ESSENTIAL DYNAMICS OF PROTEINS USING GEOMETRICAL SIMULATIONS AND SUBSPACE ANALYSIS

by

Charles Christian David

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2012

Approved by:

____
Dr. Donald J. Jacobs

____
Dr. Dennis R. Livesay

____
Dr. Anthony Fodor

____
Dr. Yuri Nesmelov

ABSTRACT

CHARLES CHRISTIAN DAVID. Essential dynamics of proteins using geometrical simulations and subspace analysis. (Under the direction of DONALD J. JACOBS)

Essential dynamics is the application of principal component analysis to a dynamic trajectory derived from a simulation protocol in order to extract biologically relevant information contained in the high dimensional data. In this work, we apply the methodology of essential dynamics to protein trajectories derived from geometrical simulations, which are based on the perturbation of geometrical constraints inherent in a protein. Specifically, we show that the geometrical simulation model is highly efficient for the determination of native state dynamics. Furthermore, by the application of subspace analysis to the essential subspaces of multiple sets of proteins that were simulated under multiple modeling paradigms, we show that the geometrical modeling paradigm is internally consistent and provides results that are qualitatively and quantitatively similar to results obtained from the more commonly employed methods of elastic network models and molecular dynamics. The geometrical paradigm is therefore established as a viable alternative or co-model for the investigation of native state protein dynamics with application to both small, single domain proteins as well as large, multi domain systems.

ACKNOWLEDGMENT

I am deeply grateful for the generous support and mentorship of my advisor Don Jacobs. In Don I found a partner in the pursuit of science who pushed the limits of my knowledge and skills to levels I never imagined. Science is the art of asking good questions, and Don is a master interrogator. In all the years that I spent working with him, his passion and intellect were an ever-present companion. Don went far beyond obligation in providing the essential elements for my professional growth.

The vast technical knowledge that I gained was generously supplied by my teachers. I thank Dennis Livesay for his chemical and biophysical insights into proteins and Anthony Fodor for his piercing insight and for his patience in teaching me how to program. I am also thankful for the insights into the workings of myosin V provided by Christopher Yengo and Yuri Nesmelov.

Finally, I wish to thank my family for their support for this journey. My dad originally planted the seeds of science in my heart by his never ending questions about the universe. My mom was always there to remind me that I could do anything I set my mind to do. Most dear to me was the encouragement and support from my wife who has been a fellow traveler on the road of discovery for the last twenty years.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 General Overview

Proteins are biological molecules that participate in and catalyze chemical reactions necessary for life. The basic structure of a protein is that of a polymer composed of units called residues, which are chemically identified as alpha amino acids composed of an alpha carbon that is a chiral center (except in glycine), a carboxyl group, an amino group, and a functional group commonly referred to as the side chain, as shown in Figure 1.1. The sequence of the residues is determined primarily by the genetic code contained within the cell (deoxyribonucleic acid or DNA) and post-transcriptional processing that often occurs in the cell that produces the protein. Higher order structures of these polymers emerges as interactions between the residues cause the linear chain to adopt regular conformations known as secondary structures, such as alpha helices and beta sheets as shown in Figure 1.2. More complex structures arise, called tertiary structures, which engender identifiable topologies of the molecule as a whole.

Additionally, when two or more proteins interact, a quaternary structure is formed. The levels of protein structure are summarized in the Figure 1.3. Given such a hierarchy of structural organization, the ultimate arrangement of a large protein can be quite complex. In spite of this inherent complexity, many protein structures have been determined using a wide array of techniques such as x-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electronic microscopy (cryo EM), and others.

Indeed, The Protein Data Bank (NCBI) (1,2) has acquired tens of thousands of structures for proteins, many of which play vital roles in the determination of health. While this is a noble achievement, these structures provide only a single (or a few when multiple structures exist) snapshot of a protein as it exists within a living cell. The key element missed by structure determination is dynamics and it is this dynamical behavior that ultimately determines the function of the protein in a biological sense.

The ability to determine the dynamics of a protein given only its structure still remains a Holy Grail for computational biologists and biophysicists. Each known structure provides a static snapshot of a protein, one configuration out of many possible states known as an ensemble, but it does not convey information about what other states are possible or whether those other states are probable. In order to address the determination of the dynamics of a protein, many computational approaches have been developed. All of these approaches invoke a three dimensional model of the protein structure, and the level of detail that goes into constructing that model has a great impact on how the model performs. The competing factors to be optimized in the model are the speed versus the accuracy. Including more details can improve model accuracy, but doing so may result in massive calculations that greatly reduce the model’s speed. Clearly, trade-offs must be made to account for the processing power of computational hardware. Additionally, the resolution of the description determines the time-scale at which the model functionally operates. Since a protein is a molecular system, a full description of all its chemical bonds requires quantum mechanics. However, this level of detail is not always needed to describe biologically relevant behaviors of the molecule. The key is to include the details that are needed for the behavior of interest, but no more.

Not surprisingly, many models have emerged that span a spectrum of complexity, with the level of complexity being determined by the number and resolution of the details that are included in the computations. Some of these are extremely coarse-grained models that treat the residues in the protein as “beads on a string” (Figure 1.4), while others employ a very high resolution, tracking every atom in the molecule as it moves under the influence of a set of forces (Figure 1.5). Each such model uses a set of assumptions within a biophysical paradigm that computes a set of structures, called a trajectory, which purportedly characterizes the dynamics of the given protein. After a trajectory is computed, it is critical to assess how well it represents the actual dynamics of the sampled protein. In order to determine if the model sufficiently sampled the states available to the protein, a number of statistical tools are used. The ultimate aim here is to extract the relevant information from the trajectory that indicates the primary motions that the protein samples and relate this to possible known biological functions. Care is needed in the interpretation of a trajectory as each model probes a particular set of timescales that is determined by its level of complexity. Figure 1.6 illustrates this idea. Typically, dynamics at a very short timescale is distinct from the dynamics at longer times, as atomic fluctuations in side chains are very different from collective motions of large groups of atoms that form domains. Therefore, a trajectory that samples a particular timescale with statistically significant coverage of the possible structures (macrostates) can be analyzed to reveal the essential motions that define the protein’s dynamics at that timescale, while another trajectory that insufficiently sampled the proteins macrostates when similarly analyzed will likely produce misleading conclusions (3).

1.2 Computational Methods

1.2.1 Protein Structure Preparation

In order to employ computational methods to determine the dynamics of a protein, the molecular structure of the protein must be assessed and possibly repaired. Most of the structures that exist in the PDB contain missing atoms and many have entire residues missing. Coarse-grained methods that employ an analysis of the protein backbone or the set of alpha carbons will not give reliable results unless the structure is first prepared for the analysis by filling in the missing residues. All atom methods are even more dependent on the initial structure, so some effort needs to be made to complete the structure in a biochemically sound manner. One approach to repair the protein is to first employ homology modeling (HM) to insert missing loops and flexible regions in the molecule. In this approach, one or more known structures are used as a template to inform how the missing residues should be placed in the gap. If no other structures are used, the incomplete structure may be used as its own template in a process called self-HM. In this case, no evolutionary information is used to fill the gaps. After a set of HMs is created, the structures can be minimized in terms of energy and a satisfactory structure may be chosen using a variety of criteria.

Once the backbone of the molecule is complete, the next step is to assess the structure for bad phi-psi angles on a Ramachandran Plot (Figure 1.7) and to determine if there are any clashes in the side chains. Many minor problems can be either reduced or eliminated by performing an energy minimization under an appropriate forcefield, such as AMBER 99, which will relax the structure and relieve pockets of strain. The last step in the structure preparation involves the addition of hydrogen atoms and the determination of the protonation state of the titratable residues (at a specified pH, usually close to 7). This is a critical step as it is the foundation for the hydrogen bond network within the protein. There are many programs to perform such a task, such as H++ available from Virginia Tech. Once the missing hydrogen atoms are added and the protonation state of the titratable residues determined, the structure is minimized again so that any new clashes or strains due to the additions or changes in partial charges can be relieved. One program that may be used for this entire process is called the Molecular Operating Environment (MOE) available from the Chemical Computing Group. The end result of the structure preparation recipe is a molecule that is ready to be used in either a coarse-grained or all-atom simulation.

1.2.2 Elastic Network Models

One of the simplest computational models that has been applied to proteins is the elastic network model (ENM). In this model, a single point, either the center of mass of the residue or more commonly, the alpha carbon, is used to represent each residue. Each of these points forms a node on a graph and the interactions that each residue participates in is represented by an edge on the graph. While it is possible to use different values for the interaction, typically a single value is used as in the Anisotropic Network Model (ANM) (4,5,6). The result is a graph that is a network of beads connected by springs all having the same spring constant. Figure 1.8 provides an illustration of the process. A key simplification here is that a simple harmonic potential is implemented to describe the native basin defined by the global topology of the input structure. This coarse-graining has the effect of reducing complexities in the energy landscape of the real protein as shown in Figure 1.9, and makes the model insensitive to slight variation in the protein structure.

In much the same way that coupled motion is studied in physics, the protein is analyzed using a normal mode analysis (NMA) (7). This is done by constructing the matrix of the second derivatives of the potential (the Hessian) and diagonalizing this matrix. The result of this eigenvalue-eigenvector decomposition is a set of frequencies (the eigenvalues) and a set of directions (the eigenvectors) that characterize the correlated motions of the protein. Timescale information is obtained from the frequencies, as the lowest frequency motions occur on the longest timescales and have been shown to represent biologically relevant motions of the protein. Higher modes, on the other hand, have been shown to be more local and represent events that occur on very short time scales such as side-chain fluctuations.

While this approach is both very simple and requires little computational time even for large proteins, it has been shown to be quite accurate in predicting large-scale motions of proteins. Another benefit is that no energy minimization is needed for the analysis. A limitation of this method is that no trajectory is produced and thus, one does not generate a complete ensemble of states in a thermodynamic sense. However, it is possible to calculate the vibrational energy of the modes from these models.

1.2.3 Molecular Dynamics

Molecular dynamics (MD) represents a comprehensive model for the analysis of protein dynamics (8). While the model does not use quantum mechanical calculations, it does implement an all-atom classical forcefield. The basic assumption of this model is that the way a protein behaves can be elucidated by examining how it evolves through a set of molecular steps under the influence of the specified forcefield. In this sense, the forcefield is key, as it must capture the essence of the interactions that drive the dynamics of the protein. A sample MD potential is shown in Figure 1.10. Most applications of MD to proteins use either an implicit or explicit solvent representation, as real proteins do not exist in vacuum, but rather are hydrated by water in the biological environment. Another critical aspect of the MD model is that it is a thermal simulation and proteins are not only hydrated, but the entire system of protein plus solvent is equilibrated at a specific temperature. Figure 1.11 illustrates a solvated protein in an MD simulation box. The benefit here is that the trajectory is not only a set of structures as a function of time, but it is a true ensemble in the thermodynamic sense, so that other thermodynamic quantities may be determined from such a trajectory. The downside to performing MD simulations is the computational resources and time that they require.

The critical issues here are the size of the protein to be simulated and the timescale that is to be investigated. The calculations are highly intensive and serial in nature, thus the simulation of large proteins can not be performed long enough to ensure that the ensemble has been sampled in a statistically significant manner for the timescales that are typically of interest. This problem arises due to the fact that an MD simulation can and does sample beyond the native basin of the input structure. The reason is that a protein exists within a free-energy landscape (FEL) and, for any given temperature, there will be a large number of local minima in this FEL that represent stable structures. This situation is illustrated in Figure 1.12. It is the nature of MD to allow the protein access to states beyond the single basin that represents the input structure, and the rate at which the protein samples the other minima in the FEL is dependent on the barriers that separate those basins and the temperature at which the simulation is performed. A vexing problem with MD is that one never really knows if the simulation has equilibrated (9). A simulation may appear to be nicely equilibrated in a particular basin, but if the simulation is run just a bit longer, the system jumps to another basin (macrostate) and will need to "equilibrate" again. So while there are many advantages to using MD, there are also some limitations that must be considered depending on the application.

1.2.4 Geometrical Simulation Model

A compromise between the two extreme levels of detail is achieved by using geometrical constraints (10,11). The geometrical simulation model (GSM) is a relatively new model paradigm that uses geometrical constraints to define the native basin of an input structure, while the perturbation of those constraints (with subsequent relaxation and tolerance matching) yields conformers within the constraint space of the input structure. Floppy Inclusions and Rigid Substructure Topology (FIRST) (12) implements a graph rigidity algorithm (The Pebble Game) that identifies flexible and rigid regions of the protein given the constraints that are present in the input structure. This process of recasting the protein as a set of rigid sub-graphs joined by hinges is referred to as a rigid cluster decomposition (RCD). Figure 1.13 depicts two RCDs, one at -1 kcal/mol (left) and the other at -2 kcal/mol (right). The RCDs are in color while flexible regions are shown in black. Since the Pebble Game counts the number of constraints in the network, the first requirement to perform a geometric simulation is the specification of all constraints.

After an input structure is suitably prepared by adding hydrogen atoms and determining the protonation state of the titratable residues, FIRST accomplishes this by processing the structure using constraint assignment parameters, in particular, the specified energy cutoff for hydrogen bonds (HB) and the specified distance cutoff for hydrophobic tethers (HT). HBs span an energy range form near zero to -10 kcal/mole and therefore the choice of HB energy cutoff is critical. Choosing an HB cutoff of zero results in a structure that is highly over-constrained due to the inclusion of many weak interactions. Also, specifying a very low HB cutoff (less than -5 kcal/mole) results in an under-constrained structure that is similar to an unfolded state. In a similar fashion, HTs are needed to perform the RCD and these are determined by a distance cutoff between atoms that can support HTs (carbon and sulfur atoms). Choosing to include no HTs results in an under-constrained system while setting the HT cutoff too large will lead to the inclusion of excess HTs and result in an over-constrained system. Of importance to the GSM is the fact that HBs are modeled as harmonic constraints that take 5 degrees of freedom (DOF) (5 bars in the graph model), HTs are modeled as half-harmonic constraints (distance inequalities) that consume 2 DOF (2 bars in the graph model), and covalent bonds consume 5 DOFs (5 bars in the graph model, but 6 if the bond is not rotatable) and are quenched. It has been shown through rigidity percolation analysis (13) where the number of constraints is systematically varied, that different choices of constraints yield different rigidity transitions in proteins.

Although the constraint types used in FIRST/FRODA have been selected to best match empirical characteristics of protein flexibility in the native state, there remains considerable freedom in the user-defined rules for identifying native constraints. This model parameter ambiguity is unsettling when one wants to quantitatively characterize the native state dynamics of a protein. The fact that the user can select the number of HB and HT to model as constraints may affect the outcome of a simulation, since this can substantially alter the degree of rigidity/flexibility predicted within a protein. Due to the strong dependence on constraint parameters, the structure can be characterized as being very flexible or very rigid. Presumably, both extremes are not physically correct. Unfortunately, a weakness of the GSM is that this selection is left to the user to decide, guided only by the default values, which are based on qualitative empirical results. One would ideally expect that a range of constraint assignment parameters should exist between over-constrained systems and under-constrained systems where the dynamical behavior of the system would be qualitatively and quantitatively similar. Due to this inherent ambiguity in parameter choices, the GSM has not been popular. Specifically, the GSM has not been benchmarked and therefore is not a well established model based on a thorough assessment, in contrast to ENM and MD.

In order to generate a trajectory, the GSM makes the assumption that the rigid clusters will move as a single unit, so there is no need to track each individual atom within this unit. The RCD is used as input to Framework Rigidity Optimized Dynamics Algorithm (FRODA) to generate a set of output structures that sample the native basin defined by the input structure. FRODA performs a Monte Carlo (MC) simulation in which the RCD is perturbed and allowed to relax. The acceptance criterion is that the geometric constraints must be satisfied to within a specified tolerance. The initial energy of the system is therefore zero, indicating all geometrical attributes (packing, bonding, hydrophobic interactions) are valid and native-like. Thereafter, rigid cluster center of mass coordinates and orientations are randomized slightly in each MC move. After each randomization step, the energy of the system is relaxed back to zero energy without causing any atomic clashing, and all rigid clusters are maintained within tight distance constraint tolerances. This procedure produces a trajectory that preserves all rigid clusters, while allowing relative positions of atoms in different rigid clusters to change within flexible regions without having any atom or rigid cluster pass through each other.

The efficiency of the simulation is greatly improved by activating the “momentum” feature in FRODA. This feature weights a successful MC move and then biases the simulation to continue to move in the successful direction until no acceptable conformers are produced, at which point a new random perturbation is used. When using the momentum feature, the size of the random step should be reduced from 0.1Å to 0.01Å. In spite of the order of magnitude reduction in the step size, the simulation actually explores the conformational space much more effectively than when using diffusion mode (no momentum feature), equilibrates rapidly, and generates constraint violations that are almost all less than 0.05Å.

A critical difference between the GSM and MD is that the GSM is an athermal model and the trajectory is not a function of time. In the GSM, the HB energy cutoff plays a role that is analogous to temperature in an MD simulation, but this analogy is not exact and equilibration must be assessed by studying the molecular fluctuations. The consequence of the athermal nature of the GSM is that it is difficult to extract thermodynamic quantities from the trajectory. The advantage of the GSM is that is runs very quickly, up to four orders of magnitude faster than MD, and since the RCD is not fluctuating, the GSM trajectory samples the native basin of the input structure exclusively and thus the simulation “equilibrates” very rapidly. These advantages allow the GSM to be applied to large proteins and used for investigating the dynamics at long time scales.

1.3 Essential Dynamics and Analyzing Trajectories

Essential dynamics (ED) is the application of principal component analysis (PCA) to a protein trajectory in order to extract the most important dynamics for biological function (14,15,16). The idea is that even though proteins are fully characterized by a high dimensional space defined by a set of coordinates, or a set of degrees of freedom (DOF), the critical aspects of their dynamics can be represented in a much smaller subspace (SS) of that vector space (VS), called the essential subspace (ESS). The ESS is inherently stable against noise that is captured in the lower modes and is able to extract the biologically relevant motions of the protein when atomic variance/covariance is a determining factor.

The method of ED can be applied to both single trajectories as well as sets of trajectories obtained from either different simulation models or different model conditions. The value of pooling dynamic data is that greater statistics are obtained for sampling the configuration space of the protein and meaningful comparisons may be made between not only multiple states of a protein, but also between different models that are defined by distinct physical paradigms.

1.4 Overview of the Dissertation Project

We develop a statistically sound protocol for extracting biologically relevant motions from geometrical simulations. In order to achieve this result, we tested the geometrical simulation paradigm and determined appropriate values of constraint assignment parameters. Additionally, we discovered that there is consistency within the model with respect to parameter choice. We also assessed the degree of trajectory equilibration and the statistical sampling efficiency of the model. Our first task was to benchmark the geometrical paradigm. Our second task was to compare geometrical simulation results to the well accepted and commonly employed MD and ENM methods. The outcomes of these tasks allowed us to determine that the geometrical simulation paradigm is effective and efficient for the determination of the native state dynamics of single domain proteins. Our third task was to apply the geometrical simulation methods with subsequent ED analysis to large multi-domain proteins and compare the results to experimental methods such a fluorescence resonant energy transfer (FRET). From this work we were able to determine that the geometrical paradigm is also effective and efficient for elucidating the dynamics of subsets of interest within the large protein.

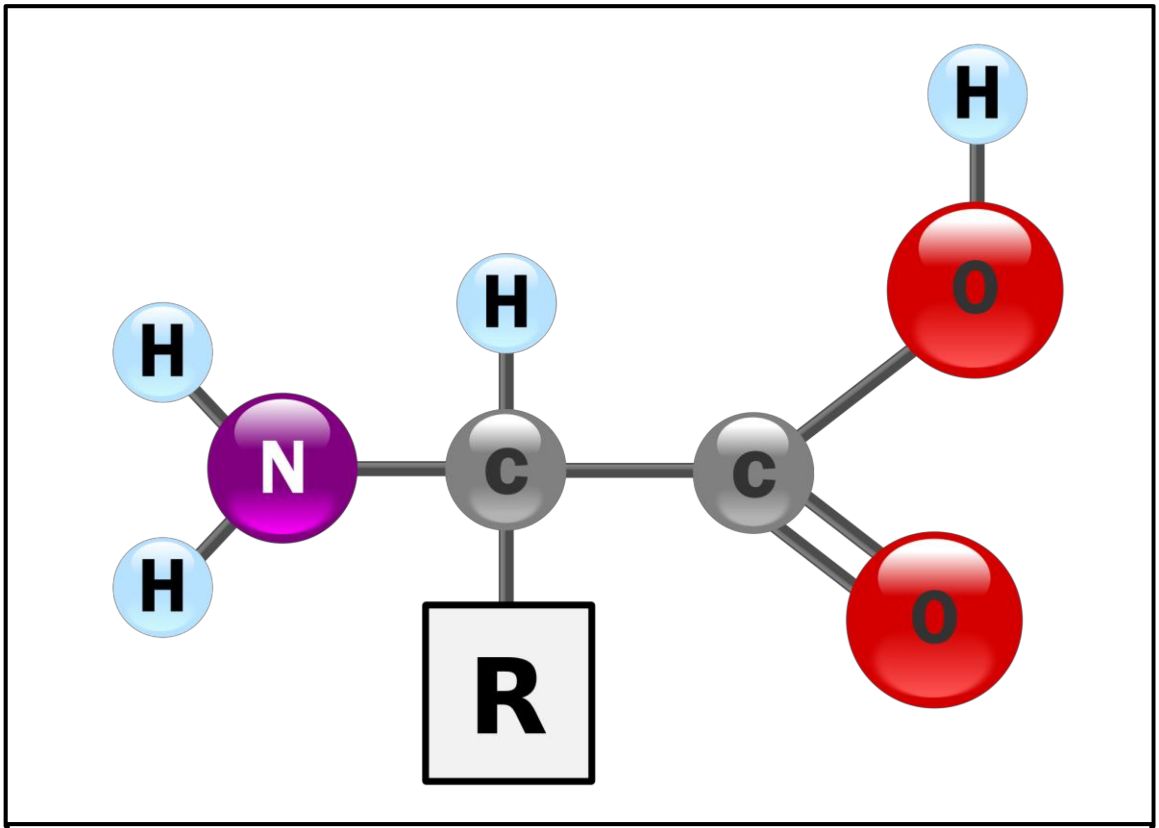


Figure 1.1: The general chemical structure of an alpha amino acid showing the functional groups and the chiral alpha carbon.
Source: http://upload.wikimedia.org/wikipedia/commons/thumb/c/ce/AminoAcidball.svg/2000px-AminoAcidball.svg.png

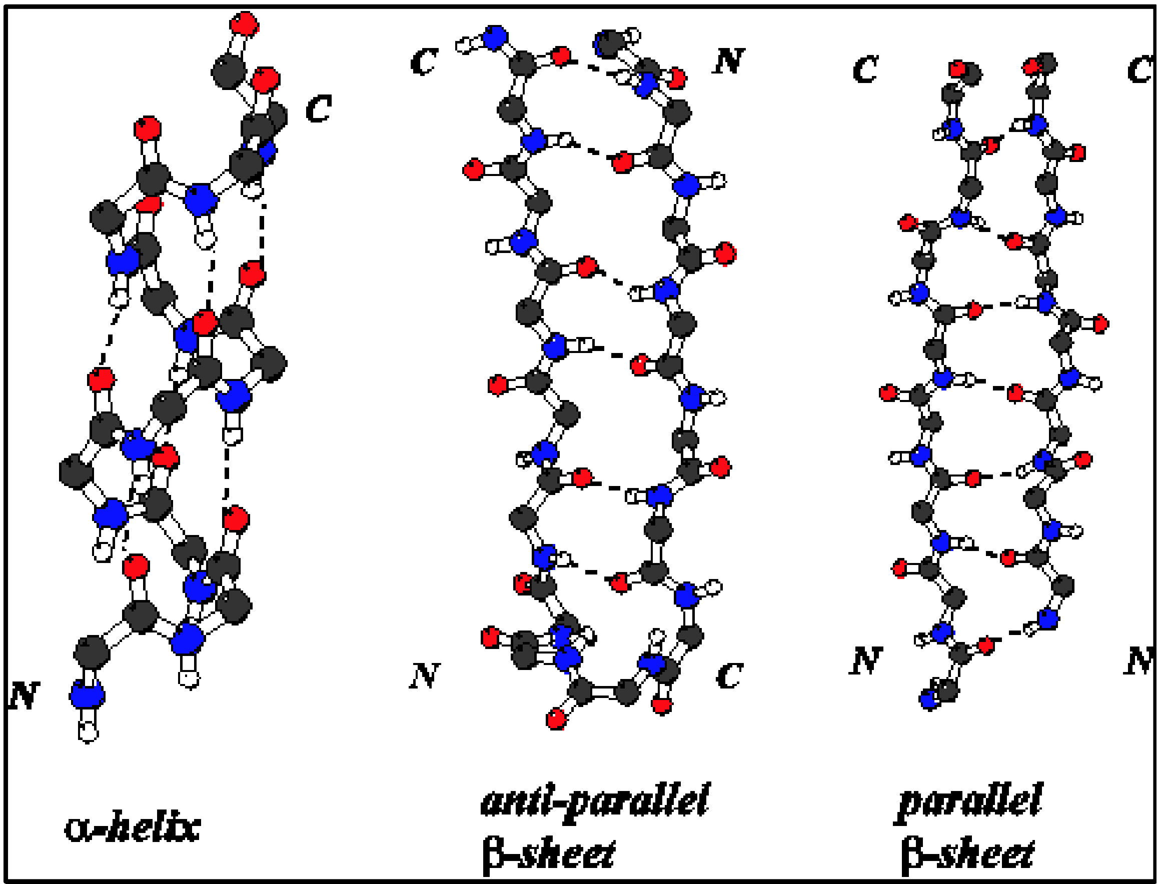


Figure 1.2: Proteins adopt secondary structures due to the formation of hydrogen bonds along the backbone. Source: http://nook.cs.ucdavis.edu/~koehl/BioEbook/Classification/images/figure3.gif

