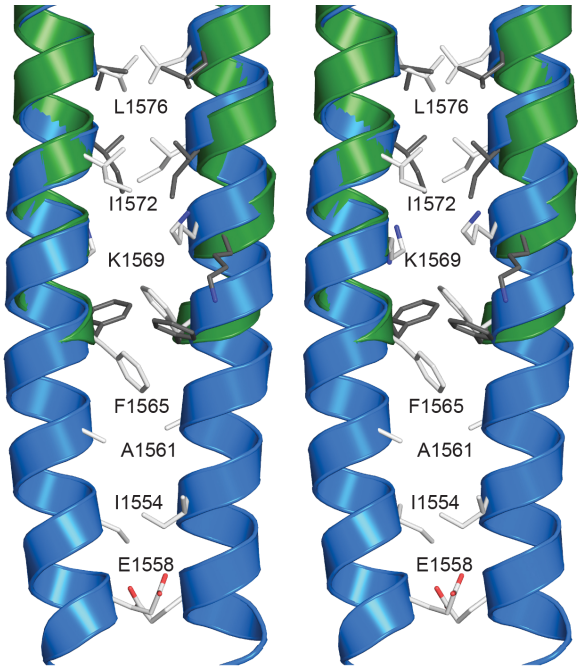


Figure 6.4: Stereo representation of the coiled coil centered on F1565 in the Xrcc4-1551-1609 and Xrcc4-1562-1622 fusion proteins. Xrcc4-1551-1609 is represented with blue cartoon helices and white stick side chains. Xrcc4-1562-1622 is represented with green cartoon helices and dark grey side chains. Only the side chains of residues along the interface are displayed. Residues from the Xrcc4 folding domain are not shown.

6.3.3 Regions of Stable Coiled-coil Interface Should be Selected for C-terminal Truncations

The most N-terminal segment described in this study, Gp7-1526-1571, contains myosin residues spanning repeat 25 and half of repeat 26 and also includes F1565. Unlike in the previous two Xrcc4 constructs, F1565 is located six residues from the C-terminus of the coiled-coil. Interestingly, the last two heptads of this fusion protein splay apart forming an antiparallel four-helix bundle between crystallographically related dimers (Fig. 6.5). Given the redundancy of structural information for this section of the coiled-coil it is safe to assume that this arrangement is an artifact of crystal packing and was excluded from consideration in the composite model. This tetramerization probably arises because of a weak dimerization interface near the C-terminus of the coiled-coil that is readily counterbalanced by crystallographic packing forces (Fig. 6.5). Interestingly, the coiled-coil prediction for this region is less than 100 % in the region leading up to F1565 [70] which suggests that a minimum of one heptad of canonical coiled-coil should be included at the C-terminus to avoid formation of antiparallel four helix bundles. The observation that truncated fragments of coiled-coils can come apart at their ends raises the question to what extent the folding domains stabilize these segments of coiled-coil. This was addressed experimentally by determining the folding and unfolding characteristics of a short section of myosin fused to three different folding domains. The same question was also examined earlier in the MD studies.

6.3.4 Fusion Proteins Modulated the Level of Expression and Increased the Total α-helical Content of the Chimera

Fusions were constructed between the short coiled-coil of GCN4, Gp7, or Xrcc4 and a region of human cardiac myosin, that encompasses the second skip residue (MyH7 residues 1361 to 1406). The fusions were compared to a construct that did not include any folding domain. A mutant form of the GCN4 leucine zipper that was designed to have greater coiled-coil stability (GCN4-pMSE hereafter abbreviated to GCN4) was used as a comparison to two other folding domains. [71] The particular region of myosin was selected because it contains a tryptophan residue enabling fluorescence-based measurements. All MyH7 fusions were purified by Ni-NTA chromatography under identical conditions and the amount of soluble protein was compared using relative band intensity on a Coomassie-stained SDS-PAGE gel. All of the constructs yielded a substantial quantity of soluble protein, where the Xrcc4 fusion yielded the greatest mass of soluble material (Fig. 6.6). Interestingly, although the native coiled-coil protein was soluble without a folding domain, the expression

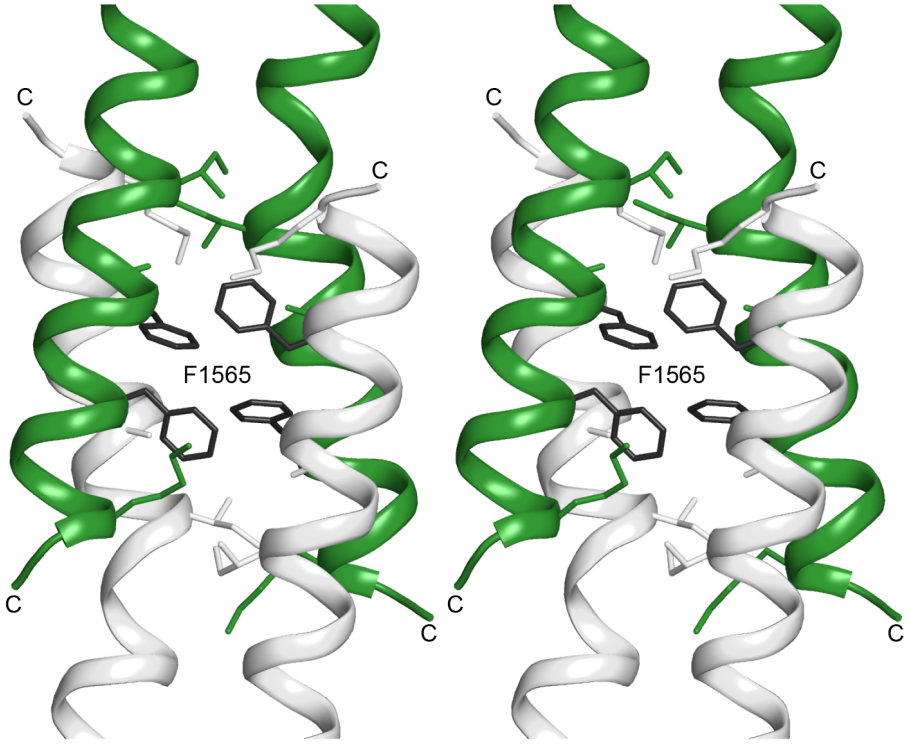


Figure 6.5: Stereo diagram showing the antiparallel four helix bundle formed by Gp7-1526-1571. One dimer is colored in green while the symmetry related dimer is colored in white. The clustering of F1565 in all four chains is critical to the formation of the antiparallel helix bundle. The C-termini of the polypeptide chains are indicated.

levels do not convey the dramatic differences in biophysical behavior exhibited by these constructs. While MyH7-1361-1406 appeared folded, as indicated by a single sharp peak in gel filtration (data not shown), it had low α-helical content as measured by circular dichroism. The mean residue molar ellipticity was -12733 $deg \times cm^(2) \times dmol^{-1}$ at 222 nm which was much lower than a value of ∼33,000 expected for a protein that is predicted to be entirely α-helical. [72] This is consistent with the observation by Wolny et al., that many isolated short sections of the myosin rod exhibit low α-helical content. [23] The addition of the folding domain greatly increases the mean residue molar ellipticity, as listed in Table 6.4, indicating that the stability conferred by the folding domain propagates into the target coiled-coil.

6.3.5 Fusion Proteins Enhance Thermal Stability

All fusions greatly enhanced the thermal stability of the target coiled-coil as shown by tryptophan autofluorescence (Fig. 6.7). The control construct lacking an N-terminal fusion had an emission maximum of 355 nm indicating the tryptophan residue is solvent exposed and did not undergo a cooperative unfolding transition in temperature scanning measurements. The emission maximum of the fusion proteins were blue shifted to 342 nm (Gp7) and 340 (GCN4) due to increased shielding from solvent. GCN4 conferred the greatest enhancement of thermal stability. The analysis of Xrcc4 constructs, based on autofluorescence, is complicated by multiple tryptophan moieties and was therefore not included. The CD (Fig. 6.7)) and autofluorescence data taken together demonstrate that the fusions increase the total α-helical content and thermal stability of the target coiled-coil suggesting that these constructs are more suitable for biophysical analysis. This was reflected in the crystallization properties of these proteins.

6.3.6 The Fusion Protein Facilitates Crystallization

Constructs were evaluated for crystallization in a 144 condition sparse matrix screen at 4 °C and 25 °C that samples a range of polyethylene glycol, hexylene glycol, and salt, based conditions. The screen was prepared in-house. No crystals have been observed for MyH7-1361-1406 alone, while the Xrcc4, GCN4, and Gp7 fusions readily crystallized, but even here among the fusions there were significant differences. GCN4 yielded visually stunning crystals, but these exhibited highly anisotropic diffraction. Conversely, the inclusion of a structurally more complex motif such as Gp7 yielded more ordered crystals that were readily amenable to structural analysis. In general, Gp7 fusions crystallize under a wider range of conditions than Xrcc4 fusions.

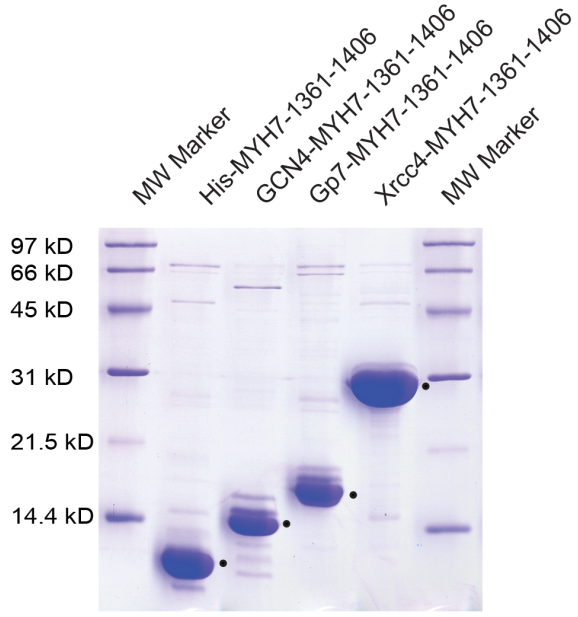


Figure 6.6: 15 % SDS-PAGE gel demonstrating soluble expression of MyH7 constructs. (A) All lanes contain 1.25 μL of a Ni-NTA purified MyH7 construct (equivalent of 0.625 μg of cells). Bands corresponding to MyH7 are marked with a dot to the right of the band. All MyH7 constructs show soluble overexpression, including fusion-less MyH7, and only minimal changes in soluble expression are noted with different fusions. Xrcc4-MyH7 shows the highest soluble expression level.

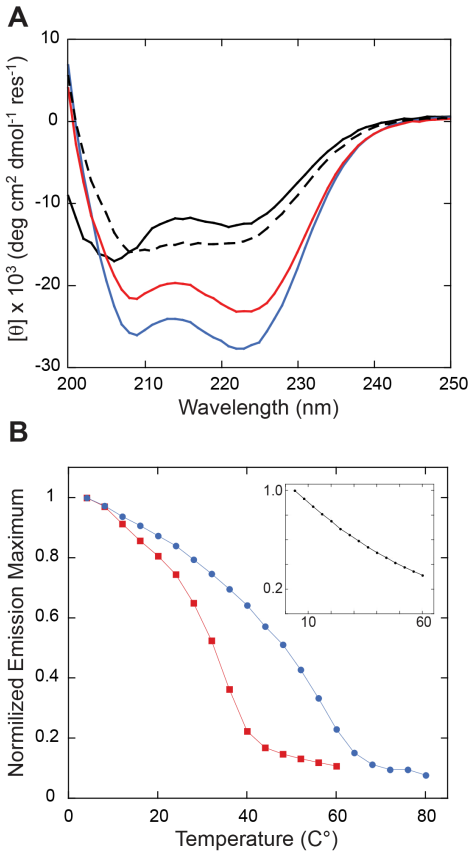


Figure 6.7: (A) Circular dichroism spectra of MYH7-1361-1406 (black) and Gp7 (Red), GCN4 (blue), Xrcc4 (black dashed line) fusions to MYH7-1361-1406. (B) Temperature scanning autofluorescence emission maximum for each construct is plotted versus temperature. Gp7 (Red) and GCN4 (Blue) fusions to MYH7-1361-1406 show temperature dependent transition. MYH7-1361-1406 (Black) without fusion is shown in the inset and does not display a cooperative temperature depend change in tryptophan fluorescence. A line connecting the measured data has been added to each trace as a guide.

Table 6.4: Circular Dichroism Theta Values for Myosin Fusion Proteins.

Construct	Θ (deg cm^2 dmol^{-1})
MyH7-1361–1406	–12,733
Gp7-MyH7-1361–1406	–23,247
GCN4-MyH7-1361–1406	–27,766
Xrcc4-MyH7-1361–1406	–14,873

The increased thermal stability of the fusion relative to the target coiled-coil coupled with a tendency for crystallization provide strong support for the use of these folding domains in the context of myosin.

6.3.7 Assembly of Structural Fragments into a Contiguous Model

There is considerable structural overlap between the five structures shown in Figure 6.1. However, a model built by simple alignment represents a crude approximation considering that overlapping regions exhibit slightly different conformations (Cα root mean square deviation of < 1 Å). Although previous studies that focus on modeling canonical coiled-coils have been reported, modeling the myosin rod based on those studies is unrealistic due to the presence the skip residues that lead to deviations from an ideal coiled-coil [73]. Our studies of the skip residues have shown that the deviation from an ideal coiled-coils surrounding a skip residue extends over ∼four heptads where the distortions is dependent on the surrounding amino acid sequence. [26] Due to the irregularities in the structure a model based on a typical canonical coiled-coil is incapable of defining the structure of the myosin rod [73]. Consequently a more systematic approach described here was developed to assemble a contiguous structure from the individual structures with the homology modeling program MODELLER. [51]

The initial model from this procedure was simulated through molecular dynamics to remove steric overlaps and allow adequate structural relaxations through three separate simulations. The trajectory length for each simulation is listed in Table 6.3. MD simulations of the composite model yielded a conformational ensemble rather than a single conformation and thus provide information on the flexibility of the molecule. This was derived from the sum of three independent trajectories (summing up to 2.2 μs).

Hierarchical clustering was performed on the ensemble of structures generated from the simulation to distinguish major conformers (Fig. 6.8). The RMSD cut off for clustering was selected after trying several values varying from 6 Å to 10 Å. Within that range, 8 Å was found to yield the most robust clustering where this best reflected the diversity of the conformations sampled by the simulations. Six clusters were formed when the ensembles were sorted with this Cα-RMSD threshold (Fig. 6.8G). The representative structure of a given cluster was taken as that which had the lowest overall Cα-RMSD to the other members of the cluster. A comparison of the six representative structures, one from each cluster, revealed that the major difference between the ensembles is the degree of supercoiling. At one extreme the Cα-RMSD increases to 10.7 Å between the 4th cluster (shown yellow) relative to the initial model, where the structure is less tightly

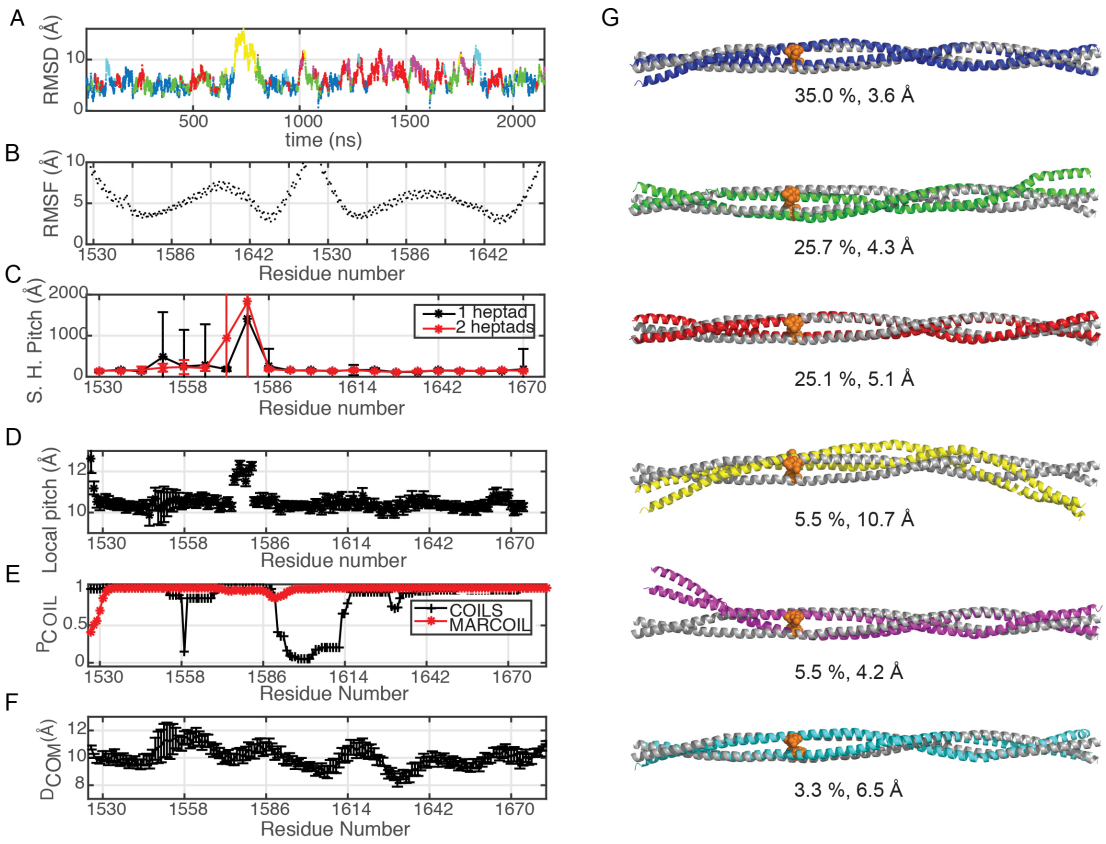


Figure 6.8: Molecular dynamics analysis of the MyH7-1526-1689 segment of the myosin rod. (A) Cα-RMSD with respect to the initial model that was assembled directly from the crystal structures. Color coding represent different clusters obtained through hierarchical clustering using the k-clust protocol in MMTSB. The Cα-RMSD cut off for clustering was set to 8 Å which gives the most robust clusters. (B) Root Mean Square Fluctuations (RMSF) based on the Cα-atoms. (C) Estimated super-helical pitch (Å) trend over each heptad-repeat (Skip 3 residue: E1582) based on a single heptad (black) and two heptad repeats (red). The values shown are calculated through averaging over conformations from all 3 trajectories. (D) The length of each heptad repeat calculated as a moving average to reveal local fluctuations more closely (E) Coiled-coil propensities are calculated with a 28 residue sliding window using COILS (red) and MARCOIL (black) servers using only the sequence information (F) DCOM for the composite model calculated through averaging over conformations from all 3 trajectories. (G) Representative members from each cluster are shown along with their overall population percentages and Cα-RMSD with respect to the initial model for MyH7-1526-1689.

wound and more flexuous than the initial model. The slight bend towards C-terminal is expected since the structure becomes more flexuous due to length. This particular cluster was only sampled in 5.5 % of the overall population.

The simulated composite model for the MyH7-1526-1689 section of the myosin rod was also compared to the simulation of MyH7-1551-1602 extracted from the crystal structure of Xrcc4-1551-1609. This analysis probed how the increased length of the composite model alters the structures relative to simulations of smaller sections. It is important to emphasize that the conformational space visited in each simulation does not change between the simulations of the smaller section and the composite model. Clusters obtained from the composite model simulation remain within 2.1 Å Cα-RMSD of the crystal structure of Xrcc4-1551-1609. The composite model is expected to be flexuous due to its length. MyH7-1526-1689 exhibits higher Cα-RMSF than the MyH7-1551-1603 segment (Fig. 6.8), which is typically in the range of 3-4 Å for the composite model and ∼2 Å for Xrcc4-1551-1602 simulation. The highest degree of RMSF is observed between residues 1582-1640, following the skip residue. This prediction from the simulations is consistent with the observed structure for Xrcc4-1590-1657 which shows that the α-helices separate between residues 1608-1615 consistent with a lower predicted coiled-coil propensity.

Since the clustering results pointed to a difference in the degree of coiling among the representative structures, the super-helical pitch of the myosin rod was calculated over the full simulation. Super helical pitch was calculated based on a sliding window of 7 (black) or 14 (red) amino acids. As expected, there is an increased super-helical pitch around the skip residue, extending 20 amino acids upstream from the skip location where the computed super-helical pitch averages over 1300 Å as compared to the value of 170 Å for other regions (Fig. 6.8C) for both approximations (single heptad vs 2-heptads). Both MyH7-1551-1602 and MyH7-1526-1689 follow the same qualitative trend (data not shown). These data taken together strongly support the feasibility of assembling a complete model for LMM and illustrate the difficulties associated with building a model based on a canonical coiled-coil.

The average length of each heptad repeat, or local pitch, was calculated over the trajectory (Fig. 6.8D), to further emphasize the local disruption due to the skip residue. While the local pitch averages to 10 Å for most of the coiled-coil, it rises up to 12.5 Å around the skip residue. DCOM (See Methods for details) also shows increased distance between the centers of masses of the two helices around the skip residue and

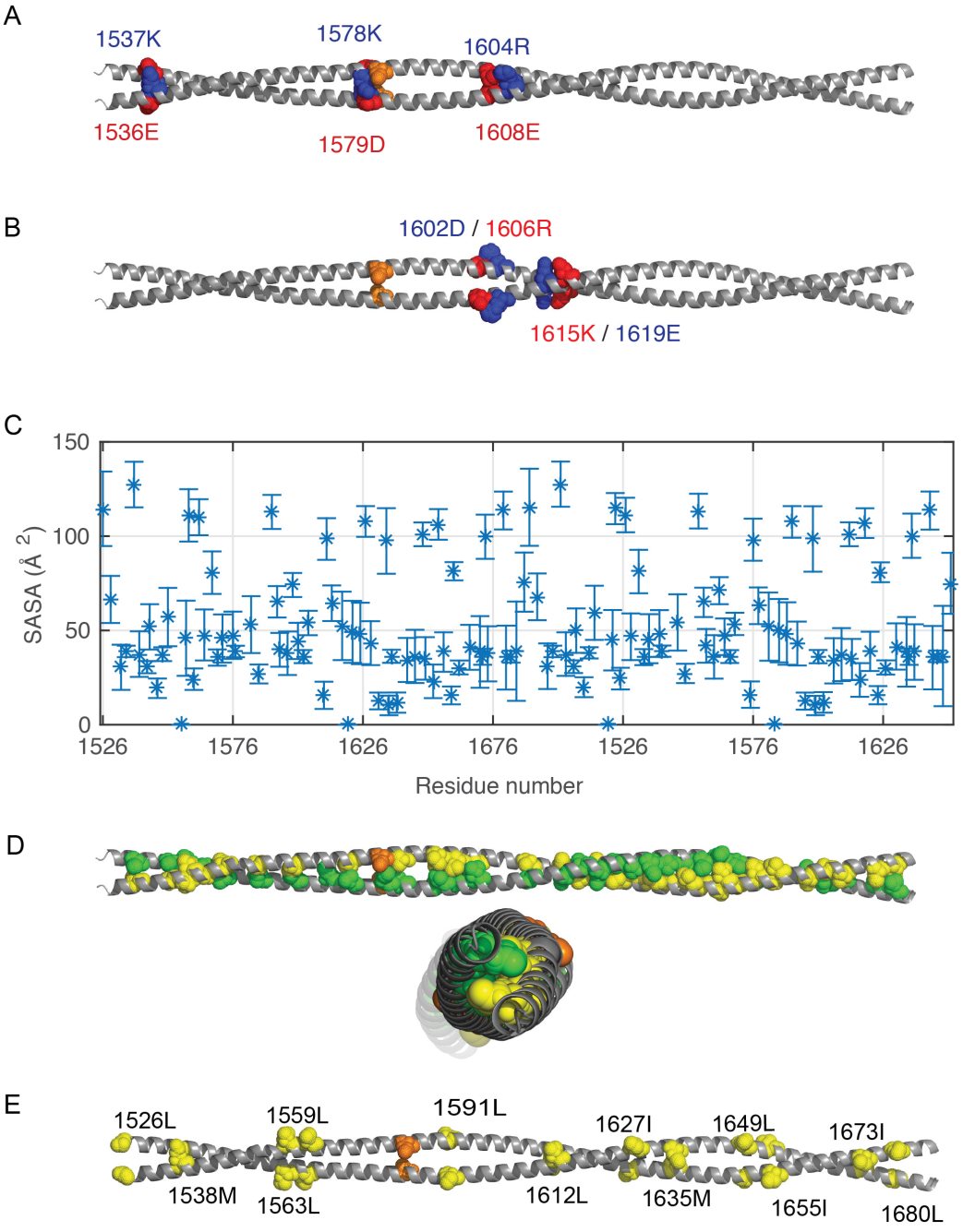

A
1537K
1578K
1604R
1536E
1579D
1608E
B
1602D / 1606R
1615K / 1619E
C
150
100
50
0
SASA (Å 2)
1526
1576
1626
1676
1526
1576
1626
Residue number
D
E
1526L
1559L
1591L
1627I
1649L
1673I
1538M
1563L
1612L
1635M
1655I
1680L

Figure 6.9 *(previous page)*: Analysis of the charged interactions and exposed hydrophobic surface of the composite model for MyH7-1526-1689. (A) Charged interactions between two helices evaluated by the average distances between center of mass of nitrogen atoms (NH1, NH2) of Arginine, nitrogen atom (NZ) of Lysine and center of mass of oxygen atoms from Glutamate (OE1, OE2) and Aspartate (OD1, OD2). The pairs of residues within 4.5 Å distance are shown (Table D1). The Skip 3 residue is shown in orange. Positively charged amino acids are displayed in blue whereas negatively charged amino acids are displayed in red. (B) Similarly, charged interactions within each chain are shown in sphere representation. (C) Solvent accessible surface area (SASA) ($Å^2$) of the hydrophobic amino acid residues in the composite model (Table D2). (D) Side and top view of MyH7-1526-1689 are displayed with the predicted hydrophobic interactions. The hydrophobic interactions within the rod were evaluated from the average minimum distances between the hydrophobic side-chains. The pairs of residues that are closer than 5 Å are shown (Table D3). To distinguish residues from different helices, the predicted residues are colored in yellow and green for different helices. (E) Hydrophobic residues that have a SASA over 90 $Å^2$ are shown on the composite model in sphere representation.

provides further evidence that the thick filament is less tightly wound around the skip residue.

Prediction software, such as that provided by servers such as COILS [74] and MARCOIL, [70] are commonly used to provide an estimate of how well a sequence matches a canonical coiled-coil. While these are outstanding at predicting potential coiled-coils they do not provide direct evidence of the structure adopted by a polypeptide chain in those regions where the sequence deviates from a canonical heptad pattern. There is also considerable discrepancy between predictions for those segments that do not exhibit a standard pattern. As shown in Figure (Fig. 6.8E), the predictions from COILS and MARCOIL are not only inconsistent with each other but do not predict the observed structural features seen in our structures and simulations. MARCOIL predicts that the entire region should be coiled-coil with over 99 % propensity whereas COILS predicts low coiled-coil propensity around residues 1558 and 1590-1613 region. This comparison demonstrates that the thick filament cannot be modeled using sequence information alone or built on a canonical coiled-coil; the presence of the skip residues disrupts the expected pattern such that structural information is required to accurately model these regions.

Finally, the interactions within the coiled-coil in the simulated composite model were examined. The electrostatic interactions were analyzed by calculating the average distances between the center of mass of the nitrogen atoms of arginine (NH1, NH2) and lysine (NZ) and the center of mass of oxygen atoms of glutamate (OE1, OE2) and aspartate (OD1, OD2). The distances are averaged over an ensemble of conformations sampled by three independent simulations rather than a static model. Residues with atoms that lie within 4.5 Å distance were selected as interacting pairs. Only a few electrostatic interactions were predicted to be occur between the two helices; those pairs are E1536–K1537, K1579–D1580 and R1604–E1608 (Fig. 6.9A and Table D1). It is important to note that the charged interactions predicted do not follow the well known e–g' interactions observed in typical parallel coiled-coils, where e type residue on one chain interacts with the preceding g' type residue from the other chain. [75] None of the pairs identified in this study match the e-g' category. There are however a few ionic interactions between residues in the same chain, notably between the 1602D–1606R and 1615K–1619E pairs, that might contribute to the stability of the assembly (Fig. 6.9B). No repulsive interactions were observed in this 164 amino acid segment of the myosin rod in the simulations.

The stability of the coiled-coil is expected to be driven by the hydrophobic effect. The hydrophobic interactions between two helices of the coiled-coil were evaluated from the average distances between the

center of masses of each hydrophobic side-chain (Fig. 6.9C and Table D2 for a full list of interactions identified). The distances were averaged over the ensemble of conformations from three independent trajectories. All hydrophobic pairs whose center of mass of the side chains fall within 5 Å were selected and shown in orange Figure 6.9D, where these interactions are strictly in the interface between the two α-helices. Almost all interactions can be classified as contacts between a–d' or a–g' residues.

Unlike most typical soluble globular protein structures, the myosin coiled-coil exhibits a large number of hydrophobic side chains on the surface. There are numerous solvent exposed hydrophobic residues as indicated by the solvent accessible surface area (SASA, $Å^2$) for each hydrophobic side chain (Fig. 6.9A) including, L1526, M1538, L1559, L1563, L1591, L1612, I1627, M1635, L1649, I1655, I1673, and L1680. These all have SASA values over 90 $Å^2$ (Fig. 6.9E and Table D3). The composite model contains 132 hydrophobic residues, of which 24 are almost completely solvent exposed. This solvent accessibility is by far higher than it is expected for hydrophobic residues in α-helices; usually hydrophobic amino acids are only 1 % to 5 % accessible, except for Trp and Tyr which are 10 % accessible. [76] We hypothesize that these residues are important for assembly of the thick filament beyond the simple ionic interactions [6] and will be important in future efforts to construction higher ordered structures. Interestingly, in most instances the hydrophobic character of these positions in the myosin rod is highly conserved across sarcomeric myosins.

6.3.8 Distribution of Cardiomyopathy Mutations

More than eighteen mutations that lead to cardio or skeletal myopathies have been identified in the segment of myosin modeled here. [2, 77] The location of these residues is shown in Figure 6.10, which reveals a broad distribution of sites. Of these, in vivo and in vitro measurements have only been reported for the effect of the E1555K substitution. [23] These show reduced incorporation of the mutant protein into the sarcomere and decreased helicity in the coiled-coil which, together with studies of mutations outside the segment investigated here, suggest that mutations that reduce the helix stability lead to lower incorporation into the sarcomere. As shown on the model, E1555 is a surface residue that does not appear to be involved in any inter or intramolecular ionic interactions in the wild-type structure. Thus, the loss of stability cannot be attributed to the loss of a stabilizing interaction, but rather must be due to gain of a destabilizing contact within the coiled-coil. The reduced incorporation into the thick filament could be due to either defects in interactions between myosin rods or by reduced stability of the protein itself. Resolution of this question can now be

approached through simulation of the mutant structure and through structural determination of segments that carry mutations.

It is noteworthy that there are a considerable number of mutations to proline that lead to skeletal myopathies. [77, 78] In vitro studies of two mutations outside the region studied here (R1500P and L1706P) suggest that prolines can be introduced into the myosin rod and yet still lead to protein that is incorporated into thick filaments even though the assembly appears is less stable [78]. It is expected that introduction of a proline into a helical backbone will lead to loss of hydrogen bonding opportunities. However, proline residues are occasionally found in α-helices where they lead to a pronounced bend in their path. It is difficult to predict whether the changes in assembly are due to alterations in the stability of the coiled-coil itself or to changes in the manner the myosin rods interact with each other in the thick filament. Here again, the computational and structural approach outlined here is well suited to investigating the biophysical consequences of these mutations.

6.4 Conclusions

This study establishes a methodology for determining the complete structure of the myosin rod. It shows that the use of fusion proteins allows restricted segments of the myosin rod to be expressed and purified which allows structural and biophysical characterization. It also demonstrates that the fusion domains themselves introduce only a small perturbation in the structure that is readily eliminated by structural determination of overlapping fragments. The study also shows that the fragments can be assembled into composite models through molecular dynamics simulations. The five fragments studied here contained a total of 292 residues of myosin which yielded a contiguous model of 164 residue fragment. This suggests that a complete model for LMM can be obtained by implementing a strategy that includes duplicate structures for every residue in the coiled-coil. The results thus far show a surprising number of exposed hydrophobic side chains which suggests they may play a role in the assembly of myosin rods into the thick filament. The methodology established here will facilitate construction of a model for the entire myosin rod, and lays the ground work for assembling the rods into a model for the thick filament. This study also creates a framework for understanding the biophysical consequences of mutations in myosin that lead to cardio and skeletal myopathies.

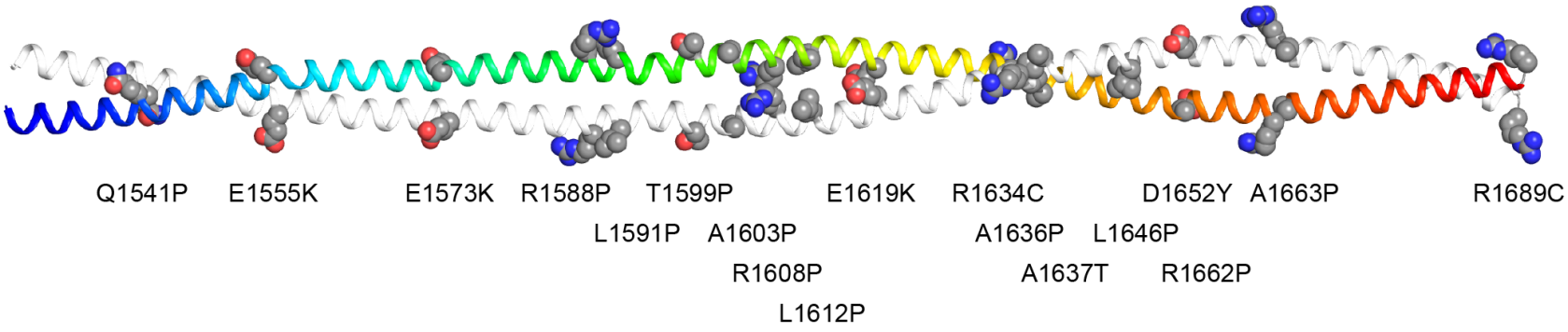


Figure 6.10: Disposition of mutations that lead to cardio or skeletal myopathies in MyH71526-1689. The mutations themselves are widely distributed along the length of the coiled-coil and occur at all heptad positions.

6.5 References

[1] S. LOWEY, H. S. SLAYTER, A. G. WEEDS, and H. BAKER, *J Mol Biol* **42**, 1 (1969).

[2] M. BUVOLI, M. HAMADY, L. A. LEINWAND, and R. KNIGHT, *Trends Cardiovasc Med* **18**, 141 (2008).

[3] J. R. MOORE, L. LEINWAND, and D. M. WARSHAW, *Circ Res* **111**, 375 (2012).

[4] A. OLDFORS, *Neuromuscul Disord* **17**, 355 (2007).

[5] A. D. MCLACHLAN and J. KARN, *J Mol Biol* **164**, 605 (1983).

[6] A. D. MCLACHLAN and J. KARN, *Nature* **299**, 226 (1982).

[7] G. OFFER, *J Mol Biol* **216**, 213 (1990).

[8] H. E. HUXLEY, *Science* **164**, 1356 (1969).

[9] J. H. BROWN, *Protein Sci* **19**, 1366 (2010).

[10] S. NICOLET, H. HERRMANN, U. AEBI, and S. V. STRELKOV, *J Struct Biol* **170**, 369 (2010).

[11] S. J. ATKINSON and M. STEWART, *J Cell Sci* **99 (Pt 4)**, 823 (1991).

[12] P. M. BENNETT, *J Mol Biol* **146**, 201 (1981).

[13] R. A. CROSS and J. VANDEKERCKHOVE, *FEBS Lett* **200**, 355 (1986).

[14] L. NYITRAY, G. MOCZ, L. SZILAGYI, M. BALINT, R. C. LU, A. WONG, and J. GERGELY, *J Biol Chem* **258**, 13213 (1983).

[15] R. L. SOHN, K. L. VIKSTROM, M. STRAUSS, C. COHEN, A. G. SZENT-GYORGYI, and L. A. LEINWAND, *J Mol Biol* **266**, 317 (1997).

[16] W. BLANKENFELDT, N. H. THOMA, J. S. WRAY, M. GAUTEL, and I. SCHLICHTING, *Proc Natl Acad Sci U S A* **103**, 17713 (2006).

[17] B. CIANI, S. BJELIC, S. HONNAPPA, H. JAWHARI, R. JAUSSI, A. PAYAPILLY, T. JOWITT, M. O. STEINMETZ, and R. A. KAMMERER, *Proc Natl Acad Sci U S A* **107**, 19850 (2010).

[18] K. A. DILL, K. M. FIEBIG, and H. S. CHAN, *Proc Natl Acad Sci U S A* **90**, 1942 (1993).

[19] R. A. KAMMERER, T. SCHULTHESS, R. LANDWEHR, A. LUSTIG, J. ENGEL, U. AEBI, and M. O. STEINMETZ, *Proc Natl Acad Sci U S A* **95**, 13419 (1998).

[20] A. N. LUPAS and M. GRUBER, *Advances in protein chemistry* **70**, 37 (2005).

[21] M. O. STEINMETZ, A. STOCK, T. SCHULTHESS, R. LANDWEHR, A. LUSTIG, J. FAIX, G. GERISCH, U. AEBI, and R. A. KAMMERER, *Embo J* **17**, 1883 (1998).

[22] K. C. WU, J. T. BRYAN, M. I. MORASSO, S. I. JANG, J. H. LEE, J. M. YANG, L. N. MAREKOV, D. A. PARRY, and P. M. STEINERT, *Mol Biol Cell* **11**, 3539 (2000).

[23] M. WOLNY, M. COLEGRAVE, L. COLMAN, E. WHITE, P. J. KNIGHT, and M. PECKHAM, *J Biol Chem* **288**, 31952 (2013).

[24] Y. WANG, R. GAO, and D. G. LYNN, *Chembiochem : a European journal of chemical biology* **3**, 311 (2002).

[25] S. A. WOLFE, R. A. GRANT, and C. O. PABO, *Biochemistry* **42**, 13401 (2003).

[26] K. C. TAYLOR, M. BUVOLI, E. N. KORKMAZ, A. BUVOLI, Y. ZHENG, N. T. HEINZE, Q. CUI, L. A. LEINWAND, and I. RAYMENT, *Proc Natl Acad Sci U S A* (2015).