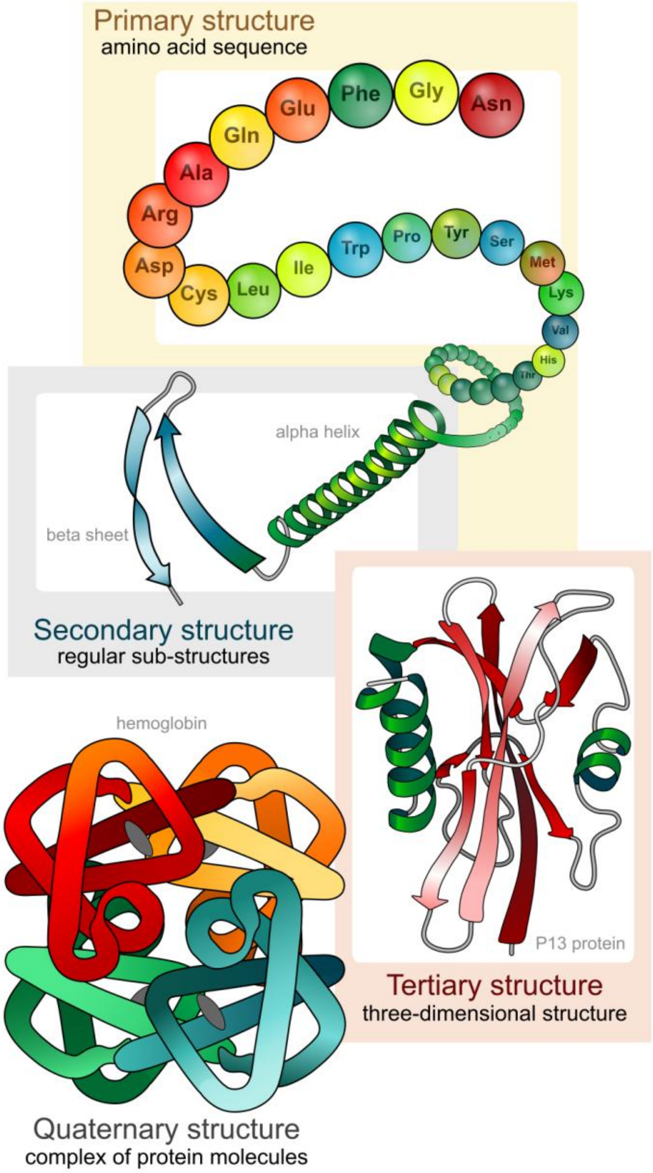


Figure 1.3: Proteins form a hierarchy of structures ranging from the linear arrangement of residues to complex interactions between multiple protein chains. Source: http://upload.wikimedia.org/wikipedia/commons/thumb/c/c9/Main_protein_structure_levels_en.svg/2000px-Main_protein_structure_levels_en.svg.png

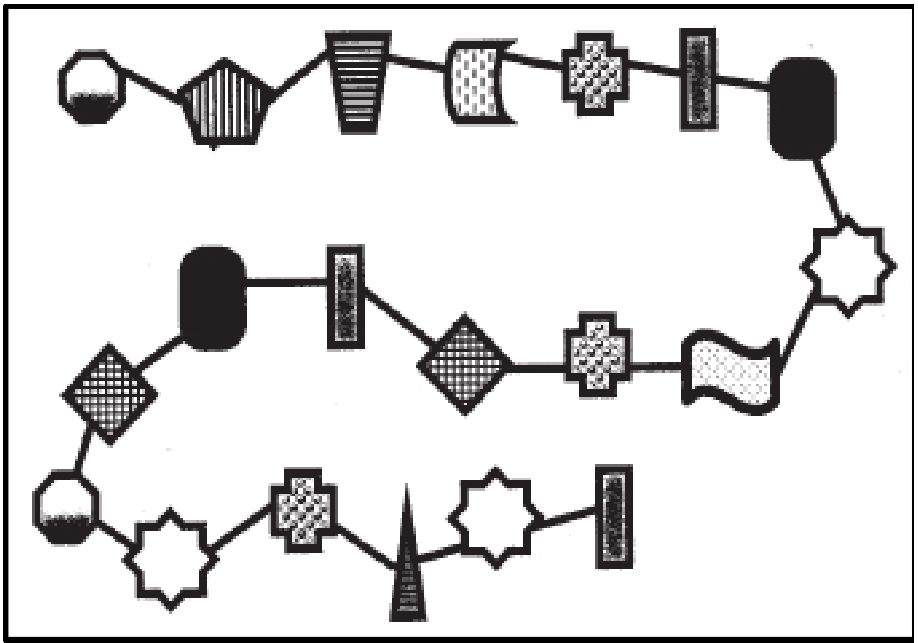


Figure 1.4: A set of residues is modeled as beads on a string. Different symbols indicate that the residues are not identical, but some models treat all residues identically. Source: http://www.waisman.wisc.edu/2mbadd/protein.jpg

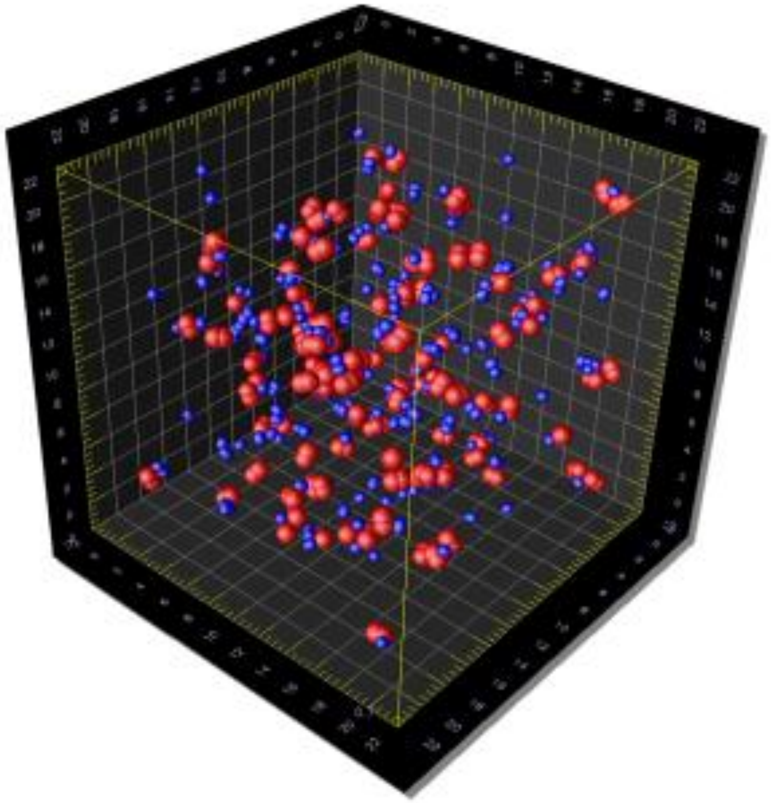


Figure 1.5: Atoms in a simulation box. All interactions supported by a specified forcefield are tracked for every particle. Source: http://cloud.gpuscience.com/wp-content/uploads/molecules.png

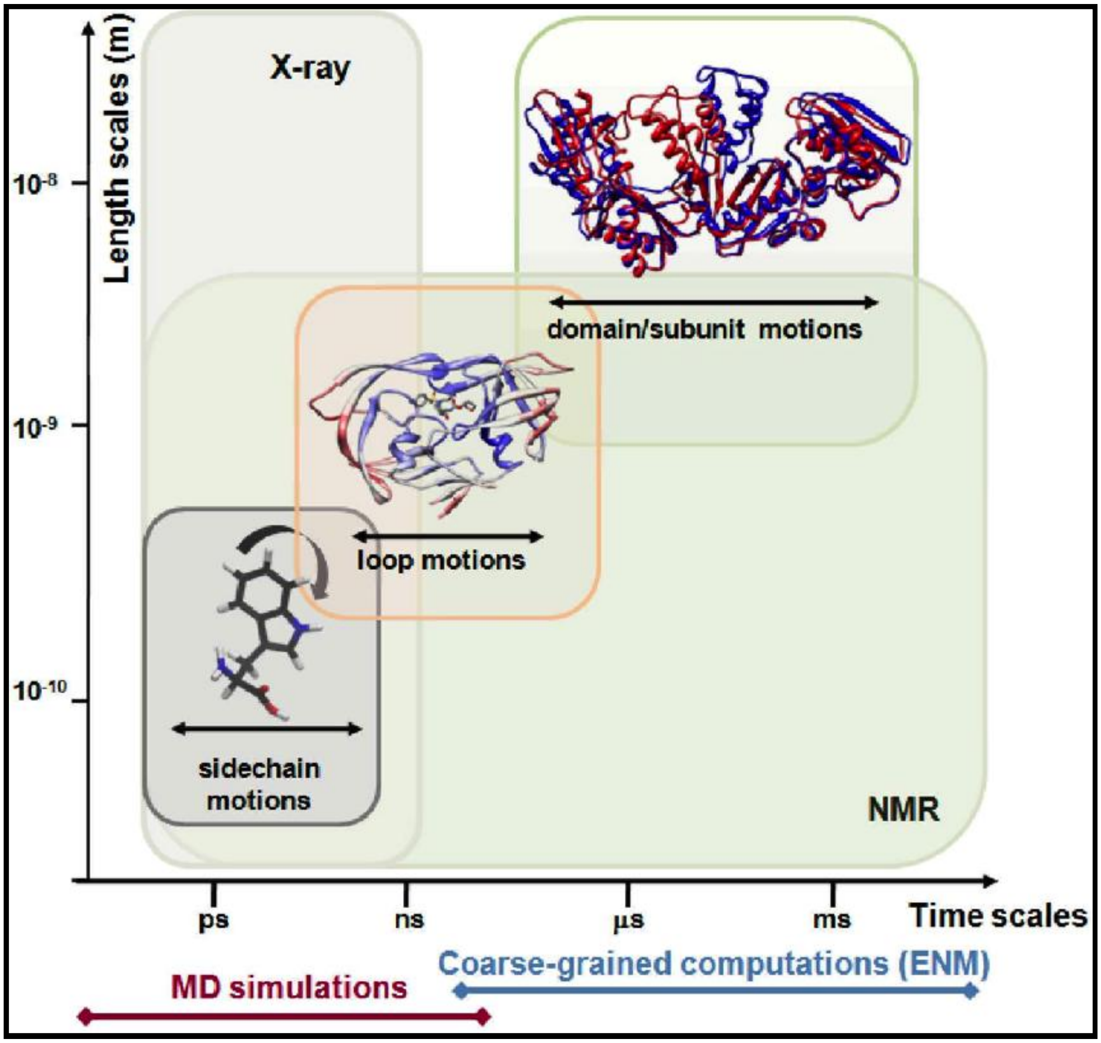


Figure 1.6 Timescales and Ranges of Equilibrium Motions.
Motions range from side-chain rotations to slower concerted domain motions. X-ray crystallography and NMR are the primary sources of information on such conformational changes at atomic resolution. Also indicated along the abscissa are the timescales of processes that can be explored by molecular dynamics simulations (MD) and coarse-grained (CG) computations. Adapted from Bahar et.al., 2009.

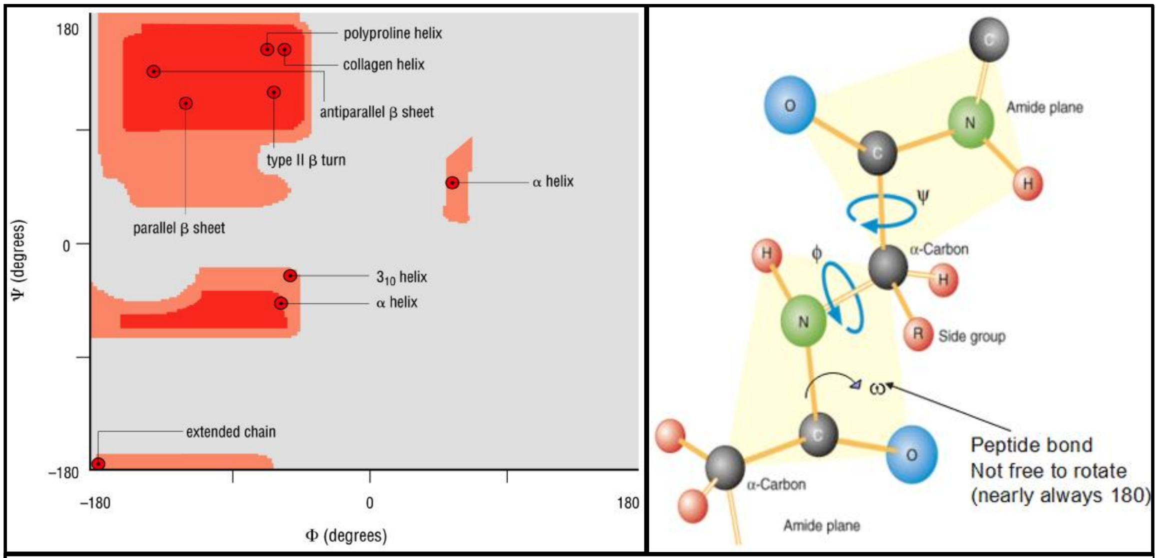


Figure 1.7: Ramachandran Plot.
This plot (left) shows the phi and psi angle pairs that correlate to typical secondary structure elements. Angle pairs that occur outside the typical range indicate there is inherent strain in the protein backbone. A description of the phi and psi torsion angles are shown in the right panel. Adapted from: http://molecularsciences.org/files/images/torsion_angles.gif

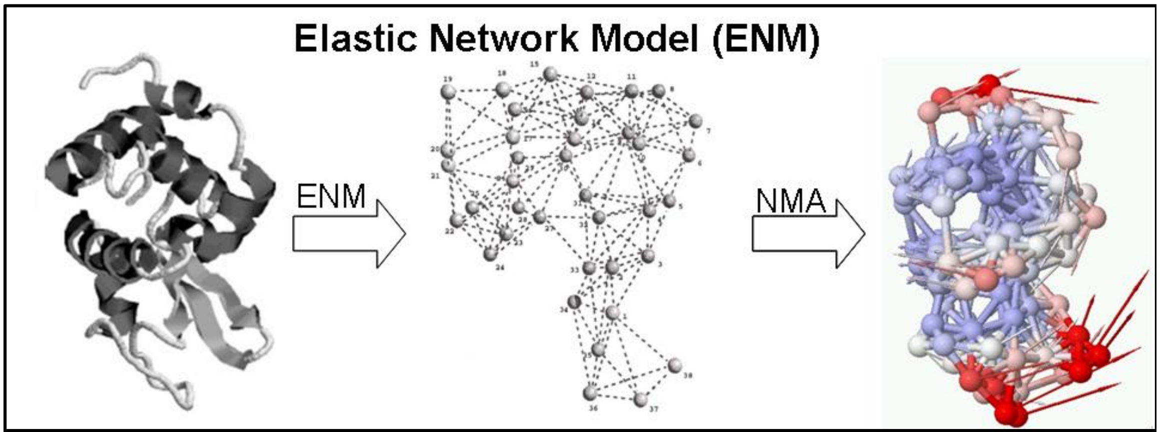


Figure 1.8: Generating a Spring Network from a Protein Structure.
The process of converting a protein structure (left panel) into a spring network is shown. Residues are first represented by the alpha carbon. Interactions are replaced by springs as shown in the middle panel. Subsequent NMA yields the modes of fluctuation as shown in the right panel. Source: http://t3.gstatic.com/images

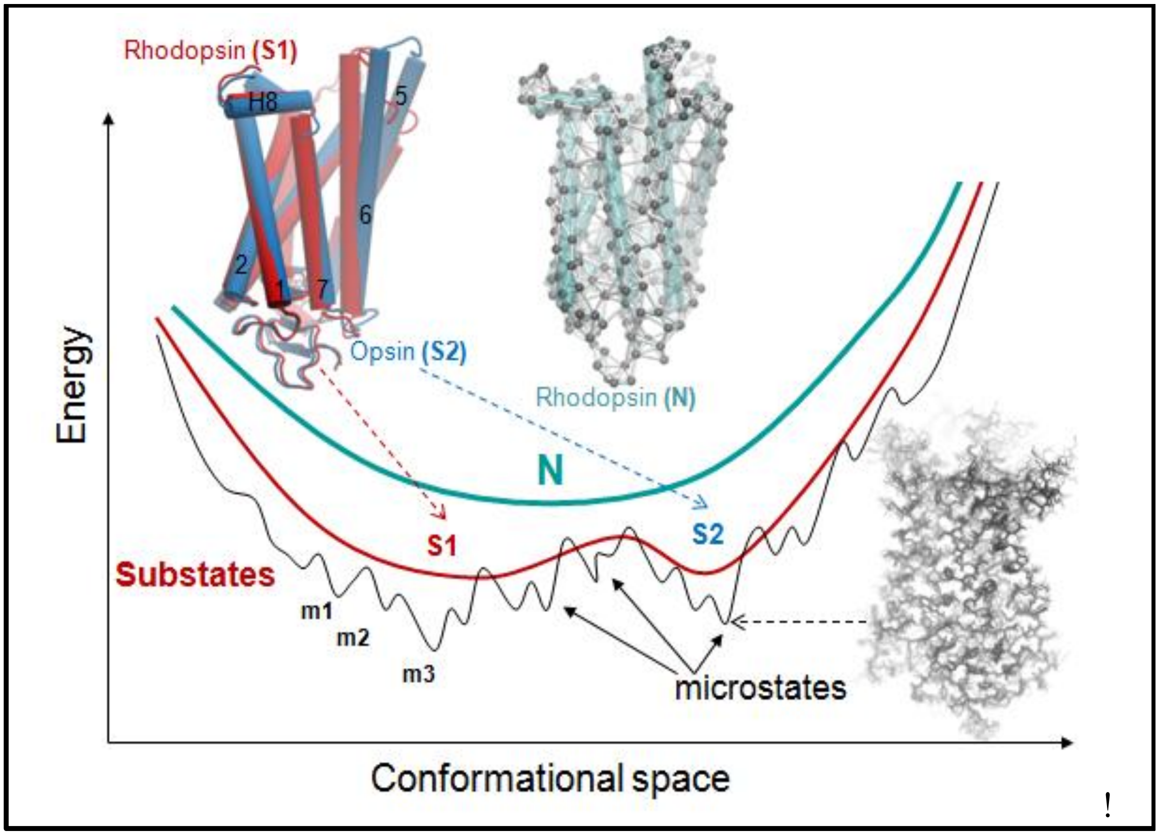
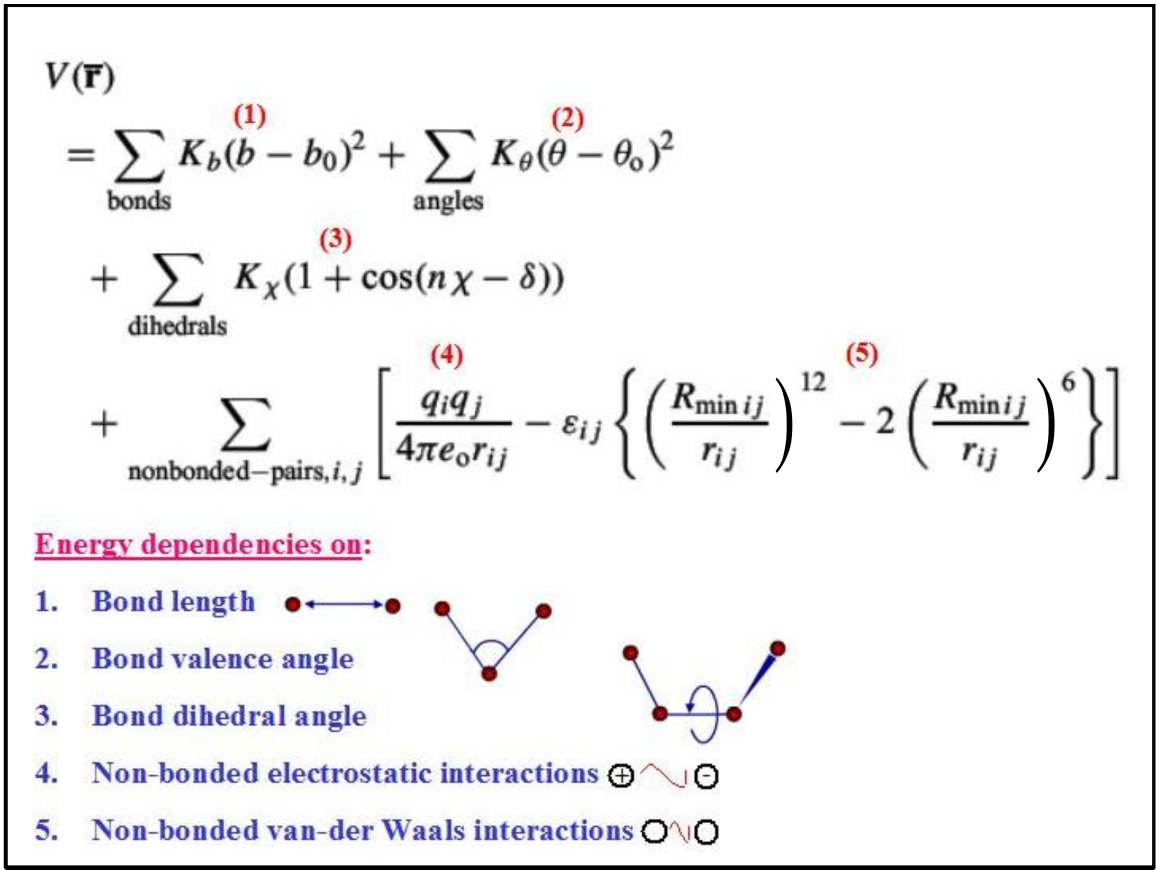


Figure 1.9 Energy Profile of the Native State.
N denotes the native state, modeled at a CG scale as a single energy minimum. A detailed examination of the structure and the energetics may reveal multiple substates (S1, S2, etc.), which in turn contain multiple microstates (m1, m2, etc.). Structural models corresponding to different hierarchical levels of resolution are shown: an elastic network model representation where the global energy minimum on a CG scaled (N) is approximated by a harmonic potential along each mode direction; two substates S1 and S2 sampled by global motions near native state conditions; and an ensemble of conformers sampled by small fluctuations in the neighborhood of each substate. Adapted from Bahar et.al., 2009.



$$V(\bar{\mathbf{r}}) = \sum_{\text{bonds}} K_b(b \overset{(1)}{-} b_0)^2 + \sum_{\text{angles}} K_\theta(\theta \overset{(2)}{-} \theta_o)^2 + \sum_{\text{dihedrals}} K_\chi(1 \overset{(3)}{+} \cos(n\chi - \delta)) + \sum_{\text{nonbonded-pairs},i,j} \left[\overset{(4)}{\frac{q_i q_j}{4\pi e_o r_{ij}}} - \varepsilon_{ij} \left\{ \left(\frac{R_{\min ij}}{r_{ij}}\right)^{12} \overset{(5)}{-} 2\left(\frac{R_{\min ij}}{r_{ij}}\right)^6 \right\} \right]$$

Energy dependencies on:

1. **Bond length**
2. **Bond valence angle**
3. **Bond dihedral angle**
4. **Non-bonded electrostatic interactions**
5. **Non-bonded van-der Waals interactions**

Figure 1.10 A sample MD Potential
The form of the potential that is used in MD simulations includes terms for both bonded and non-bonded interactions.
Source: http://amit1b.files.wordpress.com/2008/04/force-field2.jpg

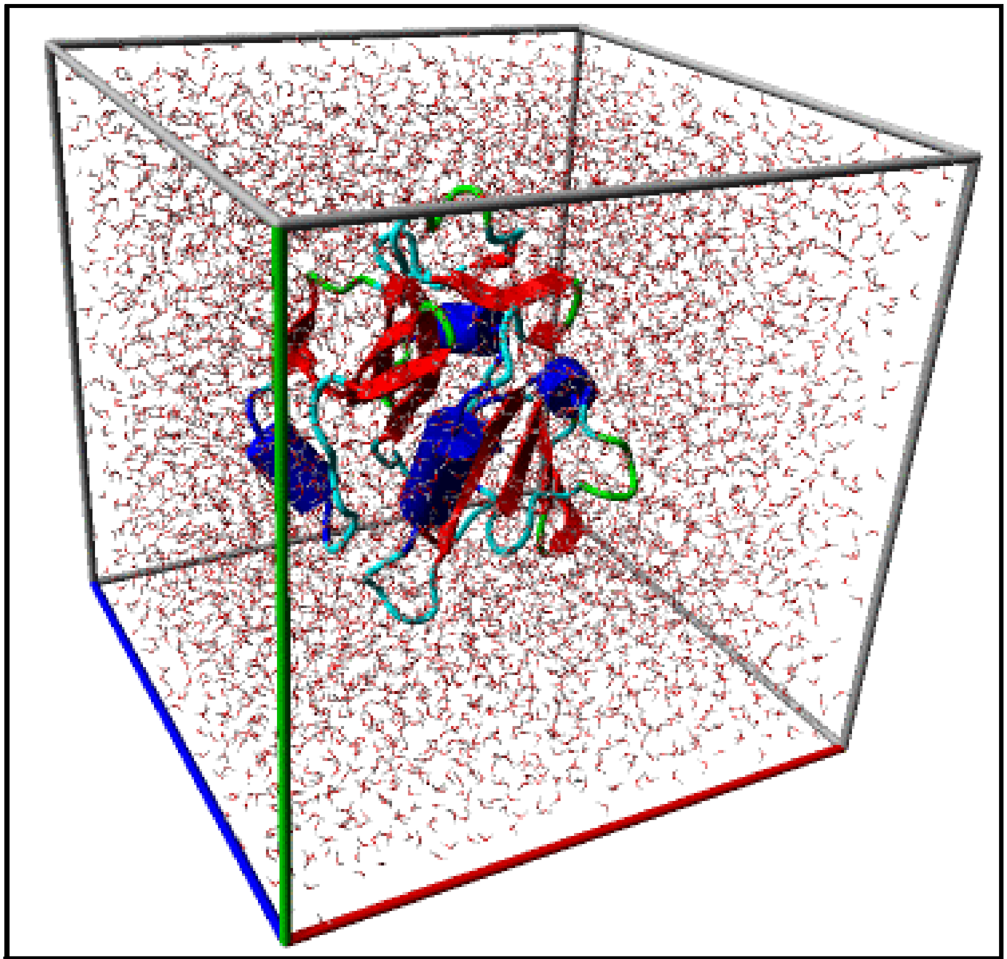


Figure 1.11 A Hydrated Protein in a Simulation Box.
While MD simulations may be run in vacuum, more often they use either an implicit or explicit solvent in the model. Source: http://www.yasara.org/dhfr.gif

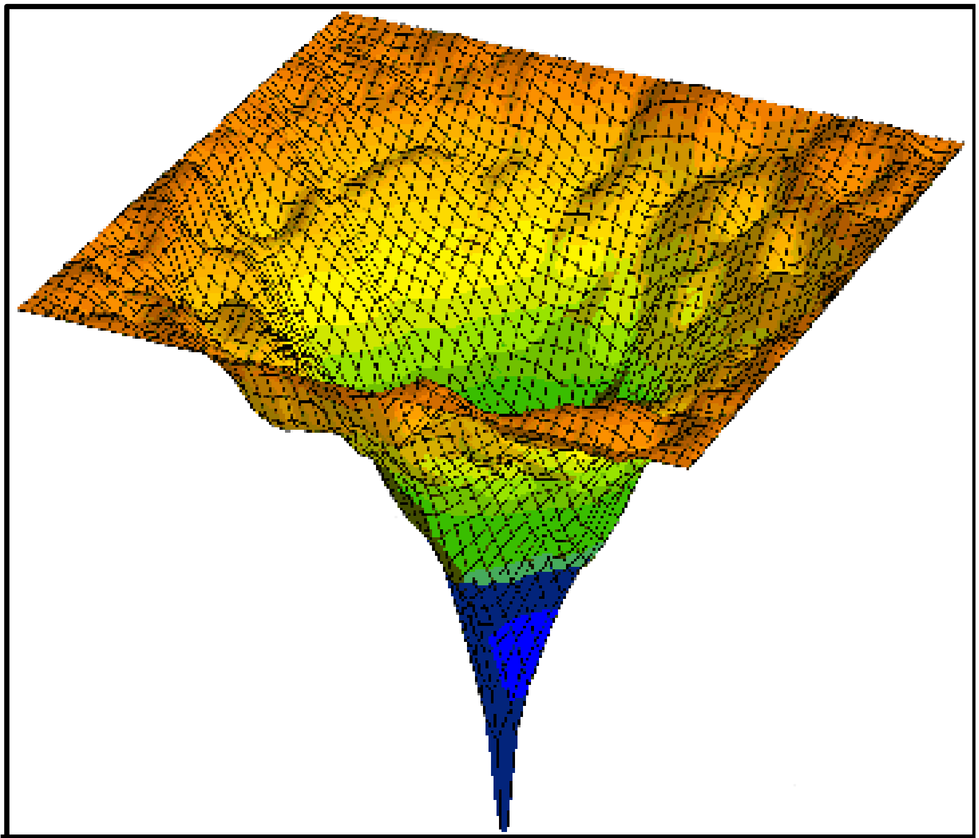


Figure 1.12 Free Energy Landscape of a Protein
The free energy landscapes of proteins tend to be rough with many local minima and a single global minimum corresponding to the native state.
Source: http://www.lsbu.ac.uk/water/images

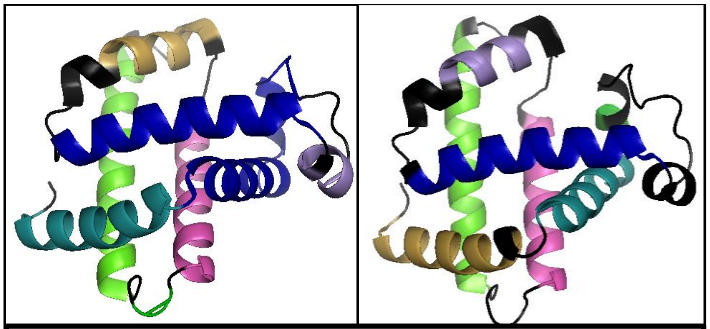


Figure 1.13 Rigid Cluster Decomposition (RCD)
The RCD identifies all atoms that are located in rigid regions joined by hinges, shown here in color. The RCD is strongly dependent on the hydrogen bond energy cutoff (HB_{Cut}). The left panel shows the RCD of myoglobin using HB_{Cut}= -1.0 kcal/mole. The right panel shows the RCD of myoglobin using HB_{Cut}= -2.0 kcal/mole. While similar, the lower HB cutoff results in a more flexible structure, as indicated by the black regions.