[27] J. FRYE, V. A. KLENCHIN, and I. RAYMENT, *Biochemistry* **49**, 4908 (2010).

[28] V. A. KLENCHIN, J. J. FRYE, M. H. JONES, M. WINEY, and I. RAYMENT, *The Journal of biological chemistry* **286**, 18240 (2011).

[29] M. S. JUNOP, M. MODESTI, A. GUARNE, R. GHIRLANDO, M. GELLERT, and W. YANG, *EMBO J* **19**, 5962 (2000).

[30] M. C. MORAIS, S. KANAMARU, M. O. BADASSO, J. S. KOTI, B. A. OWEN, C. T. MCMURRAY, D. L. ANDERSON, and M. G. ROSSMANN, *Nat Struct Biol* **10**, 572 (2003).

[31] B. L. SIBANDA, S. E. CRITCHLOW, J. BEGUN, X. Y. PEI, S. P. JACKSON, T. L. BLUNDELL, and L. PELLEGRINI, *Nat Struct Biol* **8**, 1015 (2001).

[32] G. J. CHEN, N. QIU, C. KARRER, P. CASPERS, and M. G. PAGE, *Biotechniques* **28**, 498 (2000).

[33] K. MIYAZAKI and M. TAKENOUCHI, *Biotechniques* **33**, 1033 (2002).

[34] C. J. ROCCO, K. L. DENNISON, V. A. KLENCHIN, I. RAYMENT, and J. C. ESCALANTE-SEMERENA, *Plasmid* **59**, 231 (2008).

[35] F. VAN DEN ENT and J. LOWE, *J Biochem Biophys Methods* **67**, 67 (2006).

[36] P. G. BLOMMEL and B. G. FOX, *Protein Expr Purif* **55**, 53 (2007).

[37] A. V. MCDONNELL, T. JIANG, A. E. KEATING, and B. BERGER, *Bioinformatics* **22**, 356 (2006).

[38] I. RAYMENT, *Methods in enzymology* **276**, 171 (1997).

[39] W. MINOR, M. CYMBOROWSKI, Z. OTWINOWSKI, and M. CHRUSZCZ, *Acta Crystallogr D Biol Crystallogr* **62**, 859 (2006).

[40] Z. OTWINOWSKI and W. MINOR, *Methods in Enzymology* **276**, 307 (1997).

[41] N. COLLABORATIVE COMPUTATIONAL PROJECT, *Acta Crystallogr D Biol Crystallogr* **50**, 760 (1994).

[42] A. J. MCCOY, R. W. GROSSE-KUNSTLEVE, P. D. ADAMS, M. D. WINN, L. C. STORONI, and R. J. READ, *J Appl Crystallogr* **40**, 658 (2007).

[43] K. COWTAN, *Acta Crystallographica Section D-Biological Crystallography* **64**, 83 (2008).

[44] K. COWTAN, *Acta Crystallogr D Biol Crystallogr* **66**, 470 (2010).

[45] P. EMSLEY and K. COWTAN, *Acta Crystallogr D Biol Crystallogr* **60**, 2126 (2004).

[46] G. N. MURSHUDOV, P. SKUBAK, A. A. LEBEDEV, N. S. PANNU, R. A. STEINER, R. A. NICHOLLS, M. D. WINN, F. LONG, and A. A. VAGIN, *Acta Crystallogr D Biol Crystallogr* **67**, 355 (2011).

[47] P. D. ADAMS, P. V. AFONINE, G. BUNKOCZI, V. B. CHEN, I. W. DAVIS, N. ECHOLS, J. J. HEADD, L. W. HUNG, G. J. KAPRAL, R. W. GROSSE-KUNSTLEVE, A. J. MCCOY, N. W. MORIARTY, R. OEFFNER, R. J. READ, D. C. RICHARDSON, J. S. RICHARDSON, T. C. TERWILLIGER, and P. H. ZWART, *Acta Crystallogr D Biol Crystallogr* **66**, 213 (2010).

[48] V. SCHOMAKER and K. N. TRUEBLOOD, *Acta Crystallographica Section B-Structural Science* **54**, 507 (1998).

[49] N. ESWAR, D. ERAMIAN, B. WEBB, M. Y. SHEN, and A. SALI, *Methods Mol Biol* **426**, 145 (2008).

[50] N. ESWAR, B. WEBB, M. A. MARTI-RENOM, M. S. MADHUSUDHAN, D. ERAMIAN, M. Y. SHEN, U. PIEPER, and A. SALI, *Curr Protoc Bioinformatics* **Chapter 5**, Unit 5 6 (2006).

[51] N. ESWAR, B. WEBB, M. A. MARTI-RENOM, M. S. MADHUSUDHAN, D. ERAMIAN, M. Y. SHEN, U. PIEPER, and A. SALI, *Curr Protoc Protein Sci* **Chapter 2**, Unit 2 9 (2007).

[52] M. A. MARTI-RENOM, A. C. STUART, A. FISER, R. SANCHEZ, F. MELO, and A. SALI, *Annu Rev Biophys Biomol Struct* **29**, 291 (2000).

[53] A. SALI and T. L. BLUNDELL, *J Mol Biol* **234**, 779 (1993).

[54] D. A. CASE, T. E. CHEATHAM, T. DARDEN, H. GOHLKE, R. LUO, K. M. MERZ, A. ONUFRIEV, C. SIMMERLING, B. WANG, and R. J. WOODS, *J Comput Chem* **26**, 1668 (2005).

[55] D. CASE, T. DARDEN, T. CHEATHAM III, C. SIMMERLING, J. WANG, R. DUKE, R. LUO, R. WALKER, W. ZHANG, and K. MERZ, *University of California, San Francisco* **1**, 3 (2012).

[56] D. A. PEARLMAN, D. A. CASE, J. W. CALDWELL, W. S. ROSS, T. E. CHEATHAM, S. DEBOLT, D. FERGUSON, G. SEIBEL, and P. KOLLMAN, *Computer Physics Communications* **91**, 1 (1995).

[57] R. SALOMON-FERRER, A. W. GOTZ, D. POOLE, S. LE GRAND, and R. C. WALKER, *Journal of Chemical Theory and Computation* **9**, 3878 (2013).

[58] M. FEIG and C. L. BROOKS, *Current Opinion in Structural Biology* **14**, 217 (2004).

[59] A. W. GOTZ, M. J. WILLIAMSON, D. XU, D. POOLE, S. LE GRAND, and R. C. WALKER, *Journal of Chemical Theory and Computation* **8**, 1542 (2012).

[60] J. MONGAN, C. SIMMERLING, J. A. MCCAMMON, D. A. CASE, and A. ONUFRIEV, *Journal of Chemical Theory and Computation* **3**, 156 (2007).

[61] K. A. BEAUCHAMP, Y. S. LIN, R. DAS, and V. S. PANDE, *Journal of Chemical Theory and Computation* **8**, 1409 (2012).

[62] V. HORNAK, R. ABEL, A. OKUR, B. STROCKBINE, A. ROITBERG, and C. SIMMERLING, *Proteins-Structure Function and Bioinformatics* **65**, 712 (2006).

[63] D. W. LI and R. BRUSCHWEILER, *Journal of Physical Chemistry Letters* **1**, 246 (2010).

[64] D. W. LI and R. BRUSCHWEILER, *Angewandte Chemie-International Edition* **49**, 6778 (2010).

[65] S. A. SHOWALTER and R. BRUSCHWEILER, *Journal of the American Chemical Society* **129**, 4158 (2007).

[66] S. A. SHOWALTER and R. BRUSCHWEILER, *Journal of Chemical Theory and Computation* **3**, 961 (2007).

[67] J. P. RYCKAERT, G. CICCOTTI, and H. J. C. BERENDSEN, *Journal of Computational Physics* **23**, 327 (1977).

[68] G. GRIGORYAN and W. F. DEGRADO, *J Mol Biol* **405**, 1079 (2011).

[69] D. A. KONDRASHOV, W. ZHANG, R. ARANDA, B. STEC, and G. N. PHILLIPS, *Proteins-Structure Function and Bioinformatics* **70**, 353 (2008).

[70] M. DELORENZI and T. SPEED, *Bioinformatics* **18**, 617 (2002).

[71] M. LU, W. SHU, H. JI, E. SPEK, L. WANG, and N. R. KALLENBACH, *J Mol Biol* **288**, 743 (1999).

[72] J. T. PELTON and L. R. MCLEAN, *Anal Biochem* **277**, 167 (2000).

[73] G. OFFER and R. SESSIONS, *J Mol Biol* **249**, 967 (1995).

[74] A. LUPAS, M. VAN DYKE, and J. STOCK, *Science* **252**, 1162 (1991).

[75] O. D. MONERA, C. M. KAY, and R. S. HODGES, *Biochemistry* **33**, 3862 (1994).

[76] L. LINS, A. THOMAS, and R. BRASSEUR, *Protein Sci* **12**, 1406 (2003).

[77] P. J. LAMONT, W. WALLEFELD, D. HILTON-JONES, B. UDD, Z. ARGOV, A. C. BARBOI, C. BONNEMAN, K. M. BOYCOTT, K. BUSHBY, A. M. CONNOLLY, N. DAVIES, A. H. BEGGS, G. F. COX, J. DASTGIR, E. T. DECHENE, R. GOODING, H. JUNGBLUTH, N. MUELAS, J. PALMIO, S. PENTTILA, E. SCHMEDDING, T. SUOMINEN, V. STRAUB, C. STAPLES, P. Y. VAN DEN BERGH, J. J. VILCHEZ, K. R. WAGNER, P. G. WHEELER, E. WRAIGE, and N. G. LAING, *Hum Mutat* **35**, 868 (2014).

[78] M. BUVOLI, A. BUVOLI, and L. A. LEINWAND, *J Mol Biol* **415**, 807 (2012).

Chapter 7

Concluding Remarks

To understand the function of biological macromolecules and the mechanism through which they function, structural information and native states are required. Obtaining information on the conformational ensemble from experimental approaches is challenging. Computational tools, especially when combined with experimental data, is essential for understanding the conformational dynamics of proteins. Motivated by the goal of understanding the relationship between structure and function, we employed molecular dynamics technique at various resolutions to explore of the conformational states accessible to the macromolecules in conjunction with experimental data from our collaborators. This thesis integrates the projects toward this end.

In Chapter 2, the conformational duality of lymphotactins is investigated. Human lymphotactin (hLtn) is a metamorphic protein that adopts a monomer (hLtn10) fold at 10 °C with 0.20 M salt, and a different conformation that forms a dimer (hLtn40) at 40 °C in the absence of salt. The interconversion is slow and requires complete disruption of the native contacts. Despite this interesting system, the mechanism of the interconversion and all of the important structural contacts are not well understood. Thus, my goal was, first, to identify additional physical interactions and structural motifs that govern the relative stabilities of the hLtn10 and hLtn40 species through extended simulations with both implicit and explicit solvent. Discovering these interactions can help regulate the hLtn conversion rates, through stabilization (or destabilization) of either fold. The interconversion is partially driven by salt concentration, hence the electrostatic interactions are essential for the stability. In this system, it is challenging to achieve equilibration of electrostatic interactions. This system requires extensive sampling to be able identify charged pairs that potentially play a role in maintaining the stability of the system. To address this challenge, I have designed simulations with explicit and implicit

solvation methods, varying the conditions and simulations with mutations. We have found evidence that Arg 23 and Arg 43 are repulsive to one another, with two positive charges in close proximity. My simulations indicate that the removal of the charges on these residues or alternatively Cl- binding can improve the stability the hLtn10 species. In addition to the Arg 23–Arg 43 interaction, we find several other like-charge pairs that are likely important to the stability of hLtn10. hLtn40 features a larger number of salt bridges, and several hydrophobic residues undergo major changes in the solvent accessible surface area between hLtn10 and hLtn40, pointing to their importance to the relative stability of the two folds. Free energy perturbation calculations are carried out to formally evaluate the contribution of the Arg 23–Arg 43 interaction to the stability of hLtn10. In line with the experimental findings from Prof. Brian Volkman's lab at Marquette University, we found that removal of charges from Arg 23 and Arg 43 stabilize the hLtn10 structure. We have also shown that for the hLtn10 system, using the FEP method along with the BAR protocol provides more robust results than the linear response approximation (LRA). In this work, we have also tested the applicability of Generalized Born (GB) implicit solvation method. Contrary to the expectation that the GB method does not perform well for systems mediated through electrostatic and hydrophobic interactions, our results using this method were in good agreement with the atomistic simulations at room temperature. However, at higher temperatures it appeared that the particular GB model used here does not generate an accurate representation for the stability of several structural element.

Ubiquitin (Ub) is a small protein that can be attached to target proteins as either a single molecule or as oligomers. Deubiquitinases can reverse ubiquitination or trim oligomers of Ubs. However the exact mechanism by which DUBs catalyze the removal of Ub is unknown. This lack of clarity is due in part to a lack of structural information for Ub-chains and also due to the complex structures of Ubs with DUBs. In this project, we focused on determining conformational ensembles for Ub-dimers using MD and SAXS data. Our approach entails incorporating SAXS data into MD simulations to aid filtering out the irrelevant conformational ensembles. In Chapter 3, the structural characterization of ubiquitin dimers is summarized. Here, we compare the structure and function of Ub dimers bearing native and non-native linkages. The SAXS profiles calculated over atomistic molecular dynamics simulations exhibit a wide array of overall shapes, in accordance with the known structural diversity for the K48-linked Ub-dimers. After comparing the experimental SAXS profiles for native (isopeptide) and thiolene K48-linked Ub2, the average conformation for both models did not

reflect either of the extremes: the most compact or extended structures. This finding agrees well with recent in-solution NMR and FRET measurements [1, 2] reported that K48-linked dimer exists in a multi-state energy landscape with a number of intermediate conformers in addition to the compact and extended states. [3].

We have extended this work to structures of Ub trimers (See Appendix B) to reveal the differences between various trimers with the aim to understand the differences in catalytic activity. Our trimer data will be important to further investigate the structural characteristics and specificities of complex Ub structures between DUBs and Ub moieties. These data will be compared to experimental data provided by our collaborators from Prof. Eric Strieter's group at UW–Madison. With both computational and experimental data in hand, we aim to elucidate the mechanism by which DUBs recognize substrates specifically and catalyze the removal of Ub.

The myosin rods fold into a coiled-coil repeat and direct thick filament assembly in the sarcomere. This coiled-coil is interrupted by the insertion of four skip residues that introduce discontinuities in the structure. In Chapter 4, we report that the regions surrounding the first three skip residues share high structural similarity despite their low sequence homology. Near each of these skip residues, the coiled-coil transitions to a non-close-packed structure inducing local relaxation of the super-helical pitch. Moreover, molecular dynamics suggest that these distorted regions can assume different conformationally stable states. In contrast to the first three skip residues, the last skip residue induces a true molecular hinge, providing C-terminal rod flexibility. This finding is the first evidence that the myosin rod requires a high degree of conformational flexibility to accommodate an antiparallel assembly. This flexibility ultimately leads to the construction of the bare zone.

We next asked whether the skip residues could be interchanged with each other. In cell culture, we found that each skip residue is not functionally equivalent in promoting the myosin assembly. Furthermore, the function of each skip residue is influenced by its positioning along the rod. In addition, we found that Skip 3 has a greater effect on the assembly of sarcomeres compared to Skips 1 or 2. Deletion of Skip 3 caused both aggregation of the myosin thick filament in the cytoplasm and lowered incorporation into the sarcomere. Assembly of myosin with mutated skip residues in cardiomyocytes shows that the functional importance of each skip residue is associated with rod position and reveals the unique role of the molecular hinge in promoting myosin antiparallel packing. By defining the biophysical properties of the rod, the structures and molecular dynamic calculations together provide insight into thick filament formation, and highlight the structural differences occurring between the coiled-coils of myosin and the stereotypical tropomyosin. This is the first study

reporting that these regions fold into specialized structures engaged in promoting proper myosin assembly into the thick filaments. In addition to extending our knowledge into the conformational and biological properties of coiled-coil discontinuities, the molecular characterization of the four myosin skip residues also provides a guide to modeling the effects of rod mutations associated with cardiac and skeletal myopathies.

Sarcomeric myosins have the remarkable ability to form regular bipolar thick filaments that, together with actin thin filaments, constitute the fundamental contractile unit of skeletal and cardiac muscle. This observation has been known for more than fifty years, and yet a molecular model for the thick filament has not been attained. In part this is due to the lack of a detailed molecular model for the coiled-coil that constitutes the myosin rod. In continuation of our work on the skip region of the light mero myosin (LMM) of the myosin thick filament, in Chapter 5 I summarize our work on testing the feasibility of generating a complete model for the myosin rod by combining overlapping structures of five sections of coiled-coil spanning 164 amino acid residues (20 % of the complete LMM). Each section contains $\sim$7-9 heptads of myosin. The effect of these domains on the stability and conformation of the myosin rod was examined through homology modeling, molecular dynamics and biophysical studies carried out in the Rayment lab. Here, a computational approach was developed to combine the sections into a contiguous model. The structures were aligned, trimmed to form a contiguous model, and simulated for $>$700 ns through MD to remove the discontinuities and achieve an equilibrated conformation that represents the native state. This experimental and computational strategy forms the basis for building a model for the entire LMM region of the myosin rod.

7.1 References

[1] K. BERLIN, C. A. CASTANEDA, D. SCHNEIDMAN-DUHOVNY, A. SALI, A. NAVA-TUDELA, and D. FUSHMAN, *J Am Chem Soc* **135**, 16595 (2013).

[2] Y. YE, G. BLASER, M. H. HORROCKS, M. J. RUEDAS-RAMA, S. IBRAHIM, A. A. ZHUKOV, A. ORTE, D. KLENERMAN, S. E. JACKSON, and D. KOMANDER, *Nature* **492**, 266 (2012).

[3] Y. WANG, C. TANG, E. WANG, and J. WANG, *PLoS Comput Biol* **10**, e1003691 (2014).

APPENDICES

Appendix A

Supporting Information for Human Lymphotactins

Due to the fairly large solvent box used for the explicit solvent simulations (due to the disordered N- and C-termini), it is expensive to run explicit solvent simulations at the microsecond time scale. Therefore, we explore conformational sampling by comparing RMSD values among independent explicit solvent simulations (Fig.A1) and among implicit/explicit solvent simulations (Fig. A2). These plots indicate that generally similar conformations are sampled in those simulations. For example, the pairwise RMSD value between explicit and implicit s10 simulations peaks in the middle of the implicit solvent simulations, suggesting that the trajectories do not diverge despite the fact that implicit solvent simulations are substantially longer (600 – 800 ns). This conclusion is further supported by the similarity in the dominant clusters from explicit and implicit solvent simulations for both hLtn10 and hLtn40 (compare Figs. A3, A4 with Fig. A5).

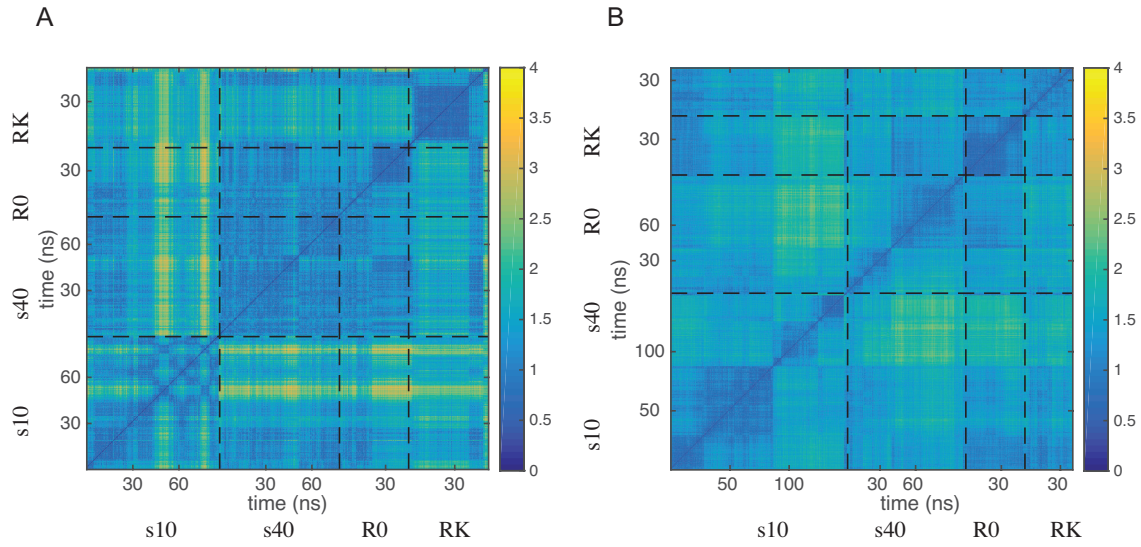


Figure A1: Pairwise Cα-RMSD of (A) hLtn10 based on residues 25-60 and (B) hLtn40 based on residues 11-51. RMSD is calculated over a pool of trajectories formed by WT at s10 condition, n40 condition, *RØ*; (s10) and RK (s10).

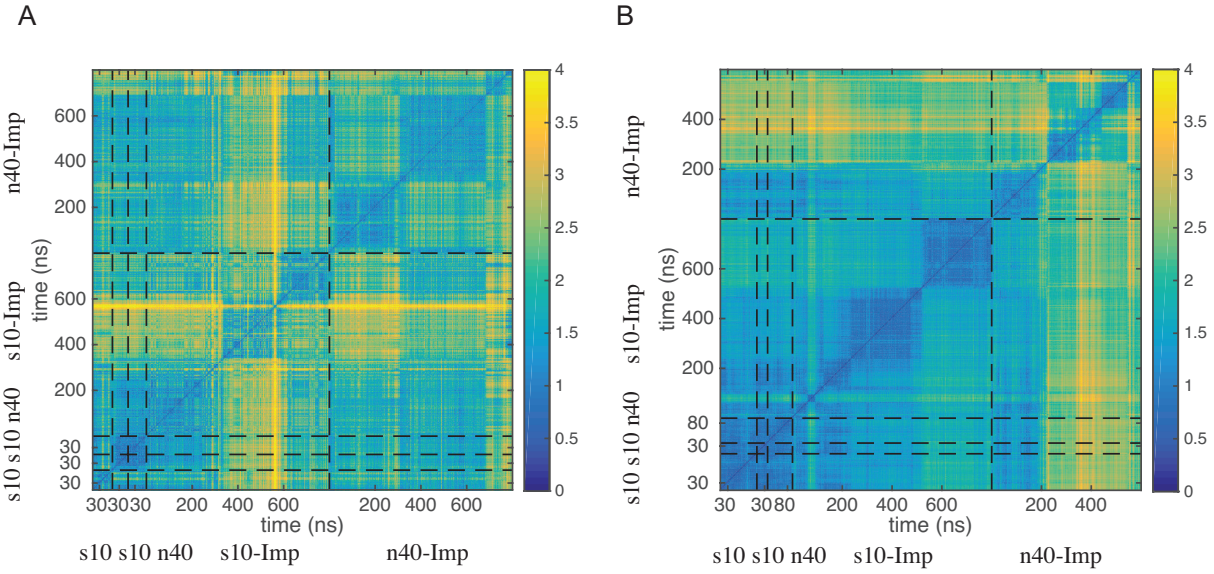


Figure A2: Pairwise Cα-RMSD of (A) hLtn10 based on residues 25-60 as the structure is truncated to residues 60 for implicit solvent simulations and (B) hLtn40 based on residues 11-51. RMSD is calculated over a pool of trajectories formed by two trajectories of WT at s10 condition, n40 condition with explicit solvent and with implicit solvent simulations under s10 and n40 conditions.

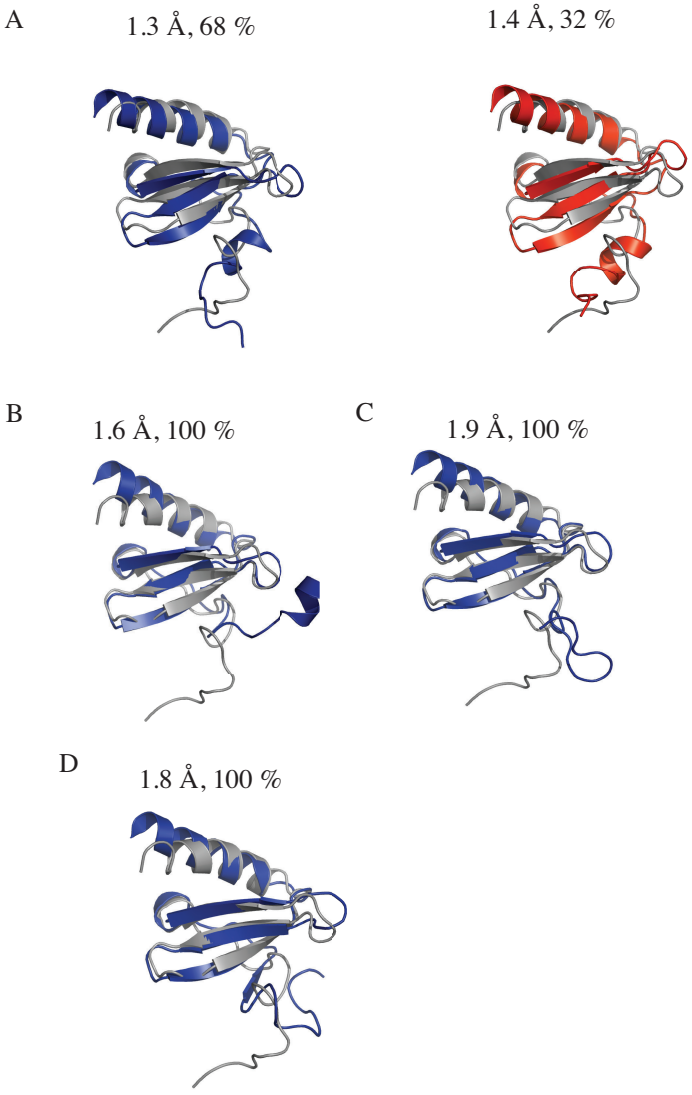


Figure A3: Representative members of each cluster aligned on the NMR structure 1J9O for explicit solvent simulations of (A) WT hLtn10, s10; (B) WT hLtn10, n40; (C) hLtn10 *RØ*, s10; (D) hLtn10 RK, s10. RMSD values with respect to the NMR structure (excluding the tail regions) is shown on the top of each structure (C-terminal tail of hLtn10 structure is truncated for the clarity of display).

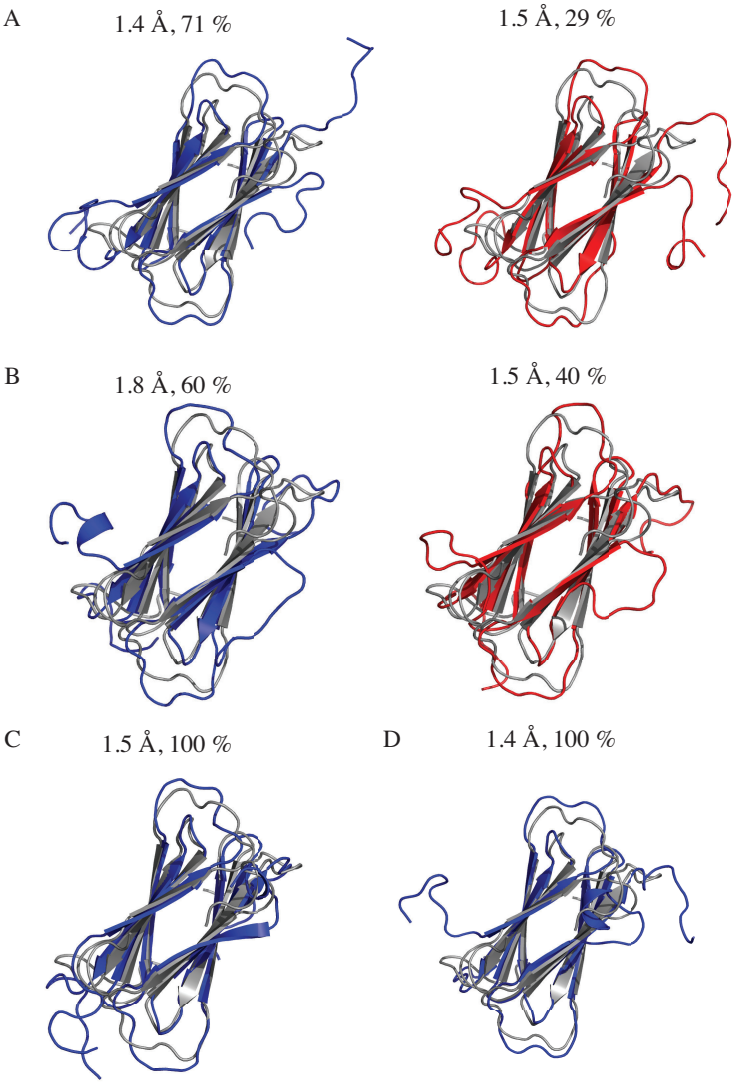


Figure A4: Representative members of each cluster aligned on the average of NMR structure 2JP1 for explicit solvent simulations of (A) WT hLtn40, s10; (B) WT hLtn40, n40; (C) hLtn40 *RØ*, s10; (D) hLtn40 RK, s10. RMSD values with respect to the NMR structure is shown on the top of each structure.

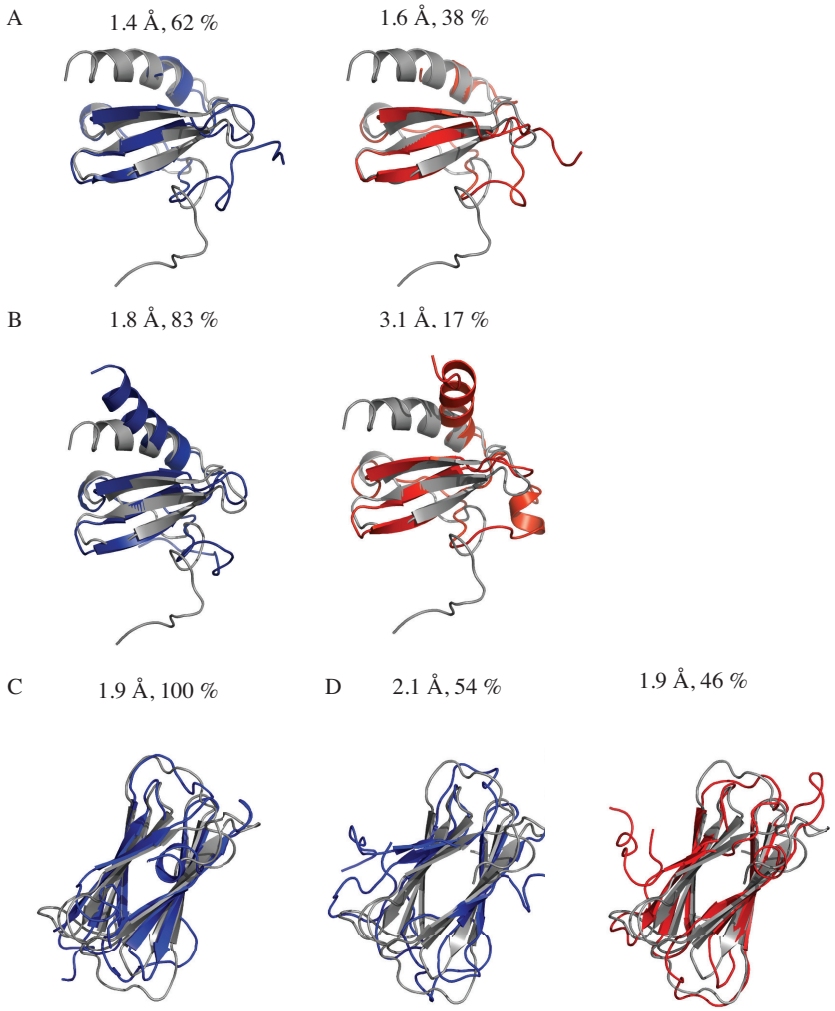


Figure A5: Representative members of each cluster aligned on the average of NMR structure 2JP1 for gb7 simulations of (A) WT hLtn10, s10; (B) WT hLtn10, n40; (C) WT hLtn40, s10; (D) WT hLtn40, n40. RMSD values with respect to the NMR structure (excluding the tail regions) is shown on the top of each structure.

Appendix B

Supporting Information for Comparison of Native and Non-native Ubiquitin Dimers through Molecular Dynamics and Small Angle X-Ray Scattering Reveals Analogous Structures

B.1 Additional Data on Determination of Conformational Ensembles via Molecular Dynamics and SAXS Data for the K48- and K63-linked Ubiquitin Dimers

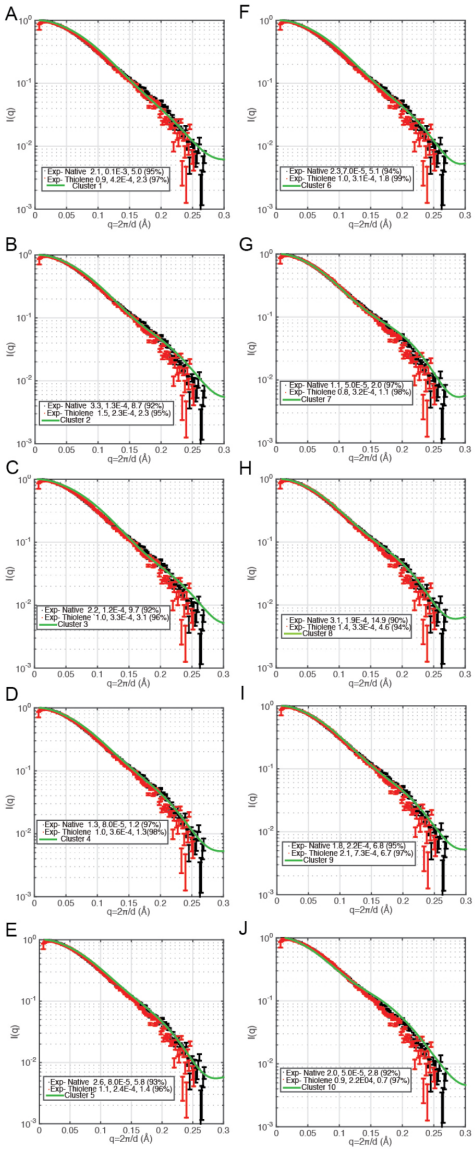


Figure B1: Comparison of SAXS profiles and Pair Distribution Functions for each cluster for K48-linked Ub2. The lusters are obtained from a pool of trajectories for the native dimer that started with 2KDF and the two thiolene based dimer simulations. Calculated SAXS profiles and $P(r)$ curves are shown (green) along with experimental SAXS profiles for the native (black) and TEC-derived (red) K48-dimer. The scores of the SAXS profile of each cluster with respect to the native and TEC-derived dimer are listed in the legend of each panel. Percentage of overlap over the $P(r)$ curves is listed in parenthesis in the legend.

Table B1: Fitted populations of the Native and TEC-derived K48-linked Ub2 based on the pool of conformations obtained from atomistic simulations and matching experimental SAXS profiles.

		Native K48 Ub2				TEC-derived K48 Ub2			
Cluster[a]	R_g (Å)	Fitted Population Population (%)	Score 1[a] (χ_1)	Score 2[b] ($\chi^2_2 <$ 1E-03)	Score 3[c] ($\chi^2_3 < 20$)	Fitted Population Population (%)	Score 1[a] (χ_1)	Score 2[b] ($\chi^2_2 <$ 1E-03)	Score 3[c] ($\chi^2_3 < 20$)
1	17.9 ± 0.2	0	2.1	1.00E-04	5.0	0	0.9	4.20E-04	2.3
2	17.1 ± 0.7	0	3.1	1.90E-04	14.9 0	1.4	3.30E-04	4.6	
3	17.1 ± 0.6	0	2.2	1.20E-04	9.7	0	1.0	3.30E-04	3.1
4	17.5 ± 0.6	0	3.3	1.30E-04	8.7	0	1.5	2.30E-04	2.3
5	18 ± 0.4	0	2.3	7.00E-05	5.1	0	1.0	3.10E-04	1.8
6	18.3 ± 0.3	0	1.1	5.00E-05	2.0	0	0.8	3.20E-04	1.1
7	18.3 ± 1.4	0	2.6	8.00E-05	5.8	0	1.1	2.40E-04	1.4
8	18.5 ± 1.2	25	1.3	8.00E-05	1.2	28	1.0	3.60E-04	1.3
9	18.5 ± 0.9	45	2.0	5.00E-05	2.8	72	0.9	2.20E-04	0.7
10	21.2 ± 1.8	30	1.8	2.20E-04	6.8	0	2.1	7.30E-04	6.7

[a] Ten clusters from K48-Ub2 MD simulations only, numbered according to R_g values. [b] Equation (2) in the main text.

[c] Equation (3) in the main text.

[d] Equation (4) in the main text.

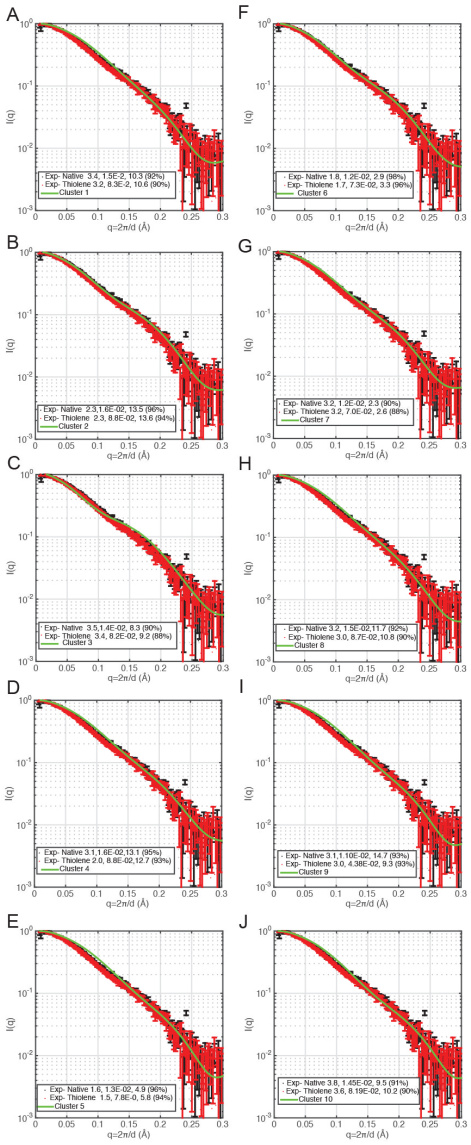


Figure B2: Comparison of SAXS profiles and Pair Distribution Functions for each cluster for K64-linked Ub2. The clusters are obtained from a pool of trajectories for the native dimer that started with 2JF5 [1], 2RR9 and 3A1Q [2] and 3H7P [3], and the two thiolene based dimer simulations. Calculated SAXS profiles and $P(r)$ curves are shown (green) along with experimental SAXS profiles for the native (black) and TEC-derived (red) K48-dimer. The scores of the SAXS profile of each cluster with respect to the native and TEC-derived dimer are listed in the legend of each panel. Percentage of overlap over the $P(r)$ curves is listed in parenthesis in the legend.

Table B2: Fitted populations of the Native and TEC-derived K63-linked Ub2 based on the pool of conformations obtained from atomistic simulations and matching experimental SAXS profiles.

		Native K63 Ub2				TEC-derived K63 Ub2			
Cluster[a]	R_g (Å)	Fitted Population Population (%)	Score 1[a] (χ_1)	Score 2[b] ($\chi^2_2 <$ 1E-03)	Score 3[c] ($\chi^2_3 < 20$)	Fitted Population Population (%)	Score 1[a] (χ_1)	Score 2[b] ($\chi^2_2 <$ 1E-03)	Score 3[c] ($\chi^2_3 < 20$)
1	17.4 ± 0.89	0	3.4	1.50E-02	10.3	0	3.2	8.30E-02	10.6
2	17.5 ± 0.65	0	3.1	1.60E-02	13.1	0	2.0	8.80E-02	12.7
3	17.6 ± 0.78	30	3.2	1.50E-02	11.7	0	3.0	8.70E-02	10.8
4	17.6 ± 0.85	0	2.3	1.60E-02	13.5	0	2.3	8.80E-02	13.6
5	18.3 ± 1.17	0	3.5	1.40E-02	8.3	0	3.4	8.20E-02	9.2
6	18.7 ± 0.37	0	1.6	1.30E-02	4.9	0	1.5	7.80E-02	5.8
7	18.8 ± 0.98	24	1.8	1.20E-02	2.9	73	1.7	7.30E-02	3.3
8	18.8 ± 0.71	0	3.8	1.45E-02	9.5	0	3.6	8.19E-02	10.2
9	19.9 ± 2.03	12	3.2	1.20E-02	2.3	0	3.2	7.00E-02	2.6
10	23.7 ± 1.65	34	3.1	1.10E-02	14.7	27	3.0	4.38E-02	9.3

[a] Ten clusters from K63-Ub2 MD simulations only, numbered according to R_g values. [b] Equation (2) in the main text.

[c] Equation (3) in the main text.

[d] Equation (4) in the main text.

B.2 Determination of Conformational Ensembles from a Library of Available Structures and Experimental SAXS Data for the K48-linked Ub2

To compare the experimental data to a library of extant structures, we compiled 74 X-ray and NMR structures of free Ub2 (PDBs: 1AAR [4], 2PE9 [5], and 3AUL [6], 2BGF [7], 3M3J [8], 3NS8 [9]), and NMR structures of Ub2 bound to UBA2 from hHR23a (PDB: 1ZO6 [10]), UIM1 and UIM2 from S5a (PDBs: 2KDE and 2KDF [11]), CUE domain from the E3 ligase gp78 (PDBs: 2LVQ and 2LVP [12]), and the DUB OTUB1 (PDB: 4DDI [13]). 2KDE and 2KDF each have seven different conformations for K48-linked Ub2. We also extracted 20 different conformations from 2LVP, 24 from 2LVQ, 10 from 1ZO6 and 3 from 4DDI. The calculated SAXS profiles for the 88 structures exhibit a range of conformations from the most compact structure 2PE9 (R_g: 16 Å) to the most extended 2KDF-6 (R_g: 23.9 Å). We organized the library of structures by R_g values and RMSD relative 2PE9, 2KDF-6, and the intermediate conformation 2LVP-12 (Fig. B4A). Differences in conformations can also be evaluated by comparing normalized $P(r)$ plots (Fig. B4B). Similar structures, such as 1AAR and 2PE9 (both compact) or 2LVP-13 and 2KDF-6 (both extended) have more than 90 % overlap in their $P(r)$ curves. A decrease in $P(r)$ overlap corresponds to structural differences, and the two most distinct structures, 1AAR and 2KDF-6, only exhibit 40 % overlap (Fig. B3).

Confident that the existing library of structures for K48-linked Ub2 samples broad conformational space, we compared the experimental SAXS patterns for native and TEC-derived Ub2 to those of the 88 extant structures. Both experimental $P(r)$ plots showed similar overlap with plots generated from the set of known structures (Fig. B4C). Minimal ensemble search (MES) analysis identified 2LVQ-24 and 4DDI-3 as the best single-state representation for native and TEC-derived Ub2, respectively. Both structures have intermediate R_g values, 19.3 Å for 2LVQ-24 and 19.0 Å for 4DDI-3 (Fig. B4A, 4.5)

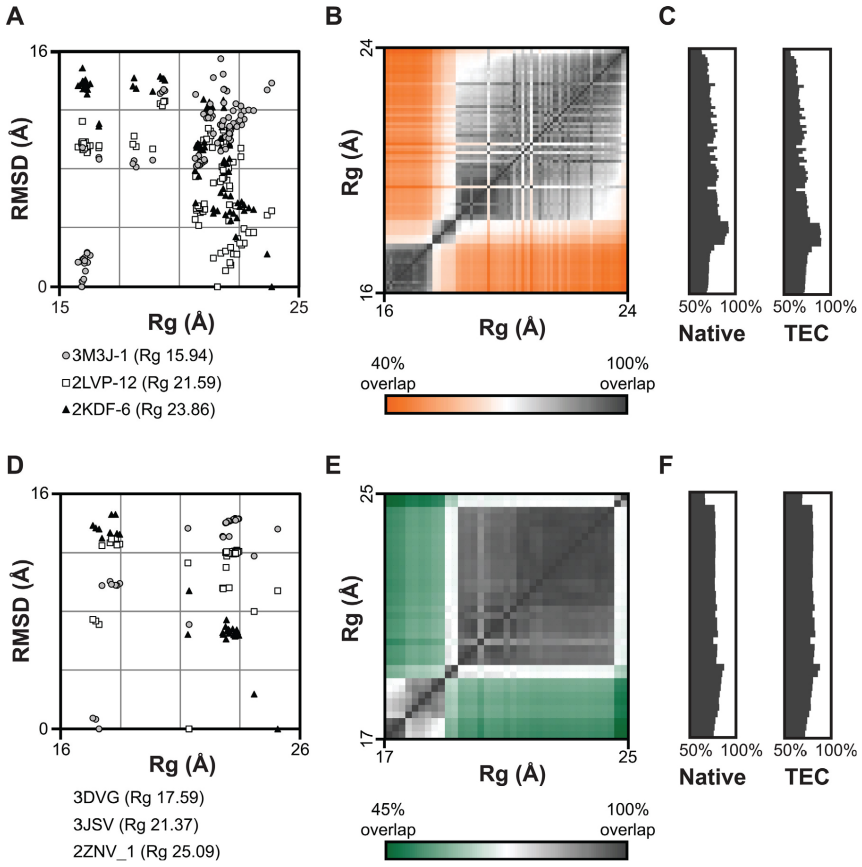


Figure B3: A library of reported K48-linked (top) and K63-linked (bottom) Ub2 structures samples a range of conformations. (A) RMSD calculated for 85 K48-linked Ub2 structures relative to 2PE9 (compact), 2LVP-12 (intermediate) and 2KDF-6 (extended); and 37 K63-linked Ub2 structures relative to 3DVG (compact), 3JSV (intermediate) and 2ZNV-1 (extended). (B) Heat map generated from absolute differences between $P(r)$ curves to show structural variability among reported structures, arranged according to their R_g values. K48-Ub2 (orange/grey), K63-Ub (green/grey). (C) Evaluation of $P(r)$ overlap between native and TEC-derived Ub2 to reported structures presented in the heat map. The extent of overlap (from 50 % to 100 %) is illustrated with bar graphs.

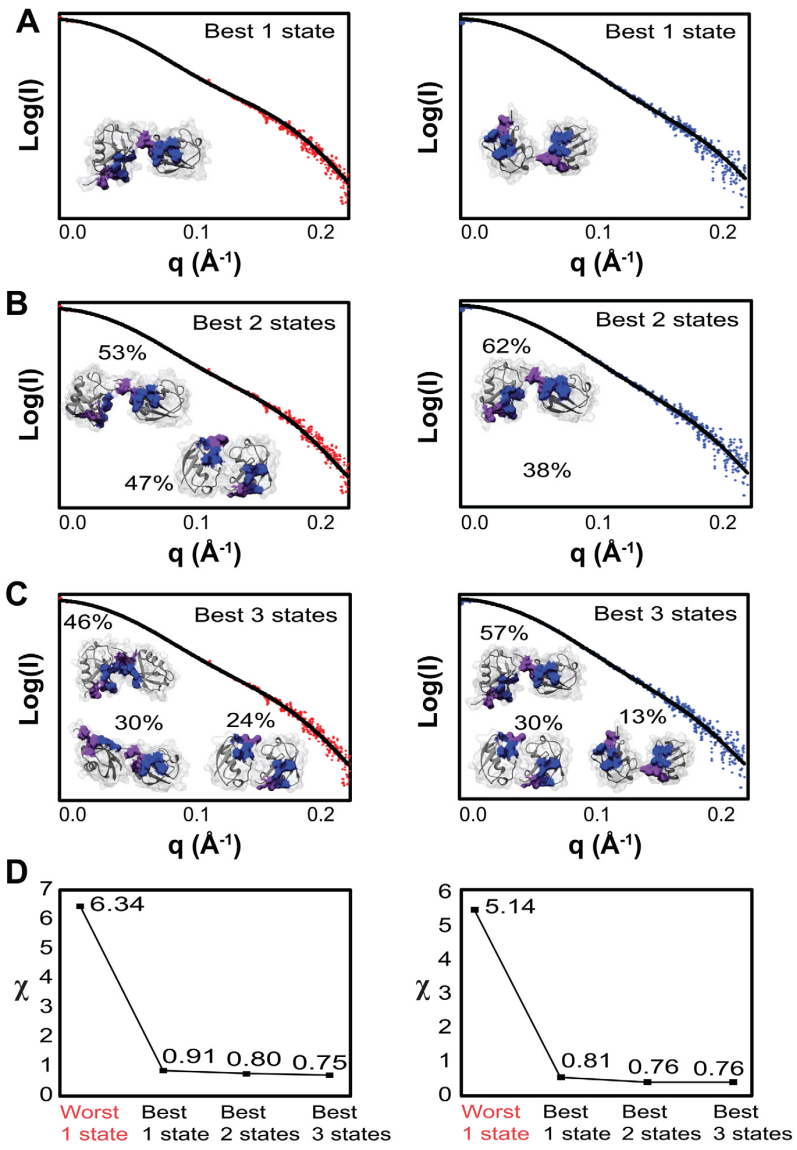


Figure B4: Minimal ensemble search for native (left) and TEC-derived (right) K48-linked Ub2. MES analysis identified (A) 2LVQ- 24 (for native) and 2LVQ-6 (for TEC) as the single best-fit structures; (B) 11 % 3NS8, 89 % 2LVQ-24 (for native) and 39 % 3AUL, 61 % 2LVQ-23 (for TEC) as the best two-state ensemble; (C) 24 % 2LVP-12, 26 % 3NS8, 5 % 2LVQ-21 (for native) and 33 % 2LVQ-24, 32 % 2LVQ-21, 36 % 3NS8 (for TEC) as the best three-state ensemble. (D) To show the range of fitting scores, χ_1 values for the worst fit one-state structure and the best MES fits are plotted. Hydrophobic patches centered on Ile44 and Ile36 are shown in blue and purple, respectively. Structures were visualized in Chimera.

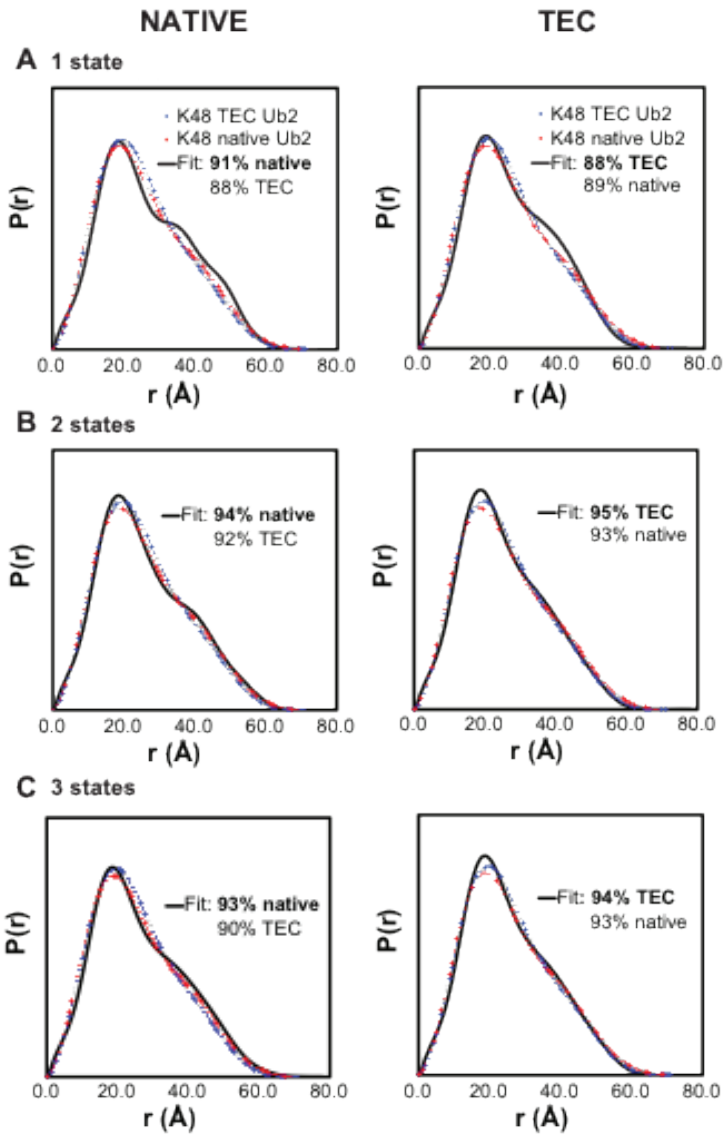


Figure B5: Distance distribution $P(r)$ comparison of K48-linked Ub2 ensembles. $P(r)$ curves for the best 1 state, 2 states and 3-states fits are shown for native (left) and TEC-derived (right) Ub2. (A) The best 1 state for native Ub2 has 91 % $P(r)$ overlap with the native experimental data and 88 % overlap with the non-native experimental data. The best 1 state for TEC-derived Ub2 has similar overlap with both experimental datasets. (B) and (C) The best 2-state and 3-state fits for native and TEC-derived Ub2 have greater than 90 % $P(r)$ overlap with experimental data.

The best 2-state ensembles for native and TEC-derived Ub2 both include a compact structure with a more extended structure (Fig. B4B): 3AUL with 2LVQ-17 (native) or 3NS8 with 2LVQ-23 (TEC). 2LVQ-17 and 2LVQ-23 have R_g values 21.0 Å and 19.4 Å respectively. 3AUL and 3NS8 are nearly identical – they have the same R_g (16.6 Å), and RMSD difference between the two structures is 0.3 Å. For TEC-derived Ub2, incorporating a third component to the ensemble did not enhance the MES fitting score (Figures 6C and 6D), whereas a modest enhancement was calculated for native Ub2 (Fig. B4C and B4D). All best-fit ensemble results have greater than 90 % $P(r)$ overlap with experimental curves for both native and TEC-derived Ub2 (Fig. B3).

B.3 Determination of Conformational Ensembles from a Library of Available Structures and Experimental SAXS Data for the K63-linked Ub2

To compare the experimental data to a library of extant structures, we compiled X-ray structures of free K63-linked Ub2 (PDBs: 2JF5, 3H7P, and 3H7S); X-ray structures of Ub2 bound to the NZF domain of TAB2 (PDBs: 2WWZ, 2WX0, 2WX1, 3A9J), the NZF domain of TAB3 (PDB: 3A9K), the CoZi domain of NEMO (PDB: 3JSV), UIMs from RAP80 (PDB: 3A1Q), a K63-specific antibody (PDBs: 3DVG and 3DVN), and the DUB AMSH (PDB: 2ZNV); and NMR structures of Ub2 bound to UIMs (PDB: 2RR9). The 37 structures range from the most compact structure 3DVN-2 (R_g: 17.3 Å) to the most extended 2ZNV-1 (R_g: 25.1 Å). We organized the library of structures by Rg values and RMSD relative to 3DVG, 2ZNV, and the intermediate conformation 3JSV (Fig. B3D). The diversity in Ub2 conformations were also demonstrated by comparing normalized $P(r)$ plots (Fig. B3E).

With the library of structures, MES analysis identified 3A1Q-1 and 3JSV as the best single-state representation for native and TEC-derived Ub2, respectively. Both structures have intermediate R_g values, 21.4 Å for 3A1Q-1 and 21.3 Å for 3JSV (Fig. B6A). The best two-state ensembles for native and TEC-derived Ub2 both include a relatively compact structure with a more extended structure (Fig. B6B) 2WX1 with 2RR9-17 (native) or 3A9J with 2RR9-15 (TEC). The 2RR9-17 and 2RR9-15 have R_g values 23.0 Å and 23.3 Å , respectively, while 2WX1 and 3A9J have Rg values 17.7 Å and 18.3 Å. For both native and non-native Ub2, incorporating a third component to the ensemble did not enhance the MES fitting score (Fig. B6C and B6D). Overall, both native and TEC-derived K63-linked Ub2 adopt similar shapes, and the non-native Ub2 is slightly more extended.

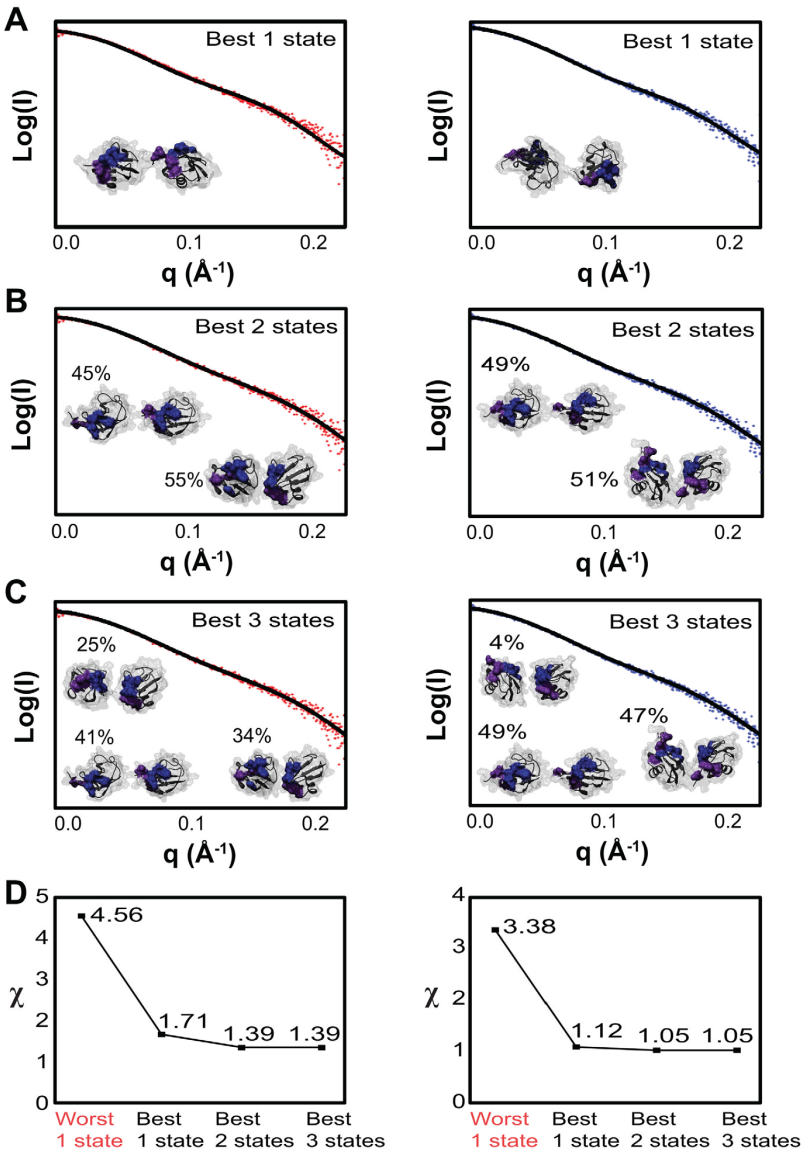


Figure B6: Minimal ensemble search for native (left) and TEC-derived (right) K63-linked Ub2. MES analysis identified (A) 3A1Q (for native) and 3JSV (for TEC) as the single best-fit structures; (B) 55 % 2WX1, 45 % 2RR9 (for native) and 51 % 3A9J, 49 % 2RR9-15 (for TEC) as the best 2-state ensemble; (C) 25 % 2WXO-1, 41 % 2RR9-17, 34 % 2WX1 (for native) and 4 % 2WXO-2, 49 % 2RR9-15, 47 % 3A9J as the best 3-state ensemble. (D) To show the range of fitting scores, χ_1 values for the worst fit 1-state structure and the best MES fits are plotted. Hydrophobic patches centered on Ile44 and Ile36 are shown in blue and purple, respectively. Structures were visualized in Chimera.

B.4 Additional Information on Experimental Methods

B.4.1 Materials

Primers were acquired from Integrated DNA Technologies (Coralville, IA). All restriction and cloning enzymes were from New England Biolabs (Boston, MA). Native K11-linked Ub2 were purchased from Life Sensors (Malvern, PA). OTUB2 was purchased from Boston Biochem (Cambridge, MA).

B.4.2 Cloning and Mutagenesis [14–21]

Lysine to cysteine mutations were introduced at specified sites in the DNA sequence of Ubiquitin (Ub 1-76) using splice overlap extension. Primers containing the TGC mutation were inserted at the desired codon position. D77 was encoded in the reverse primer to afford three constructs (UbKxC-D77; where x represents the position of native lysine residues), which were ligated into a pET22b vector (Novagen). The catalytic domains of AMSH (219-424) and OTUD7B (56-446) were cloned into pGEX-6P-1 and pDP.His.MBP, respectively. Full length USp15 from addgene was cloned into pDP.His.MBP vector. E1 was amplified from the HeLa cDNA library and cloned into pET24a(+). Cdc34 and ubc13/mms2 constructs were obtained from Addgene.

B.4.3 Protein Expression and Purification [14–21]

Ub variants were expressed and purified from RosettaTM 2(DE3)pLysS cells (Novagen) as described. All the enzymes were expressed recombinantly in BL21(DE3) E. coli. Cdc34 and E1, with 6x-histidine (His) tags were purified via cobalt affinity chromatography. Ubc13/mms2 was purified in two steps: nickel affinity chromatography followed by removal of 6xHis by TEV protease and anion exchange chromatography. GST-AMSH (219 - 424) was purified in two steps: glutathione affinity chromatography followed by removal of GST tag by TEV protease and size exclusion chromatography. 6xHis-MBP-OTUD7B was first purified by cobalt affinity chromatography, followed by removal of 6x-MBP by TEV protease, and size exclusion chromatography. 6xHis-MBP-USP15 was similarly purified.

B.5 Determination of Conformational Ensembles using the Available Structures and Experimental SAXS Data

To support our finding from MD simulations, the experimental data is compared to a library of Ub2 structures. We compiled 74 X-ray and NMR structures of free Ub2 (PDBs: 1AAR [4], 2PE9 [5], and 3AUL [6], 2BGF [7], 3M3J [8], 3NS8 [9]), and NMR structures of Ub2 bound to UBA2 from hHR23a (PDB: 1ZO6 [10]), UIM1 and UIM2 from S5a (PDBs: 2KDE and 2KDF [11]), CUE domain from the E3 ligase gp78 (PDBs: 2LVQ and 2LVP [12]), and the DUB OTUB1 (PDB: 4DDI [13]). 2KDE and 2KDF each have seven different conformations for K48-linked Ub2. We also extracted 20 different conformations from 2LVP, 24 from 2LVQ, 10 from 1ZO6 and 3 from 4DDI. The calculated SAXS profiles for the 88 structures exhibit a range of conformations from the most compact structure 2PE9 (R_g: 16 Å) to the most extended 2KDF-6 (R_g: 23.9 Å). We organized the library of structures by R_g values and RMSD relative 2PE9, 2KDF-6, and the intermediate conformation 2LVP-12 (talk). Differences in conformations can also be evaluated by comparing normalized $P(r)$ plots (Fig. B3B). Similar structures, such as 1AAR and 2PE9 (both compact) or 2LVP-13 and 2KDF-6 (both extended) have more than 90 % overlap in their $P(r)$ curves. A decrease in $P(r)$ overlap corresponds to structural differences, and the two most distinct structures, 1AAR and 2KDF-6, only exhibit 40 % overlap.

Confident that the existing library of structures for K48-linked Ub2 samples broad conformational space, we compared the experimental SAXS patterns for native and TEC-derived Ub2 to those of the 88 extant structures. Both experimental $P(r)$ plots showed similar overlap with plots generated from the set of known structures (Fig. B3C). Minimal ensemble search (MES) analysis identified 2LVQ-24 and 4DDI-3 as the best single-state representation for native and TEC-derived Ub2, respectively. Both structures have intermediate R_g values, 19.3 Åfor 2LVQ-24 and 19.0 Åfor 4DDI-3 (Fig. B4A).

The best 2-state ensembles for native and TEC-derived Ub2 both include a compact structure with a more extended structure (Fig. B4B): 3AUL with 2LVQ-17 (native) or 3NS8 with 2LVQ-23 (TEC). 2LVQ-17 and 2LVQ-23 have R_g values 21.0 Åand 19.4 Årespectively. 3AUL and 3NS8 are nearly identical–they have the same R_g (16.6 Å), and RMSD difference between the two structures is 0.3 Å. For TEC-derived Ub2, incorporating a third component to the ensemble did not enhance the MES fitting score, whereas a modest enhancement was calculated for native Ub2 ((Fig. B4C and D). All best-fit ensemble results have greater than

90 % $P(r)$ overlap with experimental curves for both native and TEC-derived Ub2.

B.6 End Point Gel-Based Assays to Measure Hydrolysis Activity of Various DUBs

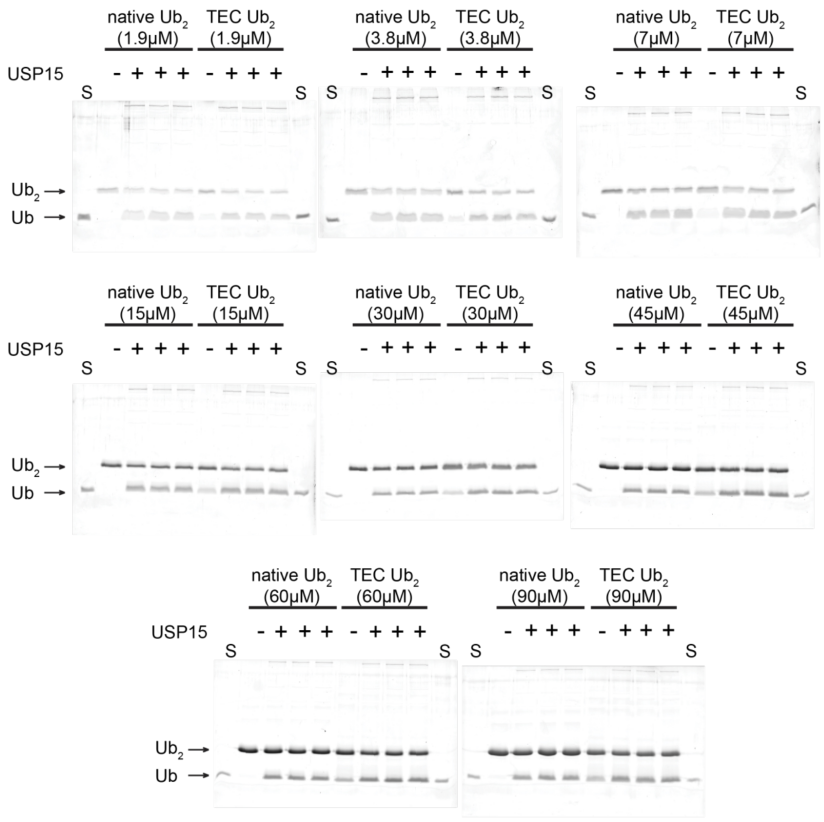


Figure B7: Hydrolysis of native and TEC-derived K6 Ub2 by USP15. End point assays ($time_{final}$ = 1.5 min) using both Ub2 at the indicated range of substrate concentrations and constant USP15 concentration of 200 nM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Two replicates of mono-Ub standard [3.7 μM], indicated as "S", were also included in every gel. (Related to Figure 4.6A.)

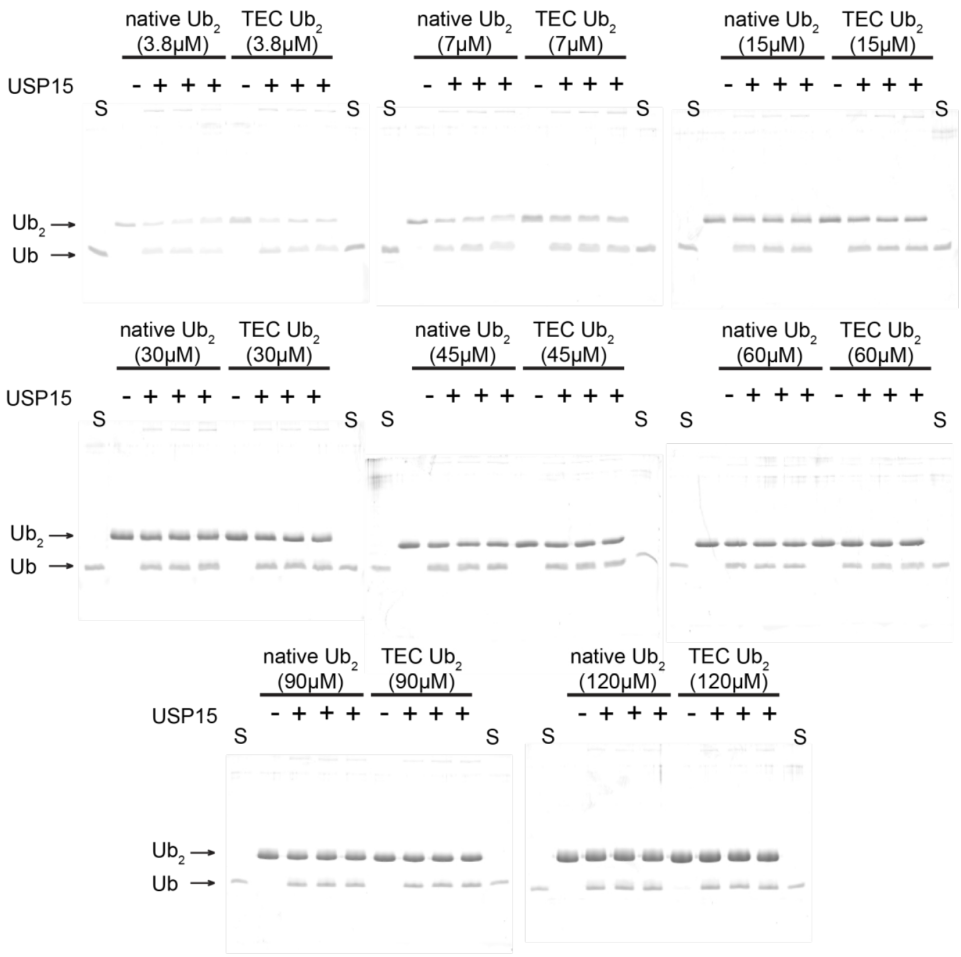


Figure B8: Hydrolysis of native and TEC-derived K48 Ub2 by USP15. End point assays ($time_{final}$ = 1.5 min) using both Ub2 at the indicated range of substrate concentrations and constant USP15 concentration of 200 nM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Two replicates of mono-Ub standard [15 μM], indicated as "S", were also included in every gel. (Related to Figure 4.6A.)

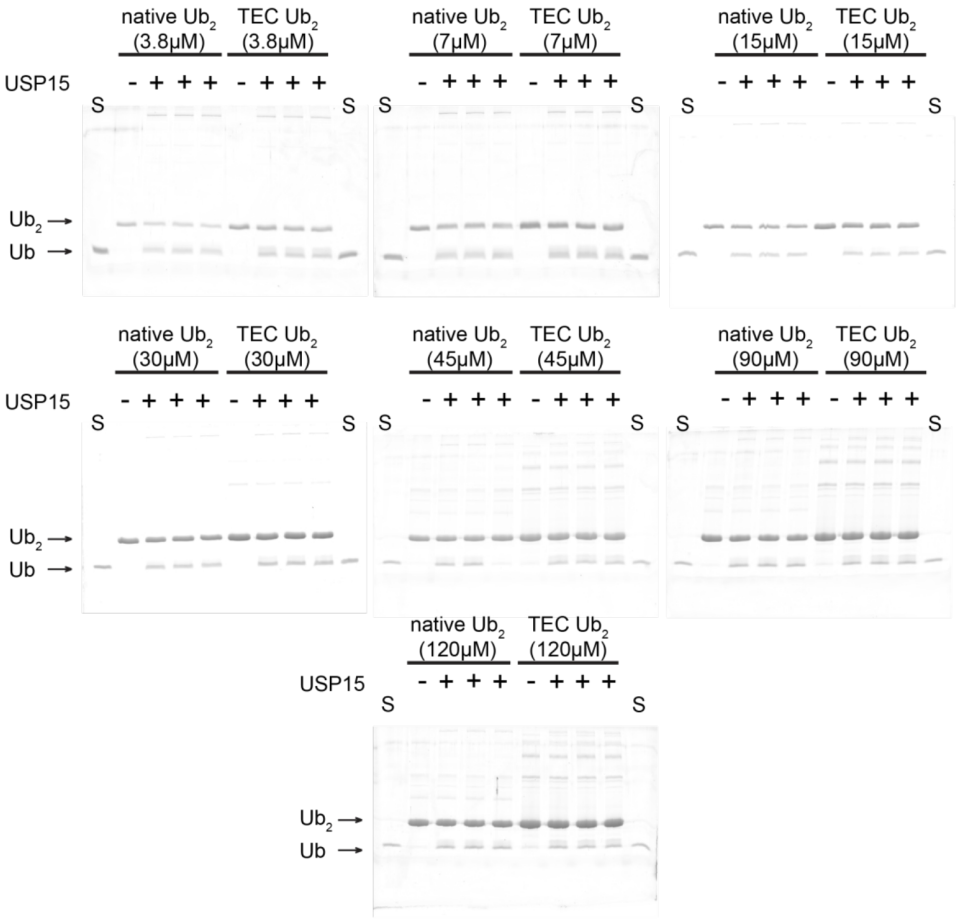


Figure B9: Hydrolysis of native and TEC-derived K63 Ub2 by USP15. End point assays ($time_{final}$ = 1.5 min) using both Ub2 at the indicated range of substrate concentrations and constant USP15 concentration of 200 nM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Two replicates of mono-Ub standard [15 μM], indicated as "S", were also included in every gel. (Related to Figure 4.6A.)

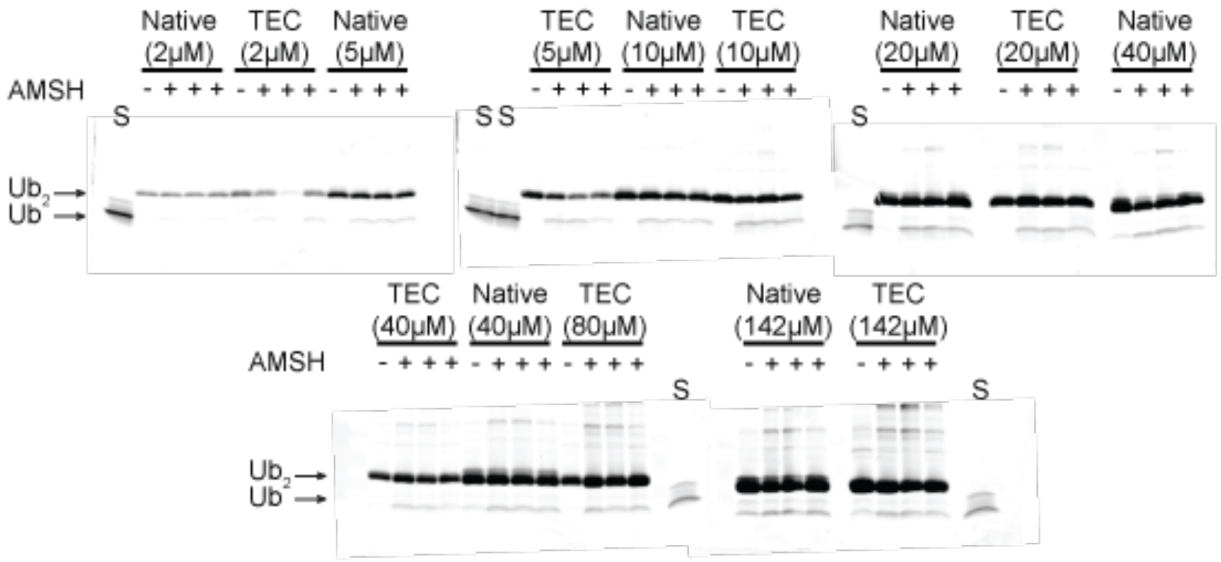


Figure B10: Hydrolysis of native and TEC-derived K63 Ub2 by AMSH. End point assays ($time_{final}$ = 5 min) using both Ub2 at the indicated range of indicated substrate concentrations and constant AMSH concentration of 500 nM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Mono-Ub standard [10.5 μM], indicated as "S", were also included in every gel. (Related to Figure 4.6B.)