CHAPTER 2: STATISTICAL METHODS

2.1 Principal Component Analysis & Essential Dynamics

Protein dynamics is manifested as a change in molecular structure, or conformation as a function of time. To describe accessible motions over a broad range of time scales and spatial scales, protein conformations are best represented by a vector space that spans a large number of dimensions equal to the number of degrees of freedom (DOF) selected to characterize the motions. Many molecular simulation techniques are available to generate trajectories to sample the accessible conformational ensemble characterized by those DOF. The interpretation of a trajectory can lead to better understanding of how proteins perform biological functions. To this end, the process of extracting information from sampled conformations over a trajectory, and checking whether the sampling is a robust representation of an ensemble of conformations accessible to the protein, are tasks well suited for statistical analysis. In particular, Principal Component Analysis (PCA) is a multivariate statistical technique applied to systematically reduce the number of dimensions (see Figure 2.1 for a simple example) needed to describe protein dynamics through a decomposition process that filters observed motions from thelargest to smallest spatial scales (17-21).

PCA is a linear transform that extracts the most important elements in the data using a covariance matrix or a correlation matrix (normalized PCA) constructed from atomic coordinates that describe the accessible DOF of the protein, such as the Cartesian coordinates that define atomic displacements in each conformation comprising a trajectory (3). When all of the atomic displacements posses similar standard deviations, a covariance matrix is typically used, otherwise it is prudent to employ the correlation matrix, which normalizes the variables to prevent rare but large atomic displacements from skewing the results. In constructing the covariance matrix or correlation matrix (henceforth C-matrix will be generically used for either matrix type), it is often assumed that the amount of sampling is sufficient, but this always requires many more observations than the number of DOF (variables) used in the matrix. An eigenvalue decomposition (EVD) of the C-matrix leads to a complete set of orthogonal collective modes (eigenvectors), each with a corresponding eigenvalue (variance) that characterizes a portion of the motion, where larger eigenvalues describe motions on larger spatial scales. When the original (centered) data is projected onto an eigenvector, the result is called a principal component (PC).

While PCA can be performed on any high dimensional dataset, for the analysis of a protein trajectory, a C-matrix associated with a selected set of atomic positions must be constructed. Often, a coarse grained description of the protein motion is made at the residue level by using the alpha carbon atom as a representative point for the position of a residue. In this case, the C-matrix will be a $3m \times 3m$ real, symmetric matrix, where m is the number of residues. Performing an EVD results in $3m$ eigenvectors (modes) and $3m-6$ non-zero corresponding eigenvalues, provided that at least $3m$ observations are used. When the eigenvalues are plotted against mode index that are presorted from highest to lowest variance, a “scree plot” typically appears as a function of mode index. When such a scree plot forms, a large portion of the protein motions can be captured with a remarkably small number of modes that define a small dimensional subspace. The top set of modes typically has a higher degree of collectivity (22), meaning the PCA modes have many appreciable components distributed quite uniformly. Conversely, a low degree of collectivity indicates there are a small number of appreciable components, although they are not necessarily tied to a localized region of space. When analyzing proteins, 20 modes are usually more than enough (even for large proteins) to define an "essential space" that captures the motions governing biological function, thus achieving a tremendous reduction of dimension.

The process of applying PCA to a protein trajectory is called Essential Dynamics (ED) since the "essential" motions are extracted from the set of sampled conformations (23-25). Of course, a linear combination of the $3m$ orthogonal PCA modes can be used to describe exact protein motions (at the selected coarse grained level). In practice, the presence of large-scale motions makes it difficult or impossible to resolve small-scale motions because the former has much greater relative amplitude in atomic displacements. Indeed, it is for this reason that the large-scale motions are often the most biologically relevant. Therefore, only a small number of PCA modes having the greatest variances are used to characterize large-scale protein motions. When small-scale motions are of interest, the method of PCA can still be used successfully by applying it to sub-regions of a protein as a way to increase the resolution for describing the dynamics within those sub-regions.

An alternative method to quantify large-scale motions of proteins is to use a Normal Mode Analysis (NMA) (7,26) derived from an Elastic Network Model (ENM) (5,6). In the ENM, one typically considers nearby alpha carbon atoms to interact harmonically, where the connectivity is determined from a single structure to extract an elastic network. Typically, the large-scale motions quantified by a small set of lowest frequency modes of vibration are in good agreement with the same corresponding number of PCA modes when direct comparisons of subspaces are made (27-29). One advantage of performing PCA to obtain the ED of a protein is that information from any selected set of atoms can be used to obtain the PCA modes associated with that subspace. While it is true that ED is often applied to the analysis of alpha carbons, this is not required. The spatial resolution of PCA analysis can be coarser than the resolution of the structures that comprise the trajectory, which, for example, may come from an all-atom based simulation. Another advantage of ED is that statistics from many trajectories may be pooled allowing a great deal of flexibility in the way data from different simulations can be combined. The overall large-scale motions and any number of selected small-scale motions can be determined in a post-simulation phase of research as the nature of the protein motions are being interrogated.

Perhaps the most important difference between NMA and PCA is in the assumption of harmonicity. The premise of NMA requires the molecular motion is confined near the local minimum in the free energy landscape where residues in close proximity (i.e. atomic packing) respond as harmonic pairwise interactions (i.e. springs). Since proteins display a significant amount of anharmonicity in their behavior, this assumption is not always suitable (30-32). PCA makes no assumption of harmonicity, and thus is not limited to harmonic motions. Indeed, because PCA is independent of the model invoked during the simulation to generate the trajectory, the resulting conformational changes that can be explored can deviate far from the harmonic assumption. On the other hand, the limitations of PCA stem from using a linear transform that is based on second moments (covariance), and the fact that subsequent factorization yields eigenvectors that are orthogonal. While a linear transform of the data is always possible, if the variables are not intrinsically linearly related, any non-linear relationships present will not be properly described. Nonetheless, in practice, standard PCA is similar to the standard ENM approach. In other words, relying on covariance implies higher-order correlated motions related to higher moments are missed.

Non-linear generalizations of PCA are available such as kernel PCA (33) that can be applied directly, or employed after the most relevant subspace is identified first using standard PCA. A disadvantage of kernel PCA is that the choice of kernel is not obvious because it is problem dependent, although we show below that some common choices work well for protein trajectories. Also problematic is that the reconstruction of data is difficult to interpret because the mapping involves feature space, which is distinctly different than conformational space that has a geometric interpretation despite being of high dimensionality. The reason for employing kernel PCA is to differentiate conformations within an ensemble beyond that possible using standard PCA, which may give insight into structural mechanisms governing protein function. Our work suggests that the simplest PCA, which follows from the C-matrix, offers a validated method to describe the dominate correlations present in atomic motions found in proteins, and it provides an effective dimension reduction scheme that can be used for subsequent analysis to capture non-linear (or higher order correlations) affects when they are of interest. Nevertheless, in practice it is always important to ensure and test the robustness of the PCA modes.

Keep in mind that individual PCA mode directions are subject to errors related to finite sampling of conformations to construct the empirical C-matrix. The empirical C-matrix should be a good estimate for the actual population C-matrix (infinite samples). In practice, PCA can be strongly influenced by the presence of outliers in a dataset. The main concern is that the outliers may skew the first few mode directions. While there are robust algorithms that are useful in stabilizing PCA in the presence of outliers (34-41), it is often effective to remove identifiable outliers or simply consider a sufficiently long trajectory for which the results are significant. Generating a large number of conformational samples and removal of outliers before the C-matrix is calculated mitigates concerns about robustness of the results. Moreover, this type of intrinsic error does not pose much of a problem as long as biologically relevant motions are described using a superposition of a small set of dominant modes (instead of focusing on one mode). As the mode number increases the core part of this subspace becomes stable against sampling noise. However, only the top several modes tend to be useful.

The choice of which modes to include is often made by examining the scree plot for a visible “kink” (the Cattell criterion) (42,43), such that all modes up to the kink are important. Although a kink does not have to exist, it typically does in the study of protein dynamics. In fact, a kink will generally appear for any high dimensional dataset. Hence the name scree (geological debris at the bottom of a cliff) plot has been tied to PCA. Other criteria are commonly used for the choice of essential modes. For example, the top set of modes associated with greatest variances when added should reach some fraction (say 80%) of the total variance possible given by the trace of the C-matrix. The problem with this method is that some a priori set fraction is arbitrary, and for fractions greater than 50% one tends to end up with many more modes than are truly relevant to the problem. The scree plot provides an objective criterion. Figure 2.2 shows the scree plots for PCA of two protein simulations and a random process created from independent and identically distributed variables. Notice there is a rapid decrease in the eigenvalues for the proteins that is not present in the random process.

When PCA is applied to Cartesian coordinates that describe the positions of atoms, an alignment step is necessary prior to the process of constructing the C-matrix because the intent is to capture the internal motions of a protein. The structural alignment step requires the center of masses to coincide as well as a global rotation to optimally align the structures. We implemented a quaternion rotation method to obtain optimal alignment defined by the minimum least-squares error for the displacements between corresponding atoms (See Appendix D). PCA is not limited to the analysis of a Cartesian coordinate-based C-matrix. Any set of dynamic variables that describe the protein motion can be used. For example, one may choose to use internal dihedral-angle coordinates such as the (Φ, Ψ) angles or interatomic distances, which eliminates the need to optimally align conformations. However, in the former case, it has been realized there is an intrinsic non-linear effect that is not well described using standard PCA, suggesting kernel PCA should be employed or an alternative internal coordinate system that is naturally linear should be chosen. In the latter case, internal atomic distances offer the possibility of an all-to-all distance C-matrix for the alpha carbons, which has a row dimension equal to the number of structures in the trajectory and a column dimension equal to $m(m-1)/2$, where m is the number of residues considered. A distance based C-matrix can be created, which is a square matrix with dimension $m(m-1)/2$, and therefore requires much more sampling. In this case, the PCA modes reveal the coordinated changes in distances between all residue pairs. Despite the advantage of working directly with internal coordinates, performing all-to-all distance PCA quickly becomes computationally prohibitive due to the need to diagonalize very large non-sparse matrices. More importantly, the interpretation of the eigenvectors becomes difficult when the number of residues is greater than ten. Nevertheless, this approach has proven useful when studying a small subset of atoms where the interpretation is clear (44,45).

The task of applying PCA to a conformational ensemble (CE) requires that a CE be generated. There are multiple ways to create a CE including molecular dynamics (MD) and geometrical simulations such as FIRST/FRODA (10-12). A CE may be generated by experimental methods such as using protein structures from X-ray crystallography or nuclear magnetic resonance (NMR) techniques. For certain applications it is prudent to combine multiple CEs together that define a single dataset. One reason for combining different CEs is to boost statistics, where each CE has the same characteristics. This is convenient, as the simplest way to apply parallel computing occurs when multiple simulations are run simultaneously and independently. However, the CEs that are combined could represent different conditions, such as different temperatures in MD simulation, fixing a different set of distance constraints in geometric simulation or contrasting mutant structures. Clustering different CEs in the subspace defined by the most relevant PCA modes provides insight into the effect of varying conditions. In some cases, a protein may undergo large-scale (anharmonic) conformational changes that bridge two distinct basins of low free energy. The combined CEs will allow these basins to be clearly identified, as well as the paths connecting them.

Similarly, different CEs that represent a set of mutant structures, or apo and holo forms of a protein, possibly with different ligands bound, allow one to differentiate the conformations easily by clustering in a small dimensional subspace.

The most appealing and intuitive way to investigate the nature of protein motions is to project the displacement vectors (DV) defined in the original high dimensional space that characterize different conformations onto a pair of PCA modes. It is even possible to project onto higher dimensions as one visualizes multiple PCA modes simultaneously using specialized software such as R or XL-STAT ™ , which is a plug-in for Microsoft Excel developed by Addinsoft™. Such plots are indispensible for assessing how well certain parts of the subspace are sampled, especially in comparative studies where differentiation in dynamics can have functional consequences. The results of such an analysis show how each state occupies a region of the conformational space defined by the first two PCA modes.

2.2 Subset Selection and Analysis

A substantial advantage to ED over other methods is that one may choose to examine a subset of a protein to determine the large-scale motions within that particular set of residues under the influence of all the entire set of residues contained in the protein. This last constraint is the key: Running a smaller simulation will inform about a subset of a protein, but it will NOT inform about how that subset behaves under the influence of all the other residues in the protein that interact with it. Subset selection involves choosing a subset of residues/atoms from a protein and analyzing their dynamics from a trajectory derived from the entire protein. Since PCA will always reveal the correlated motions of the atoms with greatest variance in the first mode, zooming in on a protein allows an investigator to determine the functional motions of subsets of a protein. It is critical to realize that the motions of a subset may be washed out in a simulation of the entire protein, but become prevalent once the subset is analyzed. One very useful subset of atoms is the set of all alpha carbons. Analysis of this subset reveals global residue-level correlated motions within the protein and is a coarse-grained analog of the all-atom NMA, which allows one to better interpret the large-scale motions that are often relevant to protein function.

More detailed analysis can be performed using the entire backbone or even all heavy atoms. As the level of detail increases, the results of the analysis portray emergent correlated motions of more complexity. This could be taken to the extreme limit of including all atoms, but interpreting the results of such an analysis would be difficult at best. Other subsets of special interest involve active sites and possible allosteric pathways. Focusing on particular sets of atoms often reveals details that are critical to overall protein function, which while latent within the overall simulation, are often washed out in the subsequent statistical analysis. The result of zooming in on a subset of the protein helps to elucidate such washed out features and provide insight into biological mechanisms. However, determining an appropriate subset of atoms is non-trivial in a brute-force way and is best assisted by biological/biochemical/physical intuition and from experimental insight. Such insight may be derived from using probes, as in the application of fluorescence resonance energy transfer (FRET) analysis.

2.3 Metrics for Assessing Subspace Similarity

Once ED has been performed on a number of trajectories, it is useful to assess how similar the dynamics from each trajectory are to one another. Given that the ED of a protein is characterized using a small vector space defined by PCA modes that reflect different CEs and a combined CE, it becomes necessary to benchmark how similar these subspaces are to one another. When subspaces are sufficiently similar, this implies that the different ensembles capture the same type of protein dynamics. Conversely, when subspaces overlap poorly, different types of motions are being captured, which may have biological consequences tied to the different conditions analyzed. As such, it is necessary to define a measure to quantify the overlap of vector subspaces, as a natural generalization to the concept of a projection (dot product) of one vector onto another.

That said, note that a set of n PCA modes forms an orthogonal n dimensional subspace (SS) within the full vector space (VS) defined by the size of the C-matrix. Common metrics that quantify SS similarity include cumulative overlap (CO), root mean square inner product (RMSIP), and principal angles (PA) (23, 46-50). The CO metric quantifies how well one SS is able to capture the PCA modes of the other SS. The RMSIP metric is a single number that quantifies the SS similarity in terms of multiple inner products between the two. The PA method provides a quantification of the optimal alignment between the two SS that is based on the singular value decomposition (SVD) of a matrix of overlaps (inner products) between the two SS. The result is a sorted (monotonically increasing) set of n angles, where n is the dimension of the compared subspaces, that quantify how well the two SS can be aligned.

When comparing essential subspaces, keep in mind that all of the subspace metrics described above depend on both the dimension of the SS and the dimension of the full VS as shown in Figure 2.3. One way to assess PCA modes is to compare them to the modes of a random process to obtain a baseline for determining the significance of the subspace comparisons as the dimensions for the SS and full VS change. With these baselines, a Z-score can be calculated to assess the statistical significance of the scores, for example when using RMSIP:

$$\mathrm{Z} = \frac{\mathrm{RMSIP_{obs}} - \mathrm{RMSIP_{rand}}}{\mathrm{stdev(RMSIP_{rand})}} \qquad \text{Eq. (2.1)}$$

However, the essential SS of a random process has very different characteristics than the essential SS constructed from a protein trajectory as Figures 2.2-4 clearly show. Randomly shuffling the indices for the components of modes produces a new set of modes that have essentially the same character as the modes determined by PCA on a purely random process. Consequently, any two same-sized proteins share much more in common than would be expected by a random process, making large Z-scores not very useful in practice. This is due to the fact that compared to a completely random process all proteins share much more common dynamics because they share common structural features such as a covalent backbone even if their fold topology is very different. What this means in practice is that any of the metrics described above for any two proteins will show much more overlap compared to a random process. In fact, using two different trajectories under the same conditions, we found that the scores for overlap between two identical proteins can be lower than the overlap between two different proteins when the number of residues is small (<100). This result is augmented when using a coarse-grained approach that prunes many discriminating features (to reduce DOF). To obtain a more stringent criterion for z-score determination, the data presented strongly suggests that a comparison to other proteins, possessing the same number of DOF, that define a decoy set should be used to define the random baseline in Eq. (1), rather than a generalized random process. However, to the best of our knowledge, baselines from decoys have not been done.

Figures 2.5-6 show the risk of comparisons made for small proteins using a coarse-grained model. For this analysis, four proteins having distinctly different folds were simulated under the same conditions using geometrical simulation and then subjected to PCA as a combined set, where only the first 75 residues were included in the covariance matrix starting from the N-terminus and always remaining within the N-terminal domain. Figure 2.5 shows the Z-scores for the comparisons in Figure 2.4. Here it is critical to note the similarity between the random process and the decoy comparisons. When 1WIT is compared to itself (using different simulation conditions), RMSIP saturation suggests that the proper essential subspace dimension is 9 modes. However, the random process and the decoy comparisons do not reach a saturation point within the first 30 modes. When working with larger proteins, such comparisons are much safer, as shown in Figures 2.6-7 with myosin V (MV). The moral here is that extra care must be taken to claim significance of PCA results on small proteins when coarse-graining is used.

Another way to assess how stable the PCA results are can be made by looking for cosine content within the top few PCs. It has been noted that MD trajectories, which insufficiently sample the conformational space of the protein, yield PCs that resemble cosine functions with periods equal to half the mode number, which is what occurs for PCs derived from random diffusion (51). The resemblance is determined by finding the correlation between the set of T values of PCi and the function $\cos\left(2\pi t / bT\right)$ where $0 < t < T, \quad b = i/2$. We note that CEs derived from geometrical simulation do not produce PCs that resemble cosines due to the restriction of conformational space imposed by the locking in of the distance constraints at the beginning of the simulation. However, it can occur in MD simulations, which would be an indication that sampling is limited. Thus, if the CE is derived from an MD simulation, it is prudent to assess the first two or three PCs to determine how much they resemble a cosine function with a period equal to half of the mode index.

2.4 Sampling

A final concern with assessing the PCA output is the significance of the results. While PCA is robust when there is sufficient sampling, the questions that remain are what constitutes sufficient sampling and how trustworthy are the modes. Since PCA relies on the factorization of the C-matrix, the condition number of the C-matrix indicates the numerical accuracy that can be expected within the solution of the associated set of equations. For a given process, more sampling reduces the condition number. Therefore, if the condition number for a C-matrix is high, this could be an indication that there is not enough statistics. If possible, the number of independent samples should be at least ten times the number of variables. Two direct measures for sampling significance are known as the Kaiser-Meyer-Olkin (KMO) score given as:

$$KMO = \left(\sum_j \sum_{k \neq j} r_{jk}^2 \right) \Big/ \left(\sum_j \sum_{k \neq j} r_{jk}^2 + \sum_j \sum_{k \neq j} p_{jk}^2 \right) \qquad \text{Eq. (2.2)}$$

and the associated measure of sampling adequacy (MSA) given as:

$$MSA_j = \left(\sum_{k \neq j} r_{jk}^2 \right) \Big/ \left(\sum_{k \neq j} r_{jk}^2 + \sum_{k \neq j} p_{jk}^2 \right) \qquad \text{Eq. (2.3)}$$

where r is the standard correlation coefficient and p is the standard partial correlation coefficient (52). These statistics can take values between zero and one. If all the partial correlations are zero, then the MSA score is 1. The KMO score indicates the amount of partial correlations between the sampled variables and provides an indicator for when applying PCA is appropriate. The MSA provides a metric for each variable. KMO and MSA should ideally be greater than 0.5. It is worth noting that the MSA scores for each variable are related in a non-trivial way to the protein environment. Specifically, there tends to be a small to medium negative correlation between the MSA scores when averaged for each residue and the residue RMSD.

2.5 Kernel Methods

When choosing to work in the sample space, either due to a small number of samples or to implement a non-linear method, one must construct the kernel matrix (K), which is an $n \times n$square symmetric matrix, where $n$is the number of observations. Each element of $K$is formed by computing $K(i, j)$, where $i$and $j$represent two observations from the centered data set, using the definition for the specific kernel function of interest, k. Essentially, the kernel function maps N dimensional vectors in $\Re^N$ from the sample space to a new high dimensional (possibly infinite) vector space referred to as feature space. Working in the high dimensional feature space can often detect features that are not apparent in sample space. The “curse of dimensionality” is avoided by constructing the feature space from a collection of inner-products so that the actual mapping function is never calculated. Calculating inner products over the sampled data is not by itself an intensive operation. This method of avoiding the difficulties normally associated with high-dimensional spaces is known as the “kernel trick”. It is worth noting that using this approach, only a subset of feature space is being explored, which is limited by the range of the data of the original sample space.

The kernels that can be employed must yield positive-definite symmetric square matrices (33). When the kernel is defined simply as the inner product of the input data (linear kernel), then the results of the analysis are identical to the standard PCA. Specifically, one will recover the same set of non-zero eigenvalues as that from the covariance matrix based PCA. In this sense, kernel PCA (kPCA) subsumes standard PCA. Additional features may be detected by using other types of non-linear kernels, such as a Gaussian kernel, a Neural Net kernel (i.e., a tanh function), a kernel that maps the data to a set of degree n polynomials (either homogeneous or inhomogeneous), or a mutual information kernel. There are no rigorous guidelines for which kernel to apply to the data of interest and thus the method of kPCA requires intimate knowledge of one's data (or based on trial an error) as well as how a particular kernel might or might not affect the resolution of multiple states. Furthermore, most kernel functions have adjustable parameters that need to be set to obtain the best resolving power within feature space.

Unfortunately, there is no a priori formula for parameter optimization because this process is highly dependent on the data used. Lastly, unlike standard PCA where the PCs are generated by taking the dot product of the DVs and the appropriate eigenvector, the process for kPCA is more involved. First, the eigenvectors must be normalized in the sample space to reflect the fact that their magnitude in the feature space is unity, and then the PCs (for the training set) are calculated by determining the sum of the inner products of the normalized eigenvectors with the kernel columns. Having used both standard PCA and kPCA, we note that when the parameters are suitably tuned, the ability of kPCA to discriminate multiple states from a trajectory is impressive.

If kPCA is to be used, we note that an ideal approach for computationally intensive kernels is to first use PCA to reduce the dimension of the data and then apply the kernel methods to the top set of PCs. In this approach, we have found that as few as five PCs may be used as input to kPCA with no substantial loss in numerical accuracy. This filtering process greatly reduces the computational intensiveness of the kPCA, although it does not reduce the size of the kernel matrix.

2.6 Additional PCA Variations

For completeness, we briefly consider the method of Independent Component Analysis (ICA) (53). ICA is a method for performing blind source separation, as when one wishes to decompose a mixed signal into two signals or a signal plus noise. The underpinning mathematics of the method is to detect non-Gaussian processes by looking at higher order correlations than second degree. To achieve this, ICA is typically implemented using either kurtosis or an information theoretic quantity like mutual information (FastICA) as a contrast function (54).

To apply ICA, one must first center the data and then whiten it. Whitening is the process of transforming an observed data vector linearly so that one obtains a new vector, which is white, i.e. its components are uncorrelated and their variances equal unity. In other words, the covariance matrix of a whitened data vector equals the identity matrix. One method for whitening data involves an EVD of the covariance matrix and is given by $\tilde{X} = ED^{-1/2}E^{T}X$ where X is the centered data, E is the matrix of eigenvectors from the EVD of the covariance matrix, with E^{T} its transpose, and D is the diagonal matrix of eigenvalues from the EVD of the covariance matrix. Once the data has been centered and whitened, the ICA algorithm essentially computes the optimal rotation of the data using higher order statistics (e.g., fourth moments), thereby determining the independent components (ICs). We note that the algorithm can be computationally expensive for high dimensional data when a large number of ICs are to be extracted.

In order to make ICA amenable to large, high-dimensional datasets like protein CEs, PCA is first applied to perform a dimensionality reduction and whitening pre-processing step. Similar results to ICA may be obtained from kPCA by choosing to work with a kernel that maps the data to inner products of degree two polynomials. Such kernels have the property of detecting fourth moments, i.e. kurtosis. Alternatively, we note that one may perform post hoc analyses of the PCs derived from either standard PCA or kPCA to determine which ones have the highest amount of kurtosis. Choosing to examine such PCs will allow the investigator to see if non-Gaussianity, as measured by kurtosis, leads to the detection of a biological signal. The real criterion for assessing the usefulness of ICA is determining if the assumptions of the model are met. We find that for investigating native state dynamics, where proteins are described by a large set of DOF and are not undergoing large conformational shifts, ICA does not provide greater insight than what PCA (or kPCA) provides because most of the variables in the CEs are Gaussian.

PCA is a multivariate statistical approach, and there is almost no limit to the variants available to an investigator. For example, one may perform sparse PCA (SPCA) in which one attempts to form linear combinations that are sparse, meaning that they are combinations of less than all the variables. This is done in an attempt to make the interpretation of the PCA more manageable as is the case of standard PCA the linear combinations include all the variables and in high dimensional data, rendering an interpretation as non-trivial at best. Typically this is done by using a thresh-holding method such as any component less than c is mapped to zero, where c is an ad hoc chosen number between 0 and 1 or by solving an optimization criterion as in the case of SPCA (55). The effect of such a sparsification is the reduction of complexity in interpretation of correlated motions and often better cluster separation. The problem with the approach is that there is no guarantee that the sparse variables are the important ones.

Another approach combines PCA and ICA methodologies in a process called Independent Principal Component Analysis (IPCA) (56), based on the assumption that biologically meaningful components can be obtained if most noise has been removed from the associated loading vectors. In IPCA, PCA is used as a pre-processing step to reduce the dimension of the data and to generate the loading vectors. The FastICA algorithm is then applied to the previously obtained PCA loading vectors to generate the Independent Principal Components (IPCs). In this method, the kurtosis measure of the loading vectors is used to order the IPCs. There is also a sparse variant with a built-in variable selection procedure implemented by applying soft-thresholding on the independent loading vectors (sIPCA).

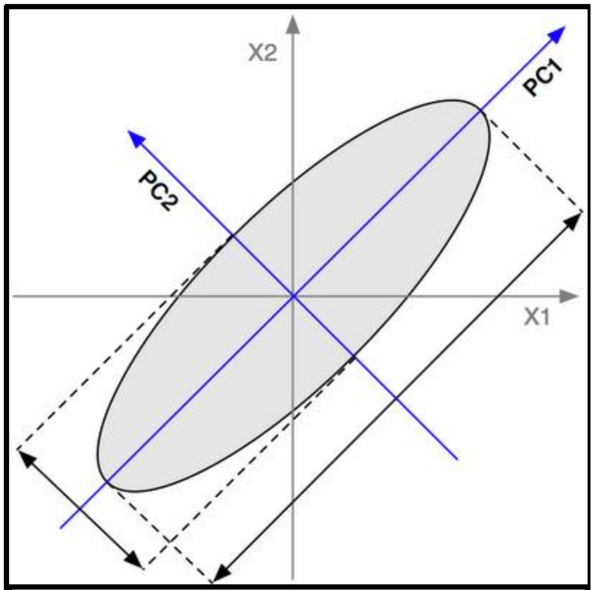


Figure 2.1 PCA
PCA identifies the directions of highest variance in a data set. PC1 is aligned with the direction of highest variance in the data. While PC1 and PC2 together represent the original data perfectly, most of the original information is captured by PC1.

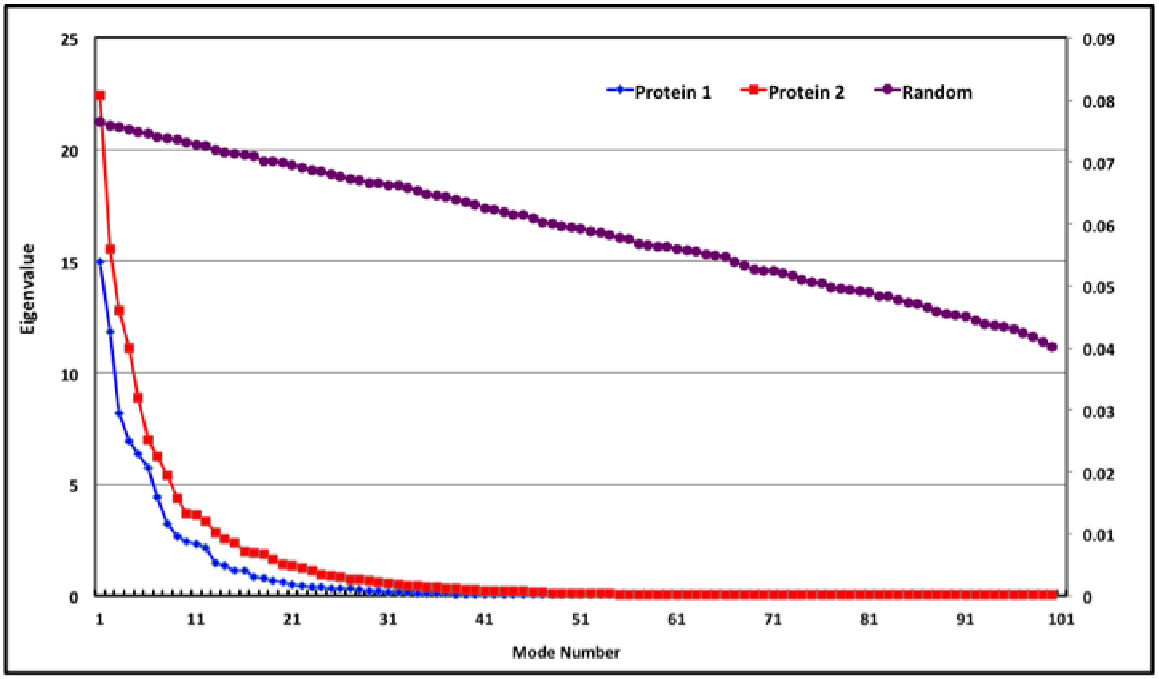


Figure 2.2: Eigenvalue Scree plot for first 100 modes of two example protein simulations and a random process, each having 225 dimensions. The random process is shown on the secondary y-axis.

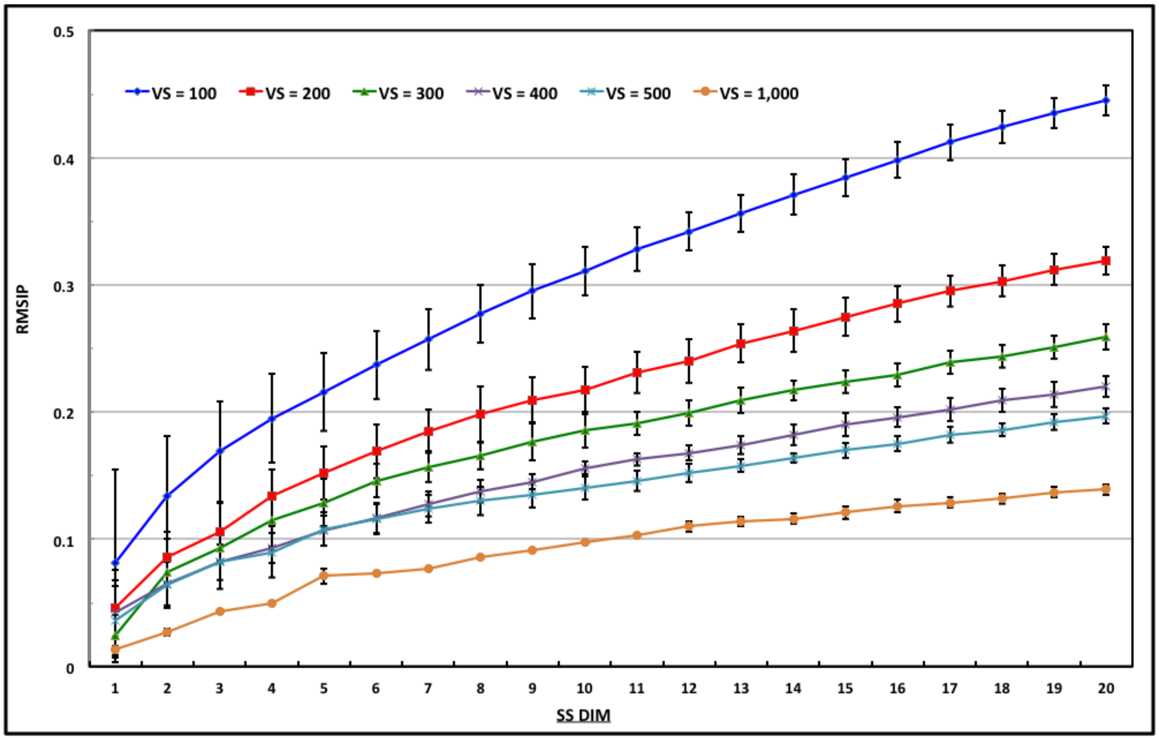


Figure 2.3: Average RMSIP scores for a random process in different vector space dimensions as a function of subspace dimension. Error bars show plus and minus one standard deviation.

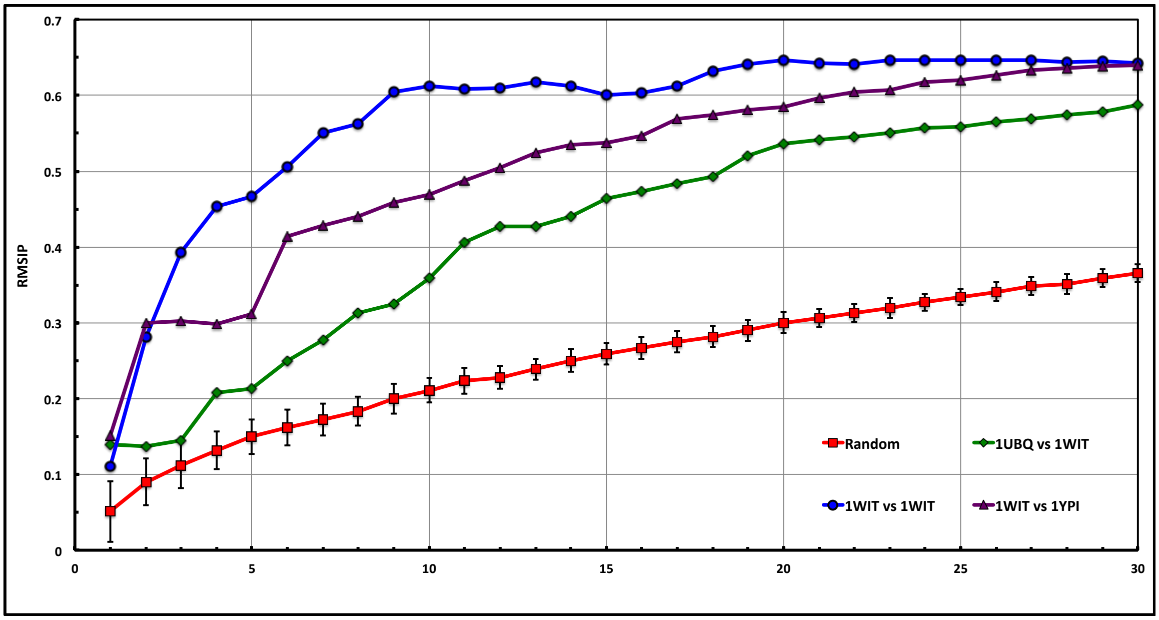


Figure 2.4: RMSIP scores for inter-comparisons between 3 proteins each having 75 residues and a random process with 225 DOF. Only the true self-comparison yields a curve that saturates rapidly within a small essential space defined by the first 9 modes. The decoy plots have much in common with the random process as shown by the general increase over the first 30 modes.

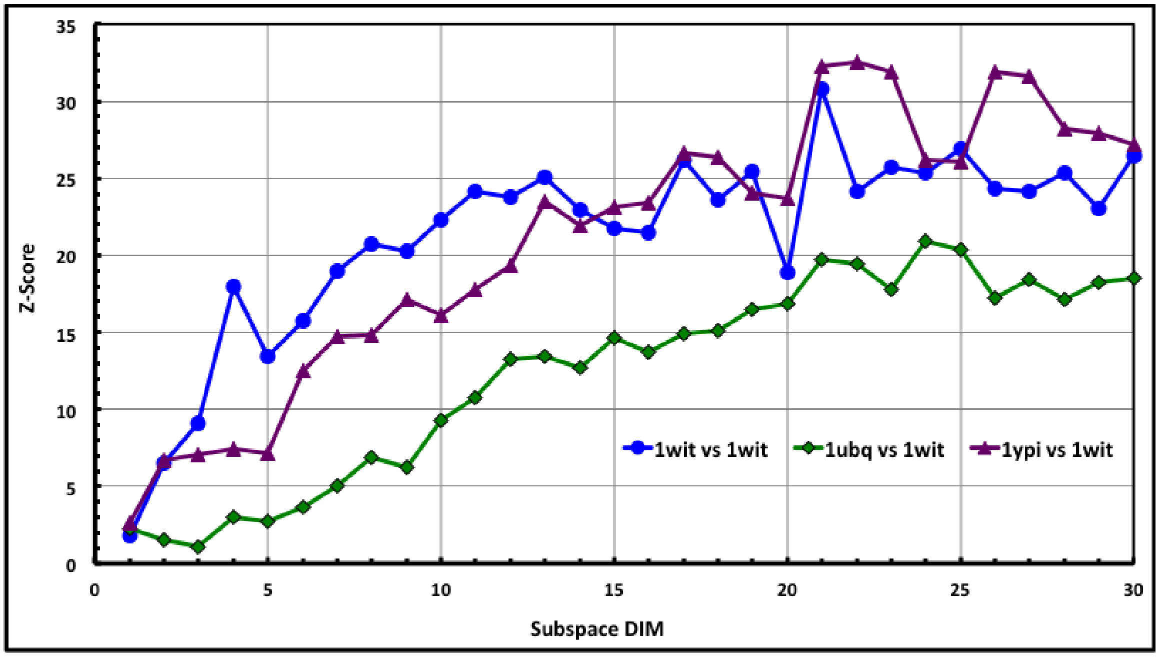


Figure 2.5: The Z-scores for the RMSIP scores shown in Figure 2.4.

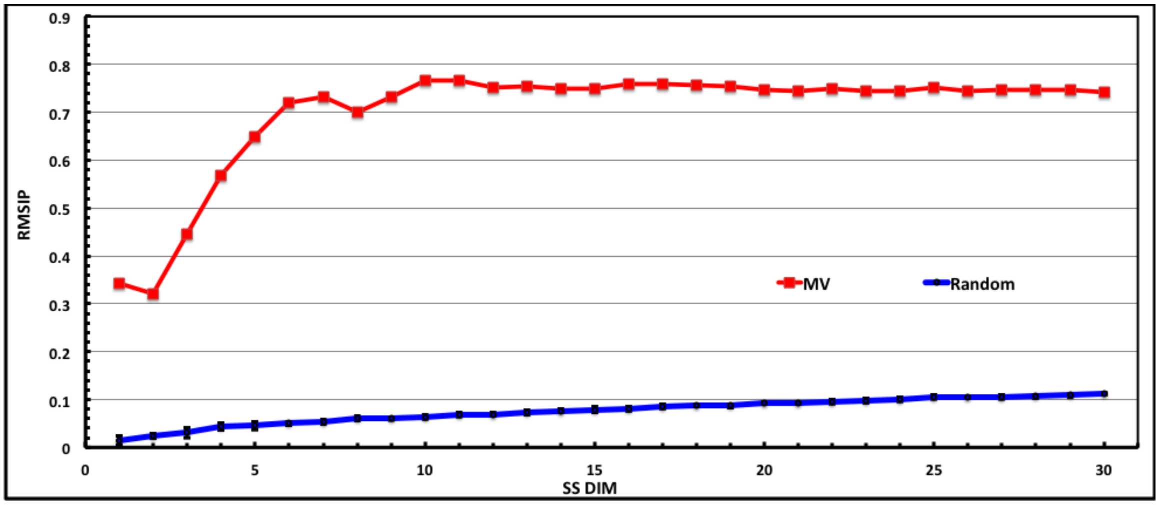


Figure 2.6: Comparison of two myosin V (795 residues) CEs run under different simulation conditions and a random process with 2,385 DOF. Again, note the rapid saturation of the RMSIP scores in an essential subspace defined by the first 10 modes.

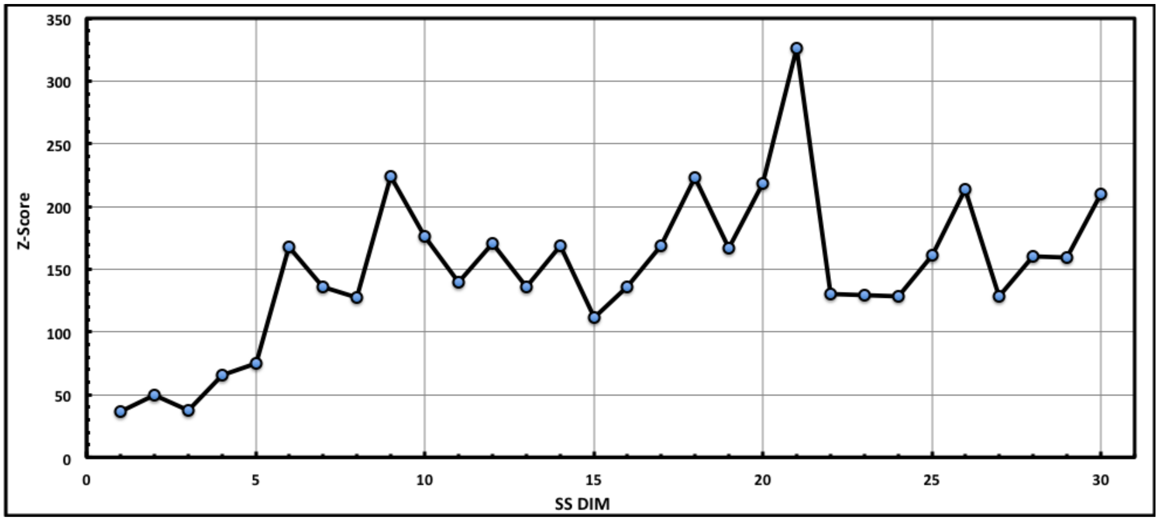


Figure 2.7: The Z-scores for the RMSIP scores in Figure 2.6.

CHAPTER 3: BENCHMARKING THE GSM

3.1 Introduction

Since the GSM requires user input to determine the number and type of constraints that will be used to determine the RCD, we ran many simulations on small single domain proteins in order to determine if the default parameters were consistent and appropriate. We also assessed how critical the choice of the constraint assignment parameters was for determining the outcome of the model. A systematic selection of HB energy cutoffs and HT distance cutoffs were assessed. Additionally, a number of assessments were made regarding how rapidly the simulations equilibrated, how many output structures were needed to determine a statistically significant trajectory, and what the optimum output frequency should be. These assessments were made from trajectories that were run in diffusion mode and using the momentum perturbation feature. Finally, measurements were made concerning the how much diversity exists in the sampled structures and if there was any concern for repeat structures due to periodicity in the simulation.

3.2 Methods

All simulations were performed on UNCC computing resources using FIRST/FRODA version 6.2 using software downloaded from http://flexweb.asu.edu/

We designed a Java software package for the analysis of trajectories produced from the GSM and MD. Key steps in the process are:

1. Read in the trajectory(s)
2. Select a set of atoms
3. Align the structures defined by the atom set to remove overall translation and rotations using quaternion operations (see APPENDIX B)
4. Compute the conformational RMSD for the trajectory
5. Compute the residue RMSD for the protein over the trajectory
6. Calculate the distance of each trajectory structure to a reference structure
7. Calculate the all-to-all distances between alpha carbon atoms
8. Construct the coordinate based C-matrix
9. Construct the distance based C-matrix
10. Perform an EVD of the C-matrix
11. Select the "essential modes"
12. Generate output files for the eigenvalue scree plots
13. Generate the top specified RMSD PCA modes
14. Construct the top specified weighted-RMSD PCA modes
15. Construct an iterated set of structures for PCA mode visualization of specified modes
16. Output an edited PDB file for the structure with B-factors replaced with simulation residue RMSD for visualization
17. Construct the displacement vectors for the simulation using a specified structure
18. Construct the principal components
19. Compare sets of specified PCA modes (subspaces) from multiple simulations using CO, RMSIP, and PA
20. Calculate the collectivity of the eigenvectors

21. Construct various kernels for non-linear feature analysis

We also relied on standard statistical packages such as XLStat, an add-on for Microsoft Excel developed by Addinsoft™ for additional statistical scores like the KMO and MSA.

3.3 Results

FRODA yields trajectories very quickly. Additionally, the trajectories equilibrate rapidly for small, single domain proteins and large multi-domain proteins when using the momentum perturbation. Longer simulations are needed when using FRODA in diffusion mode (Figure 3.1). Minimal sampling is achieved when there are more samples than variables. Some variables are sampled better than others due to their environment (e.g., the secondary structure) as measured by the MSA scores. Obtaining very high KMO scores and/or reducing the condition number of the C-matrix requires tremendous sampling. A best practices guideline suggests that obtaining a number of observations that is at least ten times the number of variables, where the samples are somewhat similar to each other (i.e., using smaller output frequencies) to enhance inter-variable correlations and reduce partial correlations. Output frequency is generally optimized when it is between 25 and 50 (i.e., every 25th or 50th frame) (Figure 3.2) but should be individually assessed per protein. Reducing the frequency means that the time of the simulation is shortened, as less conformers will need to be generated to produce a large statistical sample. While the conformational RMSD tends to fluctuate, it does so about an average and reaches that average very quickly. Although similar RMSD values are seen with regular periods, when the distances between the trajectory structures and the initial structure were examined, the distances were not close to zero, in fact they were not very different from the average distance, suggesting that the structures were nicely randomized. We also assessed the degree of constraint violation to determine the size of the randomization step to optimize the simulation. It turns out that the recommended value of .01A for using the momentum perturbation was viable and, as seen in the histogram of constraint violations (Figure 3.3), most violations were smaller than 0.05A.

Both conformational RMSD and residue RMSD are highly dependent on the HB cutoff and the HT cutoff as seen in Figure 3.4. In addition, rigidity transitions are very sensitive to the HB cutoff, as seen in the Figure 3.5. It is interesting to note that the plot of number of iDOF per residue (Figure 3.6) does not have the abrupt transitions seen in the rigidity plot (Figure 3.5). Based on movies created from the simulations, it did appear that some trajectories were over-constrained, while others were under-constrained. Furthermore, a range of parameters emerged wherein the protein behaved in similar dynamical ways. This was assessed by performing subspace analyses on the top 20 modes and discovering that there was indeed a “range of physicality” for protein dynamics in the GSM, as seen by the similar RMSIP and first PA scores, as well as robust CO scores. We chose the 20 dimensional subspace after assessing how both RMSIP and the first PA depend on the size of the subspace. Typical representations of RMSIP, PA, and CO are shown in Figures 3.8-10.

One very surprising result was that the dynamical behavior of the single domain proteins was not appreciably affected by the addition or subtraction of constraints within this range of physicality. Specifically, altering the HT cutoff did not significantly alter the mode spaces. Closer inspection of the movies of these trajectories revealed that the complete removal of the HTs yielded unphysical behavior. Moreover, the number of independent degrees of freedom (iDOF) is determined by the values of the HB cutoff and the HT cutoff. It is possible to produce two simulations with the same number of iDOF using contravariant mixtures of the HB cutoff and the HT cutoff. In all cases, the dynamical similarity persisted across a wide range of the constraint assignment parameters and only deviated notably at the extremes. Additionally, we saw that the choice of a 20 dimensional subspace, similar to what is done in NMA for the anisotropic network model (ANM), is sufficient to capture most of the large-scale long-timescale dynamics of the proteins.

Finally, we assessed how well the GSM samples the conformational space of the native state by projecting the simulation displacement vectors onto the top two PCA modes (DVP). As shown in Figure 3.11, the GSM generates conformers that very effectively spans the phase space defined by the top two PCA modes. While such plots are not always circularly symmetric, substantial coverage of the mode space is a good indication that the simulation has sampled the native basin in a statistically significant manner and has reached equilibration.

3.4 Conclusions

The GSM is an all atom model that uses geometrical constraints assigned though rigidity analysis derived from graph theory. The model is very fast and yields trajectories that equilibrate rapidly provided that one uses the momentum perturbation. Sampling of the native basin is thorough as was seen by projecting the DVs on the top few PCA modes and the mode spaces are highly conserved over a wide range of constraint assignment parameters as measured by RMSIP and PA. The simulations did not become irrevocably jammed nor did they yield structures containing large constraint violations. In fact, the GSM as implemented in FIRST/FRODA is very stable, fast, and efficient at producing trajectories that well characterize the native basin defined by the input structure.

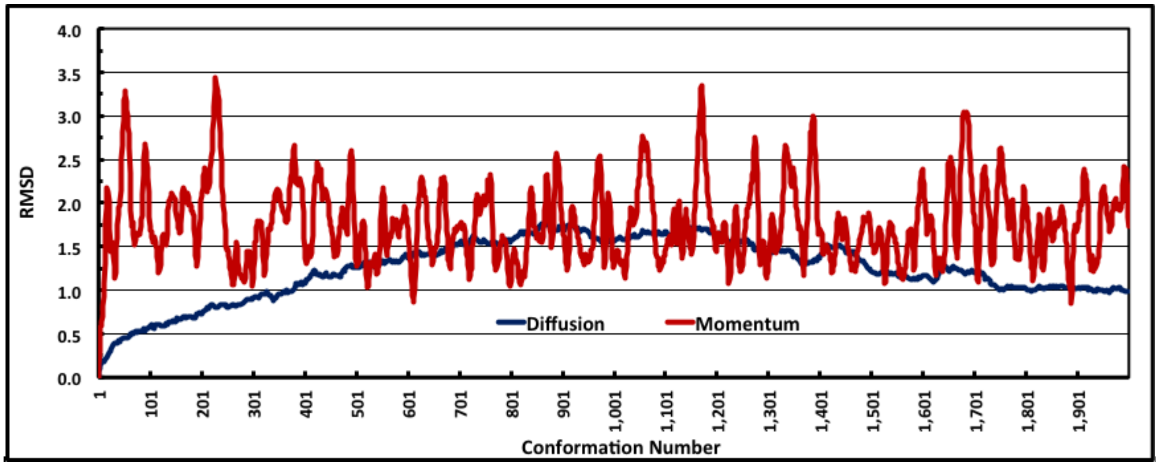


Figure 3.1 Conformational RMSD using FRODA
Running FRODA in diffusion mode yields trajectories that do not equilibrate rapidly due to low efficiency in how configuration space is sampled. These results are for MV in rigor state, which contains 946 residues.