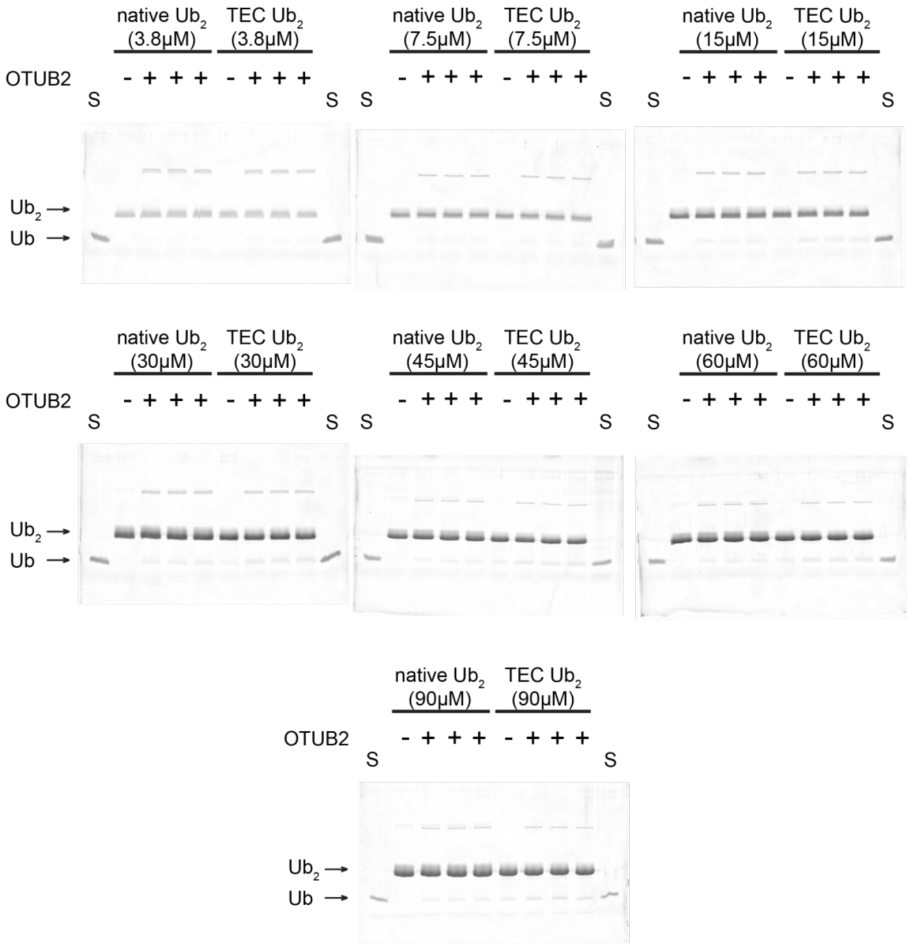


Figure B11: Hydrolysis of native and TEC-derived K48 Ub2 by OTUB2. End point assays ($time_{final}$ = 1.5 min) using both Ub2 at the indicated range of indicated substrate concentrations and constant OTUB2 concentration of 740 nM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Mono-Ub standard [15 μM], indicated as "S", were also included in every gel. (Related to Figure 4.7A.)

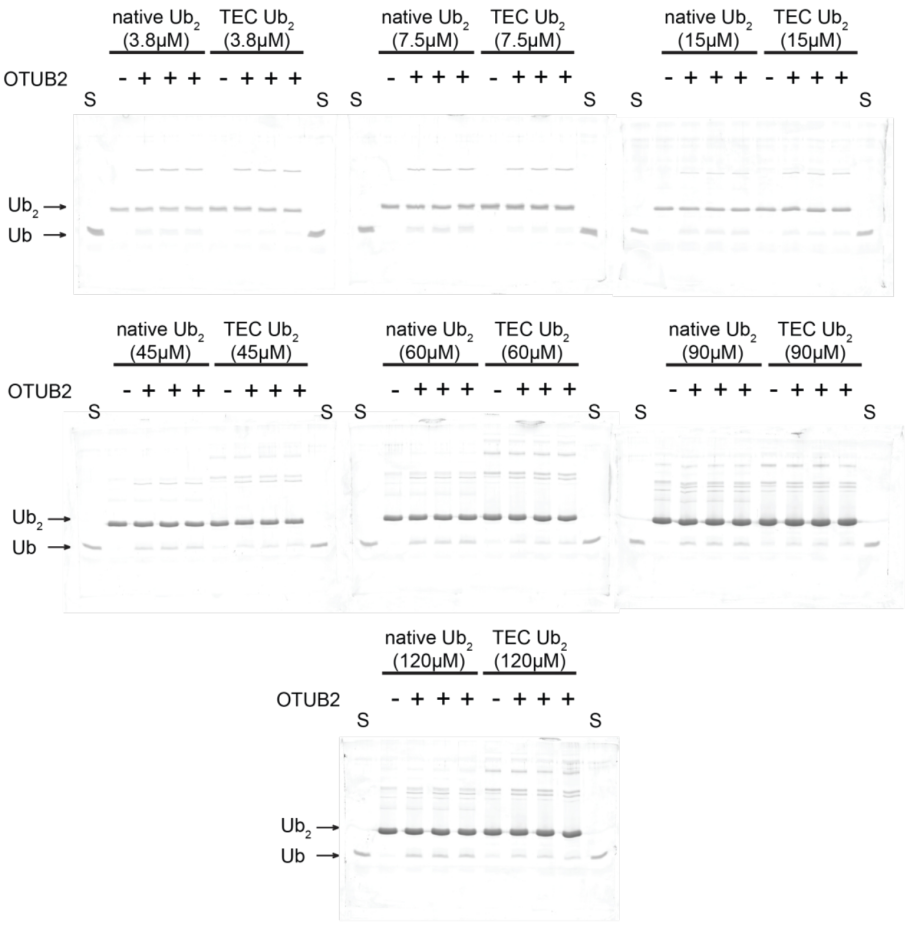


Figure B12: Hydrolysis of native and TEC-derived K48 Ub2 by OTUB2. End point assays ($time_{final}$ = 1.5 min) using both Ub2 at the indicated range of indicated substrate concentrations and constant OTUB2 concentration of 740 nM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Mono-Ub standard [15 μM], indicated as "S", were also included in every gel. (Related to Figure 4.7A.)

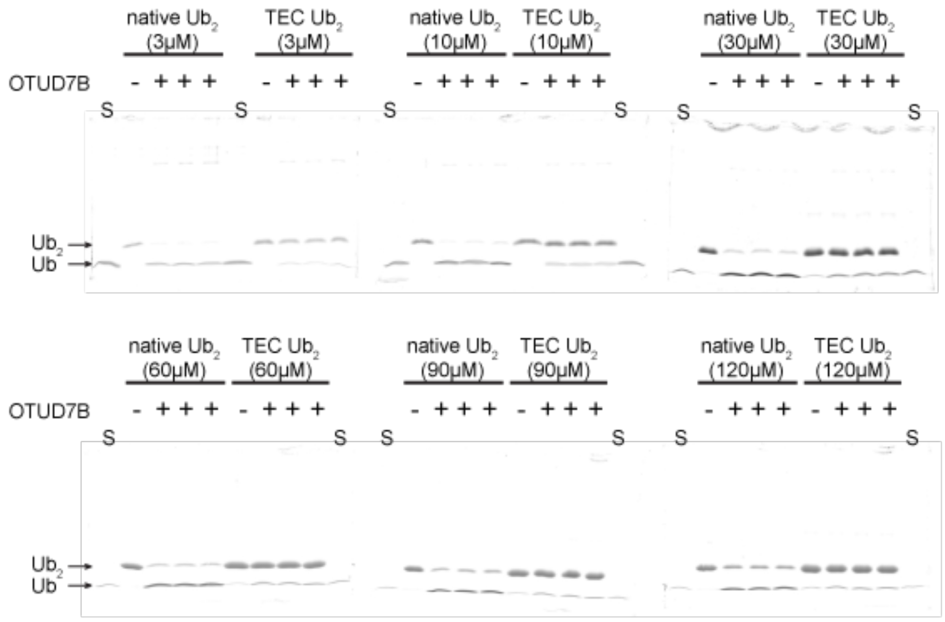


Figure B13: Hydrolysis of native and TEC-derived K11 Ub2 by OTUD7B. End point assays ($time_{final}$ = 2 min) using both Ub2 at the indicated range of substrate concentrations and constant OTUD7B concentration of 200 nM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Two replicates of mono-Ub standard [13.4 μM], indicated as "S", were also run in every gel. (Related to Figure 4.7A.)

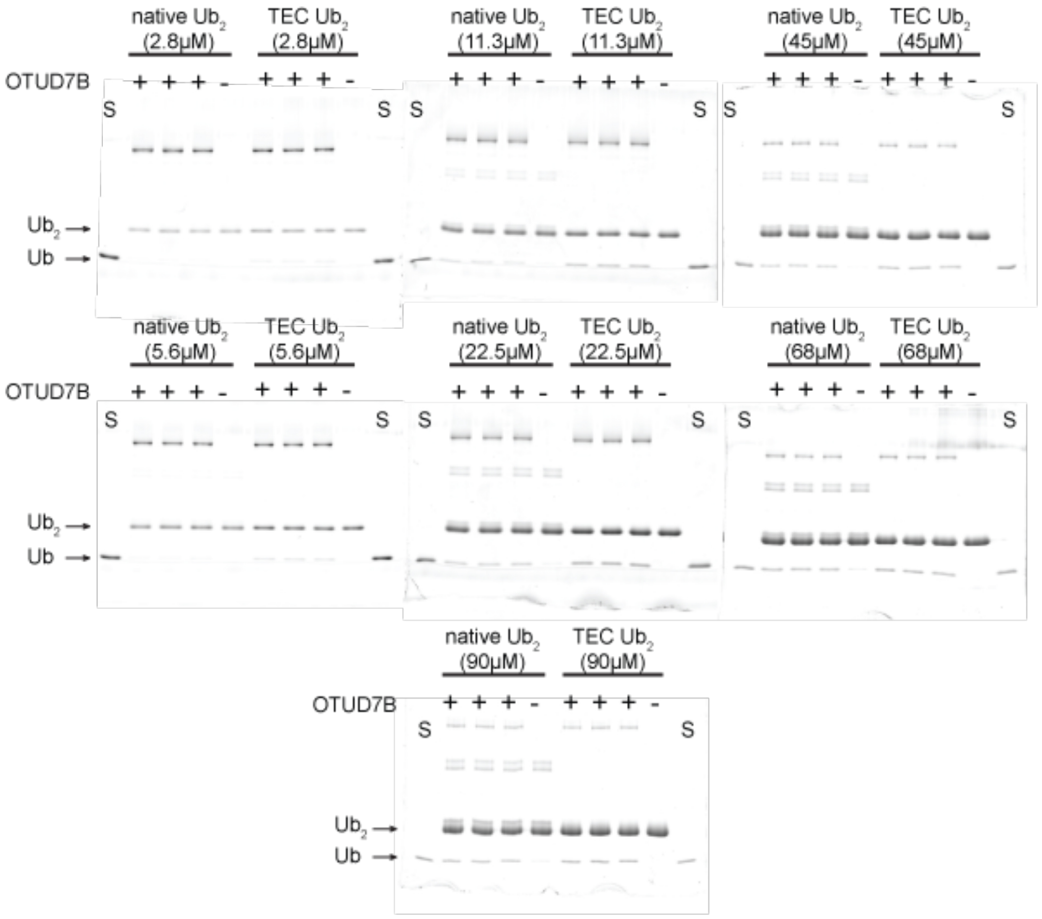


Figure B14: Hydrolysis of native and TEC-derived K48 Ub2 by OTUD7B. End point assays ($time_{final}$ = 2 min) using both Ub2 at the indicated range of substrate concentrations and constant OTUD7B concentration of 2 μM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Two replicates of mono-Ub standard [15.7 μM], indicated as "S", were also included in every gel. (Related to Figure 4.7B.)

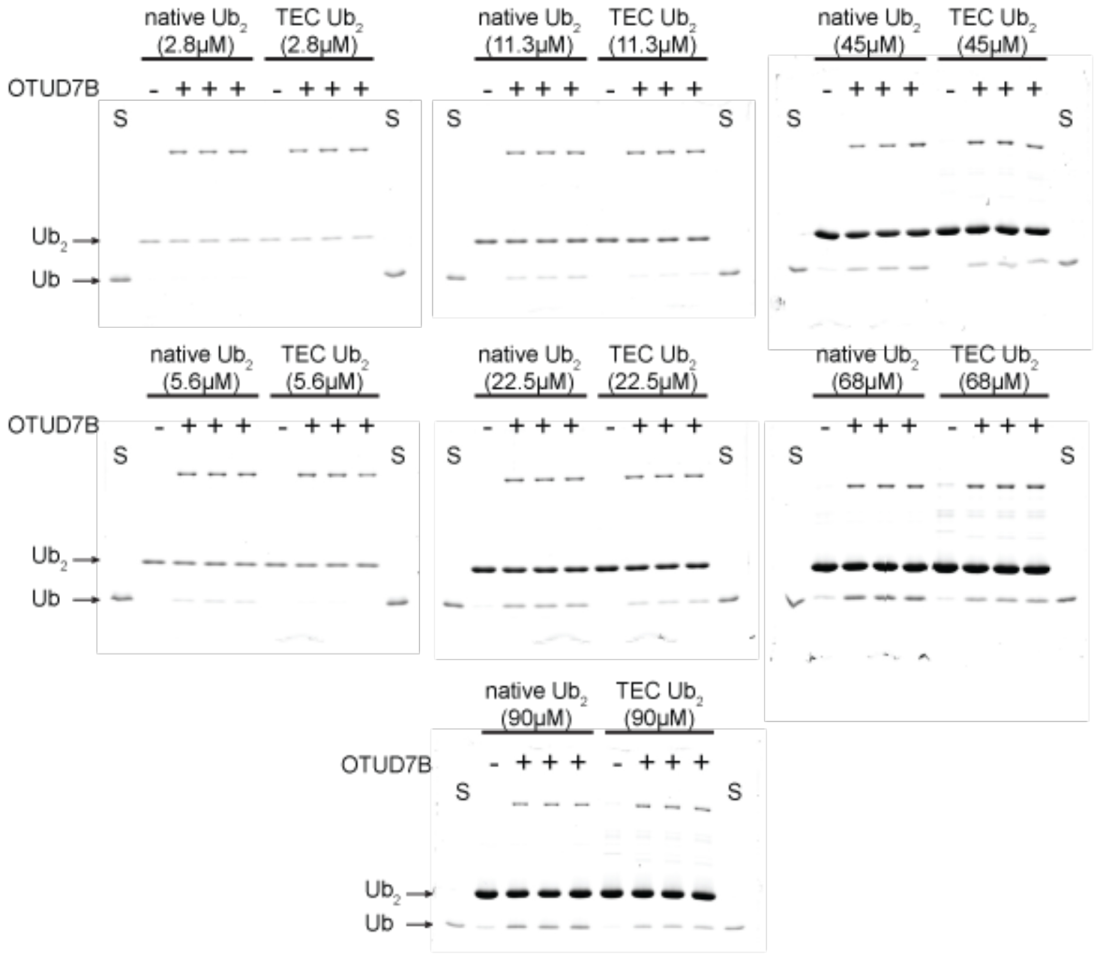


Figure B15: Hydrolysis of native and TEC-derived K63 Ub2 by OTUD7B. End point assays ($time_{final}$ = 2 min) using both Ub2 at the indicated range of substrate concentrations and constant OTUD7B concentration of 2 μM. All experiments were run in quadruplicates: enzyme was added to three replicates and only buffer was added to the remaining sample. Two replicates of mono-Ub standard [15.7 μM], indicated as "S", were also included in every gel. (Related to Figure 4.7B.)

B.7 Characterization of Conformational Ensembles for Ubiquitin Trimers

Ubiquitins can be attached to target proteins as a single molecule and oligomers. Deubiquitinases can reverse ubiquitination or trim oligomers of Ubs. Ubs known to recognize their binding partners with the hydrophobic patch formed by formed by Leu8, Ile44 and Val70 [22, 23]. However the exact mechanism through which DUBs act is not known. This is mainly due to lack or structural information for Ub-oligomers and their complex structures with DUBs. My aim is to determine conformational ensembles using MD and SAXS data.

My approach here for the trimers entails the same procedure applied to dimer; incorporating SAXS data into MD simulations to aid filtering out the irrelevant conformation ensembles. To serve for this purpose, initial structures are created in Modeler Homology Modeling Package [24–29] with the previously-solved structures of Ub dimers of the desired linkages. Next, I perform MD simulations at time scales long enough to ensure sampling diverse conformations ($\sim$1 μs). The list of structures modeled and the simulations carried out are listed below (Table B3).

After having collected a pool of conformations, I apply an Cα-RMSD based clustering technique to distinguish the main conformers. SAXS profiles for each cluster centroid are calculated [30, 31] and scored with the experimental SAXS profile. SAXS fitting consists of two scoring schemes used in previous chapters. Both schemes are derived from mean square error approximation as a measure of deviation from the experimental. Thus, a lower score means a better match to experimental data.

It is important to underline that for chains longer than a dimer, the construction of the Ub chain is important. The Ub chains can form in linear or branched configuration, causing the DUB specificity. In linear configuration, consecutive Ub monomers tag (ubiquinate) the next one, however in branched configuration the central Ub monomer is ubiquinated by two other Ub monomers (Fig. B16). Our hypothesis is that different sysines to link through; hence different configurations allow Ub chains to reveal different hydrophobic surfaces and allow specific recognition by the DUB enzymes.

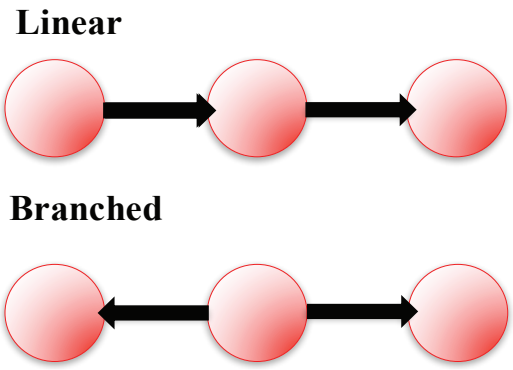


Figure B16: Configuration of linear and branched Ub chains.

Table B3: Lengths of atomistic implicit solvent molecular dynamics simulations for Ub3.

Configuration	Linkage	Lengths (ns)
Linear Trimers	K6 – K6	1000
	K6 – K48	1000×2
	K11 – K11	1000
	K11 – K63	1000
	K27 – K27	1000
	K29 – K29	1000
	K33 – K33	1000
	K48 – K48	1000
	K48 – K6	1000
	K63 – K63	1000
Branched Trimers	K11 – K48	995
	K11 – K63	1000
	K48 – K63	995
	K6 – K48	1000

B.7.1 Linear Trimers

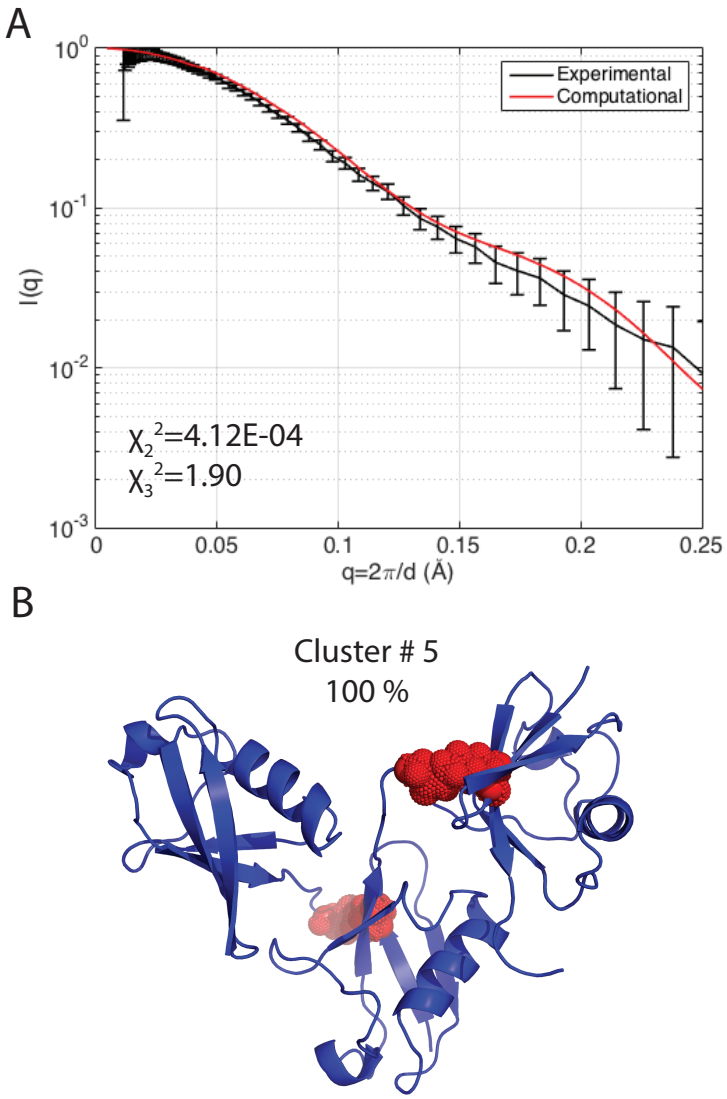


Figure B17: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K6 – K6 linked linear Ub trimer. (B) Predicted conformational ensemble for K6 – K6 linear trimer.

Table B4: Properties calculated over clusters obtained from K6 – K6 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K6 – K6 linked linear trimers.
1	2.53	19.38	36.09	0.00
2	2.74	19.33	34.32	0.00
3	3.44	19.83	34.94	0.00
4	4.14	20.00	34.45	0.00
5	3.05	19.03	34.08	0.00
6	6.25	19.22	35.15	0.00
7	2.74	19.04	34.69	0.00
8	2.48	19.07	35.47	0.00
9	3.51	19.85	34.59	0.00
10	2.95	19.26	33.72	0.00
11	5.32	19.69	33.37	0.00
12	5.54	20.79	36.20	1.00
13	3.13	19.74	34.90	0.00
14	6.28	19.84	34.21	0.00
15	6.68	19.45	34.02	0.00
16	6.47	20.11	36.92	0.00
17	9.45	19.92	34.99	0.00
18	8.81	19.87	34.74	0.00
19	7.87	19.76	33.42	0.00
20	8.68	20.08	34.38	0.00
21	6.27	19.62	35.02	0.00
22	8.25	19.36	34.48	0.00
23	7.12	19.78	35.00	0.00
24	6.97	19.69	34.82	0.00
25	4.15	20.06	35.22	0.00
26	2.99	19.09	32.85	0.00
27	5.17	20.16	34.91	0.00
28	6.86	20.04	34.37	0.00
29	9.99	19.98	34.73	0.00
30	9.26	19.46	33.79	0.00
31	5.65	19.49	37.13	0.00
32	7.98	19.87	38.89	0.00

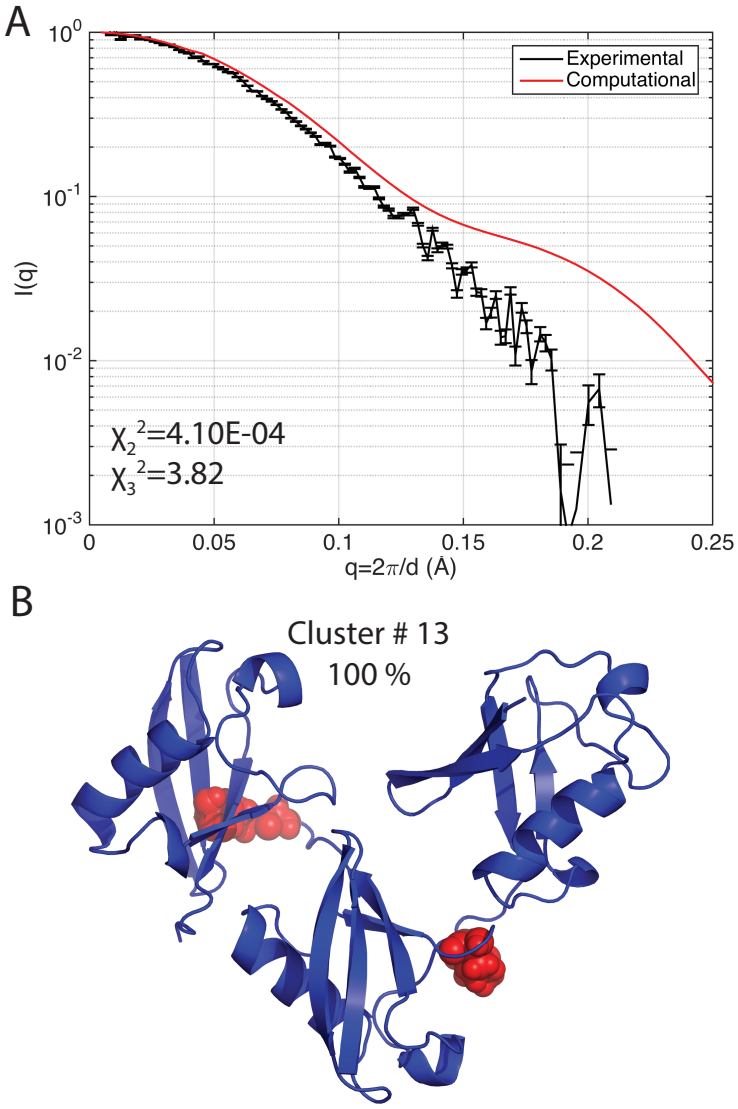


Figure B18: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K6 – K48 linked linear Ub trimer. (B) Predicted conformational ensemble for K6 – K48 linear trimer.

Table B5: Properties calculated over clusters obtained from K6 – K48 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K6 – K48 linked linear trimers.
1	9.25	18.90	32.08	0.00
2	8.33	19.65	31.95	0.00
3	9.60	18.62	32.14	0.00
4	10.16	19.31	32.97	0.00
5	11.12	18.92	33.97	0.00
6	13.62	18.91	30.99	0.00
7	10.47	18.51	31.32	0.00
8	9.70	19.30	32.40	0.00
9	10.14	19.32	32.37	0.00
10	13.99	19.19	34.21	0.00
11	10.09	18.92	33.26	0.00
12	12.41	18.85	32.99	0.00
13	5.03	20.99	36.05	1.00
14	6.05	20.10	34.10	0.00
15	7.54	19.30	32.76	0.00
16	10.30	19.79	33.70	0.00
17	8.56	19.00	32.50	0.00
18	8.32	19.48	33.95	0.00
19	9.49	19.27	34.27	0.00
20	8.49	18.87	32.80	0.00
21	9.45	19.67	32.39	0.00
22	11.93	19.76	33.05	0.00
23	9.57	20.00	34.94	0.00
24	10.72	20.91	39.78	0.00
25	9.34	19.87	38.11	0.00
26	10.68	20.55	36.02	0.00
27	11.12	18.48	32.97	0.00
28	9.69	19.30	32.29	0.00
29	9.47	19.30	32.87	0.00
30	12.66	18.89	31.72	0.00
31	6.44	19.76	36.28	0.00
32	8.27	19.61	32.80	0.00
33	14.38	19.33	34.26	0.00
34	12.21	18.60	31.33	0.00
35	8.60	19.48	32.78	0.00
36	14.64	20.51	37.28	0.00

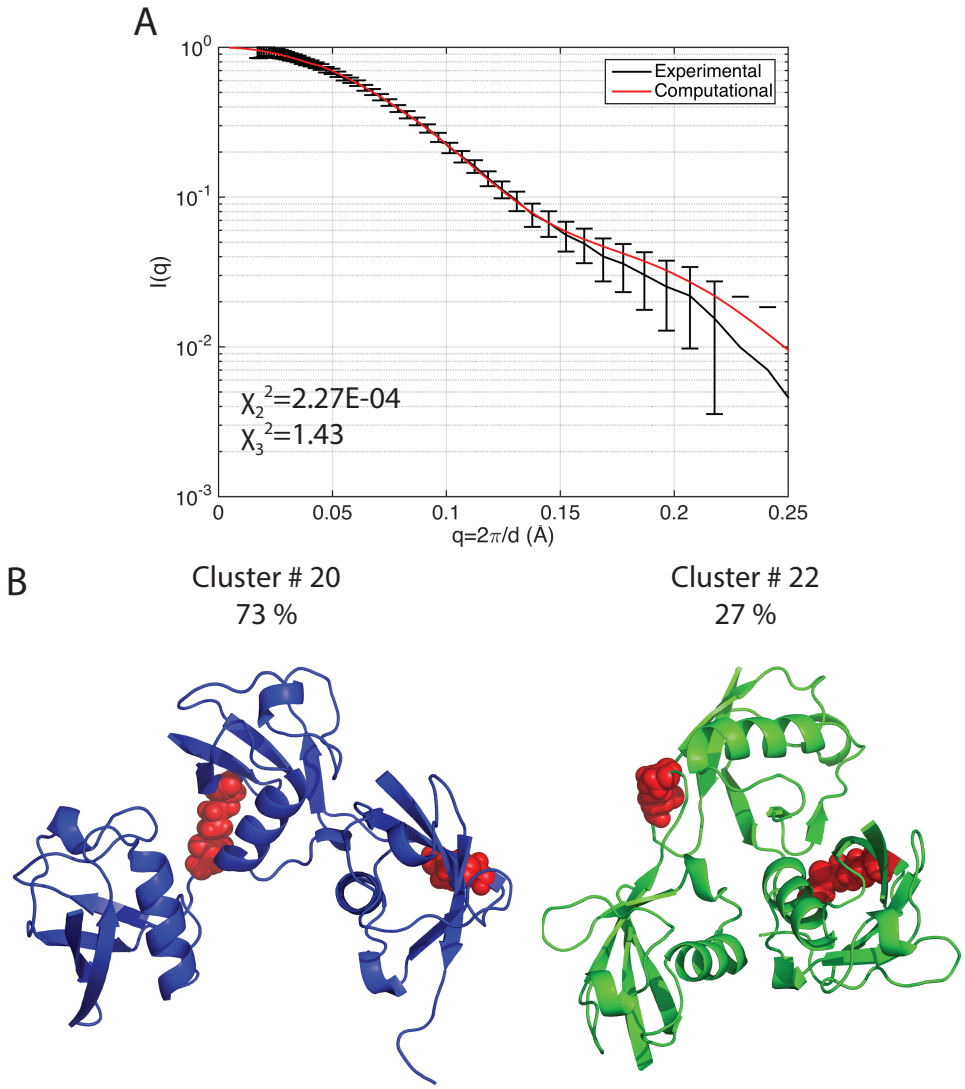


Figure B19: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K11 – K11 linked linear Ub trimer.(B) Predicted conformational ensemble for K11 – K11 linear trimer.

Table B6: Properties calculated over clusters obtained from K11 – K11 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K11 – K11 linked linear trimers.
1	5.06	21.38	38.49	0.00
2	7.25	22.77	40.76	0.00
3	9.14	20.74	34.99	0.00
4	10.92	20.53	35.35	0.00
5	9.16	21.85	39.25	0.00
6	9.60	20.72	37.07	0.00
7	11.37	20.31	37.42	0.00
8	12.15	20.39	37.50	0.00
9	12.93	20.84	35.86	0.00
10	14.03	21.06	38.15	0.00
11	12.67	20.21	36.81	0.00
12	13.72	20.56	35.60	0.00
13	14.43	20.46	34.34	0.00
14	15.40	21.02	36.81	0.00
15	15.04	20.28	35.33	0.00
16	15.53	20.74	35.95	0.00
17	14.98	20.99	36.96	0.00
18	14.24	20.69	36.11	0.00
19	15.86	20.91	36.96	0.00
20	2.35	21.47	37.15	0.73
21	8.55	21.69	36.62	0.00
22	13.30	20.59	38.42	0.27
23	15.69	20.95	37.98	0.00

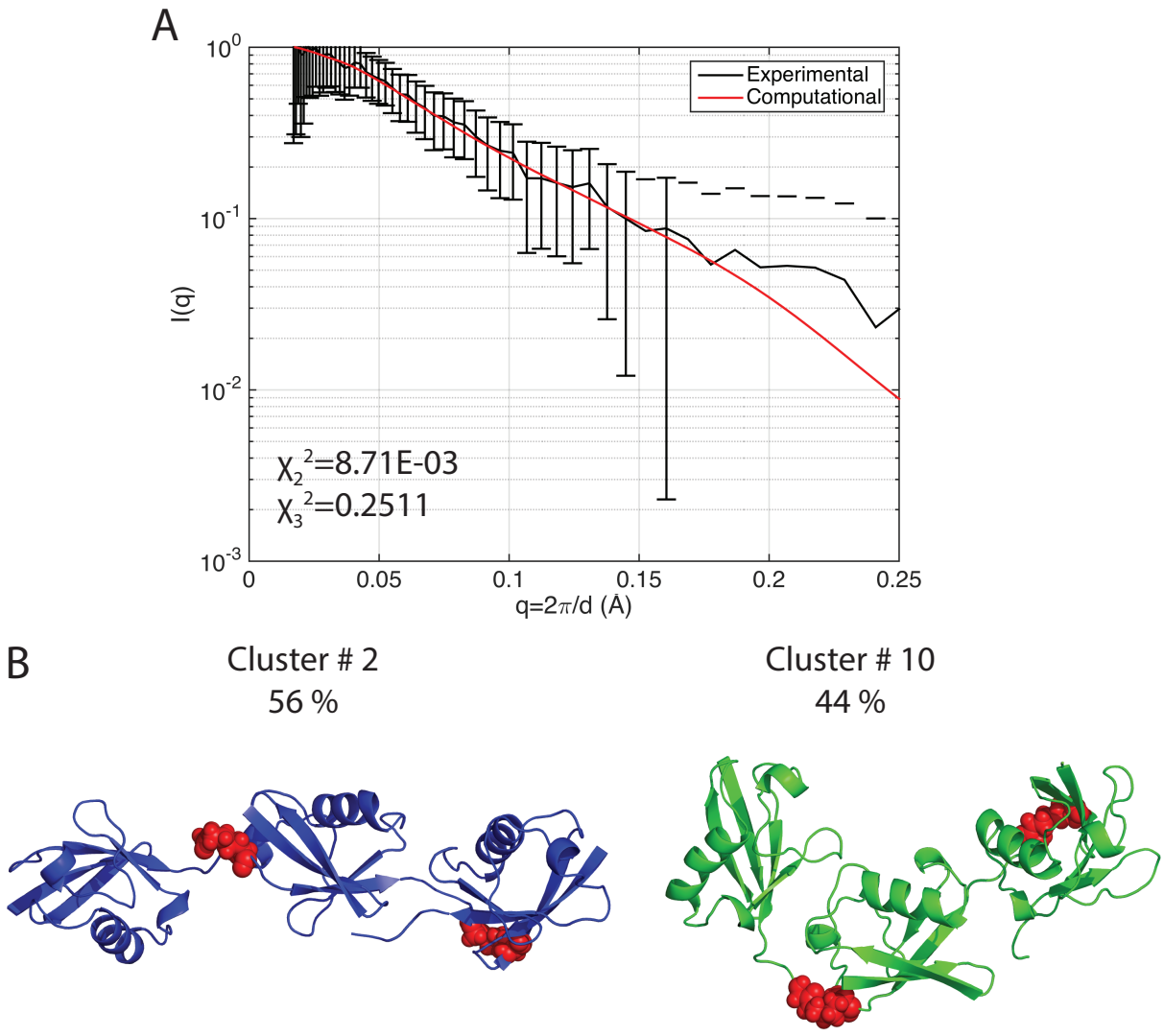


Figure B20: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K11 – K63 linked linear Ub trimer. (B) Predicted conformational ensemble for K11 – K63 linear trimer.

Table B7: Properties calculated over clusters obtained from K11 – K63 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K11 – K63 linked linear trimers.
1	4.17	28.94	53.53	0.01
2	7.41	28.15	50.89	0.56
3	8.91	28.06	49.16	0.00
4	8.85	26.45	46.83	0.00
5	14.14	23.50	41.05	0.00
6	9.96	25.64	45.42	0.00
7	13.83	24.60	42.03	0.00
8	11.52	25.76	46.21	0.00
9	11.00	25.25	44.42	0.00
10	14.96	23.60	41.43	0.44
11	11.35	25.47	44.46	0.00
12	14.38	24.40	42.71	0.00
13	11.18	26.10	45.37	0.00
14	8.58	26.47	47.76	0.00
15	10.92	25.60	46.58	0.00
16	11.98	26.05	45.17	0.00
17	10.23	26.00	44.83	0.00
18	9.55	25.67	47.08	0.00
19	9.46	26.61	48.83	0.00
20	12.46	25.15	44.49	0.00
21	12.96	24.93	42.92	0.00
22	13.12	25.07	45.05	0.00
23	13.83	24.53	41.98	0.00
24	12.74	24.26	41.65	0.00
25	13.86	25.79	46.09	0.00
26	12.13	25.71	44.08	0.00
27	10.15	26.90	47.35	0.00
28	11.77	25.55	45.45	0.00
29	13.40	24.50	44.78	0.00
30	13.73	25.87	49.24	0.00
31	13.48	24.71	42.65	0.00

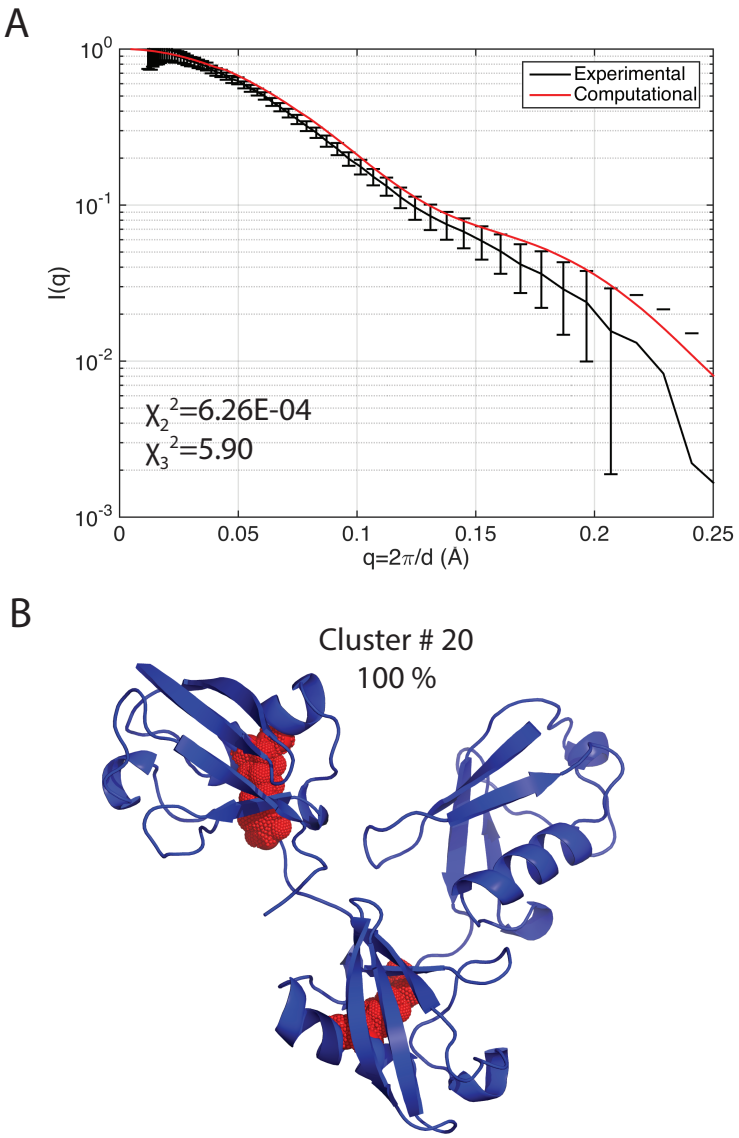


Figure B21: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K27 – K27 linked linear Ub trimer. (B) Predicted conformational ensemble for K27 – K27 linear trimer.

Table B8: Properties calculated over clusters obtained from K27 – K27 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K27 – K27 linked linear trimers.
1	4.12	19.93	34.64	0.00
2	9.57	20.69	36.63	0.00
3	10.32	21.35	36.97	0.00
4	11.84	20.59	37.37	0.00
5	12.29	20.42	37.72	0.00
6	8.25	19.76	33.26	0.00
7	11.44	21.23	38.68	0.00
8	13.13	20.36	37.84	0.00
9	13.54	20.29	37.37	0.00
10	12.76	20.34	38.76	0.00
11	12.14	21.03	39.54	0.00
12	14.31	20.44	37.02	0.00
13	14.18	20.44	37.48	0.00
14	14.39	20.34	37.41	0.00
15	14.38	20.29	37.60	0.00
16	14.66	20.62	37.79	0.00
17	12.26	20.41	38.12	0.00
18	14.29	20.14	36.94	0.00
19	11.71	20.94	38.48	0.00
20	11.53	21.60	40.30	1.00
21	12.15	21.49	40.35	0.00
22	12.13	20.81	37.95	0.00
23	11.90	21.07	38.85	0.00
24	10.25	19.62	35.92	0.00
25	10.20	20.12	36.92	0.00
26	11.60	20.75	36.27	0.00
27	11.29	20.78	36.38	0.00
28	1.48	19.42	33.77	0.00
29	7.39	21.43	38.28	0.00
30	12.45	20.83	39.03	0.00
31	12.21	20.76	37.84	0.00
32	10.43	20.65	37.46	0.00
33	11.24	21.11	37.55	0.00
34	10.72	20.05	36.84	0.00
35	8.05	19.58	33.34	0.00

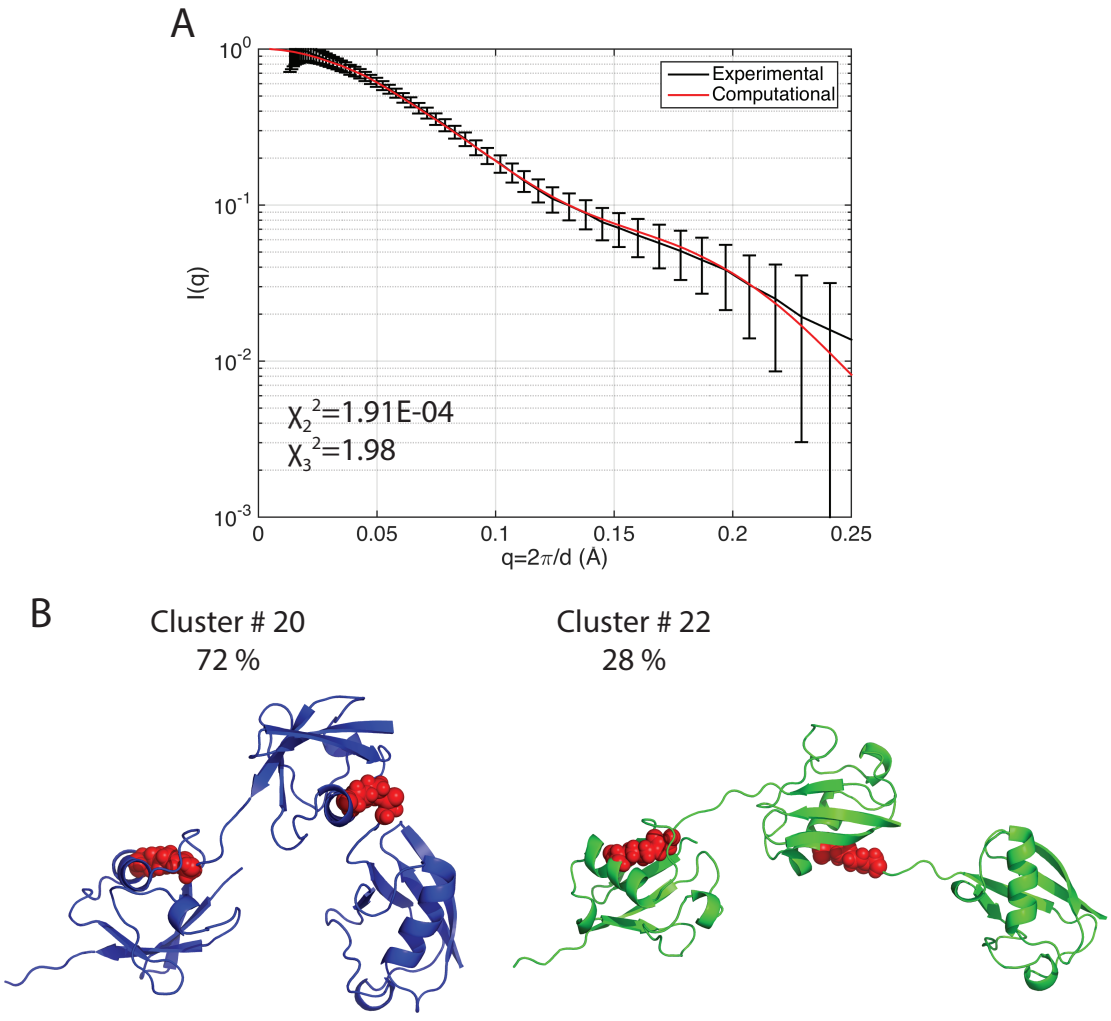


Figure B22: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K29 – K29 linked linear Ub trimer. (B) Predicted conformational ensemble for K29 – K29 linear trimer.

Table B9: Properties calculated over clusters obtained from K29 – K29 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K29 – K29 linked linear trimers.
1	7.07	29.96	59.30	0.00
2	7.83	28.27	54.55	0.00
3	8.80	27.77	54.47	0.00
4	14.64	23.09	42.04	0.00
5	15.60	23.12	41.40	0.00
6	15.74	22.40	38.98	0.00
7	16.80	21.81	39.27	0.00
8	17.91	21.09	42.18	0.00
9	16.39	22.67	44.04	0.00
10	18.66	20.23	42.28	0.00
11	17.84	22.05	38.80	0.00
12	17.44	21.71	43.98	0.00
13	16.88	22.02	44.98	0.00
14	15.14	23.51	44.67	0.00
15	15.73	22.95	42.17	0.00
16	17.52	22.13	41.34	0.00
17	15.45	22.10	45.06	0.00
18	13.43	23.74	49.15	0.00
19	18.40	20.57	36.50	0.00
20	16.36	22.75	51.07	0.72
21	19.15	20.14	36.62	0.00
22	3.94	30.10	57.46	0.28
23	19.65	20.28	39.91	0.00
24	15.44	22.76	44.83	0.00
25	18.94	20.17	39.69	0.00
26	18.08	21.58	41.97	0.00

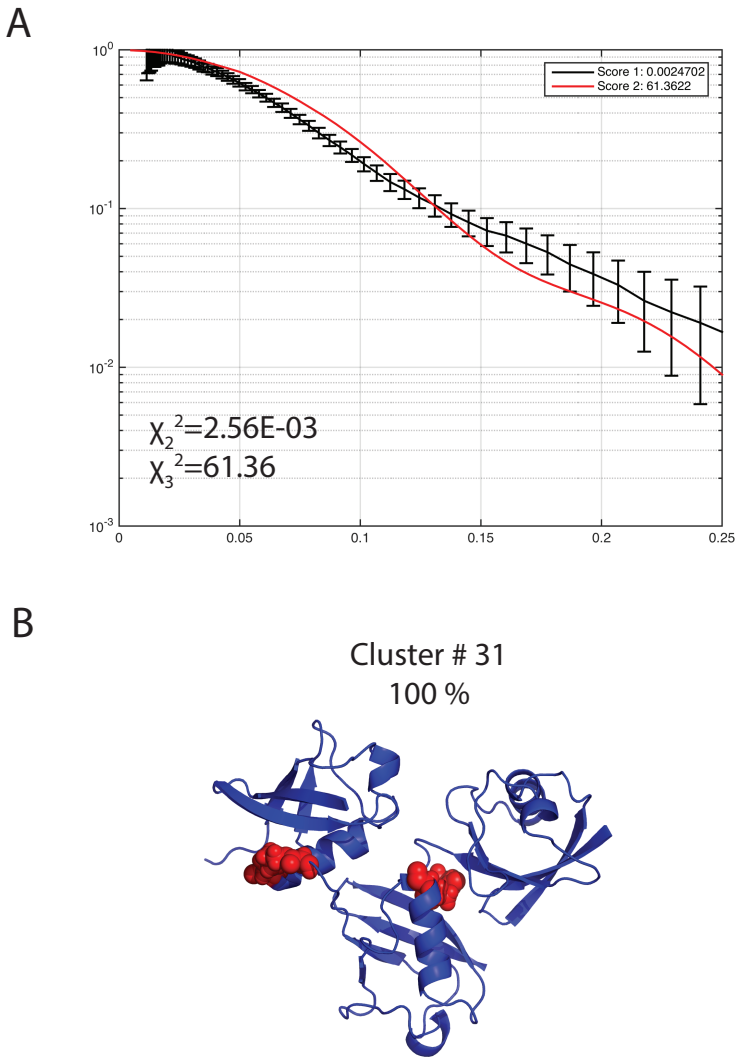


Figure B23: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K33 – K33 linked linear Ub trimer. (B) Predicted conformational ensemble for K33 – K33 linear trimer.

Table B10: Properties calculated over clusters obtained from K33 – K33 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K33 – K33 linked linear trimers.
1	1.74	30.77	54.10	0.00
2	6.46	29.64	53.62	0.00
3	9.79	31.06	54.56	0.00
4	10.87	30.63	56.79	0.00
5	12.01	28.79	55.18	0.00
6	11.83	29.39	54.97	0.00
7	12.77	28.02	52.69	0.00
8	13.59	26.47	50.56	0.00
9	13.64	26.74	49.35	0.00
10	13.65	26.46	45.68	0.00
11	12.99	25.50	43.64	0.00
12	12.97	26.05	43.79	0.00
13	12.54	25.50	44.69	0.00
14	13.96	23.65	40.89	0.00
15	15.03	23.32	41.26	0.00
16	17.33	22.04	39.63	0.00
17	19.78	19.27	33.92	0.00
18	20.09	19.45	33.65	0.00
19	20.39	19.09	34.31	0.00
20	19.20	19.71	36.20	0.00
21	19.54	19.23	33.56	0.00
22	20.40	19.22	33.47	0.00
23	19.97	19.28	33.81	0.00
24	19.10	19.56	35.11	0.00
25	19.76	19.44	33.67	0.00
26	18.68	19.79	35.87	0.00
27	18.59	19.74	33.19	0.00
28	18.35	19.83	37.42	0.00
29	18.34	19.69	34.72	0.00
30	18.69	19.96	34.74	0.00
31	18.81	19.80	35.75	1.00
32	20.51	18.92	32.02	0.00
33	19.16	19.50	36.35	0.00
34	19.32	19.36	32.40	0.00
35	18.86	19.78	33.58	0.00
36	3.77	29.97	54.12	0.00
37	18.08	20.48	35.63	0.00
38	17.67	21.46	38.02	0.00
39	19.69	19.43	34.80	0.00
40	18.06	19.98	40.52	0.00
41	18.23	20.57	37.14	0.00
42	19.10	20.07	36.97	0.00
43	19.94	19.51	33.23	0.00
44	17.31	20.27	37.07	0.00

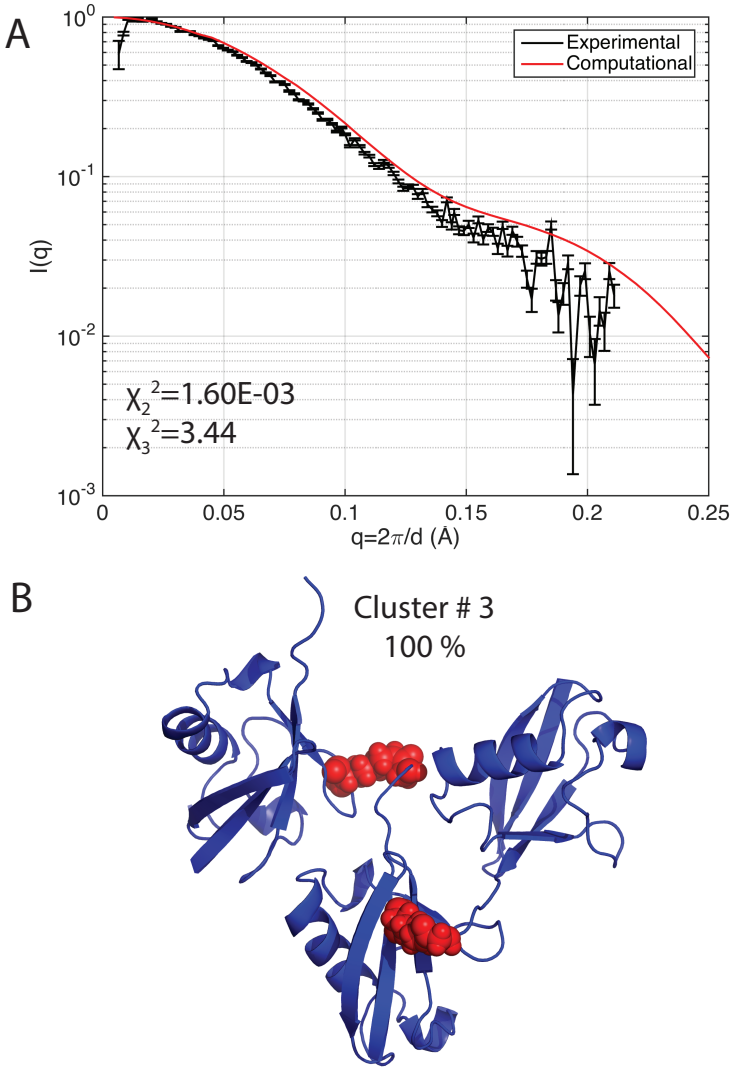


Figure B24: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K48 – K6 linked linear Ub trimer. (B) Predicted conformational ensemble for K48 – K6 linear trimer.

Table B11: Properties calculated over clusters obtained from K48 – K6 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K48 – K6 linked linear trimers.
1	3.19	19.63	36.28	0.00
2	5.74	20.41	35.92	0.00
3	6.78	20.79	36.12	1.00
4	5.47	20.39	34.68	0.00
5	4.86	19.84	34.32	0.00
6	5.29	19.42	38.47	0.00
7	7.25	19.26	33.97	0.00
8	6.70	19.36	33.50	0.00
9	6.69	19.40	33.49	0.00
10	8.69	19.03	33.52	0.00
11	8.68	19.70	33.88	0.00
12	9.03	19.57	34.35	0.00
13	6.85	19.38	33.90	0.00
14	6.14	19.95	35.93	0.00
15	7.13	19.72	34.67	0.00
16	7.63	19.75	35.08	0.00
17	8.91	19.33	32.69	0.00
18	8.19	19.28	32.57	0.00
19	8.78	19.73	36.44	0.00
20	9.18	19.98	36.51	0.00
21	8.23	19.14	34.16	0.00
22	7.25	19.27	34.08	0.00
23	8.83	19.45	34.79	0.00
24	9.20	19.54	35.03	0.00
25	9.84	19.22	33.90	0.00
26	7.93	19.35	34.27	0.00
27	9.41	19.43	33.07	0.00
28	7.27	19.01	32.63	0.00
29	6.56	18.76	32.92	0.00
30	2.12	19.26	36.10	0.00
31	8.86	19.24	34.31	0.00
32	6.55	19.06	33.62	0.00
33	9.50	19.36	34.33	0.00
34	9.35	19.46	34.28	0.00
35	9.95	19.02	33.16	0.00
36	9.94	19.40	33.46	0.00

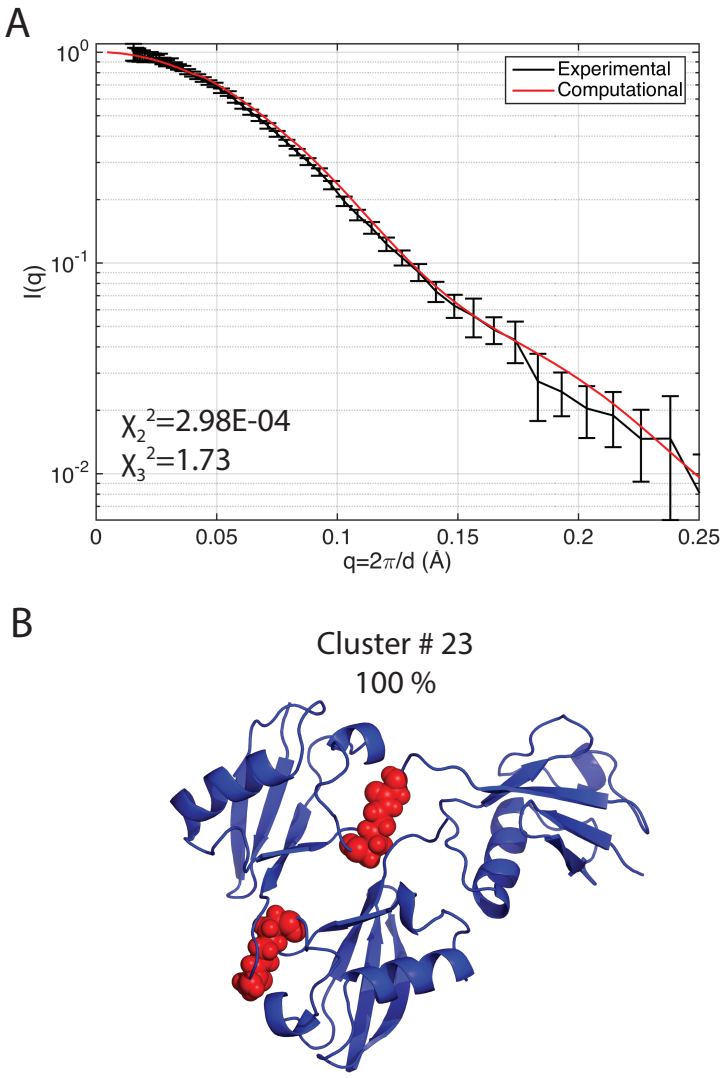


Figure B25: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K48 – K48 linked linear Ub trimer. (B) Predicted conformational ensemble for K48 – K48 linear trimer.

Table B12: Properties calculated over clusters obtained from K48 – K48 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K48 – K48 linked linear trimers.
1	2.31	19.81	37.19	0.00
2	4.39	20.09	37.61	0.00
3	8.72	19.27	37.10	0.00
4	10.14	19.42	34.56	0.00
5	3.29	20.01	36.83	0.00
6	3.94	19.84	36.19	0.00
7	4.03	19.74	36.55	0.00
8	6.46	19.44	34.72	0.00
9	8.55	18.95	33.82	0.00
10	10.12	19.51	35.29	0.00
11	9.79	19.75	35.02	0.00
12	8.94	19.69	35.56	0.00
13	8.82	19.28	34.38	0.00
14	9.69	19.08	34.05	0.00
15	8.54	19.02	34.30	0.00
16	7.78	20.41	36.83	0.00
17	8.15	19.75	37.09	0.00
18	8.72	19.51	36.39	0.00
19	9.32	19.30	33.65	0.00
20	10.18	19.53	33.96	0.00
21	9.39	19.53	36.13	0.00
22	2.92	19.75	37.15	0.00
23	4.69	20.56	39.06	1.00

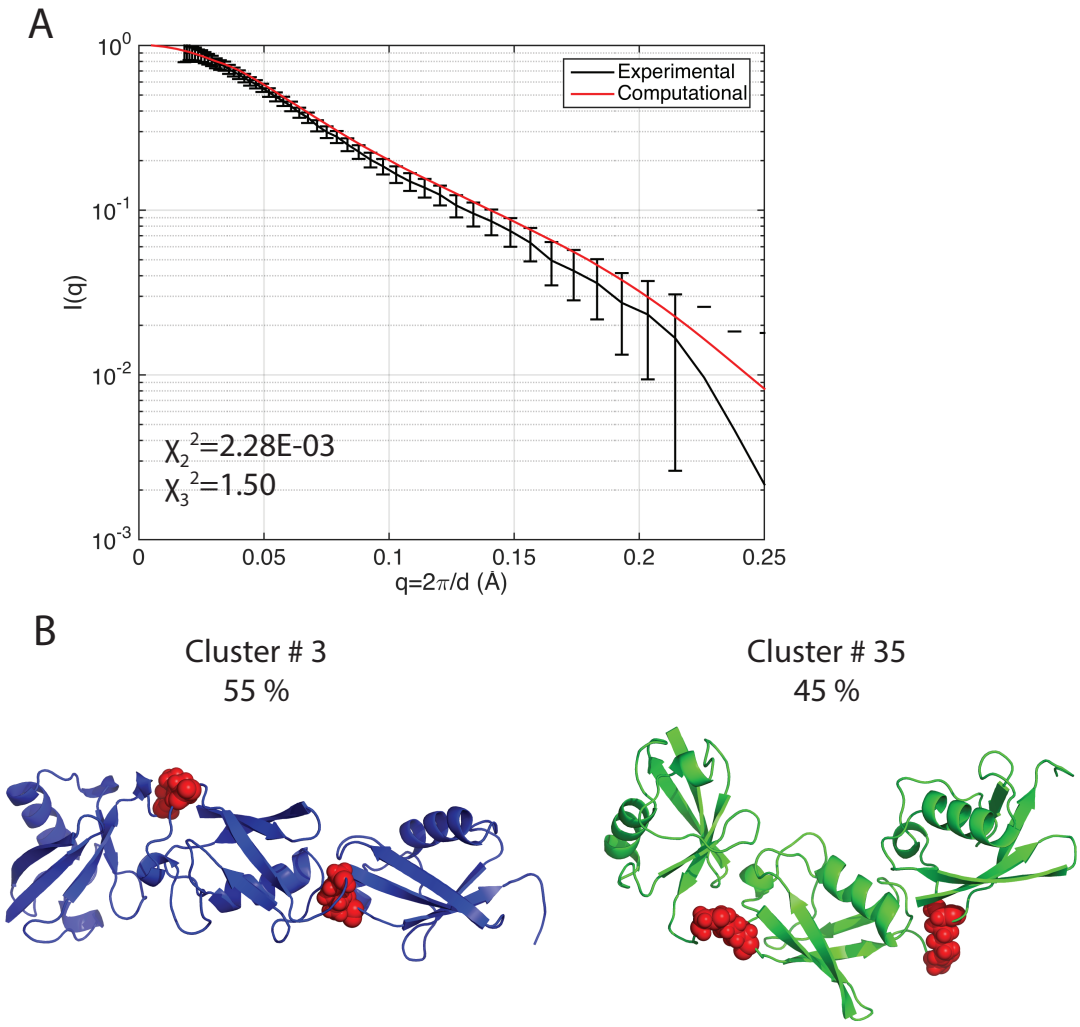


Figure B26: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K63 – K63 linked linear Ub trimer. (B) Predicted conformational ensemble for K63 – K63 linear trimer.

Table B13: Properties calculated over clusters obtained from K63 – K63 linked linear trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K63 – K63 linked linear trimers.
1	5.81	31.80	52.31	0.00
2	12.33	28.94	57.77	0.55
3	16.37	25.44	49.41	0.00
4	9.29	31.76	59.09	0.00
5	9.68	33.76	63.70	0.00
6	11.76	33.02	59.57	0.00
7	16.65	25.30	51.25	0.00
8	12.35	31.66	57.87	0.00
9	12.33	31.99	56.73	0.00
10	12.60	30.04	52.25	0.00
11	11.43	31.13	62.61	0.00
12	12.87	28.16	52.92	0.00
13	12.23	28.81	53.03	0.00
14	13.02	27.43	54.34	0.00
15	17.52	24.88	50.40	0.00
16	15.36	25.09	47.16	0.00
17	13.09	27.25	56.76	0.00
18	12.69	29.00	59.89	0.00
19	12.98	30.35	59.02	0.00
20	12.86	30.50	56.14	0.00
21	16.47	25.99	48.84	0.00
22	15.20	26.36	50.87	0.00
23	15.49	26.23	53.75	0.00
24	19.22	24.02	51.50	0.00
25	23.88	21.18	39.95	0.00
26	16.10	25.10	49.02	0.00
27	24.48	20.78	39.82	0.00
28	17.39	24.77	46.59	0.00
29	16.03	25.49	47.36	0.00
30	22.54	22.07	36.59	0.00
31	12.38	28.48	58.04	0.00
32	15.71	24.96	46.28	0.00
33	17.94	24.91	45.26	0.00
34	19.69	23.95	41.86	0.45
35	19.97	23.31	44.16	0.00
36	17.15	25.38	49.25	0.00
37	16.83	25.05	51.37	0.00
38	16.72	25.51	48.00	0.00

B.7.2 Branched Trimers

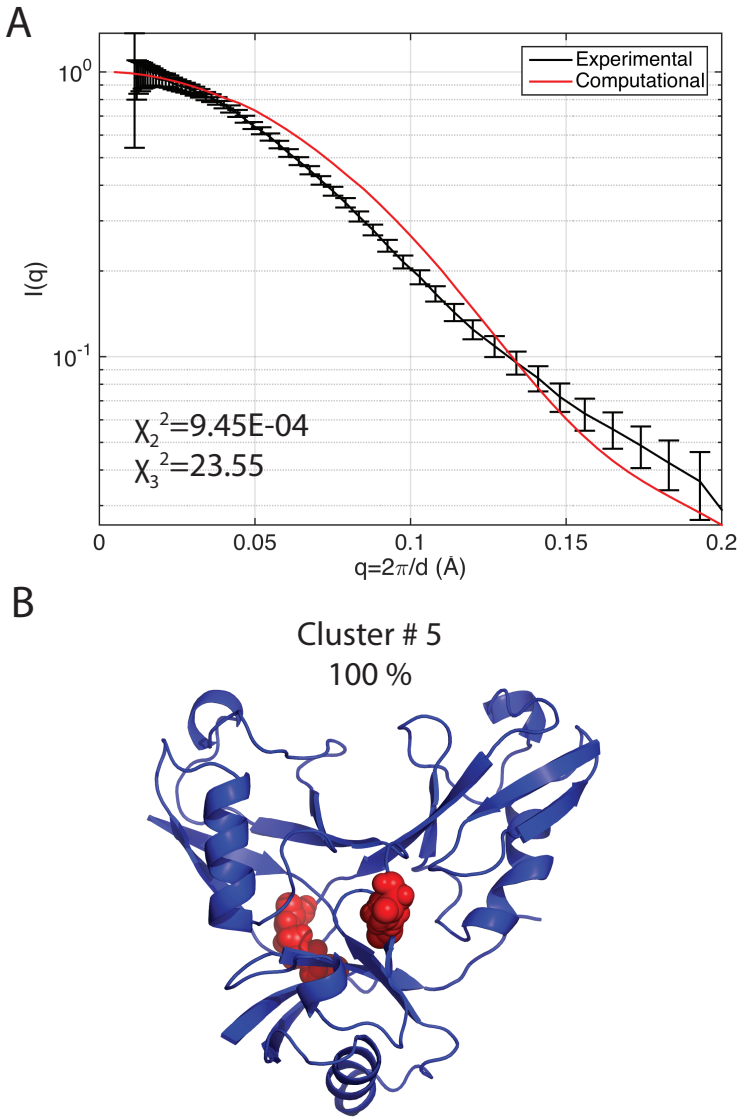


Figure B27: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K6 – K48 linked branched Ub trimer. (B) Predicted conformational ensemble for K6 – K48 branched trimer.

Table B14: Properties calculated over clusters obtained from K6 – K48 linked branched trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K6 – K48 linked branched trimers.
1	2.59	18.53	32.06	0.00
2	2.45	18.80	34.32	0.00
3	2.51	18.71	34.32	0.00
4	2.62	18.80	34.65	0.00
5	2.77	19.02	34.62	1.00
6	2.52	18.59	31.62	0.00
7	2.71	18.52	31.82	0.00
8	2.70	18.48	31.64	0.00
9	2.81	18.71	32.46	0.00
10	2.74	18.62	32.97	0.00
11	2.95	18.59	31.77	0.00
12	3.71	18.79	32.57	0.00
13	1.04	18.31	32.91	0.00

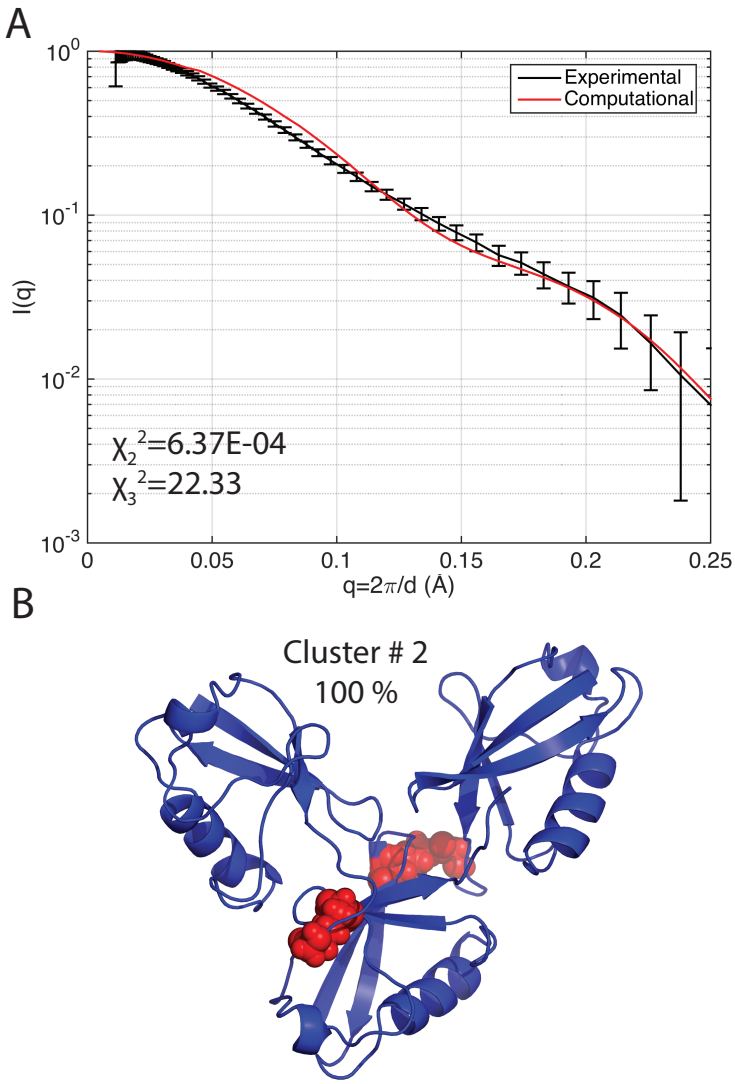


Figure B28: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K11 – K48 linked branched Ub trimer. (B) Predicted conformational ensemble for K11 – K48 branched trimer.

Table B15: Properties calculated over clusters obtained from K11 – K48 linked branched trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K11 – K48 linked branched trimers.
1	6.59	19.69	34.23	0.00
2	4.21	20.00	35.44	1.00
3	4.71	20.00	35.52	0.00
4	7.52	19.81	33.91	0.00
5	7.60	19.81	34.25	0.00
6	5.91	19.39	34.70	0.00
7	6.87	19.72	34.36	0.00
8	7.99	19.54	32.70	0.00
9	7.62	19.58	32.16	0.00
10	7.44	19.42	33.21	0.00
11	7.02	19.39	34.13	0.00
12	7.65	19.67	33.95	0.00
13	7.21	19.78	34.31	0.00
14	6.17	19.08	34.26	0.00
15	8.12	19.27	33.42	0.00
16	5.45	19.92	35.15	0.00

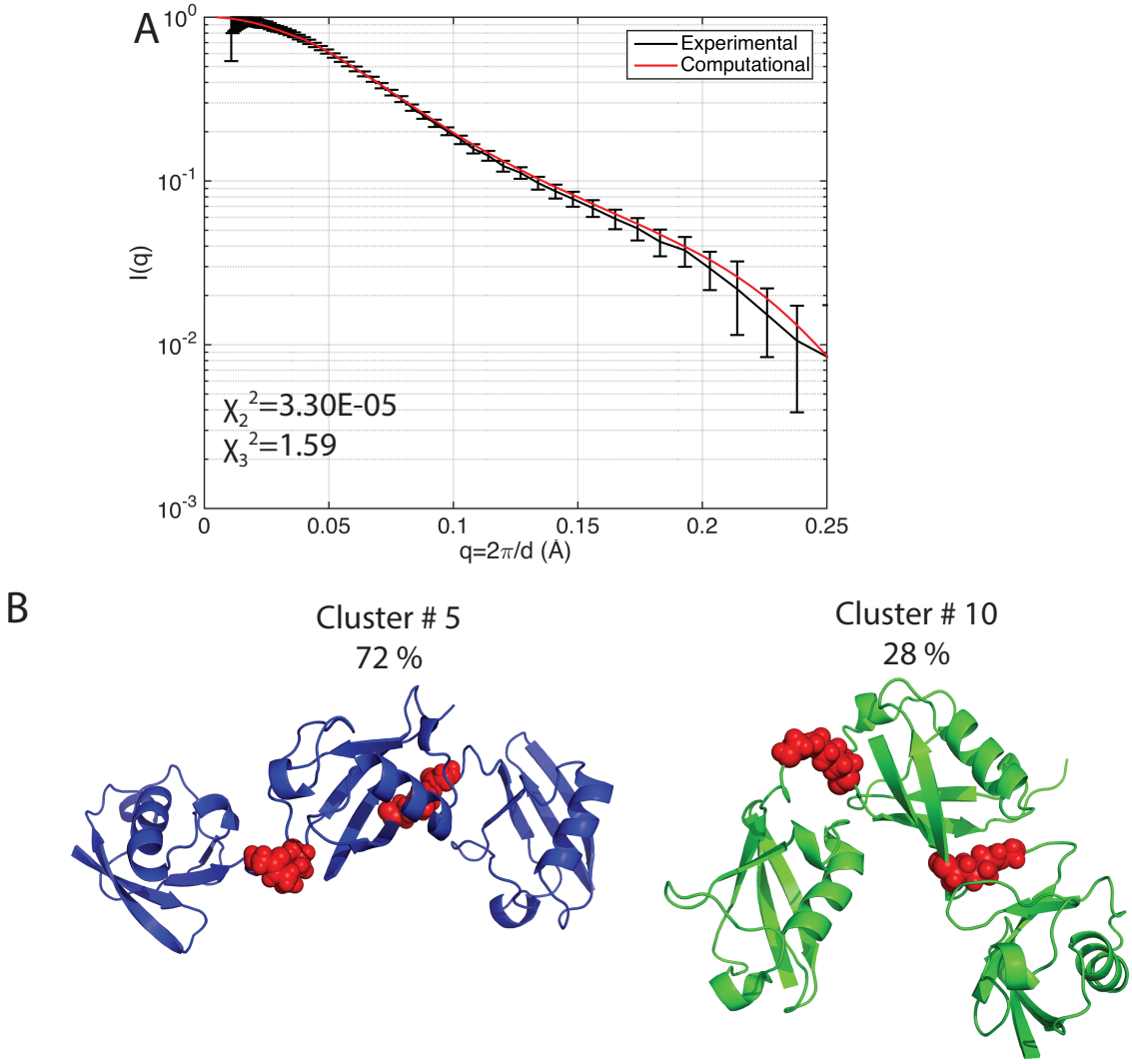


Figure B29: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K11 – K63 linked branched Ub trimer. (B) Predicted conformational ensemble for K11 – K63 branched trimer.

Table B16: Properties calculated over clusters obtained from K11 – K63 linked branched trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K11 – K63 linked branched trimers.
1	6.42	26.98	47.31	0.00
2	7.10	25.71	45.56	0.00
3	10.06	26.61	47.36	0.00
4	11.89	25.50	45.16	0.00
5	13.24	26.15	47.73	0.72
6	13.36	23.48	42.88	0.00
7	14.51	22.29	39.88	0.00
8	17.24	21.10	36.94	0.00
9	20.93	19.40	33.45	0.00
10	17.47	20.91	35.22	0.28
11	15.45	21.65	37.69	0.00
12	21.24	18.78	30.42	0.00
13	22.45	18.64	30.12	0.00
14	21.79	19.07	32.44	0.00
15	15.56	21.65	37.82	0.00
16	16.86	21.08	37.10	0.00
17	17.29	20.69	35.11	0.00
18	19.09	19.87	34.29	0.00
19	16.11	21.59	38.83	0.00
20	17.68	20.96	39.77	0.00
21	15.70	21.78	38.05	0.00

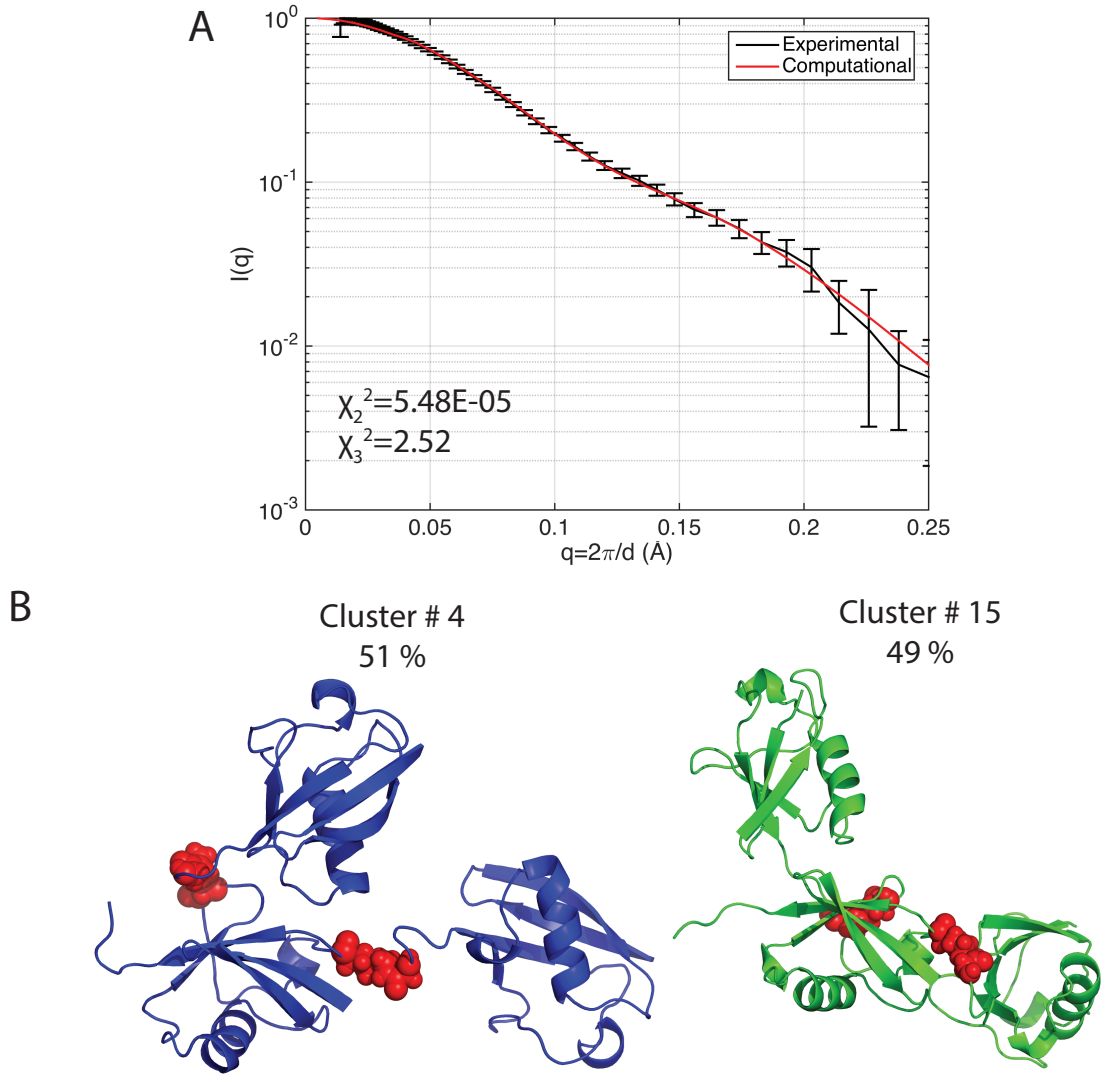


Figure B30: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K48 – K63 linked branched Ub trimer. (B) Predicted conformational ensemble for K48 – K63 branched trimer.