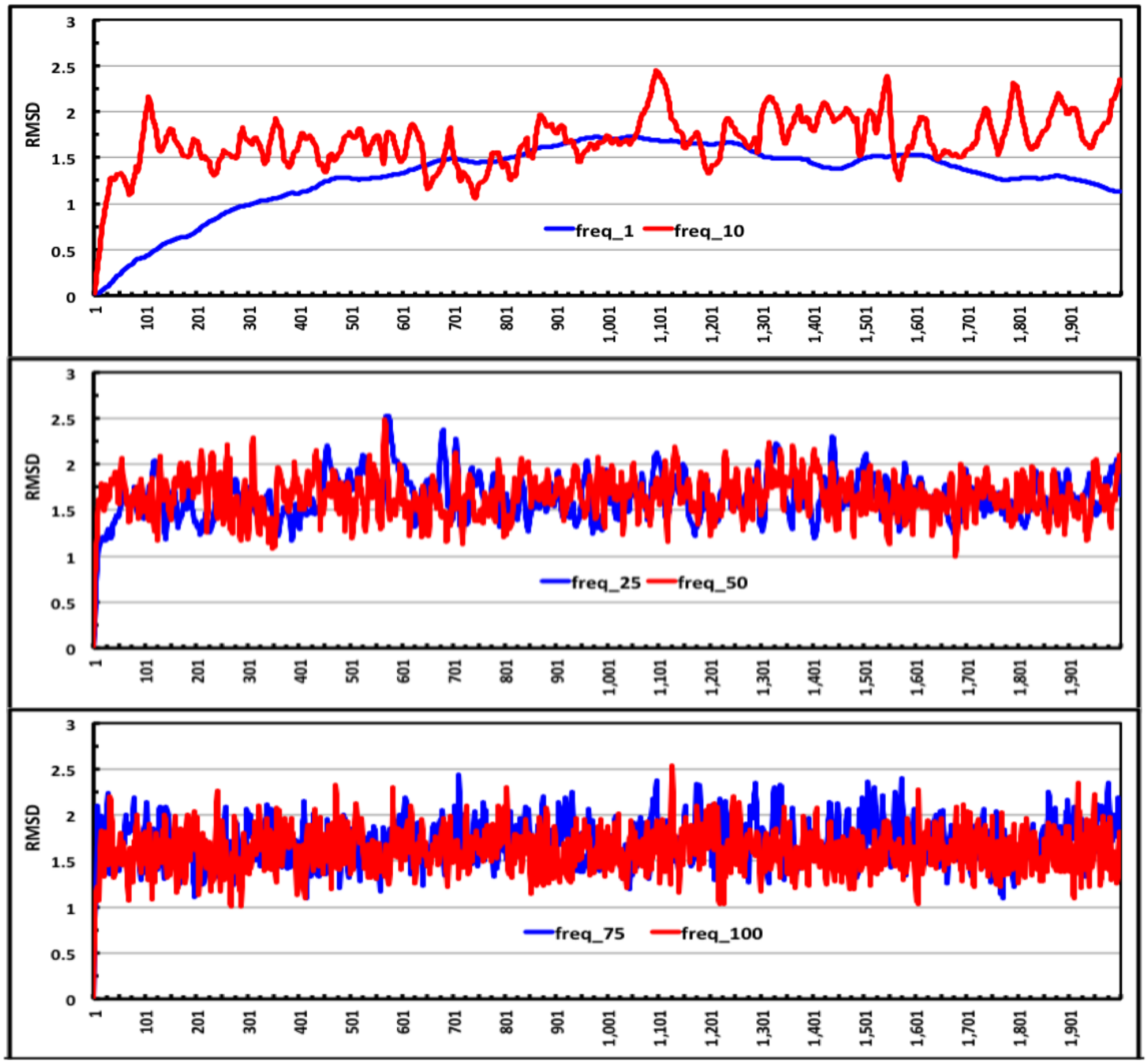


Figure 3.2 The dependency of equilibration rate on output frequency.
There is a rtrade-0ff between sampling every conformation and sampling every 100th conformations. Here, the conformational RMSD is plotted as a function of FRODA output, with the abscissa showing conformations from1 to 2,000. Using an output frequency of 25 to 50 balances the time needed to sample more of the configuration space with the requirement to achieve sufficient statistical sampling. There results are for myoglobin, which contains 1512 residues.

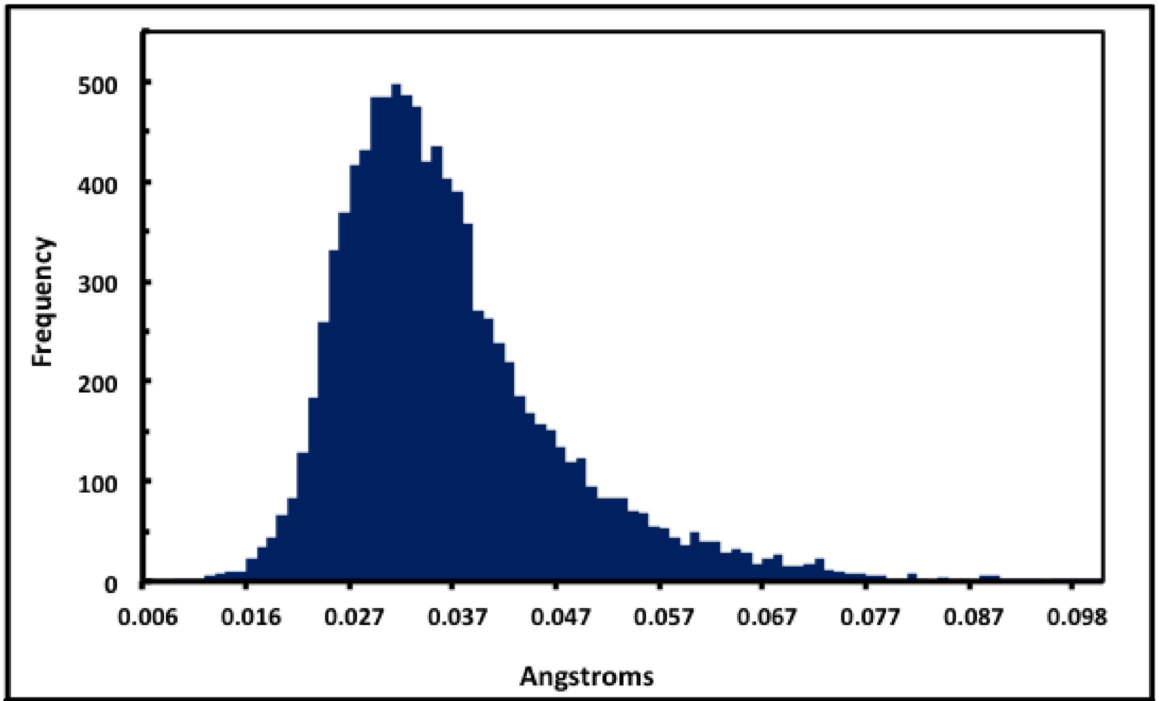


Figure 3.3 Constraint Violations
When using the momentum feature or FRODA, it is important to verify that the output structures are of good quality with very small errors in the geometric constraint matching. With a small perturbation (step size) of 0.01A, very small errors are present in the structures and the simulation is able to explore the conformational space much more effectively than when using a step size of 0.1A and the diffusion option.

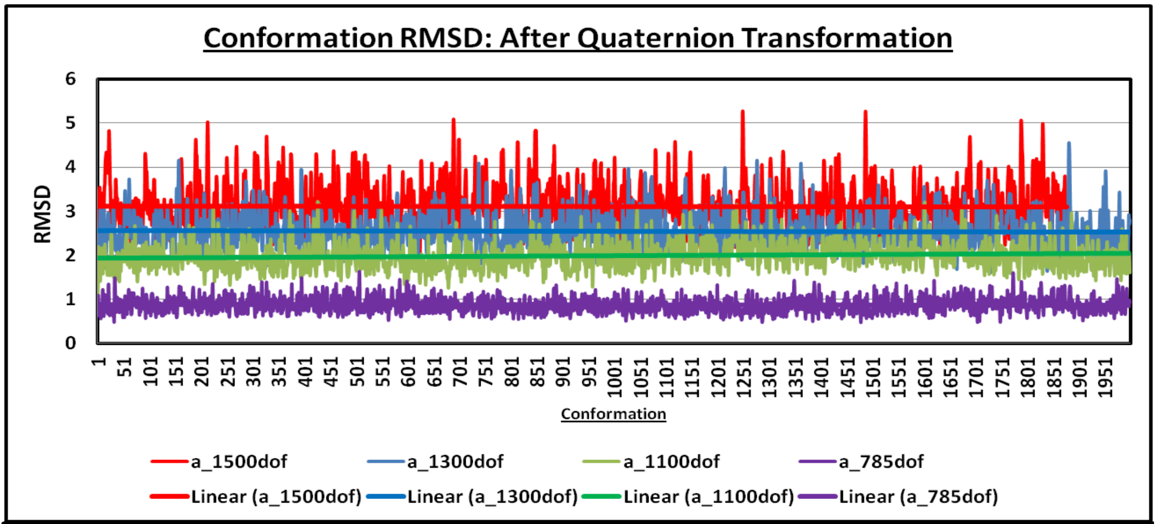


Figure 3.4 Conformation RMSD
The difference in RMSD to the initial structure is plotted for four FRODA trajectories consisting of 2,000 structures. All four show equilibration, and illustrate how the number of iDOF (and by extrapolation, the HB energy cutoff) affects the geometrical simulation.

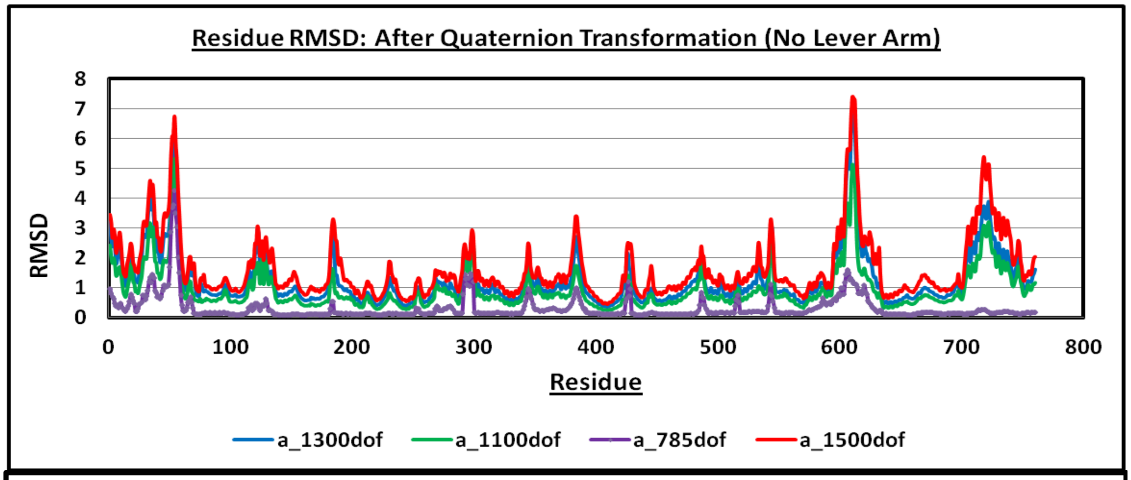


Figure 3.5 Residue RMSD
Variations in residue fluctuations as a function of the number of iDOFs. Here 763 residues of the rigor state of myosin V are plotted (the lever are was removed in this analysis).

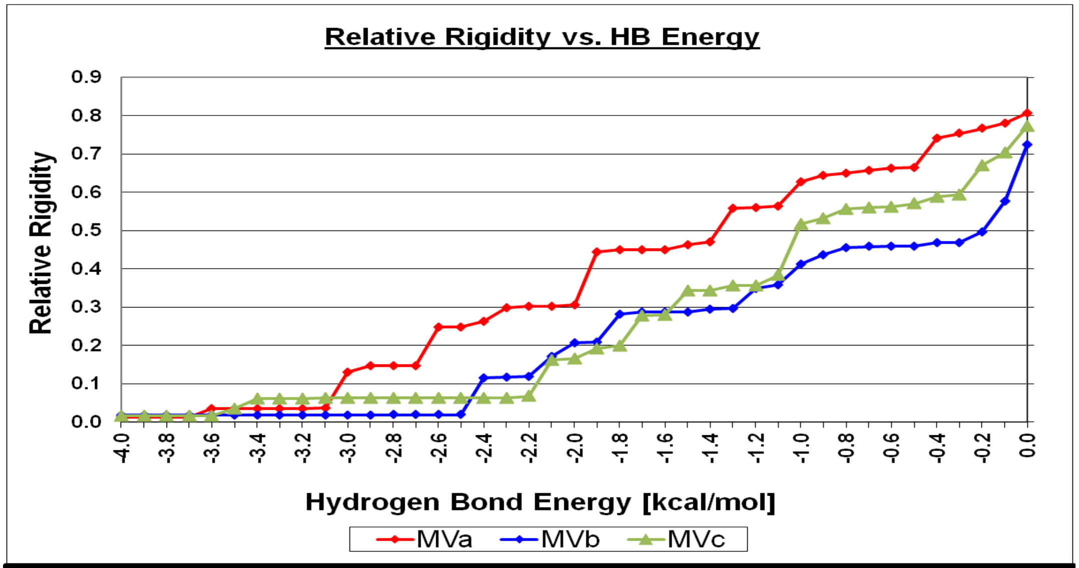


Figure 3.6 Rigidity Percolation
An analysis of how rigidity transitions occur in the three forms of myosin V as a function of hydrogen bond (HB) energy cutoff in FIRST. Due to variations in HB networks, each structure undergoes stepwise jumps in rigidity as HBs are removed.

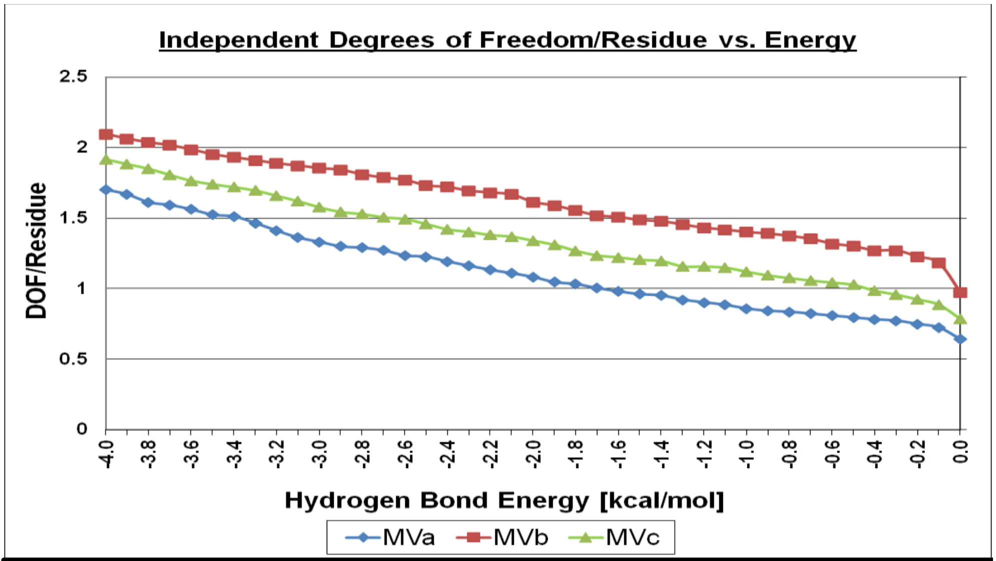


Figure 3.7 iDOF as a function of HB Energy.
The trends in changes in the number of iDOFs per residue as the number of HBs are changed. Here a smooth transition is seen, unlike the case for the rigidity transitions.

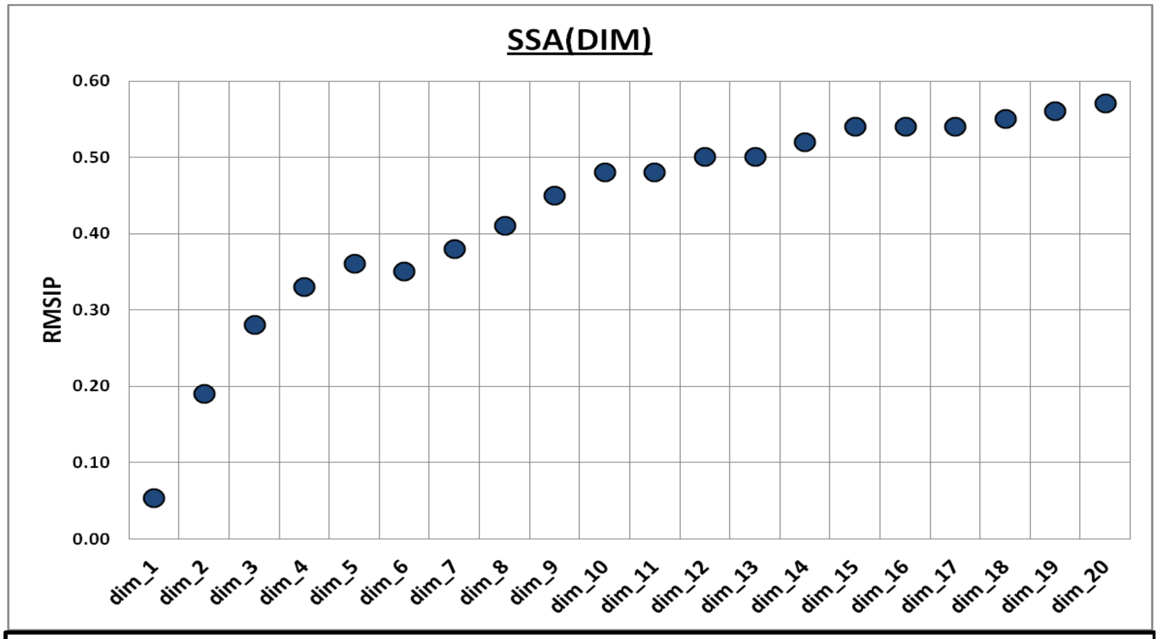


Figure 3.8 Subspace Analysis using RMSIP as a function of Dimension.
RMSIP values are heavily dependent on the size of the subspaces that are compared. The RMSIP score tends to saturate as higher dimensional subspaces within a vector space are compared. Balancing the fact that the top few modes are relevant for biological motions with the fact that better comparisons result from larger subspaces, it appears that the 20 dimensional subspace is a good compromise.

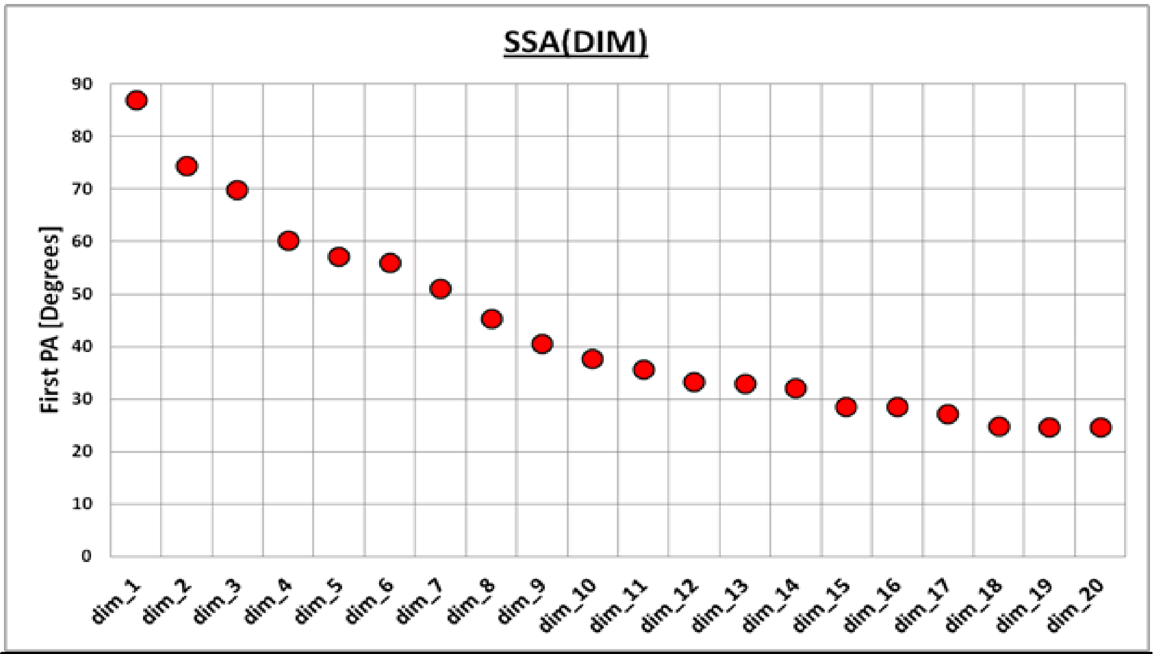


Figure 3.9 Subspace Analysis using First Principle Angle as a function of Subspace Dimension.
A similar trend is seen for the first PA analysis as is seen for the RMSIP analysis. Again, saturation occurs for larger subspaces. The choice of a 20 dimensional space is reinforced here.

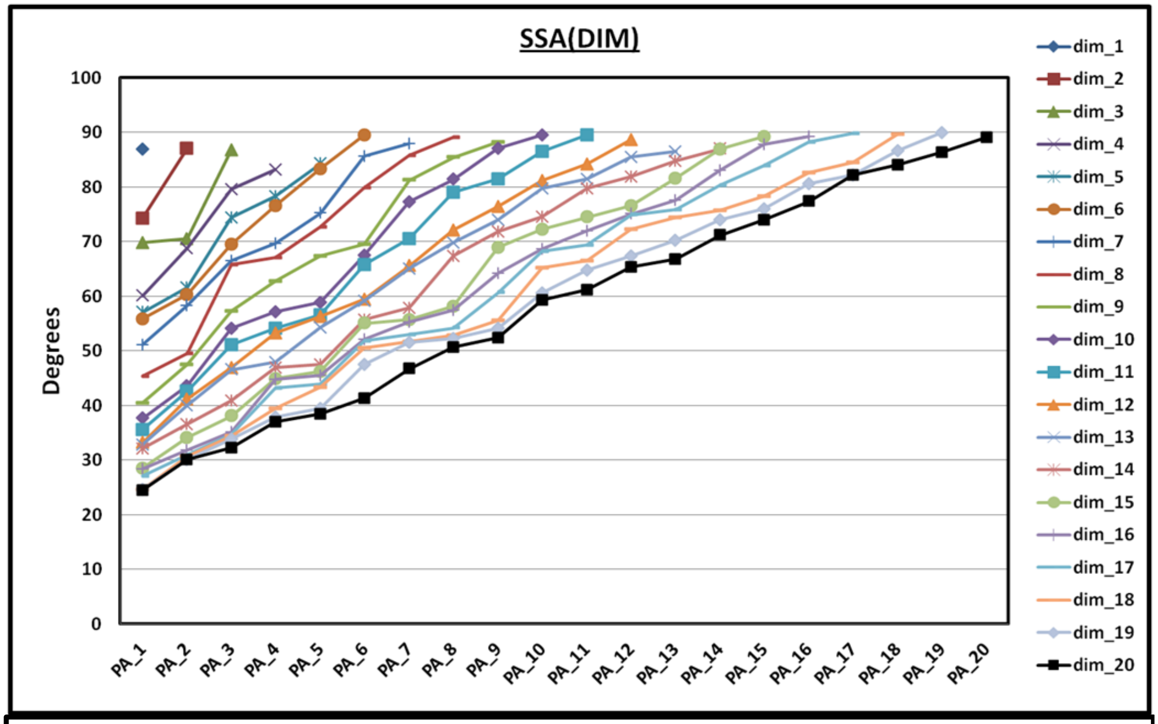


Figure 3.10 The set of Principle Angles as a Function of Subspace Dimension.
In this analysis, not only is the trend for the decrease in first PA seen, but the optimization effect of the SVD used to generate the set of Pas becomes evident too. Again, choosing a 20 dimensional subspace provides a good foundation for examining modes that are relevant to biological function while also providing optimal alignments of the two subspaces.

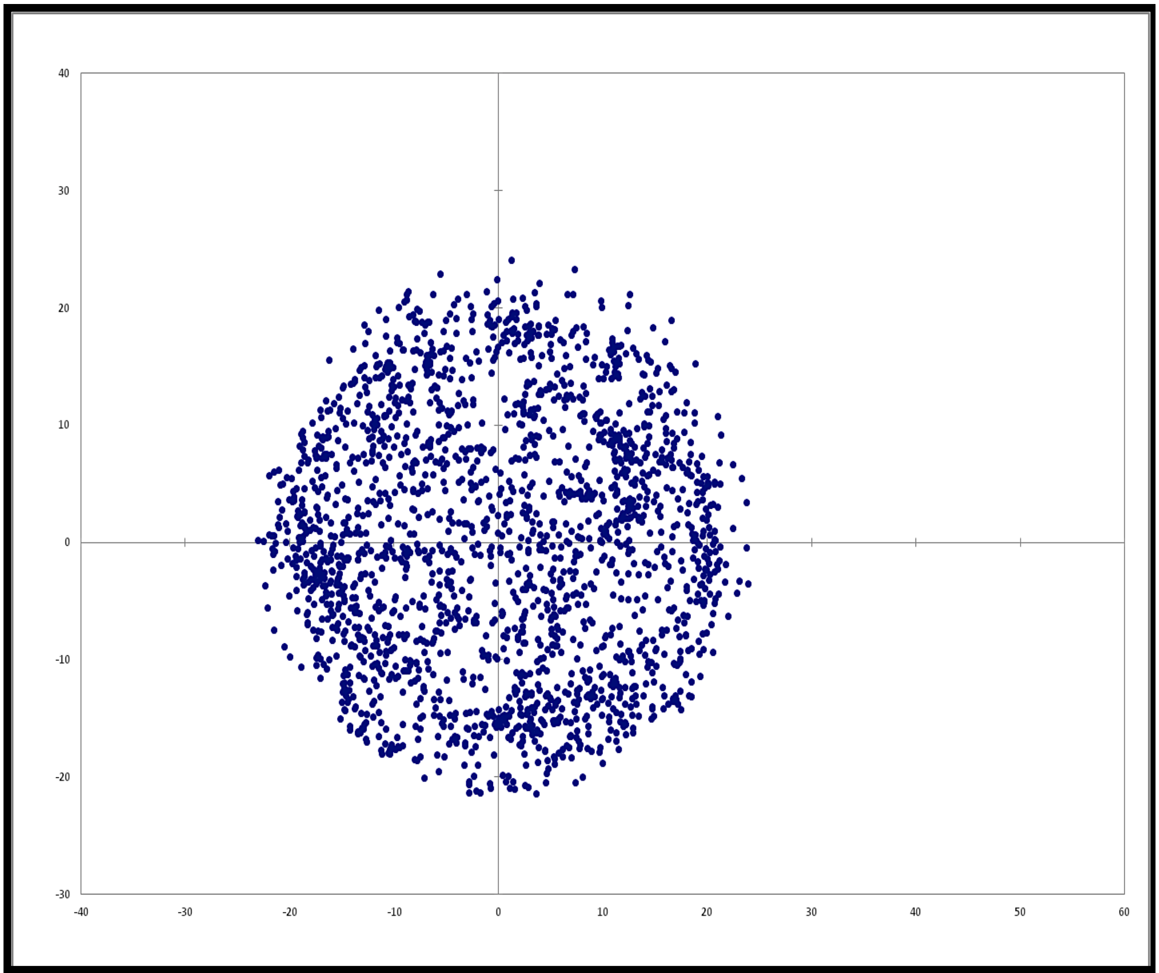


Figure 3.11 Scatterplot of the top 2 PCs. Notice that the sampling of the phase space is fairly uniform and thorough. This indicates good sampling and equilibration in the essential space.

CHAPTER 4: MODEL-TO-MODEL COMPARISONS

4.1 Introduction

In order to assess the degree to which ENM, MD, and GSM provide equivalent information, we selected four sample proteins chosen from distinct structural classes (SCOP) (57). These proteins were simulated using the three models and the results compared using subspace analysis of the top 20 mode spaces. We used MD trajectories archived by the Daggett Group available at www.dynameomic.org and we used the ANM web server available at http://ignmtest.ccbb.pitt.edu/cgi-bin/anm to generate the normal modes. We constructed the GSM trajectories in-house. All analysis was done using our Java based analysis package: Essential Dynamic of Proteins with Subspace Analysis in Java (in preparation for publication in BMC Bioinformatics).

4.2 Methods

4.2.1 Geometrical Simulation

GSM trajectories were created using FIRST/FRODA version 6.2 using software downloaded from http://flexweb.asu.edu. For each protein, trajectories were created using the command line[1] for a variety of constraint assignment parameters (the parameters x and y in the footnote) that modify how many of the possible constraints in the input structure are used for the RCD. Each trajectory was obtained by selecting every 50th structure in a simulation that generated 100,000 structures, yielding 2,000 sample

[1] FIRST –FRODA –froda2Hybrid –froda2Momentum –totconf 100000 –freq 50 –step 0.01 –body –E *x*–H 3 –c *y*–ph_tol 2.50 –non –v 0 1A6N.PDB

structures. The variations span the spectrum of rigidity from highly over constrained, with all H-bonds modeled as distance constraints, to completely flexible with no H-bonds modeled as a distance constraint.

The two main parameters that were varied were the number of H-bonds, controlled by the H-bond Ecutoff, and the number of hydrophobic tethers, controlled by the hydrophobic (HP) tether cutoff using the default and recommended “H3” hydrophobic tether assignment scheme. Note that it is expected that the H-bond Ecutoff will be changed when using FIRST/FRODA, but normally the hydrophobic interactions are left at default values. In this work, we explored modifying the HP tether cutoff criteria in addition to the H-bond Ecutoff. Within the H3-hydrophobic rules, distances between certain pairs of carbon atoms are restrained using inequalities that are implemented as a half harmonic potential function. That is, if the distance between a pair of carbon atoms exceeds a maximum value (an example command line specification of this value is given as –ph_tol=2.50 in angstroms), then a restoring force is applied to reduce their separation. However, in addition, if the distance between a pair of carbons atoms is less than a minimal value (an example command line specification of this minimum value is given as –c=0.50 in angstroms), two distance constraints are placed between these two carbon atoms, where each atom is modeled as a rigid body.

For completeness, the switch froda2Hybrid activates the newest version of the FRODA engine (no more ghost templates as in the older version), the switch froda2Momentum introduces a perturbation that allows movements that were successful to be weighted so that the next Monte Carlo step is in that direction, and the switch -body instructs the program to move entire rigid clusters as a unit rather than perturbing each individual atom.

4.2.2 Displacement Vectors

The input structure used to derive each trajectory was also used as a reference to construct a set of displacement vectors by subtracting it from each of the generated output structures.

4.2.3 Principal Component Analysis

PCA was done using the covariance matrix of the alpha carbon positions from each trajectory. Since the objective was to identify global collective motions, the sensitivity cutoff for PCA was coarse-grained by using only the alpha carbons. The structures comprising each trajectory were appropriately aligned to remove overall translation and rotation from the intrinsic atomic fluctuations prior to the PCA. After diagonalization of the covariance matrix, the top 20 PCA modes were selected for subspace comparisons and the projection of displacement vectors.

4.2.4 Overlaps of Displacement Vectors and Modes

The overlap between a displacement vector and a principal mode was determined using the inner (dot) product of the two vectors, normalized by their respective magnitudes. The NIP as defined here is the same quantity as the overlap between two vectors as defined by Sanejound (23), where O_{ij} represents the overlap of the ith vector in subspace one with the jth vector in subspace two.

$$O_{ij} = \frac{\left| u_i \cdot v_j \right|}{\left\| u_i \right\| \left\| v_j \right\|} \qquad \text{Eq. 4.1}$$

4.2.5 Comparing Subspaces by Cumulative Overlap

A cumulative overlap was calculated to assess how well a given model eigenvector is represented in another model's principal motion subspace. This value was obtained by successively determining the overlap between the given eigenvector and each of the eigenvectors in the other model's subspace (26).

$$CO(k) = \left(\sum_{j=1}^{k} O_{ij}^{2} \right)^{\frac{1}{2}} \qquad \text{Eq. 4.2}$$

For our analysis, k was always equal to twenty and O_{ij} represents the overlap of the ith vector in subspace one with the jth vector in subspace two. Since this calculation is not symmetric, the analysis was performed twice, first for vectors in subspace one on subspace two, second for vectors in subspace two on subspace one. The average of these two values for each vector was reported as the average CO.

4.2.6 Comparing Subspaces by Root Mean Square Inner Product

The mode subspaces were globally compared using root mean square inner product (RMSIP) (27,28).

$$RMSIP(I,J) = \left(\tfrac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(u_i \cdot v_j \right)^{2} \right)^{\frac{1}{2}} \qquad \text{Eq. 4.3}$$

In our analysis, I and J were both equal to twenty, u_i is the ith vector in subspace one, and v_j is the jth vector in subspace two. RMSIP scores range from zero for mutually orthogonal subspaces to one for identical subspaces. The RMSIP score is effectively the correlation between the vectors in subspace one with the vectors in subspace two. A value of 0 or 1 respectively indicates no or full correlation. A score of 0.70 is considered an excellent correspondence while a score of 0.50 is considered good (46). We note that the RMSIP score is dependent on the size of the subspaces compared, such that for a given RMSIP score the result is more significant for larger subspaces compared to smaller subspaces. This size dependence is not encountered when using principal angles, as this process will always yield a set of angles in the range of [0,90 degrees].

4.2.7 Comparing Subspaces by Principal Angles

Two mode subspaces F and G with $\dim(\mathrm{F}) = \dim(\mathrm{G}) = 20$ were assessed using principal angle analysis, also called canonical correlations (48-50). These angles were obtained by computing the singular value decomposition (SVD) of the matrix, M, constructed from the product of the two orthonormal bases $\mathrm{Q_F}$ and $\mathrm{Q_G}$, where $\mathrm{M} = \mathrm{Q_F^T Q_G}$ and $(\mathrm{Q_X})_{\mathrm{ij}} = \mathrm{x_j^i}$ where $\mathrm{x_j^i}$ is the j-th component of the i-th normalized eigenvector defining an orthogonal direction in subspace X. Following the process of $\mathrm{SVD}\left(\mathrm{Q_F^T Q_G}\right)$ produces 20 singular values $\{\sigma_\mathrm{k}\}$ where the k-th principal angle is given by $\theta_\mathrm{k} = \arccos(\sigma_\mathrm{k})$. PA values within the small angle approximation (< 23°) are considered excellent for similarity, while values near 90° indicate orthogonality or complete dissimilarity. A value of 45° corresponds to about a 71% correlation. For equidimensional subspaces, the largest PA is related to the geometric notion of distance, where the "gap" between the subspaces is:

$$gap(F,G) = \sin(\theta_k) = \sqrt{1-\cos^2(\theta_k)} \qquad \text{Eq. 4.4}$$

In our analysis, the first principal angle θ_1 provides the most stringent measure of subspace similarity as it indicates how well the two spaces can be aligned. The value of k for which the principal angles $\{\theta_k\}$ surpass the small angle approximation informs as to how many principal axes the subspaces share with a high correlation. Monitoring the rapid increase in PA provides an ideal way to quantify the most relevant size of subspaces when intra consistency is compared.

4.2.8 Datasets

Datasets were constructed for each of the four proteins investigated in this paper:

PDB ID: 1A6N (58) deoxy-myoglobin: SCOP class α, 151 residues

PDB ID: 1WIT (59) twitchin immunoglobulin: SCOP class β, 93 residues

PDB ID: 1UBQ (60) ubiquitin: SCOP class α+β, 76 residues

PDB ID: 1YPI (61) triosephosphate isomerase: SCOP class α/β, 247 residues

Each dataset contained the following:

1. One MD simulation trajectory obtained using explicit solvent at 298 K for at least 31 ns consisting of 2,000 structures.
2. 31 FRODA trajectories each consisting of 2,000 sample structures each, derived from simulation runs using a H-bond Ecutoff range of 0.0 to -10 kcal/mol and a HP tether cutoff range of 0.0 to 0.5 Å
3. PCA modes from each of the 31 FRODA trajectories.
4. One set of PCA modes derived from the combination of eight individual FRODA runs using a H-bond Ecutoff range of 0.0 to -5 kcal/mol and a HP tether cutoff of 0.5 Å. This set is referred to in the analysis as FRODA-8.
5. One set of PCA modes derived from the combination of twenty individual FRODA runs using a H-bond Ecutoff range of 0.0 to -5 kcal/mol and a HP tether cutoff range of 0.0 to 0.5 Å. This set is referred to in the analysis as FRODA-20.
6. Twenty-one sets of normal modes derived from ANM analysis on the original structure and twenty FRODA-generated structures.

7. Additionally, the 1A6N dataset contained a group of 95 structures of myoglobins with sequence identity > 98.7% and RMSDα < 1Å to 1A6N.

4.3 Results

FIRST uses a set of parameters that determine how constraints are identified, which is ultimately responsible for outcomes in determining the number of iDOF and the predicted rigid and flexible regions of a protein. Based on the RCD, a geometric simulation using FRODA is very efficient. The advantage of FIRST/FRODA is that the generation of output structures is by some comparisons four orders of magnitude faster than MD. However, this tremendous gain in speed comes at the price of model-dependent limitations. Only intra-molecular interactions are modeled (no solvent molecules are considered), and the set of distance constraints is chosen before the geometrical simulation begins. The geometrical simulation is an athermal simulation, where atoms are randomly moved without creating any atomic clashes while the RCD remains fixed for the entire simulation. In such a scenario, a substitute for temperature, or pseudo temperature, is based on the energy cutoff used for selecting H-bonds (62,63). Conversely, the identified rigid and flexible regions can fluctuate between frames within a MD simulation.

FRODA produces datasets composed of multiple structures that capture conformational changes and latent cooperativity in the high dimensional configuration space of the protein. In order to identify those conformational changes and visualize the latent cooperativity, a reduction of dimension is performed by the application of principal component analysis (PCA) (64-66) to the atomic fluctuations of the alpha carbons in the protein. The application of PCA to MD trajectories has a long history and the computation of the essential dynamics of a protein is now well accepted. PCA transforms a set of correlated variables in the original space to a new set of variables that are uncorrelated, similar to normal modes. Furthermore, the original data may be projected onto a small set of principal components that retain a large fraction of the original information, even though the data is represented in a low dimensional subspace. The reduction in dimension can be tremendous, moving the data from a space of tens or hundreds of thousands of variables to one that typically contains less than twenty.

4.3.1 Conformation and Residue RMSD for MD and FRODA

Both MD and FRODA generate trajectories that sample the native basin of a protein when provided a structure. As illustrated in Figure 4.1A for myoglobin, the conformational rmsd for all four proteins investigated indicate good equilibration in exploring the native state conformations in both methods. The MD run was performed at 298K, while different H-bond Ecutoff values and HP parameters were used for FRODA. As more H-bond constraints are removed in the geometrical simulation (FRODA), qualitatively similar results are obtained with progressively larger rmsd. The comparisons given in Figure 4.1A show the correspondence between MD and FRODA. The amount of mobility that the residues experienced in the MD simulations is bounded by the FRODA trajectories using H-bond Ecutoffs between -1 and -3 kcal/mol. In between this range, there exist a H-bond Ecutoff that yields results with high similarity between the MD and FRODA runs, where the residue rmsd is qualitatively consistent (Figure 4.1B) and robust (virtually identical) distributions in residue rmsd (Figure 4.1C) are generated in both cases. It is important to note that even for the most similar case, there is not a one to one correspondence between the two methods because MD allows interactions (both native and non-native) to fluctuate while FRODA keeps the number and identity of all native interactions initially modeled as distance constraints fixed (or constant) throughout a simulation run. The similarity and differences between the residue rmsd generated by MD and FRODA is shown in Figure 4.2 as cartoon representations at the special H-bond Ecutoff that yields maximum correspondence. It is evident that the overall similarity is outstanding, although some key differences within loop regions and helices are frequently detected. These differences are not surprising, as at no time during a FRODA simulation is solvent taken into account, and there is no ability to form non-native contacts.

The level of conformational rmsd can be varied by changing temperature in MD or the H-bond Ecutoff in FRODA. As demonstrated in Figure 4.1, as the H-bond Ecutoff is lowered, the effective temperature is raised and the physicality (or lack thereof) of the simulation for biological conditions must be considered. For example, using a H-bond Ecutoff of 0.0 kcal/mol predicts a globally rigid protein that is severely over-constrained while a H-bond Ecutoff lower than -5 kcal/mol erroneously predicts the protein to be extremely flexible characteristic of an unfolded state. This situation suggests that when using FRODA, there exists a range of physicality (ROP) for which the conformational ensemble that is generated can be considered valid. Although the precise range of the H-bond Ecutoff may vary between proteins, values between -1 kcal/mol and -3 kcal/mol provide a safe ROP, which was demonstrated by each of the 4 proteins studied here. While both H-bonds and HP constraints reduce the number of iDOF, the H-bond constraints are more plentiful. We find that acceptable values for the HP parameters are broad, and the FRODA default values work well in all cases. We include the results of a variety of HP parameter variation for completeness of our analysis. It is worth noting that the removal of all HP tethers is non-physical and yields overly flexible structures. Presumably there is a minimum number of tethers, but do to the insensitivity of the range, we focused on the H-bond criteria, which is the usual way to control the degree of flexibility in the structure. These observations were shared across all proteins studied. No differences in dynamics due to tether parameter adjustments were detected because of differences in beta sheets or alpha helices.

Crosslinking patterns in the H-bonds and the spatial distribution of both HP tethers and HP constraints are dependent on secondary structure. Nevertheless, the FRODA parameters that support a ROP are found to be insensitive (if not independent) to local secondary structural motifs in all the proteins studied here. Since the HP interactions are more often modeled as distance inequalities, it is expected there will be less sensitivity in HP parameters to secondary structures than the H-bonds. As described below, the existence of this ROP was verified by analyzing the generated mode spaces using a wide range of H-bond Ecutoffs and HP parameter variation that control the assignment of constraints/tethers.

One of the most interesting results from this work is that PCA modes generated from a large range of FRODA runs (using different user-defined parameter settings) are consistent in spanning the same subspace describing low frequency and large-scale conformational changes of the protein. It seems counter-intuitive that simulations with very large differences in rmsd could yield principal motions that are quantitatively similar. That is, once the outliers were identified (H-bond Ecutoffs that are greater than -0.50 kcal/mol or less than -5 kcal/mol) all FRODA runs produced robust results using PCA. Apparently, reducing the number of native interactions modeled as distance constraints in FRODA allows exploration of conformations with larger amplitudes of atomic displacements, but the essential dynamics (described by the eigenvectors or modes, not the eigenvalues) remain markedly consistent over a large range in flexibility/rigidity. Therefore, multiple H-bond Ecutoffs within the ROP can be determined by plots like Figure 4.1 to identify the FRODA runs that produce a mean value in conformational rmsd in the range of 1.25 to 2.75 Angstroms. On such a plot, a non-physical over-constrained protein shows a “flat line” in which expected backbone motions in loop regions are absent, while the non-physical under-constrained proteins show unfolding. These rmsd-based criteria for a range of physicality for FRODA simulations are supported by our subspace analysis of those simulations.

4.3.2 PCA for MD and FRODA

Another quantitative comparison between MD and FRODA using PCA is given in Figure 4.3A. The trace of the covariance matrix, or sum of the eigenvalues of all PCA modes, quantifies the total mobility of the protein explored by a simulation. Sorted from largest to smallest, the scree plot shows that only a relatively small number of PCA modes capture most of the mobility. Clearly, increasing the number of iDOF by lowering the H-bond Ecutoff in FRODA leads to a dramatic increase in the mobility ascribed to each PCA mode as shown in Figure 4.3A. When comparing the MD run to the most similar FRODA run (using an Ecutoff of -1 kcal/mol) in terms of raw variance of atomic positions, it is seen that the fall of eigenvalues from the MD run is similar to the FRODA run over the first ten modes. However, as the number of modes increase beyond 10, the eigenvalues of the MD simulation rise in comparison to this FRODA run, approaching the FRODA run performed using an Ecutoff of -2 kcal/mol. This data suggests that the scree plot does not fall off as fast when the protein is intrinsically more flexible, because there will be more collective modes that get mixed up in describing the protein motion. When FRODA adds constraints, many of the motions get frozen out, and a greater distinction in overall amplitudes (variance) occurs. The top two PCA modes by residue shown in Figures. 4.3B-C show similarities and differences in the modeling paradigms. While there are qualitative similarities, a number of regions exist where individual residue motion is differentially assigned. This comparison identifies the regions of the protein that each model addresses in distinct ways, where the key differences arise due to the context of the particular model assumptions.

Once a scree plot is obtained, it is desired that the mode eigenvalues drop off rapidly allowing most of the motions to be reconstructed using the reduced dimensional space. Depending on FRODA parameters, the MD eigenvalues can drop-off faster or slower than the FRODA eigenvalues. In some cases, when the scree plot falls off fast, it may be inferred that the lowest few modes are sufficient to describe biologically relevant information. However, as is the case shown, it often happens that the scree plot falls rapidly but without a distinct kink (as the name would imply). In these cases, one must make a selection based on the relative ratio of the smallest eigenvalue used compared to the largest eigenvalue. A ratio of 0.1 may be large, but may suffice. In general, it is not possible to set a fixed number such as 85% for the cumulative variance. The reason is that some proteins have most of their large-scale motions contained in just a few modes, and to reach some arbitrary pre-determined value will result in including many modes that are associated with motions that are not large-scale. In other words, looking at the ratio of how fast the eigenvalues drop off using a scree plot is always the best way to determine the dimensionality to use. In this work, we selected 20 dimensions in all cases to facilitate the comparative analysis across FRODA runs, MD simulation and ANM. In some cases, it would be better to have used more than 20 dimensions based on the scree plot. Nevertheless, the PA analysis shows that 20 dimensions still captures the majority of the overlap between runs that require less than 20 dimensions and those that may be better to use more than 20 dimensions. The disadvantage of using a higher dimension is that the benefit in reduction of dimensionality is not as great. However, in practice, this would indicate that the protein motions are distributed over more collective modes, and one cannot force the outcome to be low dimensionality.

4.3.3 Range of Physicality for FRODA Simulations

When FRODA is employed, a “range of physicality” (ROP) must be established by adjusting the selection rules that determine which interactions are modeled as distance constraints to obtain quantitatively reasonable conformational and/or residue rmsd as Figure 4.1 shows. The ROP can be further quantified by the number of independent degrees of freedom per residue (iDOF/res), as determined by FIRST. In other related work, a range of [0.5, 1.2] for iDOF/res is considered appropriate for globular proteins under biological conditions (64-66). Our FRODA analysis indicates that this range falls safely within a parameter set that generates a robust subspace analysis (explained below). The onset of atypical dynamical behavior based on quantitative comparisons over ranges of FRODA parameters and MD results, and visually identified as having unphysical characteristics as being too rigid or too flexible) is only apparent when using a H-bond Ecutoff greater than -0.5 kcal/mol or less than -5 kcal/mol. Thus, quantitatively, it appears that even a wide disparity in the assignment of constraints, which yields very different flexibility/rigidity profiles, behave in a coherent and consistent manner in terms of large-scale, low frequency motions as determined by PCA mode extraction. Since coherent and consistent results are reliably obtained by using H-bond Ecutoffs between [-0.5, -5] kcal/mol, a recipe for best practices involves performing a number of runs within this range of H-bond Ecutoffs using FRODA default settings for hydrophobic interactions, and then combining the trajectories to improve the statistical sampling of native conformations. It is important to note that the ROP does not have absolute hard boundaries for all proteins. The ROP will in general slide somewhat depending on protein, and resolution of the input structure. The main point that we found surprising is the ROP is very broad. Moreover, the PCA analysis of a series of FRODA runs provides a protocol that is easy to use to determine the ROP.

4.3.4 Projection of the Displacement Vectors on Model Modes

An important concern is how well dynamical simulation methods such as MD sample the configuration space of a protein. One approach to address this question is to project the displacement vectors obtained from a dynamical simulation onto its principal modes (3). This method makes no assumptions about the underlying distribution and allows one to explore how well the actual simulation events project on the top modes. Moreover, if both the eigenvectors and displacement vectors are normalized, then the projections will all be in the range [-1, 1] due to the normalized inner product (NIP), allowing for intuitive and consistent comparison.

Figure 4.4 presents the results for the projection of the displacement vectors on the FRODA-8 PCA modes. All the displacement vector projections from the FRODA-8 displacement vectors cover the combined FRODA-8 mode space much better than the

MD displacement vectors. For the case of 0 H-bond Ecutoff, an ellipse emerges near the origin, which shows a very different signature than any of the other runs comprising the FRODA-8 group. Interestingly, this over constrained FRODA run produces a quasi simple harmonic motion, as revealed by the hyperdimensional ellipsoidal plot. The MD run shows a much more confined clustering on the projection plot for FRODA-8 PCA modes 1 and 2 within the 95% confidence ellipse, but it does not coincide with most of the FRODA generated displacements. As a result, MD is probing a different type of motion, in a similar way that the highly over constrained FRODA run demarked by the ellipse near the origin is atypical. We can infer that in mode 1, MD is probing much less conformational diversity than FRODA. Nonetheless, beyond mode 1, there is an increasing degree of overlap in the conformatioinal space defined by high PCA modes where the atomic displacement amplitude is rapidly decreasing.

Comparisons within two-dimenisonal projections depend on the two modes used to define a plane. Therefore, to get a better picture of the similarities and differences between the three models, in Figure 4.5 we plot the displacement projections of the MD and FRODA simulations onto PCA modes obtained from MD. The distribution of the MD displacement vectors is tri-modal in the projection on modes 1 and 2, where FRODA and MD coincide within only one of the three “lobes”. This multimodality suggests that the MD simulation spends much of the run time in a few basins, some of which are not being sampled by FRODA. The reason for this is most likely that during the MD simulation, two slight rearrangements of the residues occurred. Because MD allows native contacts to break and reform, the result of these fluctuations is that the MD run is sampling beyond the native basin defined by the input structure. The evidence for this is that there were significant changes in the number of native and non-native contacts along the MD trajectory. These variations in the number and type of contacts support the claim that there was a structural rearrangement. These alterations result in the “lobes” seen in the plots. Nevertheless, using the MD modes as a metric, the FRODA runs intersect with a small portion of the MD displacement vectors. Interestingly, the rapid containment that was seen in Figure 4.4 for projections onto the combined FRODA modes is not seen in Figure 4.5, showing there is still clear cluster separation in modes 4 and 5.

To complete the picture of displacement projections, we also consider projecting onto ANM modes. The ANM modes serve as a metric by which the range of dynamical motion can be effectively measured and compared between the two dynamical models. As is evident in Figure 4.6, the FRODA displacement vector projections cover much more of the mode space defined by the top 20 ANM modes than do the MD displacement vector projections. Once again, in mode 1, it can be seen that FRODA and MD are sampling somewhat different dynamics, but the MD projections are nearly completely contained within the FRODA runs as early as mode 1. Based on the much greater coverage, FRODA appears to be probing the same dynamics that is captured using ANM. FRODA produced trajectories that covered much more conformational space than the MD simulation. We have compared the amplitudes of conformational dynamics by projecting the model displacement vectors on the modes. The FRODA subspace is larger in that it accomodates more of the dynamical displacements generated by MD. The distinction in the projections highlights the fact that the principal motions captured in the different models have some differences.

4.3.5 Comparison of Model Mode Spaces

The question of how to compare low dimensional subspaces that are derived from PCA or ANM has been answered in a number of ways. That the overlap between two vectors can be determined by the inner product of those two vectors is generally well known, however, the overlap between subspaces of high dimensional vector spaces is less intuitive. One approach to measuring the co-incidence of two subspaces is to assess how well each vector in one subspace overlaps all the vectors in the other subspace. This cumulative overlap (CO) method quantifies how well all of a given subspace captures a given vector. Another approach is to determine an average of the inner products of all the vectors in both subspaces. Such a method provides insight into how well each vector in one subspace aligns globally with the vectors in the other subspace, and it is called the root mean square inner product (RMSIP). A more detailed assessment of the interpenetration of two high dimensional subspaces can be made by measuring principal angles and the corresponding principal vectors that describe how one subspace can be rotated/scaled for optimal alignment with the other.

In the comparison of two subspaces, we start with a set of normalized eigenvectors (either from the covariance matrix or from the Hessian matrix) that define subspaces embedded within the high dimensional space equal to 3 times the number of residues in the protein (only alpha carbons were used in the covariance matrix). The process of finding principal angles involves computing the singular value decomposition (SVD) of a matrix of overlaps. The SVD factorization of the matrix of overlaps (inner products) yields a matrix of vectors (left principal vectors) that describe a high dimensional rotation, a diagonal matrix of singular values that describe a scaling, and another matrix of vectors (right principal vectors) that describe another rotation. In this work, the SVD process is applied to a 20 by 20 square matrix so that the right and left vectors describe the same rotation. The singular values are the cosines of the Principal angles and are ordered from largest to smallest. The whole process is an optimization that determines the best possible alignment between the two subspaces.

The interpretation of a PA for two 2-dimensional subspaces within a 3-dimensional space is straightforward, because two planes with different orientations that pass through the origin always coincide or intersect along a single line. In this latter case, the first PA is zero and the second is the acute angle between the two planes. For 2-dimensional subspaces within a 4-dimensional space the situation is more complicated because the two planes may intersect only at a single point (the origin), yielding two non-zero PA. Although geometric visualization fails for 20-dimensional subspaces within a 600-dimensional space, the notion of an angle between two axes remains comprehensible. While each individual PA ranges from 0 to 90 degrees, it is often useful to compute a single value for the angle between the two subspaces, similar to the RMSIP value. The geodesic distance between the two subspaces can be determined by calculating the Euclidean norm of the vector of principal angles. This means that the largest angle between two M-dimensional subspaces derived from a N-dimensional space is not ninety degrees, but rather $\sqrt{M} \cdot 90^{\circ}$ for $M < N$ and $N > 3$, making the maximum possible angle between two 20-dimensional subspaces for all the proteins considered here equal to 402.5°. The largest PA may be interpreted as the “gap” between the subspaces, so determining for which mode the angle between the two spaces leaves the small angle approximation (less than 23 degrees) indicates the number of modes that can be considered similar.

The SVD process that generates the mapping to align two subspaces always orders the set of principal angles from smallest to largest. When a PA is small in the sense that the sine of the angle is approximately equal the angle when measured in radians, this indicates that there is a high degree of overlap of the two subspaces relative to a particular rotation axis. The individual value of a PA is viewed as small or large based on the value of PA itself, independent of the dimension of the spaces being compared. In higher dimensions there are more principal angles, and typically the greatest PA will be close to 90 degrees indicating the part of the subspaces that are orthogonal. Even without using the principal vectors derived from the SVD process, the set of principal angles alone gives a more critical assessment of how much space is in common between the two subspaces. Unlike RMSIP, which tends to increase with the size of the subspaces, the ordered list of principal angles quantifies where the increases comes from. For example, if comparing a 20 dimensional space, if the first 12 principal angles are small, and the last 8 principal angles grow rapidly, we know that the most congruent part of the two subspaces actually lives in 12 dimensions, which cannot be obtained from the RMSIP measure. Since multiple spectra of principal angles can lead to the same RMSIP, the former is a more powerful method for analyzing the similarity of subspaces. It is worth noting that the average of the cosines over all the principal angles gives a qualitatively similar measure as the RMSIP. In summary, if a single number is desired to discriminate similarity, RMSIP is a good measure, but the method of PA gives a much more critical assessment, and carries with it additional information within the principal vectors that inform on how the subspaces are spatially related.