Table B17: Properties calculated over clusters obtained from K48 – K63 linked branched trimer simulations.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K11 – K63 linked branched trimers.
1	8.01	21.04	36.55	0.00
2	8.75	20.65	35.39	0.00
3	8.91	20.60	34.96	0.00
4	5.87	23.16	42.16	0.51
5	6.71	21.72	37.91	0.00
6	9.23	20.70	35.76	0.00
7	12.85	19.33	32.45	0.00
8	13.42	18.85	31.91	0.00
9	9.46	20.94	36.60	0.00
10	9.73	20.54	34.98	0.00
11	8.51	21.03	36.19	0.00
12	9.95	21.27	37.38	0.00
13	12.49	18.68	31.87	0.00
14	9.10	21.89	37.89	0.00
15	10.15	23.42	43.06	0.49
16	8.55	22.01	38.83	0.00
17	8.42	20.91	35.97	0.00
18	9.77	19.96	35.77	0.00
19	8.00	21.91	38.86	0.00
20	11.30	19.96	32.80	0.00
21	9.03	21.95	37.18	0.00
22	14.28	18.66	32.70	0.00
23	14.79	18.96	34.04	0.00
24	9.31	21.30	38.35	0.00

B.8 References

[1] D. KOMANDER, F. REYES-TURCU, J. LICCHESI, P. ODENWAELDER, K. WILKINSON, and D. BARFORD, *Embo Reports* **10**, 662 (2009).

[2] Y. SATO, A. YOSHIKAWA, M. YAMASHITA, A. YAMAGATA, and S. FUKAI, *EMBO J* **28**, 3903 (2009).

[3] S. D. WEEKS, K. C. GRASTY, L. HERNANDEZ-CUEBAS, and P. J. LOLL, *Proteins* **77**, 753 (2009).

[4] W. J. COOK, L. C. JEFFREY, M. CARSON, Z. J. CHEN, and C. M. PICKART, *Journal of Biological Chemistry* **267**, 16467 (1992).

[5] Y. RYABOV and D. FUSHMAN, *Journal of the American Chemical Society* **129**, 7894 (2007).

[6] T. HIRANO, O. SERVE, M. YAGI-UTSUMI, E. TAKEMOTO, T. HIROMOTO, T. SATOH, T. MIZUSHIMA, and K. KATO, *The Journal of biological chemistry* **286**, 37496 (2011).

[7] A. D. J. VAN DIJK, D. FUSHMAN, and A. M. J. J. BONVIN, *Proteins-Structure Function and Bioinformatics* **60**, 367 (2005).

[8] J. F. TREMPE, N. R. BROWN, M. E. M. NOBLE, and J. A. ENDICOTT, *Acta Crystallographica Section F-Structural Biology and Crystallization Communications* **66**, 994 (2010).

[9] M. Y. LAI, D. N. ZHANG, N. LARONDE-LEBLANC, and D. FUSHMAN, *Biochimica Et Biophysica Acta-Molecular Cell Research* **1823**, 2046 (2012).

[10] R. VARADAN, M. ASSFALG, and D. FUSHMAN, *Ubiquitin and Protein Degradation, Pt B* **399**, 177 (2005).

[11] N. ZHANG, Q. WANG, A. EHLINGER, L. RANDLES, J. W. LARY, Y. KANG, A. HARIRINIA, A. J. STORASKA, J. L. COLE, and D. FUSHMAN, *Molecular cell* **35**, 280 (2009).

[12] S. LIU, Y. H. CHEN, J. LI, T. HUANG, S. TARASOV, A. KING, A. M. WEISSMAN, R. A. BYRD, and R. DAS, *Structure* **20**, 2138 (2012).

[13] Y. C. JUANG, M. C. LANDRY, M. SANCHES, V. VITTAL, C. LEUNG, D. F. CECCARELLI, A. R. F. MATEO, J. N. PRUNEDA, D. MAO, R. K. SZILARD, S. ORLICKY, M. MUNRO, P. S. BRZOVIC, R. E. KLEVIT, F. SICHERI, and D. DUROCHER, *Molecular Cell* **46**, 549 (2012).

[14] E. M. VALKEVICH, R. G. GUENETTE, N. A. SANCHEZ, Y. C. CHEN, Y. GE, and E. R. STRIETER, *J Am Chem Soc* **134**, 6916 (2012).

[15] C. M. PICKART and S. RAASI, *Methods Enzymol* **399**, 21 (2005).

[16] V. H. TRANG, E. M. VALKEVICH, S. MINAMI, Y. C. CHEN, Y. GE, and E. R. STRIETER, *Angew Chem Int Ed Engl* **51**, 13085 (2012).

[17] V. H. TRANG, M. L. RODGERS, K. J. BOYLE, A. A. HOSKINS, and E. R. STRIETER, *Chembiochem* **15**, 1563 (2014).

[18] J. P. JIN, X. LI, S. P. GYGI, and J. W. HARPER, *Nature* **447**, 1135 (2007).

[19] M. E. SOWA, E. J. BENNETT, S. P. GYGI, and J. W. HARPER, *Cell* **138**, 389 (2009).

[20] C. E. BERNDSEN, R. WIENER, I. W. YU, A. E. RINGEL, and C. WOLBERGER, *Nature Chemical Biology* **9**, 154 (2013).

[21] C. W. DAVIES, L. N. PAUL, M. I. KIM, and C. DAS, *J Mol Biol* **413**, 416 (2011).

[22] R. BEAL, Q. DEVERAUX, G. XIA, M. RECHSTEINER, and C. PICKART, *Proceedings of the National Academy of Sciences of the United States of America* **93**, 861 (1996).

[23] S. RAASI, R. VARADAN, D. FUSHMAN, and C. M. PICKART, *Nature Structural and Molecular Biology* **12**, 708 (2005).

[24] N. ESWAR, D. ERAMIAN, B. WEBB, M. Y. SHEN, and A. SALI, *Methods Mol Biol* **426**, 145 (2008).

[25] N. ESWAR, B. WEBB, M. A. MARTI-RENOM, M. S. MADHUSUDHAN, D. ERAMIAN, M. Y. SHEN, U. PIEPER, and A. SALI, *Curr Protoc Bioinformatics* **Chapter 5**, Unit 5 6 (2006).

[26] N. ESWAR, B. WEBB, M. A. MARTI-RENOM, M. S. MADHUSUDHAN, D. ERAMIAN, M. Y. SHEN, U. PIEPER, and A. SALI, *Curr Protoc Protein Sci* **Chapter 2**, Unit 2 9 (2007).

[27] X. LUO, Z. TANG, G. XIA, K. WASSMANN, T. MATSUMOTO, J. RIZO, and H. YU, *Nature Structural and Molecular Biology* **11**, 338 (2004).

[28] M. A. MARTI-RENOM, A. C. STUART, A. FISER, R. SANCHEZ, F. MELO, and A. SALI, *Annu Rev Biophys Biomol Struct* **29**, 291 (2000).

[29] A. SALI and T. L. BLUNDELL, *J Mol Biol* **234**, 779 (1993).

[30] K. RAVIKUMAR, W. HUANG, and S. YANG, *Journal of Chemical Physics* **138** (2013).

[31] S. YANG, M. PARISIEN, F. MAJOR, and B. ROUX, *Journal of Physical Chemistry B* **114**, 10039 (2010).

Appendix C

Supporting Information for Structure and Function of Skip Residues in Myosin

C.1 Accession Codes

Atomic coordinates have been deposited in the Protein Data Bank under accession codes 4XA1, 4XA3, 4XA4, and 4XA6 corresponding to Skip 1: (Gp7-K1173-I1238-Eb1), Skip 2: (Gp7-L1361-I1425-Eb1), Skip 3: (Xrcc4-L1551-N1609), and Skip 4: (Gp7-A1776-T1854-Eb1) respectively.

C.2 Additional Details for Materials and Methods

C.2.1 DNA Constructs

Vectors for crystallographic studies were prepared by QuikChange cloning. Sections of the MYH7 gene, purchased as an IMAGE clone from Open Biosystems, were inserted into modified pET vectors containing the desired fusion protein as previously described [1]. QuikChange cloning allowed all constructs to be made without introducing cloning artifacts while maintaining the correct coiled-coil registration between the folding domains and the target gene fragment. The wild-type EGFP and mCherry myosin rod constructs were created by fusing each fluorescent reporter gene at amino acid 841 of the rat β-cardiac myosin gene as previously described [2]. All the myosin mutants were generated by inverse PCR [3].

C.2.2 Protein Expression, Purification

Myosin fusions were expressed in an E. coli BL21-CodonPlus (DE3)-RIL cell line (Stratagene). Cells were grown in lysogeny broth (LB) medium under appropriate antibiotic selection at 37 °C with shaking to an optical density of ∼1.0 and then cooled on ice for 15 min. 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) was then added and the cells were grown for an additional 16 h at 16 °C before harvesting them by centrifugation. Cells were washed with 50 mM HEPES pH 7.6, 50 mM NaCl, 1mM ethylenediaminetetraacetic acid (EDTA) at a ratio of 1 L of buffer per 6 L of culture prior to flash-freezing in liquid nitrogen and storage at -80 °C. Selenomethionine derivatives were prepared as previously described [4].

All purification steps were carried out at 4 °C. 10 g of cells were lysed by sonication in 100 mL of lysis buffer containing 50mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.6, 50 mM NaCl, 20 mM imidazole, 0.5 mg/mL lysozyme, 1 mM PMSF, 50 nM Leupeptin (Peptide International), 70 nM E-65 (Peptide International), 2 nM Aprotinin (ProSpec), and 2 mM AEBSF (Gold BioTechnology). Lysate was clarified by centrifugation at 125,000 g for 30 min at 4 °C in a Beckman Ti-45 rotor. The supernatant was loaded onto a 5 mL nickel-nitrilotriacetic acid (NiNTA) column (QIAGEN) by gravity and washed with 25 column volumes of buffer A (25 mM HEPES pH 7.6, 300 mM NaCl, 20 mM imidazole). The column was then washed with 5 column volumes buffer A containing 40 mM imidazole. Protein was eluted in four column volumes of buffer A with 200 mM imidazole. 1mM β-mercaptoethanol was added to all buffers for purification of constructs containing cysteine residues. The His-tag was removed by incubation with a 1:40 molar ratio of rTEV protease/myosin fusion protein at 4 °C in 25 mM HEPES pH 7.6, 100 mM NaCl, 0.1 mM EDTA, 0.5 mM tris(2-carboxyethyl)phosphine. NaCl concentration was increased to 300 mM and the cleaved protein was then loaded onto a 2 mL NiNTA column equilibrated in buffer A without imidazole. Myosin constructs were eluted with four column volumes of buffer A. Fusion proteins were concentrated in an Amicon Ultra-15 centrifugal filter 30 kDa cutoff (Millipore) to 10-25 mg/mL ofprior to overnight dialysis into storage buffer (10 mM HEPES pH 7.6, 100 mM NaCl). The proteins were flash-frozen in 30 μL droplets in liquid nitrogen and stored at -80 °C.

C.2.3 Crystallization

Crystals of Gp7-K1173-I1238-Eb1 (Skip 1) were grown by vapor diffusion at room temperature (RT) from a 1:1 mixture of protein at 10 mg/mL and 20% (w/v) polyethylene glycol methyl ether 2000, 20 mM

SrCl2, 100 mM HEPES pH 7.6, 5% pentaerythritol ethoxylate (17/8 PO/OH) 797, 0.5% 3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate (CHAPS). The crystals were transferred stepwise into a solution of 25% (w/v) polyethylene glycol methyl ether 2000, 20 mM SrCl2, 100 mM HEPES pH 7.6, 10% (w/v) pentaerythritol ethoxylate (17/8 PO/OH) 797, 0.5% (CHAPS), and 150 mM NaCl. They were then flash-frozen in liquid nitrogen.

Crystals of Gp7-L1361-I1425-Eb1 (Skip 2) were grown by vapor diffusion at room temperature from a 1:1 mixture of protein at 10 mg/mL and 4.5% (w/v) polyethylene glycol 8000, 100 mM sodium acetate pH 5.0, 50 mM CaCl2, and 2.5% (w/v) 3-methoxy-3-methyl-1-butanol. The crystals were transferred stepwise into a solution of 4.5% polyethylene glycol 8000, 100 mM sodium acetate pH 4.5, 50 mM CaCl2, 2.5% 3-methoxy-3-methyl-1-butanol, 100 mM NaCl, and 20% (w/v) ethylene glycol. They were then flash-frozen in liquid nitrogen.

Crystals of Xrcc4-L1551-N1609 (Skip 3) were grown by vapor diffusion at room temperature from a 1:1 mixture of protein at 10 mg/mL 30% (w/v) polyethylene glycol 1500, 250 mM tetramethylammonium chloride, and 100 mM 3-[4-(2-Hydroxyethyl)-1-piperazinyl]propanesulfonic acid (HEPPS) pH 8.5. The crystals were transferred stepwise into 30% (w/v) polyethylene glycol 1500, 250 mM tetramethylammonium chloride, and 100 mM HEPPS pH 8.5, and 10% glycerol. They were then flash-frozen in liquid nitrogen.

Gp7-A1776-T1854-Eb1 (Skip 4) was methylated to obtain suitably diffracting crystals with formaldehyde and dimethylamine borane complex [5]. The reaction was quenched by addition of Tris-HCl buffer (pH 7.5) to a final concentration of 200 mM, and the protein was dialyzed overnight against 2 L of 50 mM Hepes buffer pH 7.6, 100 mM NaCl, 5 mM DTT. The protein concentration was then adjusted to 20 mg/mL followed by dialysis into 10 mM Hepes pH 7.6, 100 mM NaCl overnight at 4 °C. After the final dialysis, the protein was drop-frozen in liquid nitrogen. Crystals of Gp7-A1776-T1854-Eb1 were grown by vapor diffusion at room temperature from a 1:1 mixture of protein at 20 mg/mL and 18% (w/v) polyethylene glycol 2000 methyl ether and 100 mM piperazine-N,N'-bis(2-ethanesulfonic acid) (PIPES) pH 6.4. The crystals were transferred directly to a solution A of 16% (w/v) polyethylene glycol methyl ether 2000, 100 mM PIPES pH 6.4, and 200 mM NaCl and then transferred again stepwise over the course of an hour and twenty minutes into solution B composed of 10% (w/v) polyethylene glycol 2000 methyl ether, 15% (w/v) polyethylene glycol 8000, 200 mM NaCl, and 100 mM PIPES pH 6.4. The crystals were then moved from room temperature to 4 °C to soak

for twelve hours in solution B. The crystals were then transferred stepwise over the course of forty minutes from solution B into a cryo mixture of solution B plus 10% (w/v) ethylene glycol and flash-frozen in liquid nitrogen.

C.2.4 Data Collection and Structure Determination

Data were collected at 100 K for Gp7-K1173-I1238-Eb1, Gp7-L1361-I1425-Eb1, and Xrcc4-L1551-N1609 at the SBC Beamline 19-ID (Advanced Photon Source, Argonne National Laboratory, Argonne, IL) at a wavelength of 0.9792 And, 0.9790 Å and 0.9794 Å respectively. Data were collected at 100 K for Gp7-A1776-T1854-Eb1 at the SBC Beamline 19-BM (Advanced Photon Source, Argonne National Laboratory, Argonne, IL) at a wavelength of 0.9791 ÅÀ Gp7-A1776-T1854-Eb1 selenomethionine derivative crystal produced the highest quality diffraction data and was used in final data refinement. Diffraction data were indexed, integrated, and scaled with HKL3000 [6]. Data collected for Gp7-L1361-I1425-Eb1 were highly anisotropic so that ellipsoidal truncation and anisotropic scaling was performed [7] (services.mbi.ucla.edu/anisoscale/). The following reciprocal space resolutions cutoffs were applied 3.3 Å 2.5 Å and 2.6 Å for a*, b*, and c* respectively. Initial solutions were obtained by molecular replacement using PHASER with the appropriate fusion protein as a search model (PDB 1NO4, 1IK9, or 1YIB) [8]. Initial positions of the search models were optimized by rigid body refinement in Phenix Refine [9]. Structure building was done both manually in COOT through iterative cycles and with Phenix Autobuild. Phenix refine was used for refinement. 21 TLS groups were used in refinement of Gp7-A1776-T1854-Eb1 as defined by the Gp7 and Eb1 folding domains and break in the myosin coiled-coil [10]. Model quality was assessed by use of the Molprobity server [11]. The Ramachandran statistics show greater than 98% of the residues in allowed regions and less than 2% in disallowed regions for all structures. All structures have a protein geometry MolProbity score of the 97th percentile or greater. Pymol (http://www.pymol.org/) was used to generate images of the molecular models.

C.2.5 NRVM Preparation, Culturing Transfection and Immunostaining

Neonatal rat ventricular myocytes (NRVMs) were prepared as previously described [12]. Cells were electroporated using the Rat Cardiomyocyte-Neonatal Nucleofector Kit (Lonza) according to the manufacturer's protocol and plated onto 1% gelatin coated Glass Bottom Microwell Dishes (MatTeK). Following overnight recovery, cells were washed and then treated with 15 mM L-phenylephrine. Live cell imaging was

carried out at the time points specified in the figure legends. Detection of the endogenous myosin with the anti-sarcomeric myosin heavy chain monoclonal F59 was carried out as follows: NRVMs were rinsed twice with PBS, fixed with 2% paraformaldehyde (PFA) for 5 min and permeabilized with 0.1% Triton X-100 for 5 min. After 3 PBS washes, cells were blocked with 1% BSA, 10% goat serum, and 0.1% Triton X-100 at room temperature for 30 min and then incubated with the F59 antibody (1:10 dilution) at room temperature for 1 h. After 3 PBS washes, cells were incubated with the secondary antibody (Alexa Fluor 568 goat anti-mouse IgG) at room temperature for 1 h. After final washes with PBS, cells were rinsed in distilled water and mounted in FluoromountTM.

C.2.6 Confocal Microscopy

NRVMs were analyzed using the Nikon Eclipse TE 2000-U microscope coupled with an electron-multiplying charge-coupled device camera (Cascade II; Photometrics) and a Yokogawa spinning disc confocal system (CSU-Xm2; Nikon). Images were taken with the 100x Nikon Plan Apo VC NA 1.4 oil objective. MetaMorph software was used for image acquisition; image analysis was performed with ImageJ [13].

C.3 Additional Data

The results from the explicit solvent simulations support the implicit solvent results. The RMSF is higher for the ΔS3-R model than the Skip 3 WT simulation, as observed in implicit solvent simulation (Fig. 5.3). DCOM results show similar trends for Skip 3 and ΔS3-R, implying similar degree of coiling as observed in the implicit solvent simulation (Fig. 5.3). Upon k-means clustering, ΔS3-R has more clusters than Skip 3 simulation, similar to the implicit solvent simulations (Fig. 5.3). On the right hand side, representative structures of the most populates clusters are shown. ΔS3-R has remained recoiled even and Skip 3 simulation resulted in a relaxed super helical pitch, leading to the so-called parallel conformation. Additionally, Skip 3 remains within 1.2 Å Cα-RMSD with respect to the crystal structure. The most populated structures are within $<$1.0 Å Cα-RMSD with the corresponding structures from implicit solvent simulations. Regardless, of the solvation approach, a similar ensemble of conformations is achieved; hence the implicit solvent approach can be used instead of the explicit solvent for this system to improve speed and achieve longer time-scales.

The most populated structures shown on the right hand side of the page are within 2 Å Cα-RMSD of the crystal structure. Moreover, regardless of the presence of a folding domain or the starting structure, the

representative members of the most populated cluster are within $<$ 1.0 Å Cα-RMSD with the corresponding structures from implicit (Fig. C4) and explicit solvent simulations (Fig. 5.3). In addition the RMSF and DCOM results are consistent with the explicit and implicit solvent results; thus the addition of folding domains during the crystallization procedure or simulation without the folding domains does not alter the conformations. The simulations converge into recoiled or uncoiled conformations regardless of the initial structure used in the simulations.

Table C1: Crystallographic Constructs.

Gp7 – K1173 – I1238 – Eb1 (Skip 1) MGSSHHHHHHHHDYDIPTSENLYFQGAS – Gp7(1 – 49) – Myh7(1173 – 1237) – Eb1(211 – 251)
Gp7 – L1361 – I1425 – Eb1 (Skip 2) MGSSHHHHHHHHDYDIPTSENLYFQGAS – Gp7(1 – 49) – Myh7(1361 – 1425) – Eb1(215 – 251)
Xrcc4 – L1551 – N1609 (Skip 3) MSYYHHHHHHDYDIPTSENLYFQGGSG – Xrcc4(2 – 147) – Myh7(1551 – 1609)
Gp7 – A1776 – T1854 – Eb1 (Skip 4) MGSSHHHHHHHHDYDIPTSENLYFQGAS – Gp7(1 – 50) – Myh7(1776 – 1854) – Eb1(208 – 251)

Table C2: Data collection and refinement statistics.

	Gp7 – 1173 – 1237 – Eb1 Skip 1	Gp7 – 1361 – 1425 – Eb1 Skip 2	Xrcc4 – 1551 – 1609 Skip 3	SeGp7 – 1776 – 1854 – Eb1 Skip 4
Data collection				
Space group	P1	C2221	P212121	C2
Cell dimensions				
a, b, c (Å)	56.3, 59.3, 67.7	37.5, 146.0, 153.4	54.5, 90.2, 102.7	155.5, 46.9, 169.1
α, β, γ (°)	96, 110, 93	90, 90, 90	90, 90, 90	90, 94, 90
Resolution (Å)	50 – 3.2 (3.26 – 3.20)	50 – 2.5 (2.58 – 2.54)	50 – 2.3 (2.33 – 2.37)	50 – 3.4 (3.4 – 3.46)
R_{merge}	0.03 (0.07)	0.09 (0.48)	0.06 (0.31)	0.08 (0.44)
I I 15.5 (10.2)	24.1 (2.2)	32.4 (8.5)	10.1 (4.8)	
Completeness (%)	97.9 (98.7)	87.1 (52.5)	99.6 (99.9)	99.6 (100)

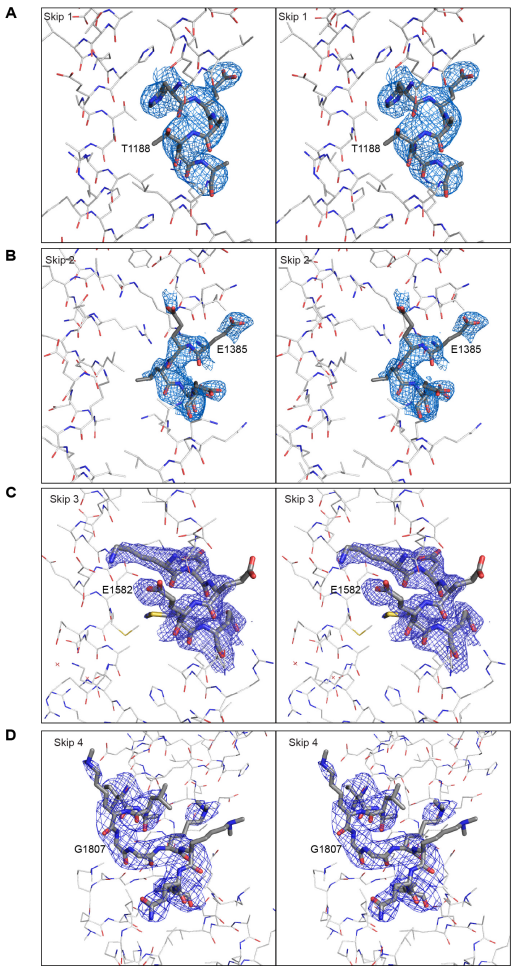


Figure C1: Stereo image of the omit electron density for each of the skip residues. Omit electron density for the residues surrounding (A) Skip 1 (T1188) H1186-A1191, (B) Skip 2 (E1385) (C) Skip 3 (E1582), and (D) Skip 4 (G1807) 1803-1812. In all cases the indicated amino acids were removed from the model and subject to cycles of refinement. The maps were contoured at 2.0 σ.

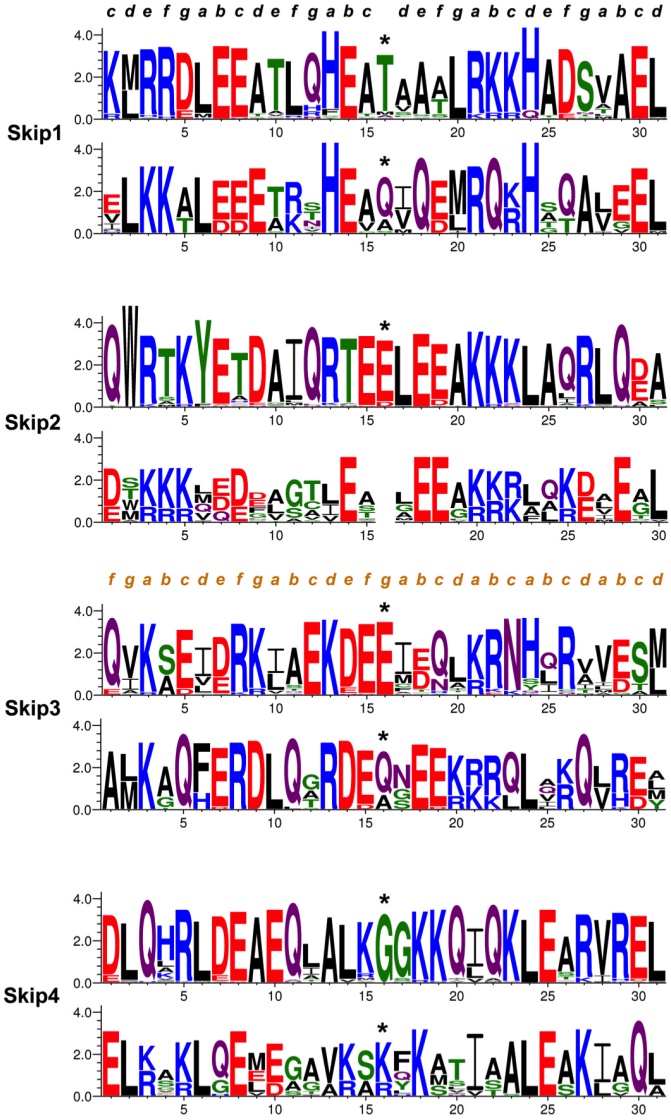


Figure C2: Sequence logos of the flanking regions of sarcomeric and non sarcomeric myosin skip residues. The sequence of 30 amino acids surrounding each skip residues is shown with the nominal heptad repeat positions reported at the top whereas the observed designation for human cardiac MyH7 Skip 3 is shown above that alignment. Skip residues are identified by an asterisk symbol. Amino acids in single letter code are color-coded according to the following scheme: polar, green; neutral, purple; basic, blue; acidic, red; and hydrophobic, black. For each skip residue, the top and bottom sequences correspond to the sarcomeric and non-sarcomeric (MYH9, MYH10, MYH11 and MYH14) myosin regions respectively. The lack of Skip 2 residue in non-sarcomeric myosins is indicated by a gap in the sequence.

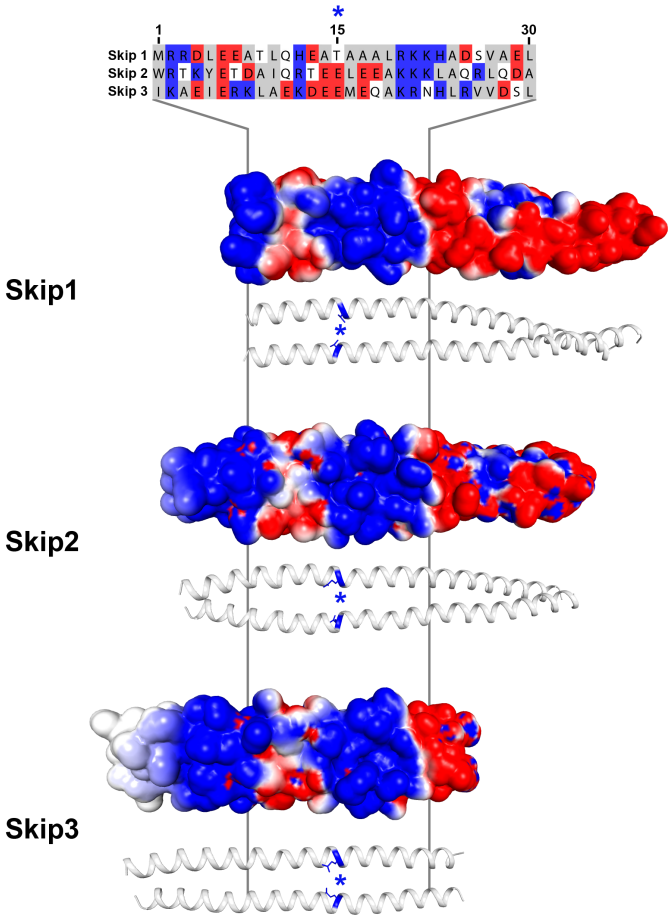


Figure C3: Sequence alignment and electrostatic surfaces for Skip 1, 2, and 3. The sequence is shown for two heptads on either side of the nominal skip residues which is depicted in blue and marked with an asterisk on the ribbon representation. The surface potential colored in blue and red for positively and negatively charged regions and was calculated with the program APBS in Pymol [14–16]

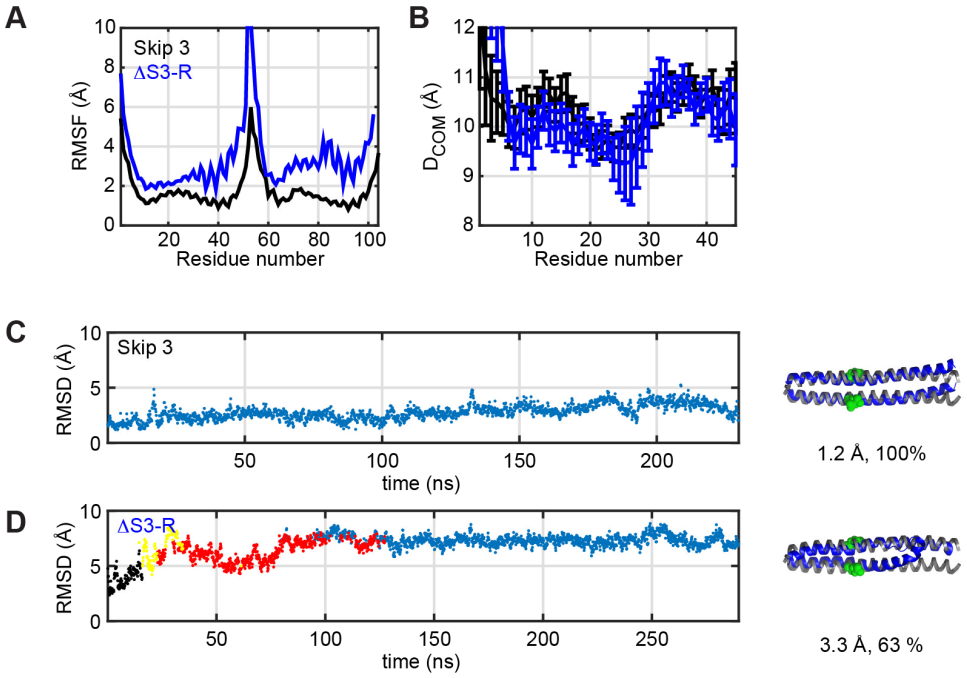


Figure C4: Implicit solvent simulations lead to similar results as explicit solvent simulations. The RMSF and DCOM analysis of the simulations of Skip 3 and recoiled Skip 3 deletion ΔS3-R with explicit solvent calculation are shown in panels (A) and (B), where Skip 3 is depicted in black and ΔS3-R is depicted in blue. Clustering analyses of the same simulations of Skip 3 and the recoiled Skip 3 deletion, ΔS3-R are shown in panels (C) and (D) where the color code differentiates between distinct clusters. The representative member of the most populated cluster is shown on the right hand side (blue), aligned on the Skip 3 crystal structure (grey). The RMSD of the representative members with respect to the crystal structure are shown underneath the structures along with their percentages in the population.

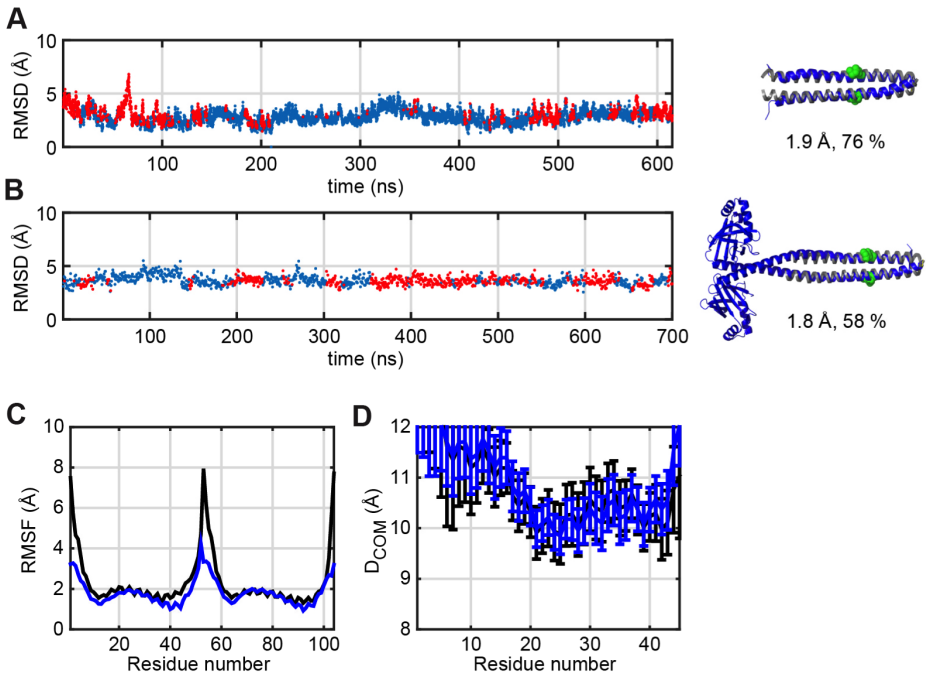


Figure C5: Folding domains do not influence the outcome of the simulations. The RMSD results for Skip 3 (black, simulation started from a uncoiled conformation rather than coiled-coil model) and Xrcc4 - Skip 3 (blue) with implicit solvent are represented in panels (A) and (B) and the color code represents different clusters. The representative member of the most populated cluster is shown on the right hand side (blue), aligned on the Skip 3 crystal structure (grey). The RMSD of the representative members with respect to the crystal structure are shown underneath the structures along with their percentages in the population. RMSF and DCOM results are shown in panels (C) and (D), respectively.

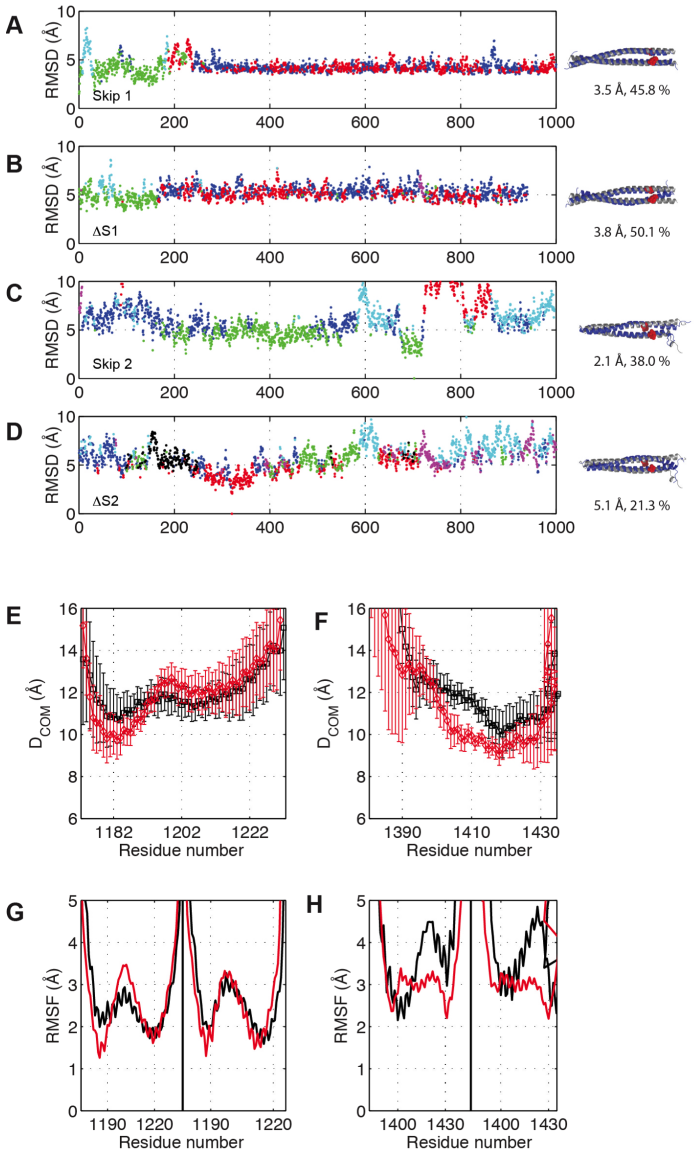


Figure C6: Analysis of the molecular dynamics simulations for Skip 1 and 2. Clustering analysis of the simulations for Skip 1, ΔS1, Skip 2, ΔS2 is shown in (A)-(D) respectively. DCOM, average distance between the center of masses of the two α-helices for Skip 1 and Skip 2 are shown in (E) and (F). The RMSF are displayed in (G) and (H) for Skip 1 (Left) and Skip 2 (Right). The wild-type and ΔS constructs in panels (E) - (H) are depicted in black and red, respectively.

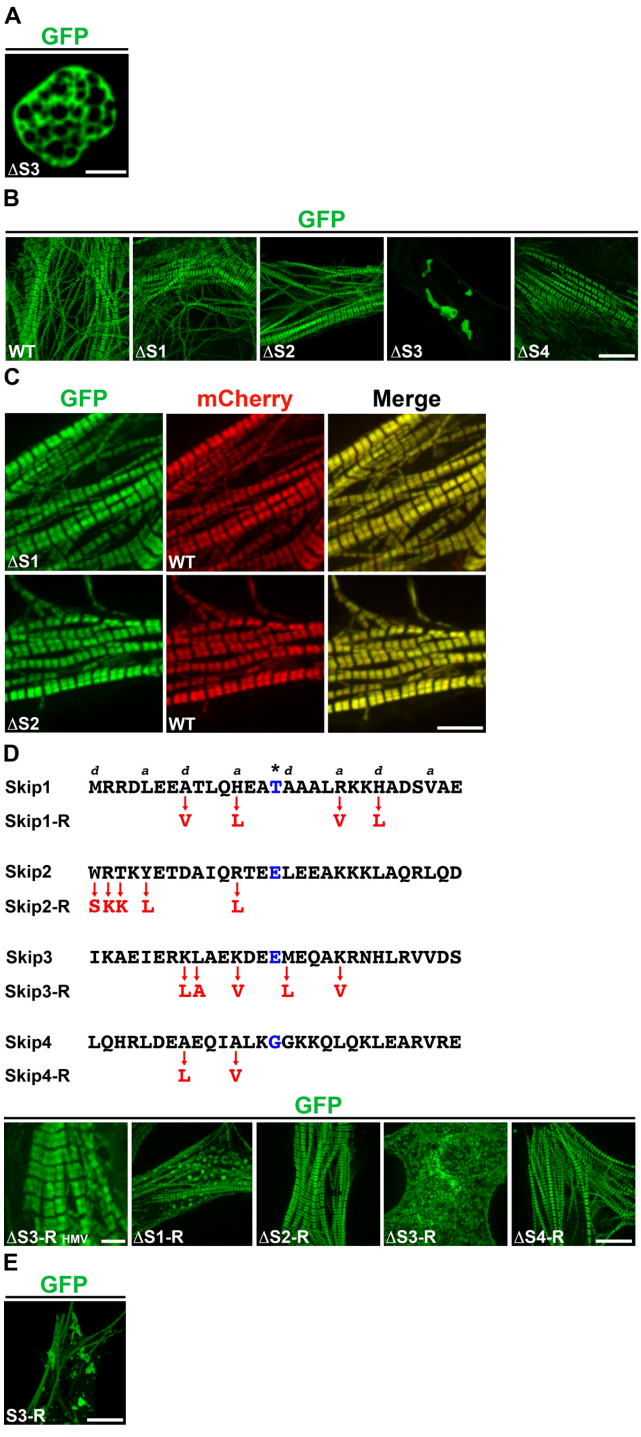

A
GFP
ΔS3
B
GFP
WT
ΔS1
ΔS2
ΔS3
ΔS4
C
GFP
mCherry
Merge
ΔS1
WT
ΔS2
WT
D
Skip1 MRRDLEEATLQHEATAAALRKKHADSVAE
Skip1-R V L V L
Skip2 WRTKYETDAIQRTEELEEAKKKLAQRLQD
Skip2-R SKK L L
Skip3 IKAEIERKLAEKDEEMEQAKRNHLRVVDS
Skip3-R LA V L V
Skip4 LQHRLDEAEQIALKGGKKQLQKLEARVRE
Skip4-R L V
GFP
ΔS3-R HMV
ΔS1-R
ΔS2-R
ΔS3-R
ΔS4-R
E
GFP
S3-R

Figure C7 *(previous page)*: Effects of skip residue deletions/recoiling and mutant overexpression on myosin incorporation into sarcomeres. (A) Structure of a cytoplasmic aggregate imaged from cardiomyocytes transfected with the ΔS3 construct; bar, 2 μm. (B) GFP-tagged skip residue deletion constructs were transfected into cardiomyocytes using double the amount of plasmid DNA (4 μg); bar, 10 μm (C) Colocalization of skip residue deletion constructs and WT myosin. Cardiomyocytes were co-transfected with mutant GFP- and WT mCherry-tagged myosin constructs as indicated; bar, 5 μm (D) The top panel shows the amino acid surrounding the four skip residues (colored in blue), the canonical a and d position of the heptad repeat, and the amino acid changes introduced in the recoiled constructs (colored in red). Bottom panel: imaging of cardiomyocytes transfected with 2 μg (ΔS3-R HMV) or 4 μg (ΔS1-R, ΔS2-R, ΔS3-R, ΔS4-R) of GFP-tagged recoiling constructs ; ΔS3-R HMV bar, 2.5 μm; HMV: high magnification view; ΔS4-R bar, 10 μm. (E) Imaging of cardiomyocytes transfected with the S3-R GFP-tagged construct; bar, 10 μm.

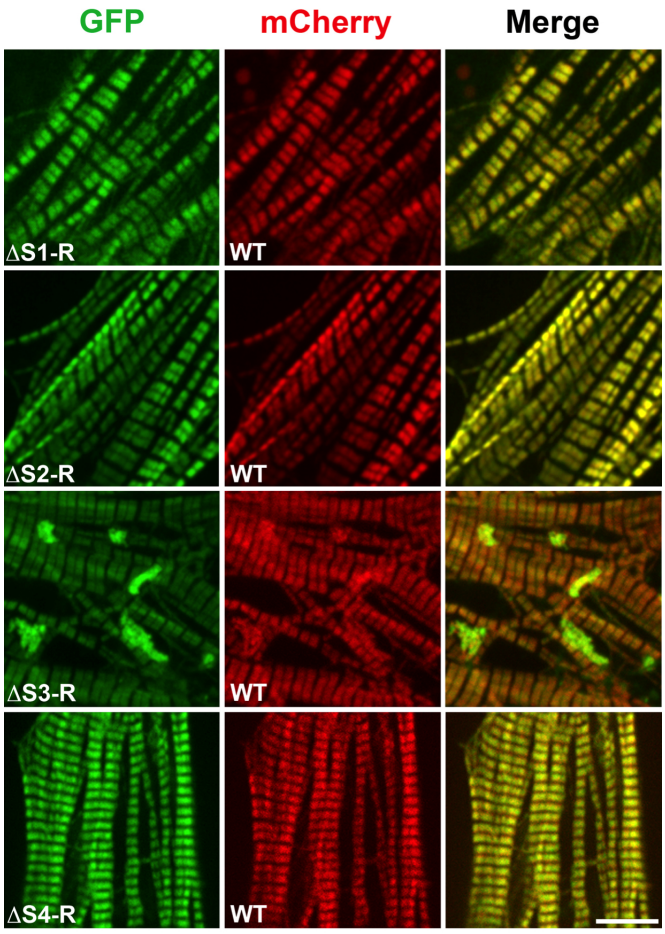


Figure C8: Colocalization of skip residue recoiling mutants and WT myosin. Cardiomyocytes were co-transfected with mutant GFP- and WT mCherry-tagged myosin constructs as indicated. Cells were imaged by confocal microscopy 96 h later.

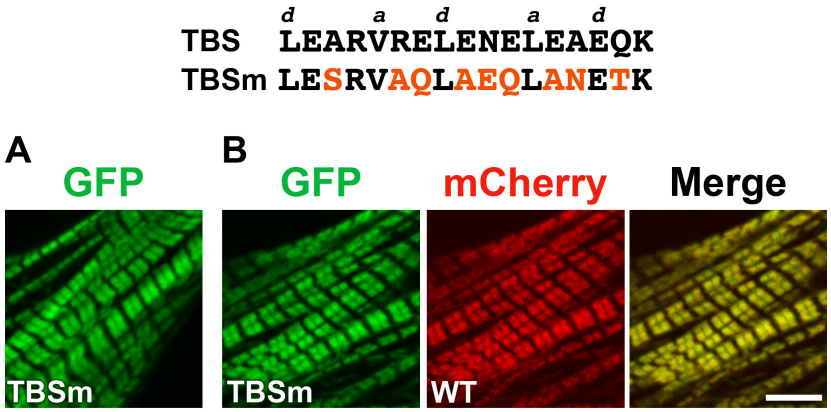


Figure C9: Mutagenesis of the titin binding site. Top: alignment the WT (TBS) and mutated (TBSm) titin binding sites. Amino acids changes are shown in red. The a and d positions of the heptad repeat are shown. (A) Cardiomyocytes were transfected with TBSm GFP-tagged myosin rod; (B) Cardiomyocytes were co-transfected with TBSm GFP- and WT mCherry-tagged myosin rod constructs. Cells in both panels were imaged by confocal microscopy 96 h after transfection. Bar, 5 μm.

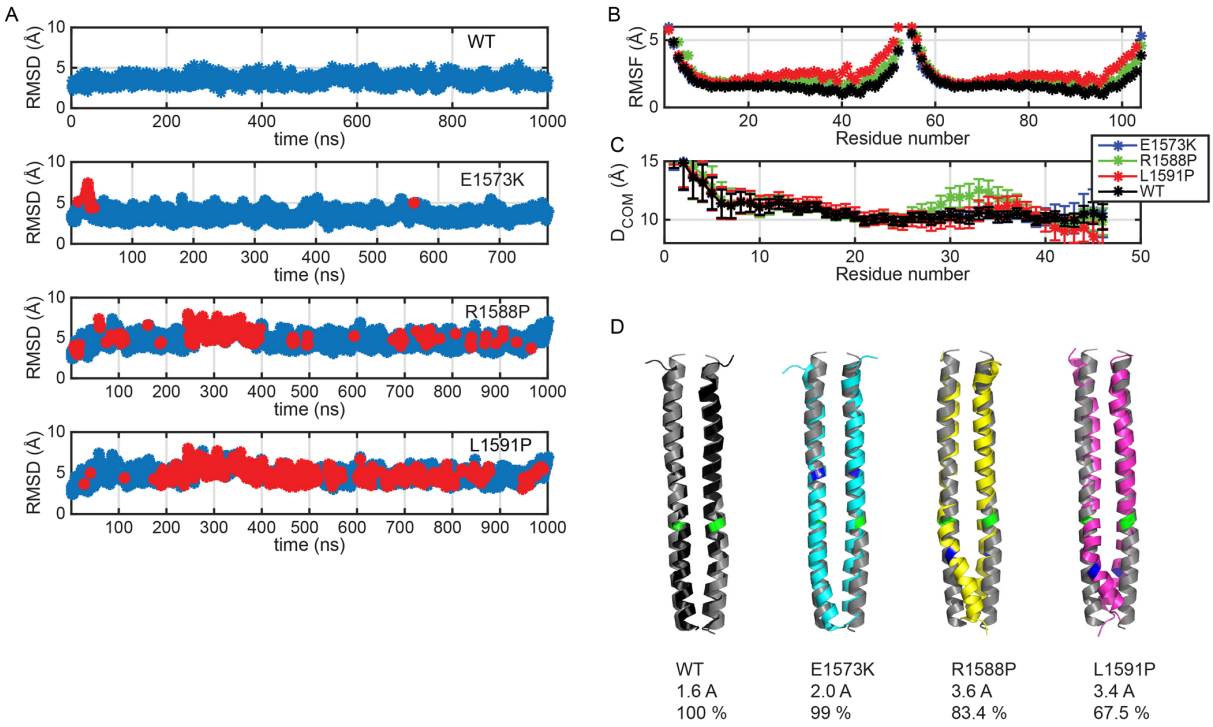


Figure C10: Simulations of disease-causing mutations surrounding Skip 3 relative to the WT. (A) Each simulation is clustered individually via k-means methods using 6 Å Cα-RMSD as the threshold and the results are shown in individual panels. Blue and Red represent different clusters formed. (B) Cα-RMSF is displayed for WT (black), E1573K (blue), R1588P (Green) and L1591P (Red), respectively. (C) Distance between center of masses of each helix (DCOM) is calculated as a moving average over Cα atoms of 7 consecutive residues for WT (black), E1573K (blue), R1588P (Green) and L1591P (Red), respectively. (D) The representative structures of each most populated cluster (WT: black), E1573K: blue, R1588P: Green) and L1591P: Red) , is aligned on the crystal structure (grey). The percentage of population that each is cluster spans and the Cα-RMSD with respect to the crystal structure is shown. The skip residue is (E1582) is colored in green, whereas the location of the mutation is shown by dark blue.

C.4 References

[1] V. A. KLENCHIN, J. J. FRYE, M. H. JONES, M. WINEY, and I. RAYMENT, *The Journal of biological chemistry* **286**, 18240 (2011).

[2] M. BUVOLI, A. BUVOLI, and L. A. LEINWAND, *J Mol Biol* **415**, 807 (2012).

[3] Y. IMAI, Y. MATSUSHIMA, T. SUGIMURA, and M. TERADA, *Nucleic Acids Res* **19**, 2785 (1991).

[4] G. D. VAN DUYNE, R. F. STANDAERT, P. A. KARPLUS, S. L. SCHREIBER, and J. CLARDY, *J Mol Biol* **229**, 105 (1993).

[5] I. RAYMENT, *Methods Enzymol* **276**, 171 (1997).

[6] Z. OTWINOWSKI and W. MINOR, *Methods in Enzymology* **276**, 307 (1997).

[7] M. STRONG, M. R. SAWAYA, S. WANG, M. PHILLIPS, D. CASCIO, and D. EISENBERG, *Proceedings of the National Academy of Sciences* **103**, 8060 (2006).

[8] A. J. MCCOY, R. W. GROSSE-KUNSTLEVE, P. D. ADAMS, M. D. WINN, L. C. STORONI, and R. J. READ, *J Appl Crystallogr* **40**, 658 (2007).

[9] P. D. ADAMS, P. V. AFONINE, G. BUNKOCZI, V. B. CHEN, I. W. DAVIS, N. ECHOLS, J. J. HEADD, L. W. HUNG, G. J. KAPRAL, R. W. GROSSE-KUNSTLEVE, A. J. MCCOY, N. W. MORIARTY, R. OEFFNER, R. J. READ, D. C. RICHARDSON, J. S. RICHARDSON, T. C. TERWILLIGER, and P. H. ZWART, *Acta Crystallogr D Biol Crystallogr* **66**, 213 (2010).

[10] G. N. MURSHUDOV, P. SKUBÁK, A. A. LEBEDEV, N. S. PANNU, R. A. STEINER, R. A. NICHOLLS, M. D. WINN, F. LONG, and A. A. VAGIN, *Acta Crystallographica Section D: Biological Crystallography* **67**, 355 (2011).

[11] V. B. CHEN, W. B. ARENDALL, J. J. HEADD, D. A. KEEDY, R. M. IMMORMINO, G. J. KAPRAL, L. W. MURRAY, J. S. RICHARDSON, and D. C. RICHARDSON, *Acta Crystallographica Section D: Biological Crystallography* **66**, 12 (2010).

[12] A. H. MAASS and M. BUVOLI, *Methods Mol Biol* **366**, 321 (2007).

[13] C. A. SCHNEIDER, W. S. RASBAND, and K. W. ELICEIRI, *Nat Methods* **9**, 671 (2012).

[14] N. A. BAKER, D. SEPT, S. JOSEPH, M. J. HOLST, and J. A. MCCAMMON, *Proceedings of the National Academy of Sciences* **98**, 10037 (2001).

[15] W. L. DELANO, (2002).

[16] T. J. DOLINSKY, P. CZODROWSKI, H. LI, J. E. NIELSEN, J. H. JENSEN, G. KLEBE, and N. A. BAKER, *Nucleic acids research* **35**, W522 (2007).

Appendix D

Supporting Information for Composite Approach Towards a Complete Model of the Myosin Rod

D.1 Accession Codes

Atomic coordinates have been deposited in the Protein Data Bank under accession codes 5CJ1, 5CJ4, 5CHX, and 5CJ0 corresponding to Gp7-1525-1571, Xrcc4-1562-1622, Xrcc4-1590-1657, and Xrcc4-1629-1692 respectively.

D.2 Additional Data

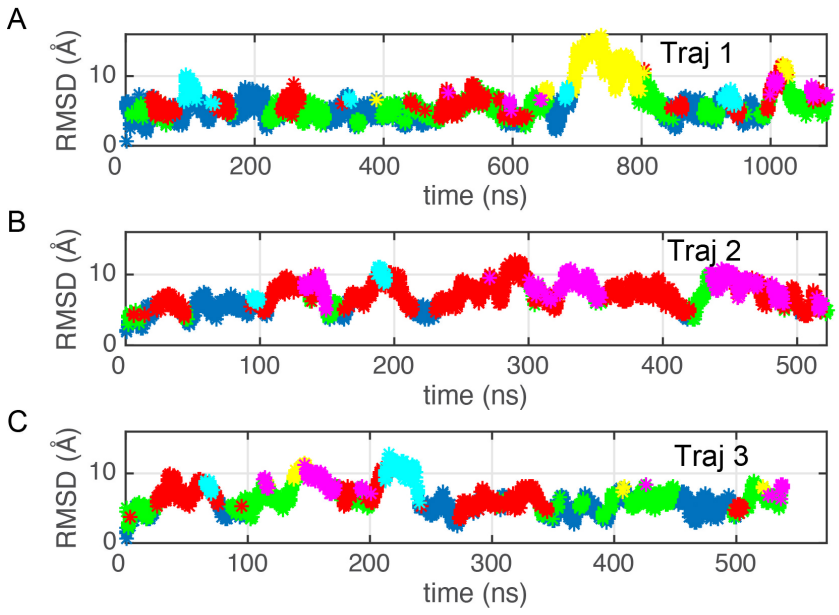


Figure D1: Distribution of clusters among three independent trajectories where each panel constitutes the conformations from a different simulation.

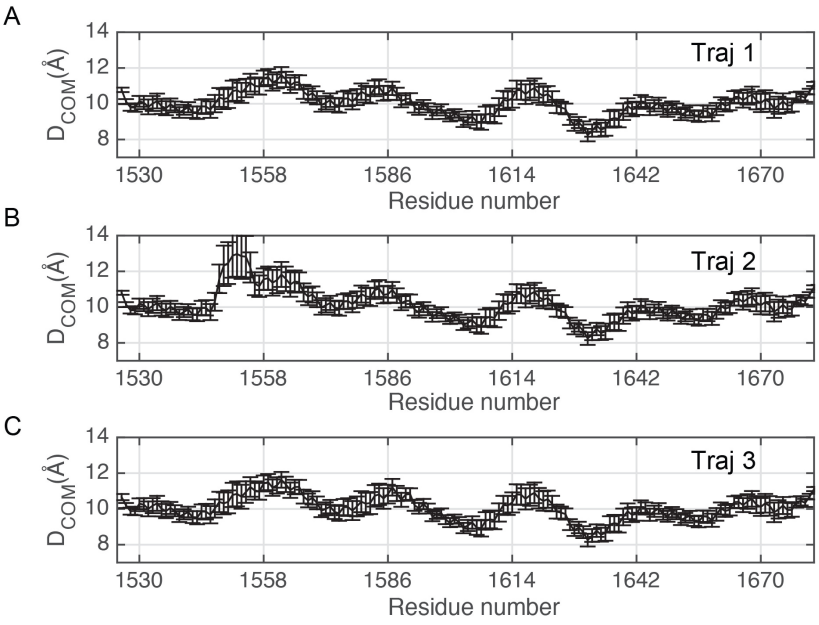


Figure D2: DCOM for the three independent simulations of the composite model. See Figure 6.8 for average over conformations from all 3 simulations.

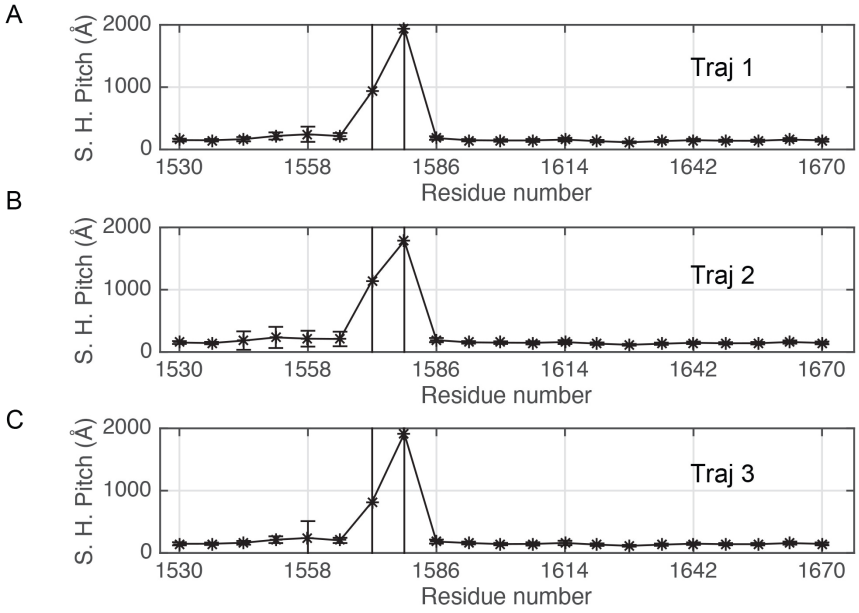


Figure D3: Super helical pitch for the three independent simulations of the composite model. These were calculated and averaged over the conformations of each individual simulation. See Figure 6.8 for average over conformations from all three simulations.

Table D1: Potential charged interactions within the rod

	Residue number and Type	Residue number and Type	Distance (Å)
Between helices	1536 E (g)	1537 K (a)	4.0 ± 0.8
	1537 K (a)	1536 E (g)	4.0 ± 0.9
	1579 K (a)	1580 D (b)	4.4 ± 0.9
	1580 D (b)	1579 K (a)	4.4 ± 1.0
	1604 E (d)	1608 R (a)	4.0 ± 0.3
	1608 R (a)	1604 E (d)	4.0 ± 0.4
Within 1st helix	1602 D (b)	1606 R (f)	4.8 ± 1.1
	1615 K (a)	1619 E (e)	5.0 ± 1.5
Within 2nd helix	1602 D (b)	1606 R (f)	4.8 ± 1.1
	1615 K (a)	1619 E (e)	5.0 ± 1.5

Residue pairs that are within 4.5 Å according to the average distance between the center of mass of nitrogen atoms (NH1, NH2) of Arginine, nitrogen atom (NZ) of Lysine and center of mass of oxygen atoms from Glutamate (OE1, OE2) and Aspartate (OD1, OD2) are listed. The distances are averaged over a combined trajectory of three independent simulations.

Table D2: Potential charged interactions within the rod

Residue number and Type	Residue number and Type	Distance (Å)
1527 E (e)	1530 E (e)	4.52 ± 2.2
1529 V (g)	1530 V (g)	2.60 ± 0.3
1530 R (a)	1527 R (a)	4.76 ± 2.1
1530 R (a)	1529 R (a)	2.60 ± 0.3
1530 R (a)	1530 R (a)	3.54 ± 0.6
1530 R (a)	1533 R (a)	2.89 ± 0.5
1533 L (d)	1530 L (d)	2.94 ± 0.5
1533 L (d)	1533 L (d)	2.32 ± 0.2
1533 L (d)	1534 L (d)	2.91 ± 0.5
1533 L (d)	1537 L (d)	2.73 ± 0.4
1534 E (e)	1533 E (e)	2.87 ± 0.4
1536 E (g)	1537 E (g)	2.16 ± 0.6
1537 K (a)	1533 K (a)	2.78 ± 0.4
1537 K (a)	1536 K (a)	2.17 ± 0.6
1537 K (a)	1537 K (a)	3.75 ± 0.7
1537 K (a)	1540 K (a)	2.56 ± 0.3
1540 L (d)	1537 L (d)	2.56 ± 0.3
1540 L (d)	1540 L (d)	2.43 ± 0.3
1540 L (d)	1541 L (d)	2.71 ± 0.4
1540 L (d)	1544 L (d)	2.58 ± 0.3
1541 Q (e)	1540 Q (e)	2.73 ± 0.4
1543 A (g)	1544 A (g)	2.73 ± 0.4
1544 L (a)	1540 L (a)	2.59 ± 0.3
1544 L (a)	1543 L (a)	2.84 ± 0.5
1544 L (a)	1544 L (a)	2.73 ± 0.3
1544 L (a)	1547 L (a)	3.24 ± 0.5
1547 A (d)	1544 A (d)	3.28 ± 0.5
1547 A (d)	1547 A (d)	2.90 ± 0.5
1547 A (d)	1548 A (d)	4.60 ± 0.7
1547 A (d)	1551 A (d)	3.97 ± 0.8
1548 E (e)	1547 E (e)	4.69 ± 0.7
1550 S (g)	1551 S (g)	2.90 ± 0.6
1551 L (a)	1547 L (a)	4.03 ± 0.9
1551 L (a)	1550 L (a)	2.95 ± 0.7
1551 L (a)	1551 L (a)	2.66 ± 0.4
1551 L (a)	1554 L (a)	2.95 ± 0.7
1554 E (d)	1551 E (d)	4.95 ± 4.1
1557 K (g)	1558 K (g)	4.18 ± 1.5
1558 I (a)	1554 I (a)	3.59 ± 1.3
1558 I (a)	1557 I (a)	4.28 ± 1.4
1558 I (a)	1558 I (a)	2.40 ± 0.3
1558 I (a)	1561 I (a)	4.80 ± 2.1

1561 A (d)	1558 A (d)	3.96 ± 1.4
1561 A (d)	1562 A (d)	4.85 ± 1.0
1562 Q (e)	1561 Q (e)	4.91 ± 1.1
1565 F (a)	1561 F (a)	4.95 ± 2.1
1565 F (a)	1562 F (a)	4.60 ± 2.1
1565 F (a)	1565 F (a)	2.81 ± 0.3
1565 F (a)	1568 F (a)	4.79 ± 2.2
1568 I (d)	1565 I (d)	4.32 ± 2.1
1568 I (d)	1569 I (d)	3.25 ± 0.8
1569 K (e)	1565 K (e)	4.49 ± 1.8
1569 K (e)	1568 K (e)	3.15 ± 0.7
1569 K (e)	1572 K (e)	3.61 ± 1.1
1572 I (a)	1569 I (a)	3.31 ± 0.9
1572 I (a)	1572 I (a)	2.34 ± 0.2
1572 I (a)	1573 I (a)	3.92 ± 1.0
1572 I (a)	1576 I (a)	3.45 ± 1.1
1573 E (b)	1572 E (b)	3.78 ± 0.9
1575 K (d)	1576 K (d)	2.67 ± 0.4
1576 L (e)	1572 L (e)	3.17 ± 1.0
1576 L (e)	1575 L (e)	2.66 ± 0.4
1576 L (e)	1576 L (e)	2.71 ± 0.3
1576 L (e)	1579 L (e)	2.88 ± 0.5
1579 K (a)	1576 K (a)	2.83 ± 0.5
1579 K (a)	1579 K (a)	2.85 ± 0.6
1579 K (a)	1580 K (a)	2.43 ± 0.8
1579 K (a)	1583 K (a)	3.38 ± 0.6
1580 D (b)	1579 D (b)	2.39 ± 0.8
1582 E (c)	1583 E (c)	2.86 ± 0.7
1583 M (d)	1579 M (d)	3.36 ± 0.7
1583 M (d)	1582 M (d)	2.89 ± 0.7
1583 M (d)	1583 M (d)	2.58 ± 0.3
1583 M (d)	1586 M (d)	3.24 ± 0.8
1586 A (g)	1583 A (g)	3.17 ± 0.8
1586 A (g)	1587 A (g)	4.67 ± 0.8
1587 K (a)	1586 K (a)	4.66 ± 0.8
1587 K (a)	1590 K (a)	4.16 ± 2.1
1590 H (d)	1587 H (d)	4.11 ± 1.9
1590 H (d)	1590 H (d)	2.63 ± 0.3
1590 H (d)	1594 H (d)	4.14 ± 0.9
1593 V (g)	1594 V (g)	2.74 ± 0.4
1594 V (a)	1590 V (a)	4.01 ± 0.9
1594 V (a)	1593 V (a)	2.69 ± 0.4
1594 V (a)	1594 V (a)	2.77 ± 0.5
1594 V (a)	1597 V (a)	2.65 ± 0.4
1597 L (d)	1594 L (d)	2.62 ± 0.4
1597 L (d)	1597 L (d)	2.32 ± 0.2
1597 L (d)	1598 L (d)	2.87 ± 0.5

1597 L (d)	1601 L (d)	2.80 ± 0.4
1598 Q (e)	1597 Q (e)	2.83 ± 0.5
1600 S (g)	1601 S (g)	2.75 ± 0.4
1601 L (a)	1597 L (a)	2.73 ± 0.4
1601 L (a)	1600 L (a)	2.67 ± 0.4
1601 L (a)	1601 L (a)	2.83 ± 0.3
1601 L (a)	1604 L (a)	2.70 ± 0.4
1604 E (d)	1601 E (d)	2.68 ± 0.3
1604 E (d)	1604 E (d)	2.36 ± 0.2
1604 E (d)	1605 E (d)	2.73 ± 0.5
1604 E (d)	1608 E (d)	1.81 ± 0.1
1605 T (e)	1604 T (e)	2.83 ± 0.5
1607 S (g)	1608 S (g)	2.68 ± 0.4
1608 R (a)	1604 R (a)	1.82 ± 0.1
1608 R (a)	1607 R (a)	2.62 ± 0.3
1608 R (a)	1608 R (a)	2.50 ± 0.3
1608 R (a)	1611 R (a)	3.87 ± 0.7
1611 A (d)	1608 A (d)	3.48 ± 0.7
1611 A (d)	1611 A (d)	2.89 ± 0.5
1611 A (d)	1612 A (d)	3.87 ± 0.9
1611 A (d)	1615 A (d)	4.72 ± 1.1
1612 L (e)	1611 L (e)	3.94 ± 0.9
1614 V (g)	1615 V (g)	2.62 ± 0.5
1615 K (a)	1611 K (a)	4.30 ± 1.1
1615 K (a)	1614 K (a)	2.64 ± 0.5
1615 K (a)	1615 K (a)	4.12 ± 1.2
1615 K (a)	1618 K (a)	3.12 ± 0.8
1618 M (d)	1615 M (d)	3.02 ± 0.7
1618 M (d)	1618 M (d)	2.64 ± 0.4
1618 M (d)	1619 M (d)	3.00 ± 0.8
1618 M (d)	1622 M (d)	2.89 ± 0.6
1619 E (e)	1618 E (e)	2.97 ± 0.8
1621 D (g)	1622 D (g)	3.37 ± 0.7
1622 L (a)	1618 L (a)	2.89 ± 0.6
1622 L (a)	1621 L (a)	3.35 ± 0.7
1622 L (a)	1622 L (a)	2.56 ± 0.4
1622 L (a)	1625 L (a)	2.78 ± 0.5
1625 M (d)	1622 M (d)	2.79 ± 0.5
1625 M (d)	1625 M (d)	2.57 ± 0.4
1625 M (d)	1626 M (d)	3.05 ± 0.8
1625 M (d)	1629 M (d)	2.67 ± 0.4
1626 E (e)	1625 E (e)	3.02 ± 0.8
1628 Q (g)	1629 Q (g)	2.67 ± 0.4
1629 L (a)	1625 L (a)	2.70 ± 0.5
1629 L (a)	1628 L (a)	2.67 ± 0.4
1629 L (a)	1629 L (a)	2.75 ± 0.4
1629 L (a)	1632 L (a)	3.23 ± 0.6

1632 A (d)	1629 A (d)	3.27 ± 0.6
1632 A (d)	1632 A (d)	2.76 ± 0.4
1632 A (d)	1633 A (d)	3.82 ± 0.6
1632 A (d)	1636 A (d)	4.18 ± 0.9
1633 N (e)	1632 N (e)	3.83 ± 0.6
1635 M (g)	1636 M (g)	2.80 ± 0.5
1635 M (g)	1640 M (g)	3.73 ± 1.2
1636 A (a)	1632 A (a)	4.22 ± 0.9
1636 A (a)	1635 A (a)	2.80 ± 0.5
1636 A (a)	1636 A (a)	2.79 ± 0.5
1639 A (d)	1639 A (d)	2.49 ± 0.3
1639 A (d)	1640 A (d)	4.10 ± 1.0
1639 A (d)	1643 A (d)	3.70 ± 1.0
1640 Q (e)	1635 Q (e)	3.76 ± 1.2
1640 Q (e)	1639 Q (e)	4.05 ± 1.0
1642 Q (g)	1643 Q (g)	2.84 ± 0.5
1643 V (a)	1639 V (a)	3.69 ± 0.9
1643 V (a)	1642 V (a)	2.86 ± 0.5
1643 V (a)	1643 V (a)	2.48 ± 0.3
1643 V (a)	1646 V (a)	2.70 ± 0.4
1646 L (d)	1643 L (d)	2.68 ± 0.4
1646 L (d)	1646 L (d)	2.40 ± 0.2
1646 L (d)	1647 L (d)	2.85 ± 0.5
1646 L (d)	1650 L (d)	2.63 ± 0.3
1647 Q (e)	1646 Q (e)	2.88 ± 0.5
1649 L (g)	1650 L (g)	2.52 ± 0.3
1650 L (a)	1646 L (a)	2.61 ± 0.3
1650 L (a)	1649 L (a)	2.52 ± 0.3
1650 L (a)	1650 L (a)	2.71 ± 0.3
1650 L (a)	1653 L (a)	2.98 ± 0.4
1653 T (d)	1650 T (d)	2.98 ± 0.4
1653 T (d)	1653 T (d)	2.49 ± 0.3
1653 T (d)	1654 T (d)	3.12 ± 0.6
1653 T (d)	1657 T (d)	2.70 ± 0.4
1654 Q (e)	1653 Q (e)	3.17 ± 0.6
1656 Q (g)	1657 Q (g)	2.57 ± 0.3
1657 L (a)	1653 L (a)	2.72 ± 0.4
1657 L (a)	1656 L (a)	2.57 ± 0.3
1657 L (a)	1657 L (a)	2.65 ± 0.3
1657 L (a)	1660 L (a)	3.23 ± 0.4
1660 A (d)	1657 A (d)	3.20 ± 0.4
1660 A (d)	1660 A (d)	2.75 ± 0.4
1660 A (d)	1661 A (d)	3.60 ± 0.6
1660 A (d)	1664 A (d)	3.96 ± 0.7
1661 V (e)	1660 V (e)	3.64 ± 0.6
1663 A (g)	1664 A (g)	3.12 ± 0.6
1664 N (a)	1660 N (a)	3.94 ± 0.7

1664 N (a)	1663 N (a)	3.13 ± 0.6
1664 N (a)	1664 N (a)	2.27 ± 0.6
1664 N (a)	1667 N (a)	3.16 ± 0.6
1667 L (d)	1664 L (d)	3.12 ± 0.6
1667 L (d)	1667 L (d)	2.39 ± 0.2
1667 L (d)	1668 L (d)	2.97 ± 0.5
1667 L (d)	1671 L (d)	2.64 ± 0.5
1668 K (e)	1667 K (e)	3.00 ± 0.6
1670 N (g)	1671 N (g)	3.11 ± 0.7
1671 I (a)	1667 I (a)	2.62 ± 0.4
1671 I (a)	1670 I (a)	3.12 ± 0.7
1671 I (a)	1671 I (a)	2.35 ± 0.2
1671 I (a)	1674 I (a)	3.28 ± 0.9
1674 V (d)	1671 V (d)	3.23 ± 0.8
1674 V (d)	1674 V (d)	2.52 ± 0.3
1674 V (d)	1675 V (d)	3.67 ± 0.8
1674 V (d)	1678 V (d)	3.40 ± 0.8
1675 E (e)	1674 E (e)	3.73 ± 0.8
1677 R (g)	1678 R (g)	3.16 ± 0.7
1677 R (g)	1682 R (g)	4.89 ± 1.8
1678 N (a)	1674 N (a)	3.42 ± 0.8
1678 N (a)	1677 N (a)	3.22 ± 0.7
1678 N (a)	1678 N (a)	2.45 ± 0.5
1678 N (a)	1681 N (a)	3.08 ± 0.5
1681 L (d)	1678 L (d)	3.08 ± 0.5
1681 L (d)	1681 L (d)	2.45 ± 0.3
1681 L (d)	1682 L (d)	2.79 ± 0.5
1681 L (d)	1685 L (d)	2.54 ± 0.3
1682 Q (e)	1677 Q (e)	4.96 ± 1.8
1682 Q (e)	1681 Q (e)	2.83 ± 0.5
1684 E (g)	1685 E (g)	3.30 ± 0.8
1684 E (g)	1689 E (g)	4.22 ± 2.9
1685 L (a)	1681 L (a)	2.56 ± 0.4
1685 L (a)	1684 L (a)	3.40 ± 0.8
1685 L (a)	1685 L (a)	2.53 ± 0.3
1685 L (a)	1688 L (a)	3.16 ± 0.8
1688 L (d)	1685 L (d)	3.12 ± 0.7
1688 L (d)	1688 L (d)	2.55 ± 0.4
1688 L (d)	1689 L (d)	3.98 ± 1.5
1689 R (e)	1684 R (e)	4.35 ± 2.9
1689 R (e)	1688 R (e)	3.97 ± 1.5

The hydrophobic interactions within the rod were evaluated by the average minimum distances between the side-chains from each helix. The pairs of residues within 5 Å are shown. The distances are averaged over a combined trajectory of three independent simulations.

Table D3: SASA (Å^2) of hydrophobic residues that are solvent exposed, averaged over a combined trajectory of three independent simulations.