Significant similarity was seen for all three models when the subspaces were compared using the RMSIP and PA metrics. There are no significant differences related to the SCOP class of the protein. When reviewing the displacement vector projections, we found that there were substantial differences between the methods in the first few modes as indicated by distinct cluster distributions. However, for all four classes of protein, the projection space derived from FRODA was greater than the mode projection derived from MD when the ANM modes were used as a metric, suggesting that the geometrical simulation samples more of the native basin than MD does. Additionally, a clear pattern is seen as the two dimensional mode spaces are defined by higher modes with increasing interpenetration of the projection spaces. This is to be expected, as the first few modes will be rather arbitrary and model specific due to statistical sampling and the nature of PCA to maximize variance in a descending fashion. This, in conjunction with the global measure of the RMSIP, is strong evidence for homogeneity within the two dynamical models.

While it is known that the rigidity of a protein is very sensitive to the H-bond Ecutoff that is used in FIRST (12), we conclusively show here that there is no such high sensitivity for the essential motions derived from the RCDs using a wide range of H-bond Ecutoffs. Each individual run from FRODA was comparatively assessed by the different model mode spaces as shown in Figure 4.7A,B. All the subspaces derived from H-bond Ecutoffs between [-1, -5] kcal/mol are essentially the same as measured by the RMSIP (Figure 4.7A) and first PA (Figure 4.7B) scores. Interestingly, variations in the number of hydrophobic tethers have almost no effect on either the RMSIP or first PA within the specified range of H-bond Ecutoffs. Taken together with the earlier analysis based on conformation and residue rmsd, this is strong evidence for a ROP within the regime of the geometrical simulation. Figure 4.7C shows the intra-consistency within the FRODA and MD trajectories. The MD comparison shows that parts one and two of the trajectory are quite similar for the top eight modes with a RMSIP value of 0.81 while the FRODA run was consistent for the top seventeen modes (PA < 23°) with a RMSIP value of 0.94. This substantial difference may be the result of non-equilibration of the MD trajectory and lends credence to the critics of MD for statistical under-sampling problems. Figure 4.7D shows the intra-consistency results for twenty ANM analyses. These comparisons of FRODA structures to the original pdb show a remarkable amount of similarity ranging from the top 13 to 17 modes with not a single RMSIP value below 0.89. Overall, this result suggests that due to the coarse-grained nature of the ANM, small perturbations of the structure do not substantially alter the normal modes obtained. This result provides additional evidence that the distance constraint perturbations that are used by FRODA remain within the native basin of the input structure.

Resolving more detail within the twenty dimensional subspace defined by the modes of different models is achieved by examining the entire set of 20 PA and the average CO. PA values of less than 23 degrees are considered to be within the small angle approximation and suggest excellent similarity. Figure 4.8 shows the comparison of the model mode subspaces using the metric of PA, with RMSIP values shown parenthetically in the legend. From Figure 4.8A, it is clear that the most similar mode subspaces are those from FRODA-8 and FRODA-20 with an RMSIP score of 0.93 and

PAs less than 23° for the top fifteen modes. The next most similar mode subspaces are from FRODA and ANM. These comparisons yield a RMSIP value near 0.75 and have PAs within the small angle approximation for six modes. The mode subspace comparisons between ANM and MD and MD and FRODA are very similar, each having an RMSIP value between 0.5 and 0.6 and PA values less than 45° for the top five modes. While these angle values are not within the small angle approximation, they do indicate a good amount of similarity, commensurate with the good RMSIP score.

It appears that the ANM and FRODA mode subspaces are best able to capture each other's essential motions. This was especially evident when looking at the top ten modes where the average CO remains greater than 0.70 (not shown). Additionally, ANM to MD as well as MD to FRODA maintain an average CO greater than 0.50 for the top twelve modes (not shown). While not excellent, these values parallel the results seen in Figure 4.8 and indicate a substantial amount of compatibility between the essential motions contained in these subspaces. Similar results were found for the other three SCOP classes of protein with the RMSIP between FRODA and MD within the range of 0.51 to 0.63, and the RMSIP between ANM and FRODA within the range of 0.69 to 0.78 for all four SCOP classes. To put these values in perspective, the RMSIP between the FRODA-8 and FRODA-20 modes spanned the range of 0.64 to 0.93 over the four studied proteins, indicating that the inter-model comparisons were on par with the intra-model comparisons. For all four classes of proteins investigated, the ANM to MD RMSIP values were the lowest, spanning the range of (0.45 to 0.57). This result shows that ANM and MD are the most divergent of the three models investigated here. There are no significant differences between the four proteins based on SCOP classification.

4.3.6 Projection of Experimental Structures on the Model Modes

An important consideration beyond the model-to-model comparisons that have been performed is how well are actual experimental structures accommodated by the three different models. To address this question, a 20-dimensional subspace was derived from the experimental dataset of myoglobins, and compared to those obtained from each model. We are specifically assessing how well the 3 models under investigation were able to access the set of experimental structures given only the initial structure. Figures 4.9A and 4.9B show the projection of the experimental displacement vectors onto the model modes. In both panels, it is evident that the ANM mode space captures the experimental displacements best, with FRODA doing almost as well in terms of both the size of conformational space defined by the top three modes and the significance of the overlaps generated therein. The MD based PCA mode space does significantly worse in terms of the amount of conformational space covered and fails to yield any significant overlap to the experimental displacements. These results are echoed in Figure 4.9C showing the RMSIP values of ANM and FRODA mode subspaces to the experimental mode subspace are about 0.60 while the RMSIP value of the MD mode subspace to the experimental mode subspace is only 0.44, a significantly lower result. Taken together, these results indicate that both ANM and FRODA are able to capture the majority of the displacements seen in the experimental structures, while MD captures significantly less of the displacements.

4.4 Conclusions

The existence for a range of physicality using FRODA has been demonstrated in this work for the first time. For the four proteins studied here: We established that the default settings for hydrophobic tethers (rule H3) combined with a H-bond energy cutoff between -1 kcal/mol to -3 kcal/mol is robust. For much larger proteins, the H-bond energy cutoff range may shift slightly lower because the decrease in surface to volume in larger proteins gives slightly higher density of H-bonds. More importantly, the range of physicality can be established on a case-by-case basis using the protocols developed here. Namely, a comparison of multiple FRODA runs should be made with respect to a common vector space derived from PCA on the combined dataset, and a subspace comparison on each separate FRODA run using RMSIP and PA metrics. In particular, the PA metric provides the most mathematically precise and sensitive measure of vector subspace overlaps, and, therefore, we note that the application of PA has much broader implications in the analysis of molecular dynamics, and other types of statistical data routinely encountered in bioinformatics and other fields.

We investigated three models, each based on very different assumptions concerning how to translate the latent information of a protein structure into essential motions within the native basin. Despite very different assumptions in their approach, all three models share marked consistency in the subspaces that describe the greatest fluctuations. The subspaces derived from ANM using a selection of structures obtained from a dynamical trajectory are robust as measured by RMSIP (> 0.85) and first PA (< 10°). Moreover, the subspaces derived from FRODA are robust across a broad range of H-bond energy cutoffs and hydrophobic tether definitions. MD trajectories are as much consistent to ANM and FRODA results, as it is consistent against itself with respect to using PCA on partial statistics. The subspace defined by an experimental set of mutant structures is well covered by both FRODA and ANM. Being less covered by MD clearly indicates longer simulations are required.

The structural basis for observed motions within single domain proteins or for proteins that do not exhibit large-scale structural rearrangement via conformational states separated by an energy barrier can be understood within the scope of a coarse-grained view of protein dynamics. The input structure imparts the information needed for the construction of a dynamical vision of the protein contained within a small subspace. However, there are differences regarding the resolution of the motions and comparing individual modes. Which model to employ will depend on what one wishes to optimize. MD allows the underlying structure to change and thus can sample beyond the native basin defined exclusively by the input structure. However, for this advantage to be fully realized very long simulation times are required. The time needed to run an all-atom geometric simulation (FRODA) on myoglobin that generates 100,000 output structures is a few hours on a modern desktop computer, compared to several seconds for ANM. On the other hand, FRODA allows one to break free of the required harmonic limitation imposed by ANM, while the choice of runtime parameters is non-critical. We note in closing that this study represents the first employment of the momentum perturbation in the GSM to a set of proteins to assess the geometrical model. Previous work used the FRODA module in diffusion mode.

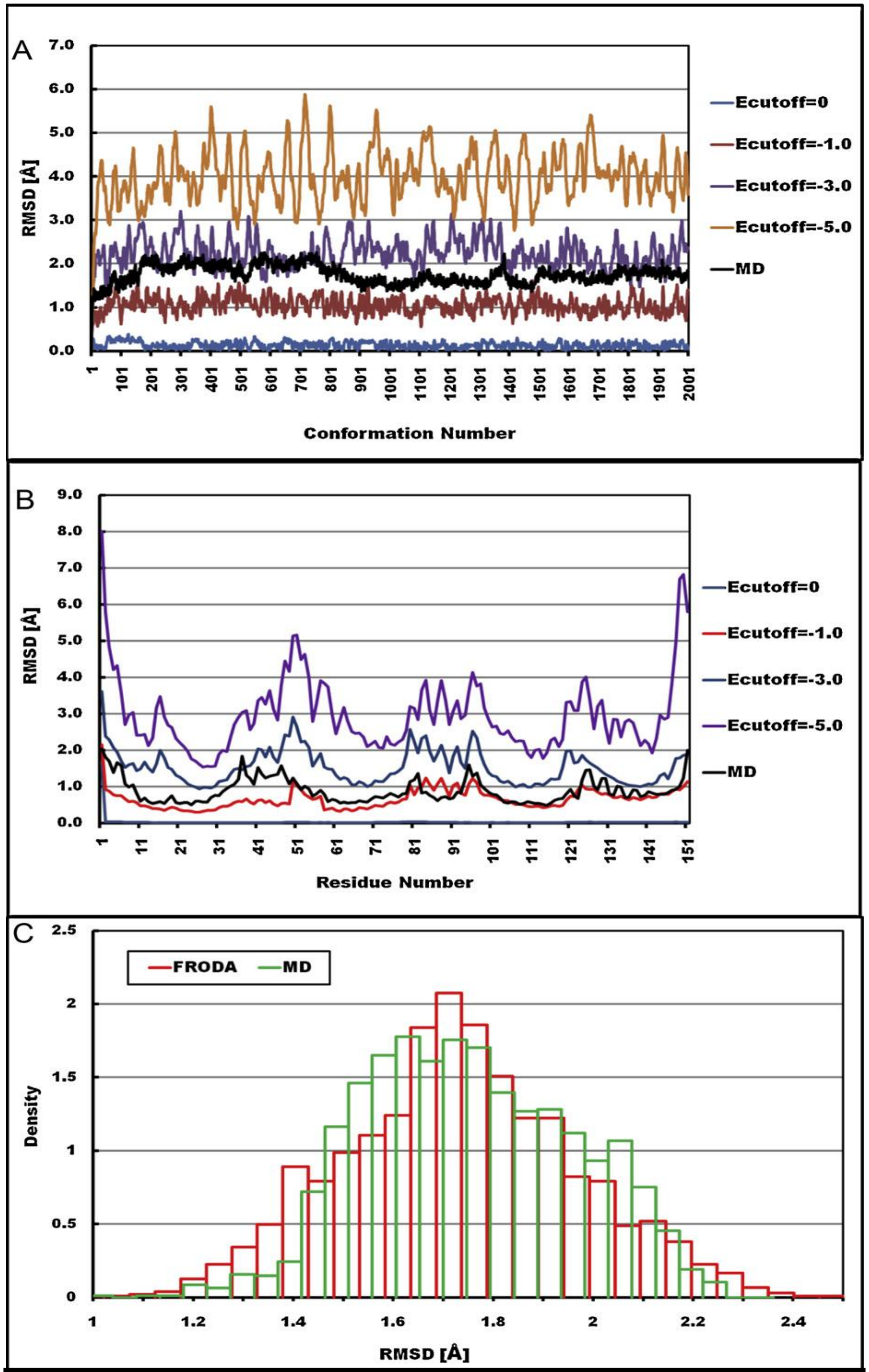


Figure 4.1 Conformational and residue RMSD from MD and FRODA runs using multiple Ecutoffs are compared for myoglobin (pdb code 1A6N). (A, Top) Conformation RMSD. (B, Middle) Residue RMSD. (C, Bottom) The distribution of residue RMSD values across the protein for MD and the most similar FRODA run using an Ecutoff of −2 kcal/mol.

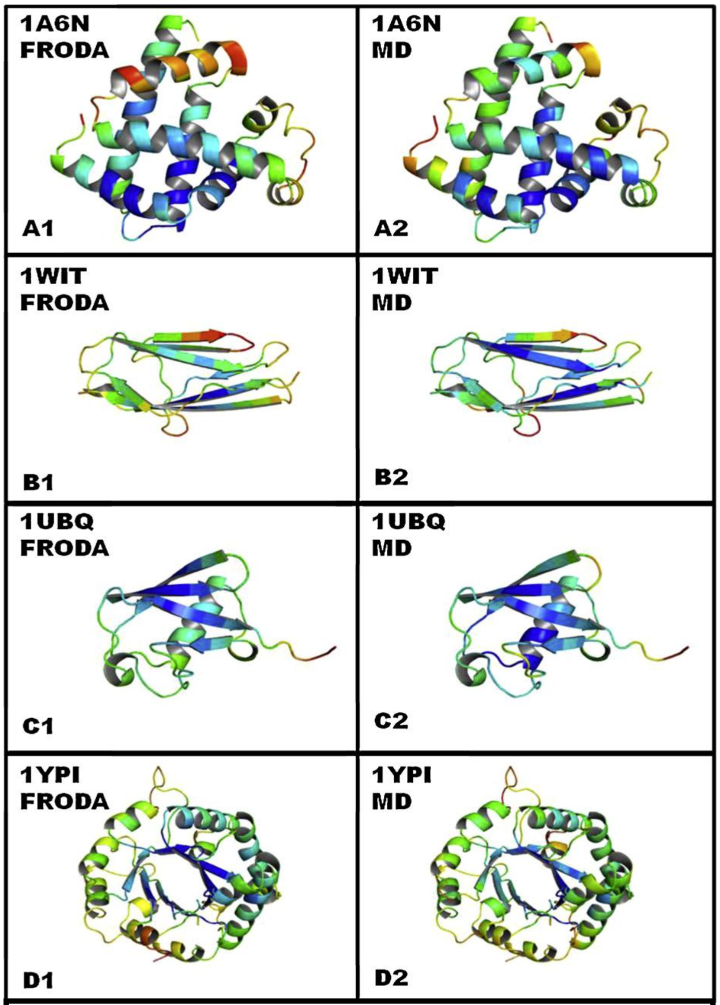


Figure 4.2 Cartoon representations of four proteins colored by residue RMSD using FRODA runs (right panels) with an Ecutoff to produce maximum similarity with MD simulation (left panels). Panels A1 and A2 show myoglobin (pdb code 1A6N). Panels B2 and B3 compare twitchin (pdb code 1WIT). Panels C2 and C3 compare ubiquitin (pdb code 1UBQ). Panels D1 and D2 compare triosephosphate isomerase (pdb code 1YPI). The cartoons are rendered by Pymol (the PyMOL Molecular Graphics System, Version 1.2, Schrödinger, LLC).

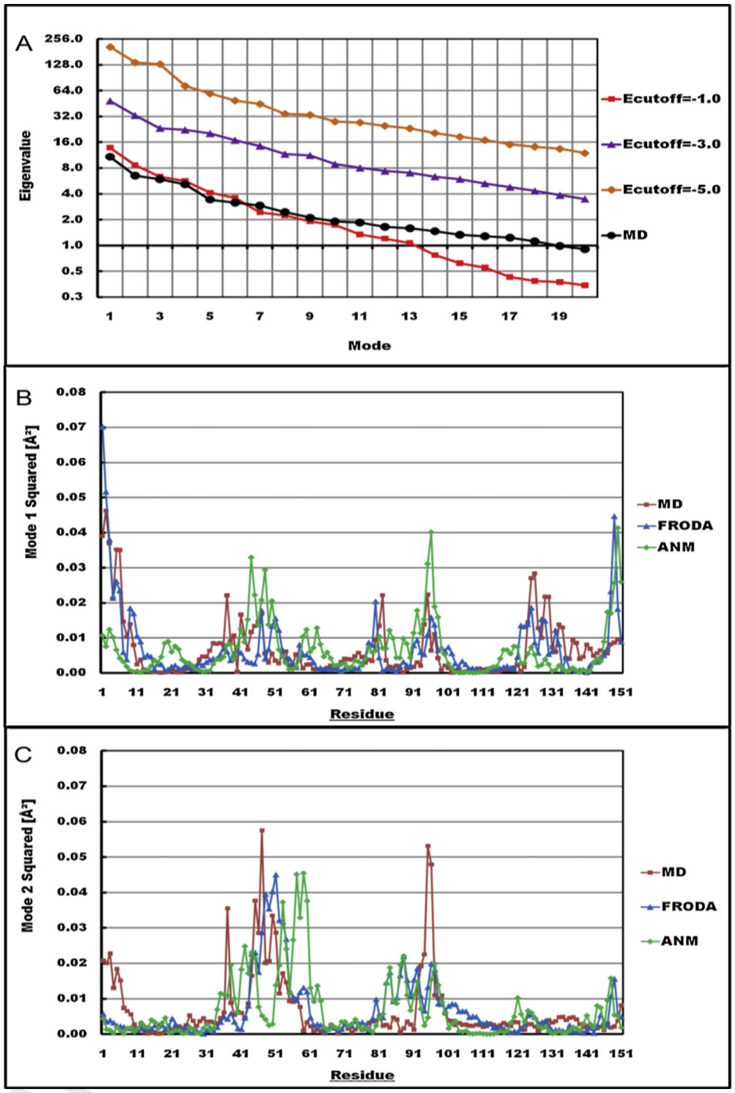


Figure 4.3 Comparison of eigenvalues and the top two modes. (A, Top)On a semi-log scale the rate of decay for the eigenvalues as the PCA modes increase is shown for a selection of FRODA runs and the MD run. (B, Middle) Comparing mode 1 from the PCA of FRODA and MD and from ANM. The FRODA modes are derived from the combination of eight runs using a range of Ecutoffs between 0 and −5 kcal/mol (FRODA-8). (C, Bottom) Comparing mode 2 for the same three models.

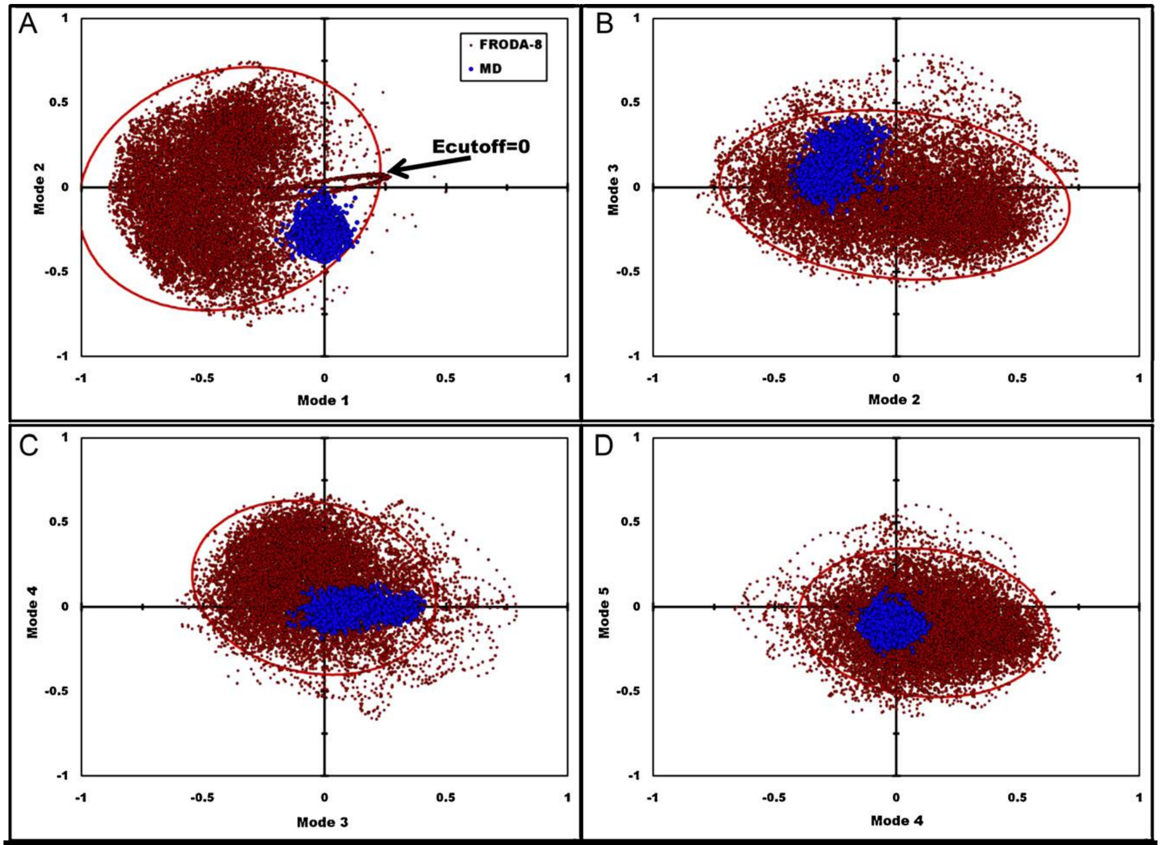


Figure 4.4 The FRODA-8 and MD displacement vectors are projected on the PCA modes derived from the combined FRODA-8 ensemble. Confidence ellipses are drawn for 95%. Projections are made on modes: 1, 2 for A (top left); 2, 3 for B (top right); 3, 4 for C (bottom left), and 4, 5 for D (bottom right). In all cases, the FRODA-8 confidence ellipse contains the MD projections. The ellipse (indicated by the arrow) seen in the top left panel is the result from a FRODA run using an Ecutoff of 0.

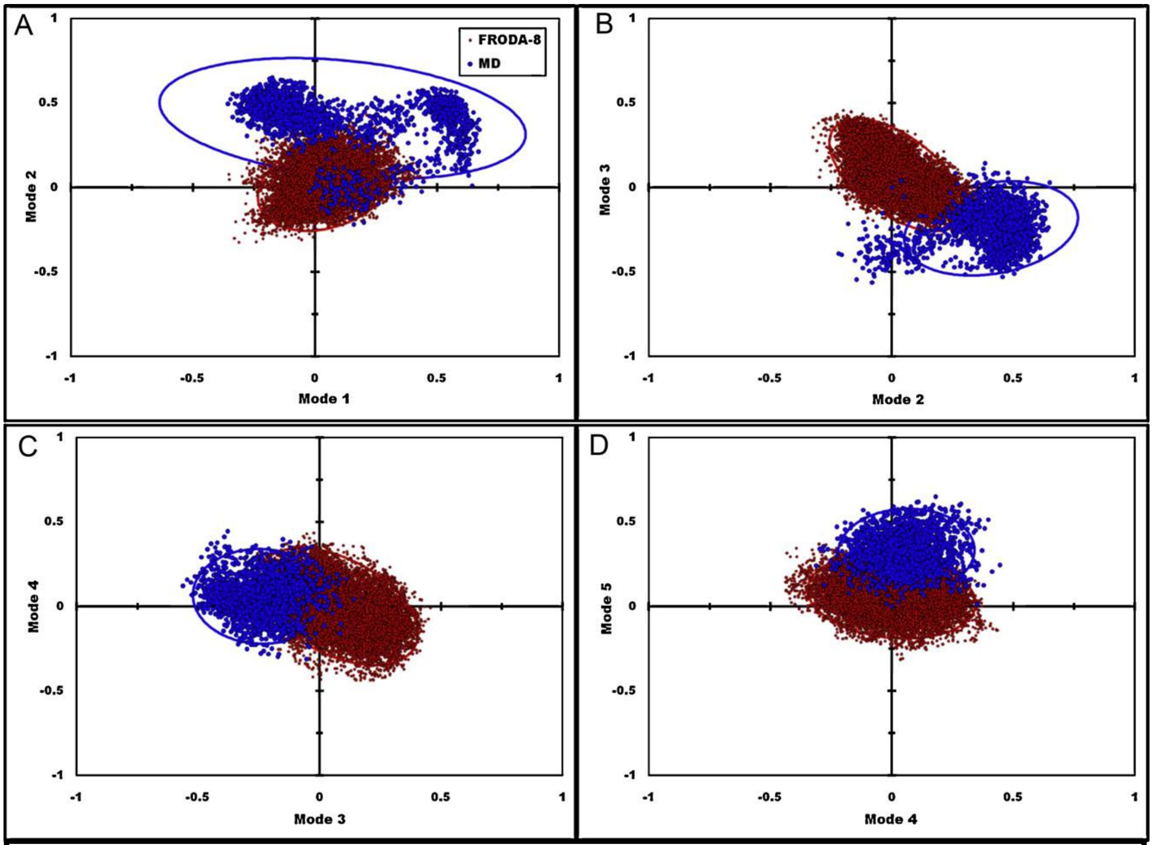


Figure 4.5 The FRODA-8 and MD displacement vectors are projected on the PCA modes derived from the MD ensemble. Confidence ellipses are drawn for 95%. Projections are made on modes: 1, 2 for A (top left); 2, 3 for B (top right); 3, 4 for C (bottom left), and 4, 5 for D (bottom right). Note the tri-modality of the MD series in the projection on modes 1 and 2. In all cases there is a clear separation between the FRODA-8 and MD displacement vectors, with only an interface between the two clusters that exhibits an overlap.