Residue number and Type	SASA (Å^2)
1526 L	114.5 ± 19.8
1538 M	127.4 ± 12.1
1559 L	111.0 ± 13.9
1563 L	110.0 ± 9.5
1591 L	112.8 ± 9.0
1612 L	98.5 ± 11.0
1627 I	107.8 ± 8.1
1635 M	97.4 ± 17.4
1649 L	101.0 ± 6.3
1655 I	106.2 ± 8.1
1673 I	99.7 ± 11.7
1680 L	113.6 ± 10.0
1690 L	115.3 ± 20.5
1702 M	127.6 ± 12.0
1723 L	114.6 ± 8.3
1727 L	111.2 ± 9.2
1755 L	113.3 ± 9.2
1776 L	98.1 ± 11.1
1791 I	107.5 ± 8.4
1799 M	98.5 ± 17.4
1813 L	101.0 ± 6.3
1819 I	106.9 ± 7.9
1837 I	100.2 ± 11.7
1844 L	113.6 ± 10.1

Appendix E

Exploration of Martini Coarse-Grained Methods Using the Ubiquitin Dimers, Trimers and the Myosin Thick Filament as Benchmarks

E.1 Exploration of Martini Coarse-Grained Methods on Ubiquitin Dimers

Table E1: Lengths of Martini simulations for Ub2.

Dimer Structure	Martini Method	Length (μs)
2KDF	MP22	1
	ELNEDYN	1
2LVP	MP22	1
	ELNEDYN	1
2LVQ	MP22	1
	ELNEDYN	1
3AUL	MP22	1
	ELNEDYN	1
3NS8	MP22	1
	ELNEDYN	1
Clust 9[a]	MP22	1
	ELNEDYN	1
Clust 10[a]	MP22	1
	ELNEDYN	1

[a] Clusters obtained from 2KDF implicit solvent simulations.

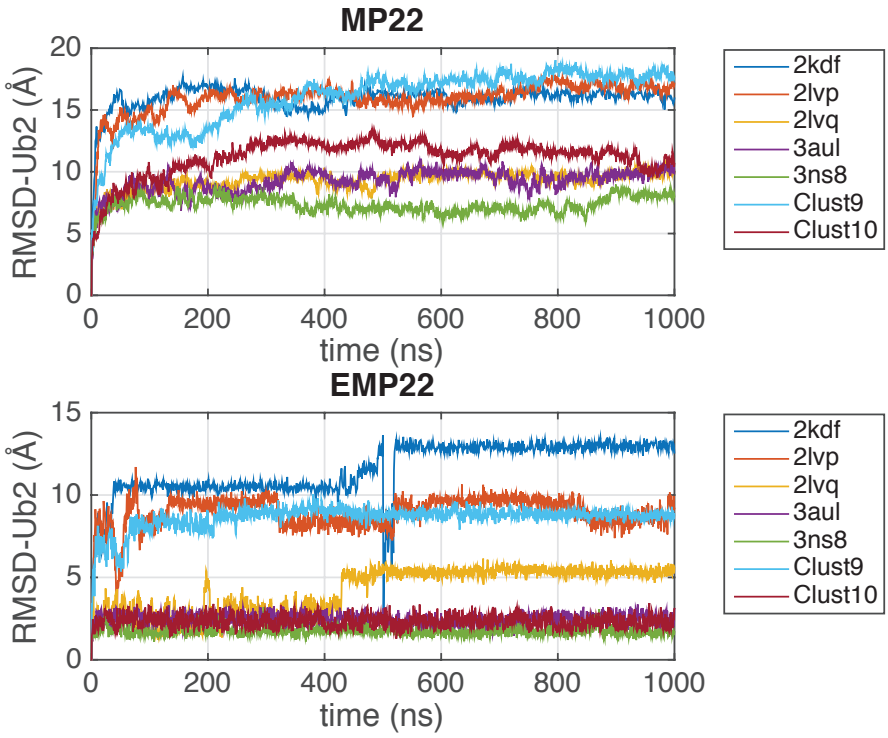


Figure E1: Cα-RMSD of the Ub2 after simulating with Martini force fields: (A) MP22 and (B) ELNEDYN for comparison.

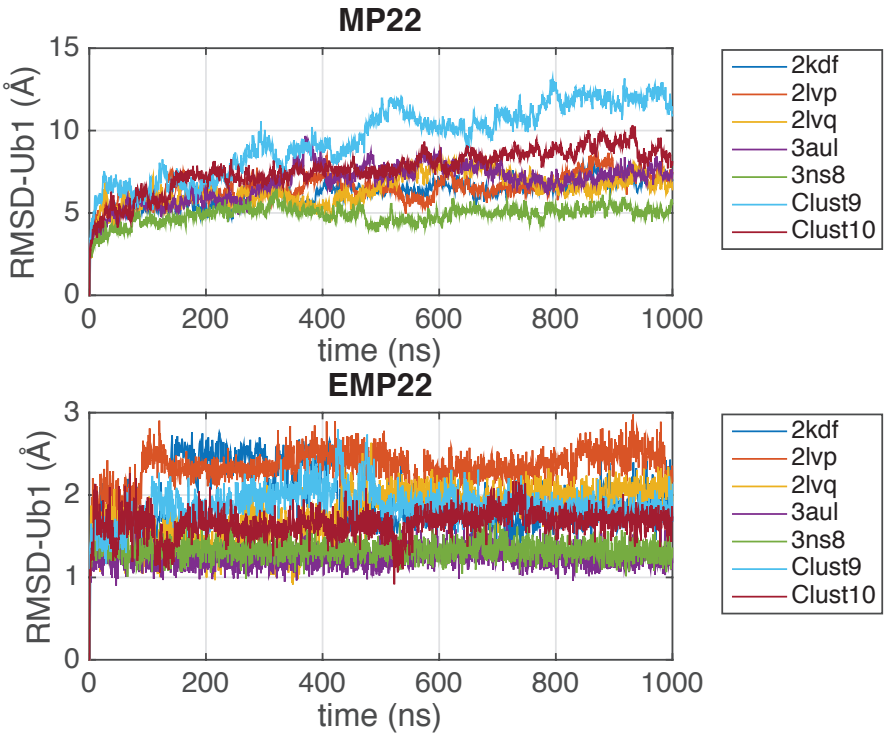


Figure E2: Cα-RMSD of the first Ub monomer alone, isolated from K48-linked dimer simulations using two protein Martini force fields: (A) MP22 and (B) ELNEDYN for comparison.

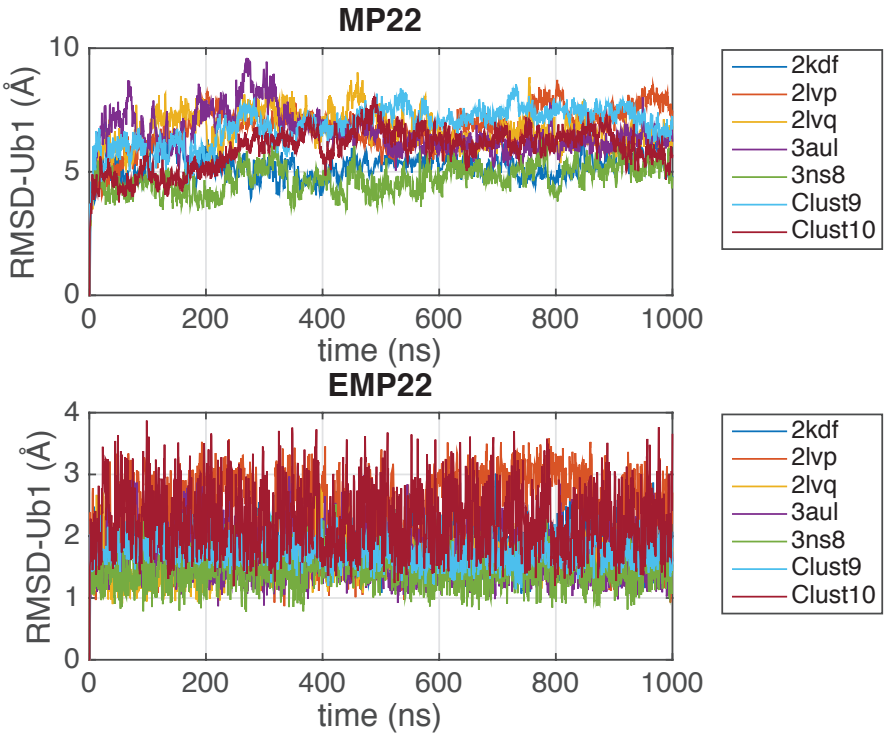


Figure E3: Cα-RMSD of the second Ub monomer alone, isolated from K48-linked dimer simulations using two protein Martini force fields: (A) MP22 and (B) ELNEDYN for comparison.

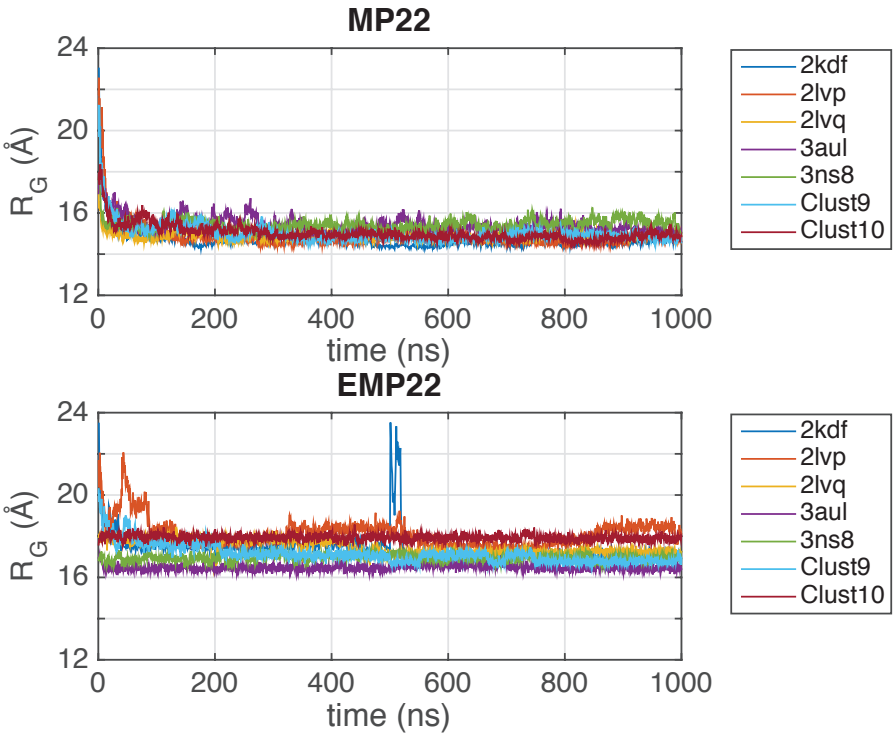


Figure E4: The radius of gyration R_g trends for the Ub2 simulations using two protein Martini force fields: (A) MP22 and (B) ELNEDYN for comparison.

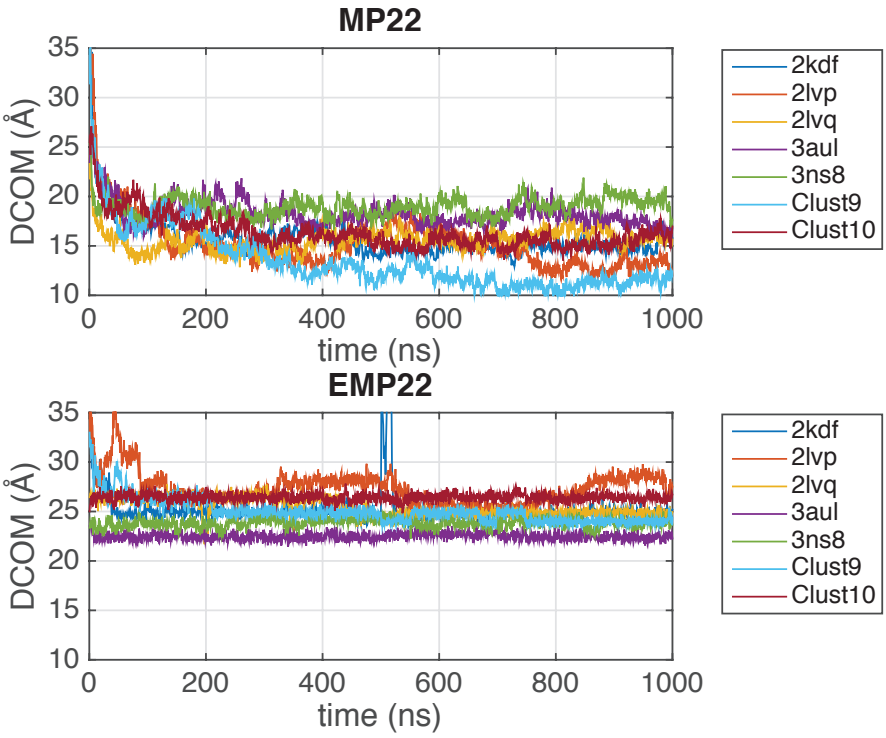


Figure E5: Distance between center of masses of the Ub monomer within Ub2 (DCOM) is calculated for different Ub2 simulations with Martini force fields: (A) MP22 and (B) ELNEDYN for comparison.

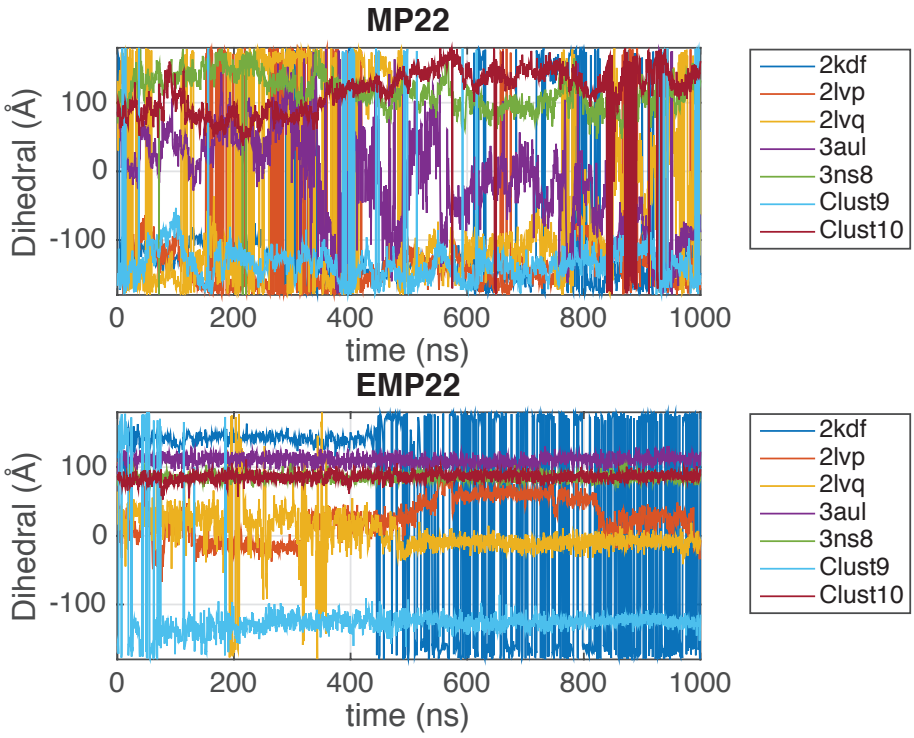


Figure E6: γ, the torsional angle created by the center of masses of the Ub monomers and the hydrophobic patches of the two subunits (residues L8, I44, H68 and V70) is calculated to quantify the degree of rotation of the subunits with respect to each other. γ is shown for different Ub2 simulations with Martini force fields: (A) MP22 and (B) ELNEDYN for comparison.

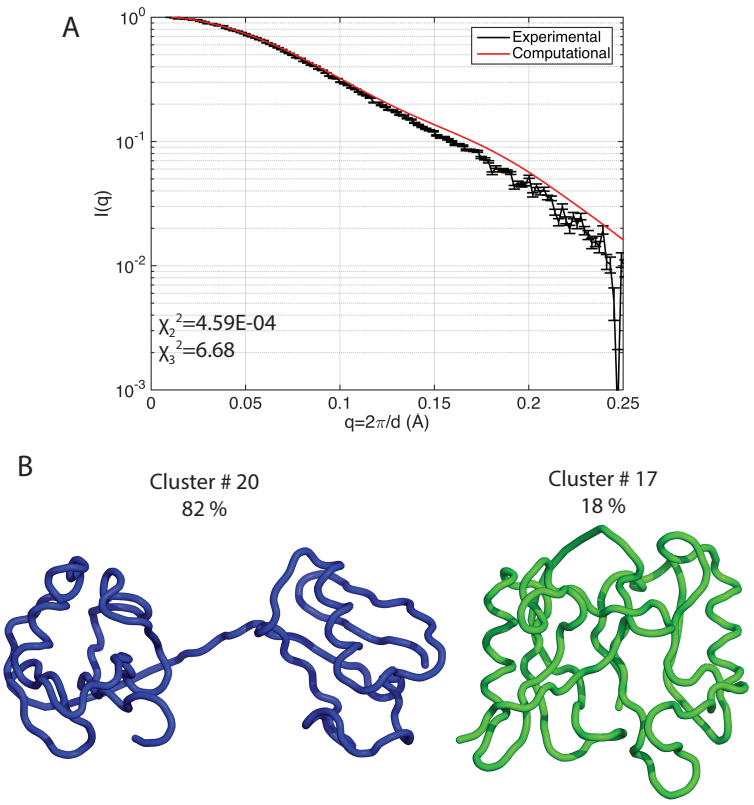


Figure E7: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K48-linked Ub2 simulated via Martini MP22 (B) Predicted conformational ensemble for K48-linked Ub2 based on Martini MP22 simulations.

Table E2: Properties calculated over clusters obtained from K48-linked Ub2 simulations with Martini MP22 force field.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K48-linked branched dimers.
1	18.60	14.65	22.27	0.00
2	17.12	14.92	27.75	0.00
3	18.51	14.51	22.31	0.00
4	17.36	14.67	22.55	0.00
5	17.91	14.35	23.03	0.00
6	17.83	14.30	21.61	0.00
7	18.32	14.56	21.74	0.00
8	17.60	15.13	26.66	0.00
9	20.12	14.59	23.65	0.00
10	20.62	14.53	22.42	0.00
11	18.42	14.61	23.91	0.00
12	14.84	15.75	30.96	0.00
13	17.29	14.67	24.04	0.00
14	17.27	14.51	22.17	0.00
15	18.12	14.62	21.09	0.00
16	8.37	20.30	31.55	0.00
17	18.61	15.19	25.70	0.18
18	17.14	15.02	23.99	0.00
19	17.94	14.52	22.44	0.00
20	6.64	20.50	30.63	0.82
21	18.59	15.11	22.50	0.00

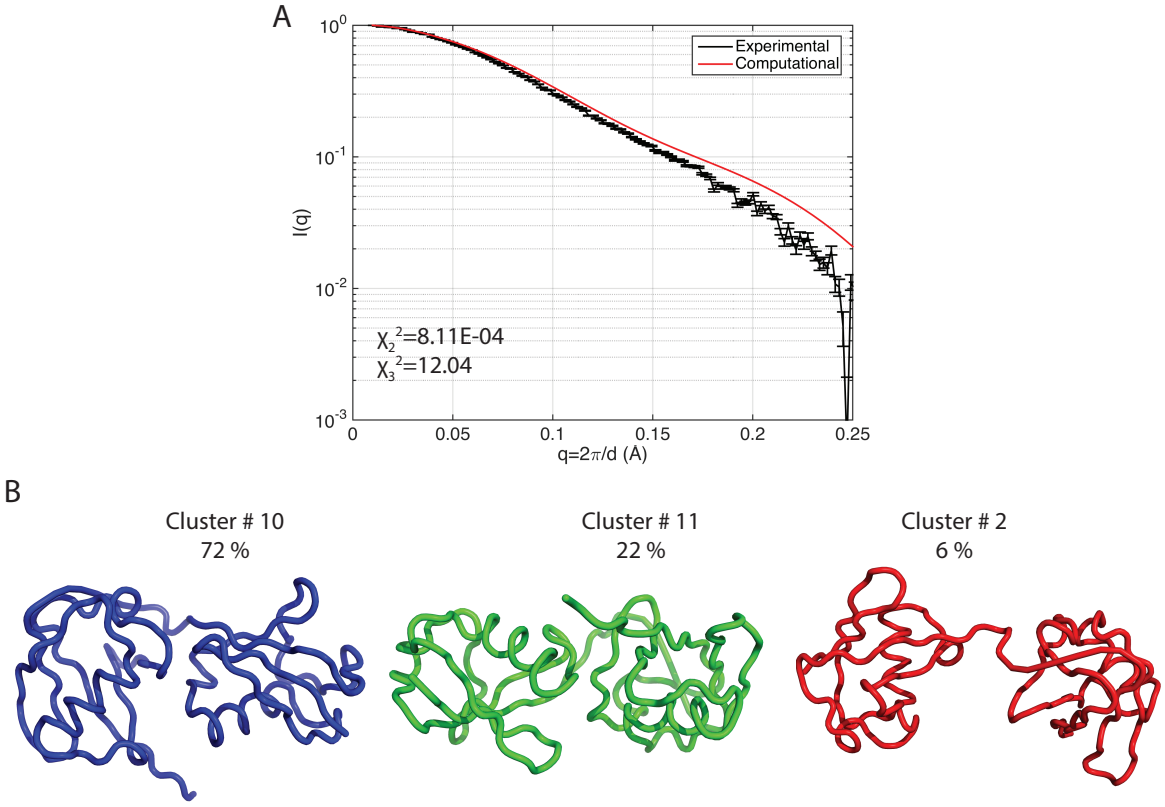


Figure E8: (A) Comparison of experimental SAXS profile with the prediction from the MD ensemble for the K48-linked Ub2 simulated via Martini ELNEDYN (B) Predicted conformational ensemble for K48-linked Ub2 based on Martini ELNEDYN simulations.

Table E3: Properties calculated over clusters obtained from K48-linked Ub2 simulations with Martini ELNE-DYN force field.

Cluster	RMSD (Å)	R_g (Å)	R_{max} (Å)	Calculated Populations for K48-linked branched dimers.
1	19.90	14.78	23.11	0.00
2	5.25	20.70	33.29	0.06
3	5.76	21.76	32.19	0.00
4	18.75	14.89	25.48	0.00
5	10.83	17.67	28.47	0.00
6	11.03	17.57	29.26	0.00
7	10.93	17.77	28.75	0.00
8	8.59	18.93	31.05	0.00
9	11.14	17.70	29.20	0.00
10	10.87	17.65	28.82	0.72
11	11.17	17.80	30.14	0.22

E.2 Exploration of Martini Coarse-Grained Methods on Segments of the Myosin Rod

Table E4: Simulations summarized in this section

Dimer Structure	Martini Method	Length (μs)
Skip 3	Implicit solvent[a]	1
	MP22	1
	ELNEDYN	1
ΔSkip 3	Implicit solvent[a]	1
	MP22	1
	ELNEDYN	1
Skip 4	Implicit solvent[a]	1
	MP22	1
	ELNEDYN	1
H1526–R1689	Implicit solvent[a]	1
	ELNEDYN	1
H1331-E1935	Implicit solvent[a]	1

[a] Generalized Born method (gb8) is used .

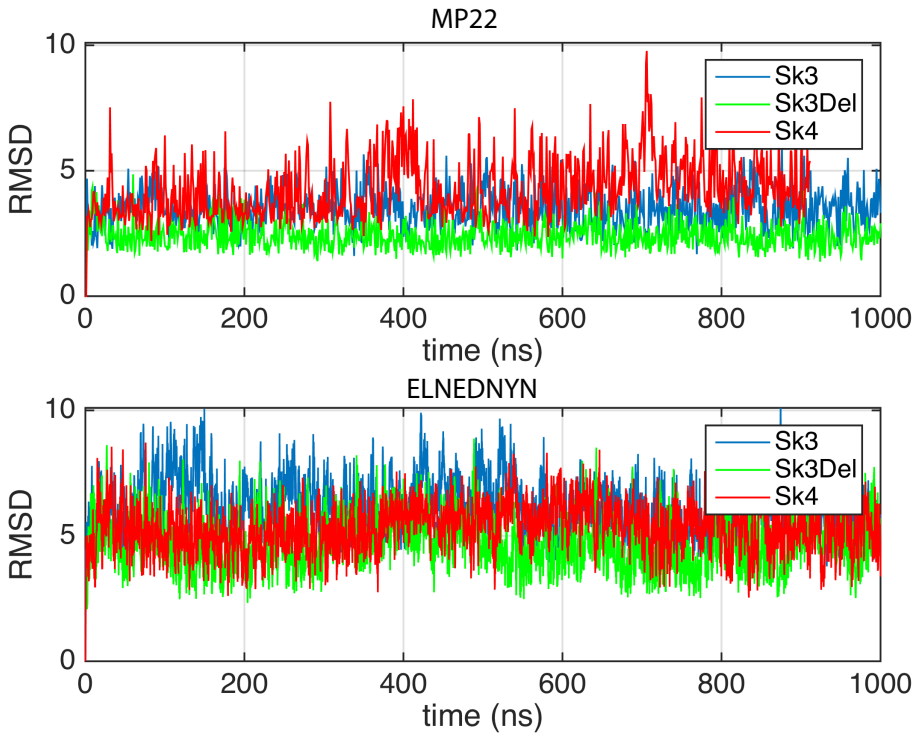


Figure E9: Cα-RMSD of the Skip 3, ΔSkip 3 and Skip 4 after simulating with Martini force fields: (A) MP22 and (B) ELNEDYN for comparison.

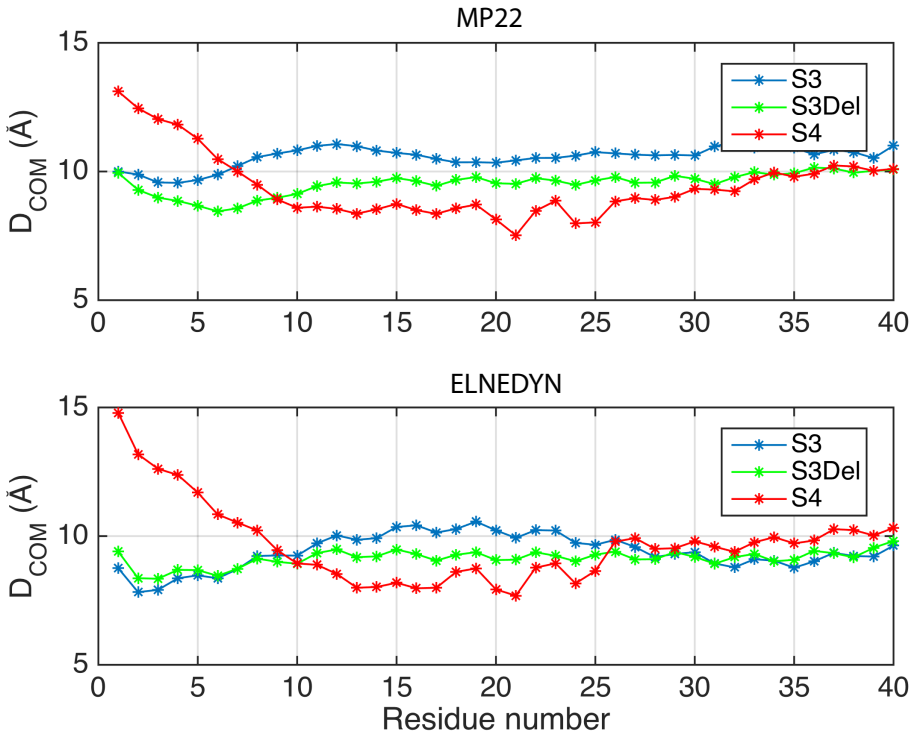


Figure E10: Distance between center of masses (DCOM) of the α-helices within Skip 3, ΔSkip 3 and Skip 4 is calculated asa a moving average of seven residues along the structures for simulation with Martini force fields: (A) MP22 and (B) ELNEDYN for comparison.

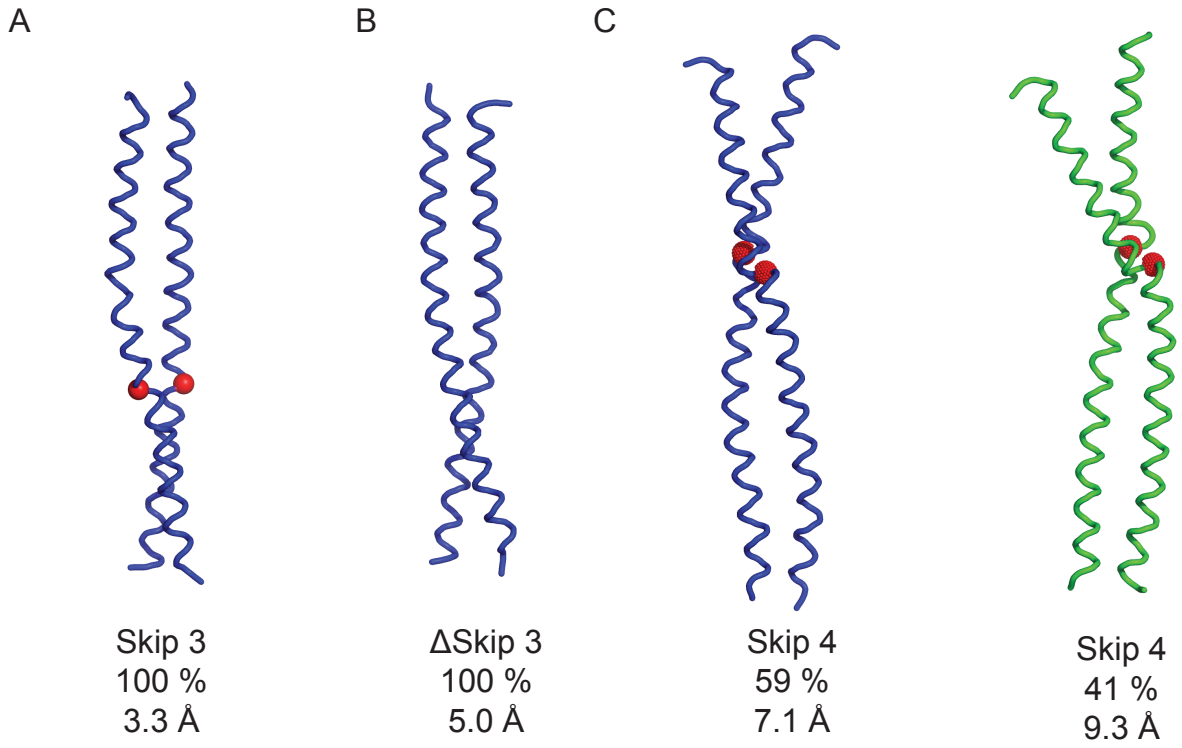


Figure E11: Representative structures for clusters obtained for Skip 3, ΔSkip 3 and Skip 4 simulations with MP22 Martini force field. Clustering is carried out at 7 Å threshold for the backbone RMSD using the *kclust* command. The percentage of each cluster is listed below the structures along with RMSD to crystal structures (PDB: 4XA4 and 4XA6).

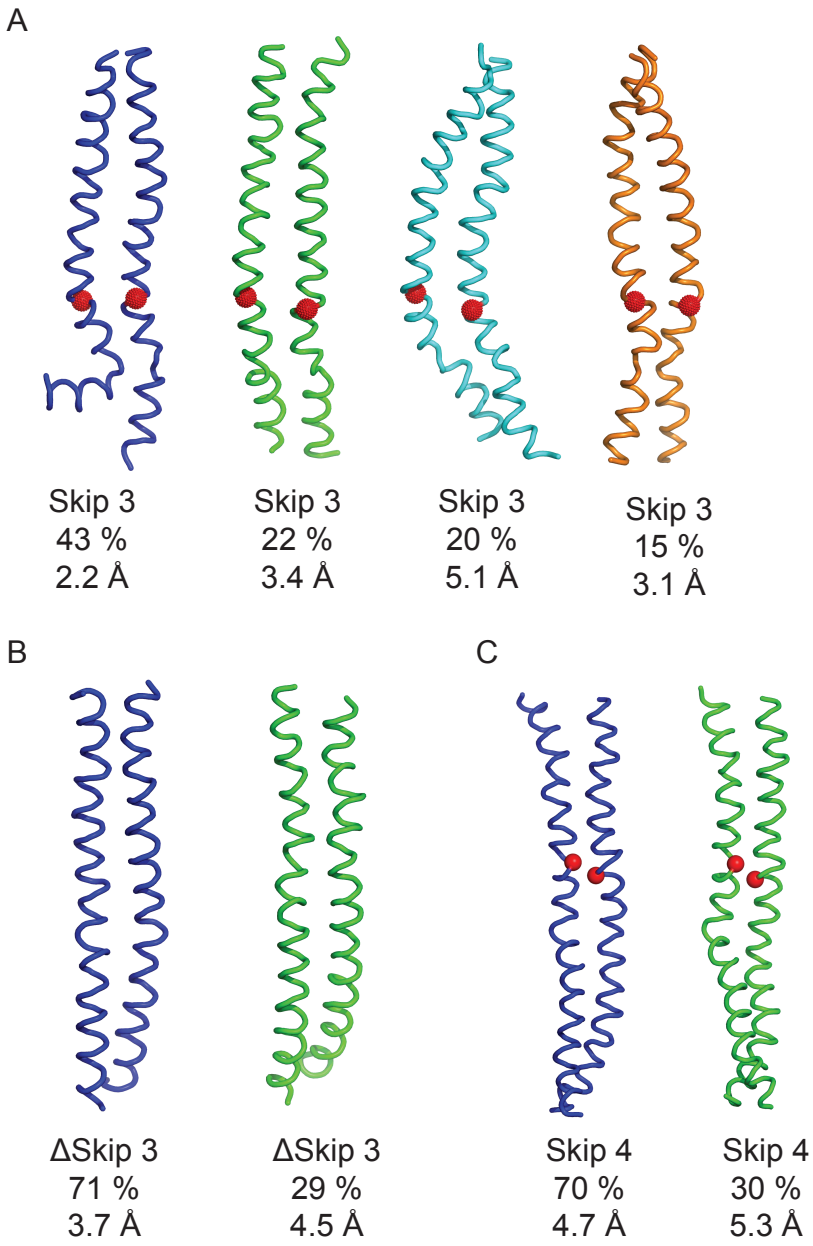


Figure E12: Representative structures for clusters obtained for Skip 3, ΔSkip 3 and Skip 4 simulations with ELNEDYN Martini force field. Clustering is carried out at 7 Å threshold for the backbone RMSD using the *kclust* command. The percentage of each cluster is listed below the structures along with RMSD to crystal structures (PDB: 4XA4 and 4XA6).

Table E5: Comparison of mechanical properties of segments of myosin rod of different lengths using different salvation techniques

Model	Solvation Method	Residue number per α-helix	θ (°)	δ (°)	s (Å)	C	$PL_{(R_G)}$ (Å)	$PL_{(Dee)}$ (Å)	PL_i (Å)	PL_d (Å)	k_f ($erg \times Å$)
Skip3	Implicit	52	15.09	8.20	70.84 ± 2.37	0.21	1226.24	2911.00	6890.82	2046.47	8.48E-11
	MP22		17.14	7.20	72.36 ± 1.13	0.24	1226.24	2911.00	9130.05	2207.25	9.14E-11
	ELNEDYN		28.22	12.52	63.28 ± 4.84	0.45	1360.58	3156.79	2629.88	1434.67	5.94E-11
ΔSkip3	Implicit	51	17.23	9.69	70.49 ± 2.24	0.24	1283.14	3044.66	4906.21	1878.76	7.78E-11
	MP22		15.04	5.68	70.96 ± 0.71	0.21	1396.77	3312.53	14415.24	2693.56	1.12E-10
	ELNEDYN		22.49	11.30	65.24 ± 1.56	0.34	1360.58	3156.79	3333.84	1621.45	6.72E-11
Skip4	Implicit	61	28.11	14.62	77.78 ± 3.90	0.36	1396.77	3312.53	2364.51	1379.68	5.71E-11
	MP22		21.66	8.40	83.39 ± 0.92	0.26	1396.77	3312.53	7729.37	2318.78	9.60E-11
	ELNEDYN		15.23	8.64	76.96 ± 1.36	0.20	1396.77	3312.53	6741.12	2221.10	9.20E-11
L1526-R1689	Implicit	164	14.88	7.79	238.95 ± 1.11	0.06	2582.88	3492.87	25751.13	3075.68	1.27E-10
	ELNEDYN		24.91	12.76	380.54 ± 30.32	0.07	2582.88	3492.87	15211.20	2840.60	1.18E-10
H1331–E1985	Implicit	605	26.96	18.27	883.39 ± 6.44	0.03	2130.44	2086.33	17073.65	1859.15	7.70E-11

[a] Generalized Born method (gb8) is used .

Appendix F

List of Publications

1. "Identification of the Ah-receptor structural determinants for ligand preferences", Y. Xing, M. Nukaya, K. Satyshur, L. Jiang, V. Stanevich, E. N. Korkmaz, L. Burdette, G. Kennedy, Q. Cui and C. A. Bradfieldi, *Toxicol. Sci.* (2012).
2. "Interplay of electrostatics and hydrophobic effects in the metamorphic protein human lymphotactin", E. N. Korkmaz, B. F. Volkman and Q. Cui, *J. Phys. Chem. B* 119 (30), 9547-9558 (2015).
3. "Skip residues modulate the structural properties of the myosin rod and guide thick filament assembly", K. C. Taylor[‡], M. Buvoli[‡], E. N. Korkmaz[‡], A. Buvoli, Y. Zheng, N. T. Heinze, Q. Cui, L. A. Leinwand and I. Rayment, *Proc. Natl. Acad. Sci. USA*. 112 (29), E3806-E3815 (2015). ([‡]contributed equally.)
4. "Comparison of native and non-native ubiquitin oligomers reveals analogous structures and reactivities", G. H. Pham[‡], A. Rana[‡], E. N. Korkmaz[‡], V. H. Trang, Q. Cui and E. R. Strieter, *Protein Sci.* 25, 456-471 (2016). ([‡]contributed equally.)
5. "A composite approach towards a complete model of the myosin rod", E. N. Korkmaz[‡], K. C. Taylor[‡], M. P. Andreas[‡], G. Ajay, N. T. Heinze, Q. Cui and I. Rayment, *Proteins* 84, 172-189 (2016). ([‡]contributed equally.)
6. "Differential Specificity and Energy Landscapes (Di-SEL) unmask sequence preferences of transcription factors that bind indistinguishably to DNA consensus motifs", D. Bhimsaria[‡], J. A. Rodriguez-Martinez[‡], E. N. Korkmaz, Q. Cui, P. Ramanathan and A. Z. Ansari, *in preparation*. ([‡]contributed equally.)

7. “Exploration of Martini coarse-grained force fields using the ubiquitin dimers, trimers and the myosin thick filament as benchmarks”, E. N. Korkmaz and Q. Cui, *in preparation*.