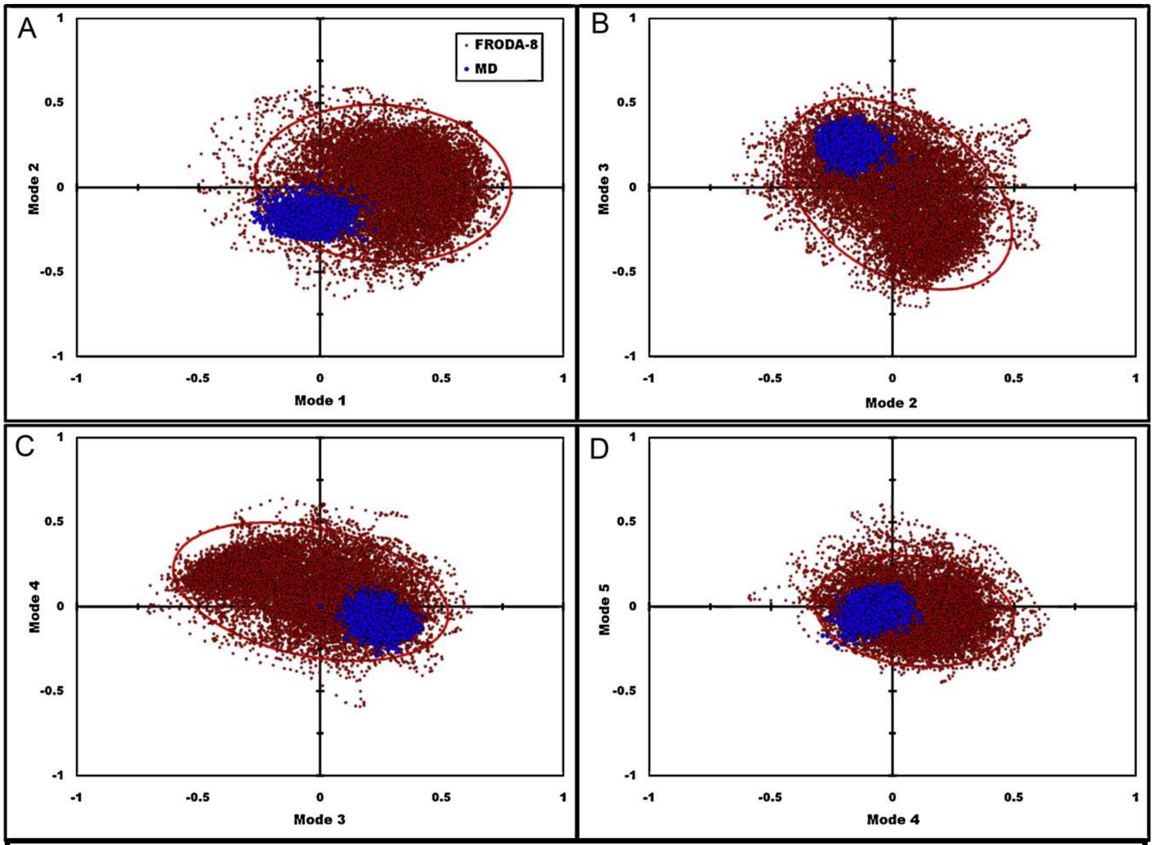


Figure 4.6 The FRODA-8 and MD displacement vectors are projected on the ANM modes. Confidence ellipses are drawn for 95%. Projections are made on modes: 1, 2 for A (top left); 2, 3 for B (top right); 3, 4 for C (bottom left), and 4, 5 for D (bottom right). The FRODA-8 confidence ellipse nearly completely contains the MD space in all cases.

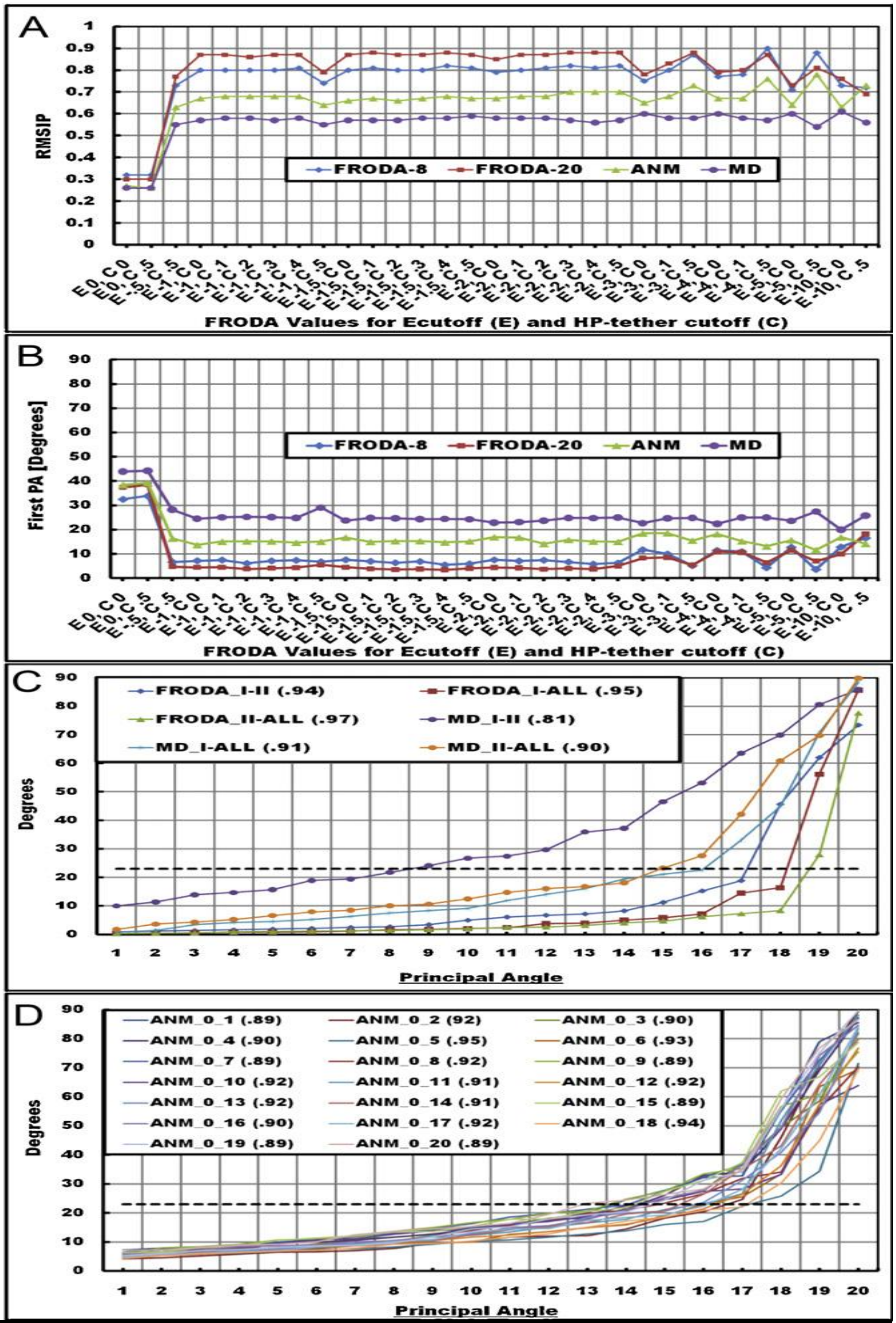


Figure 4.7 Consistency of subspaces describing essential dynamics using RMSIP and PA. (Top A) RMSIP results for individual FRODA runs compared to the four model modes: FRODA-8, FRODA-20, ANM, and MD. (Upper-middle B) PA results for individual FRODA runs to the four model modes; the horizontal axis indicates the run parameters that were used in each case. (Lower-middle C) Inter-consistency is found between FRODA and MD runs using the top 20 PA, with RMSIP values parenthetically shown. (Bottom D) Consistency in results for 20 ANM modes derived from 20 structures produced during a default run of FRODA. In C and D, a level line is drawn for the PA value 23◦.

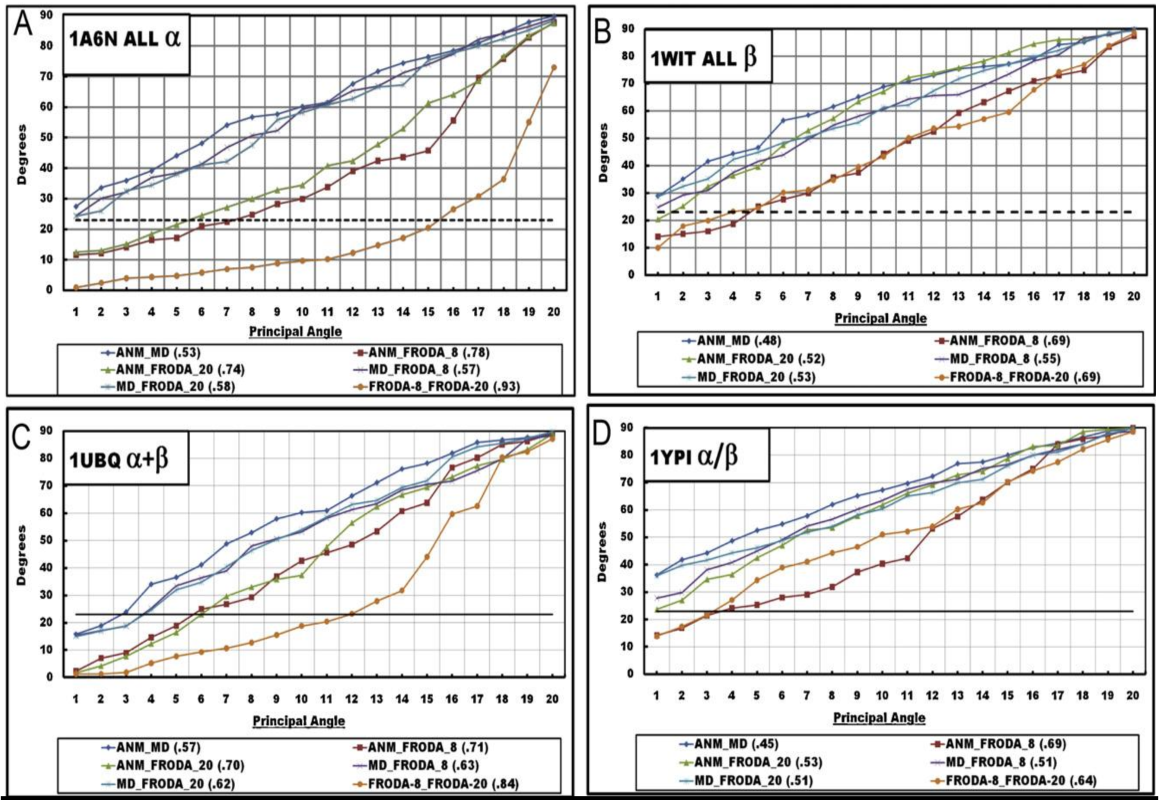


Figure 4.8 Model-to-model subspace comparisons for the 4 proteins investigated. (A)It shows the results for1A6N using the top 20PA, with RMSIP values shown parenthetically in the legend. Figure B, C, and D similarly show the results for 1WIT, 1UBQ, and 1YPI respectively. The level line is drawn for PA value 23◦ in all four panels.

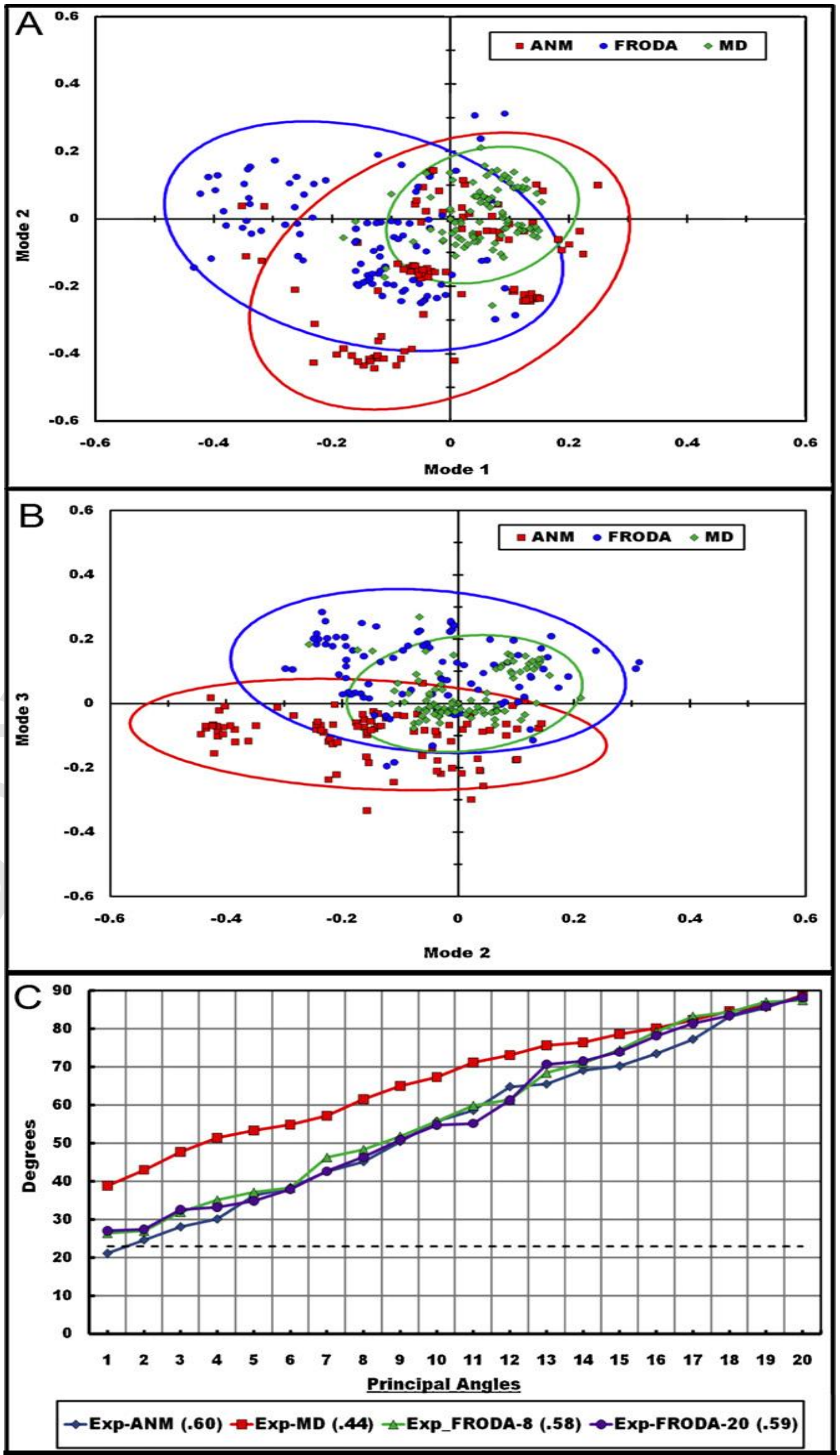


Figure 4.9 Experimental and model subspace comparisons. Projection of the experimental displacement vectors on model modes 1, 2 (Top A) and on modes 2, 3 (middle B). The PA results from comparing the experimentally derived mode space to the model mode spaces, with the RMSIP values shown parenthetically in the legend (bottom C). The level line is drawn for PA value 23◦ in C.

CHAPTER 5: APPLICATION OF THE GSM TO MYOSIN V

5.1 Introduction

We worked with collaborators who are experimentalists studying the kinetics and thermodynamics of myosins using FRET. The placement of the fluorophores allows for the determination of distance between a donor and acceptor probe, which in turn conveys information about the dynamics of the protein under various conditions. The subset of interest was defined by our collaborators and for project I was a set of 106 residues that defined the nucleotide-binding pocket (NBP) of myosin V (MV). For project II, there were three subsets: The actin-binding region (ABR), the NBP, and a proposed communication pathway (CP).

5.2 Project I: Nucleotide Binding Pocket

5.2.1 Introduction to Project I

Myosins are molecular motors capable of converting chemical energy into mechanical work through a cyclic interaction with actin filaments in what is known as the mechanochemical ATPase cycle. There is substantial evidence to support the lever arm hypothesis of force generation in which small conformational changes in the nucleotide-binding region are coupled to a large rotation in the lever arm or light chain binding region (67,68). The swing of the lever arm generates nanometer displacements of actin filaments in muscle contraction and walking of myosin along actin in non-muscle cells. In addition, coupling between the active site and the actin-binding region is critical to allow myosin to cyclically attach and detach from the actin filament. A large cleft in the 50 kDa region separates the actin-binding region into an upper and lower 50 kDa subdomain. Considerable evidence suggests this cleft favors a closed conformation in the high actin affinity states and an open conformation in the low actin affinity states (69-72). However, it is unclear how conformational changes in the nucleotide-binding pocket are communicated to the lever arm and actin-binding cleft. In addition, determining mechanisms of subdomain coupling is critical for understanding how myosin motors adapt their mechanochemical cycle to external loads, a requirement for motor function in a cellular environment.

Several studies have demonstrated the lever arm swing during the mechanical cycle of myosin, including muscle fiber studies (73,74) and single molecule processive walking experiments (75-78). High resolution crystal structures have captured the lever arm in the pre- and post-power stroke conformation while there is still debate as to what step in the ATPase cycle (before, during, or after phosphate release) the power stoke occurs. However, electron microscopy and image reconstruction studies have observed movements of the lever arm in a subset of myosins when comparing the actomyosin nucleotide-free (APO) and ADP states (79-81). The observed structural changes along with other biochemical and biophysical studies lead to the hypothesis that the ADP release step may play a role in strain sensitivity (82). Indeed, a strain sensitive step was originally proposed by Huxley to explain the non-linear force velocity relationship in muscle contraction (83) and later ADP release was found to be the step that limits contractile speed in muscle by White and coworkers (84). The ADP release step is thought to be strain sensitive in that if the lever arm is exposed to negative strain, force in the direction opposing motion, the ADP release rate will be reduced while if it is exposed to positive strain the ADP release rate is enhanced (82). In addition, gating between the two heads of the myosin V dimer is thought to occur during the ADP release step and allows for processive walking in which myosin V can take multiple steps along actin prior to detachment. Single molecule studies have provided direct evidence for strain dependent ADP release in the myosin V dimer providing support for the head-gating hypothesis (85-88). Therefore it is critical to understand the structural changes in the nucleotide binding region associated with ADP binding and release to understand this mechanism.

The current study utilizes a method of FRET between mantADP or IAEDANS actin and FlAsH labeled myosin V (70,89) to examine the conformation of the nucleotide binding region and the upper 50 kDa subdomain during the ADP binding and release steps. Our studies demonstrate that the conformational change measured by FRET correlates well with the rate-limiting step in the actomyosin V ATPase cycle. Our results allow us to propose a model in which strain dependent ADP release is mediated by the conformational change in the nucleotide-binding pocket characterized by our studies.

5.2.2 Methods for Project I

We used FRET to examine the kinetics and thermodynamics of structural changes associated with ADP release in myosin V, which is thought to be a strain sensitive step in many muscle and non-muscle myosins. We also explore essential dynamics using FIRST/FRODA starting with three different myosin V x-ray crystal structures to examine intrinsic flexibility and correlated motions. The input structures employed here are the known x-ray crystal structures (1OE9, 1W7J, 1W7I), which were processed using MOE

(Molecular Operational Environment) software from Chemical Computing Group. All missing residues and atoms where computationally added, with missing loop regions determined by homology modeling, and the energy of each structure minimized.

Samples from a FRODA trajectory are taken to form a set of conformations that represent the native state ensemble. Each conformational ensemble was constructed by selecting every 50th structure from a given simulation containing 100,000 structures, thus yielding 2,000 samples for statistical analysis. PCA was performed using alpha carbon atoms described by Cartesian coordinates. The structures comprising of each trajectory were optimally aligned to remove overall rigid body translation and rotation motions from the intrinsic atomic fluctuations. PCA analysis was performed based on individual trajectories (subject to a specific H-bond energy cutoff) and combined trajectories (using multiple H-bond cutoffs) to test sensitivity in the H-bond energy cutoff.

The results we report here are based on the application of internal coordinates based on distance fluctuations (dPCA), compatible with our focus on residues 171, 294 and 525. A covariance matrix is constructed based on the distances between the carbon alpha atoms of these three residues. This analysis yields 3 PCA modes that characterize the correlated displacements of these three selected residues. We note that no linear transformation was needed to remove overall rotations and translations of the protein because atomic pair distances are invariant under global rigid body motions.

5.2.3 Discussion for Project I

Our steady-state and time resolved FRET analysis demonstrates a temperature dependent reversible conformational change in the nucleotide-binding pocket. Our kinetic results demonstrate that the nucleotide-binding pocket goes from a closed to an open conformation prior to the release of ADP while the actin binding cleft remains closed. Interestingly, we find that the temperature dependence of the maximum actin-activated myosin V ATPase rate is similar to the pocket-opening step, demonstrating this is the rate limiting structural transition in the ATPase cycle. Thermodynamic analysis demonstrates the transition from the open to closed nucleotide binding pocket conformation is unfavorable because of a decrease in entropy. The intrinsic flexibility analysis is consistent with conformational entropy playing a role in this transition, as the MV.ADP structure is highly flexible compared to the MV.APO structure. Our experimental and modeling studies support the conclusion of a novel post-power-stroke actomyosin.ADP state in which the nucleotide binding pocket and actin binding cleft are closed. The novel state may be important for strain sensitivity as the transition from the closed to open nucleotide binding pocket conformation may be altered by lever arm position.

In order to gain insight into the closed NBP conformation of myosin V in the presence of ADP, which is more populated at low temperature, we investigated the intrinsic flexibility of each of the three crystal structures of myosin V. We measured the distance distributions between proline 294, which is near the center of the FlAsH labeling site, and lysine 171, which is near the mant motif based on aligning the structure of myosin V with the structure of myosin II bound to mantADP.BeFX (90). The distance distributions demonstrate that the 294-171 distance in the MV.ADP.BeFx structure does not overlap with the distance distributions of MV.ADP structure (Figure 5.1), which suggests the MV.ADP structure is not capable of converting to the closed pocket conformation without significantly breaking constraints. In addition, in the MV.ADP state structure we observed very few modes of pocket opening/closing that agree with experiment suggesting that this structure represents the weak open pocket MV.ADP state and that the strong closed pocket MV.ADP state has not been revealed by crystallography. Two other studies came to a similar conclusion in regard to the MV.ADP crystal structure (91,92). Our results provide support for a strong MV.ADP conformation that has an ATP-like closed NBP, while it has a closed actin-binding cleft.

5.3 Project II: Three Subsets of Myosin V

5.3.1 Introduction to Project II

Understanding the mechanism of force generation in myosins requires elucidating allosteric communication pathways that are critical for motor function. The well-established lever arm hypothesis suggests that communication between the nucleotide-binding region and light chain binding region or lever arm is critical for force generation (93-95). In addition, cyclic attachment and detachment from actin is thought to be accomplished by nucleotide-dependent conformational changes in the actin binding cleft, which has been shown to be more open in the “weak” actin binding states and more closed in the “strong” actin binding states (96). However, the details of how the active site coordinates communication between the lever arm and actin-binding regions is currently unclear.

The overall architecture of most myosins consists of a well-conserved structural core with minor modifications that lead to different mechano-chemical properties required for tuning each myosin for a specific biological function (97,98). Interestingly, myosins share sequence and structural homology with G-proteins and other NTPases (99), and utilize a similar mechanism of nucleotide binding and hydrolysis. There are three well conserved regions in the P-loop family of NTPases (100) involved in performing this task: the P-loop, switch I, switch II. However, it is unclear if the structural mechanism of communicating conformational changes from the active site to the protein effector-binding site or track is conserved in these structurally related NTPases. Studies have demonstrated that the reversible movement of switch I is coupled to the coordinated opening of the actin binding cleft and closing of the nucleotide binding pocket (101, 102). Conformational changes in switch II are directly coupled to the kink in the relay helix, which allows for positioning of the lever arm and formation of the pre-power stroke state (88,103). Hence, switch I is thought to be chiefly involved in the communication pathways between the active site and the actin binding cleft (104), while switch II mediates the communication to the converter-lever arm domain (103).

Mutation analysis studies of switch I and switch II (105-107) have provided support for this hypothesis. However, several studies have demonstrated that force generation in myosins is intimately associated with the transition from weak to strong actin binding (108-110). Thus, the switch II region is implicated for providing allosteric coupling between the actin binding, active site, and lever arm domains. Since the role of switch II in mediating the lever arm position has been well established, it is prudent to examine its role in opening/closing the nucleotide-binding pocket and in conformational communications between the active site and actin binding regions (111-113).

To address the question of how switch II affects the conformational dynamics of the nucleotide binding pocket and actin binding cleft we introduced two single site mutations in the switch II region of myosin V. We investigated the G440A mutant, which removes a highly conserved hydrogen bond to the gamma phosphate of ATP, and the E442A mutant which removes a highly conserved salt bridge between switch I and switch II (106,113). Both of these mutants inhibit the hydrolysis of ATP (scheme 1) as shown with studies of myosin V (113) as well as Dictylostelium (106,114) and smooth muscle myosin II (115-117). The conformational dynamics of the nucleotide binding pocket were examined by monitoring FRET between dmant [N-methylanthraniloyl (mant)] labeled 2'deoxy-nucleotides (dmantADP or dmantATP) and FlAsH labeled in the upper 50 kDa domain of myosin V (MV FlAsH) (118,119). Conformational changes in the actin-binding cleft were examined by monitoring FRET between IAEDANS [5-((((2-iodoacetyl)amino)ethyl)amino)-naphthalene-1-sulfonic acid] labeled actin and MV FlAsH. Our results establish a role for switch II in mediating the conformation of the nucleotide-binding pocket, which is important for the structural mechanism of ADP release from actomyosin. In addition, we establish a critical role for switch II communicating conformational change between the active site and actin-binding region.

5.3.2 Methods for Project II

The X-ray crystal structures (1OE9, 1W7J, and 1W7I) defining three conformational states were processed using MOE (Molecular Operational Environment) software from Chemical Computing Group. Different from prior work in which we only considered the motor heavy chain, we include both the myosin heavy chain and light chain, as well as the nucleotide in the active site (ATP, ADP, and no-nucleotide), where the ADP-BeF_3 moiety was modeled as ATP. All missing residues and atoms were computationally added in both the myosin heavy chain and light chain, with missing loop regions determined by homology modeling, and the energy of each structure was minimized. Once the wild type structure was prepared, single site mutants for G440A and E442A were created using the mutate residue function in MOE and the energy of each mutant structure was then minimized again. All HMs were refined using a 0.000001 gradient and had no missing atoms.

Multiple FRODA runs were generated using different H-bond energy cutoffs from −0.5 to −3.0 kcal/mol in steps of 0.5 kcal/mol. For the wild type and mutant structures in each conformational state (9 cases in all), an individual dataset consist of 2,000 frames uniformly sampled from a FRODA trajectory of 100,000 steps. These datasets were subsequently used to construct a positional covariance matrix built from the Cα atoms described by Cartesian coordinates. PCA was performed on each individual dataset (for a particular H-bond energy cutoff) and the combined dataset that includes all 6 FRODA runs. After verifying that the PCA modes from individual runs provide consistent results, the PCA mode analysis presented here is based on the combined dataset for maximum statistics. In addition to the entire protein, PCA is applied for regions of interest that include the nucleotide-binding pocket (NBP) consisting of 106 residues (156-244, 429-445), the actin-binding region (ABR) consisting of 455 residues (201-655), and a proposed communication pathway (CP) consisting of 89 residues (393-481).

5.3.3 Discussion for Project II

Our computational modeling results suggest that switch I-switch II interactions (120,121) stabilize the closed nucleotide binding pocket conformation in the absence of actin. Introducing the switch II mutations into the myosin V.ADP-BeFx structure (M.ATP - pdb) only had a minor impact on the overall flexibility as demonstrated by the RMSD plots. We also performed PCA to examine the correlated motions of the nucleotide-binding pocket, which allowed us to determine if the mutations disrupt important changes in active site dynamics. Analysis of the nucleotide-binding pocket in the ATP state reveals reduced dynamics of switch I in the G440A mutant (Figure 5.2A). The reduced dynamics of switch I in the G440A may prevent switch I-switch II interactions and reduce the stability of the closed nucleotide binding pocket conformation, which we observed at higher temperatures in presence of ATP. In addition, the mobility of the P loop, which is directly connected to the HF helix and loop 1, is increased by both switch II mutants. These results suggest mutations in the switch II region can alter the coordinated motions of the HF helix/P-loop, switch I, and switch II which appear to be involved in mediating nucleotide binding and release.

PCA on the actin-binding subset predominantly defined by the upper and the lower 50kDa sub-domains reveals that the G440A mutation alters the dynamics of the actin-binding region. In the rigor state, the G440A mutant increases dynamics of a region of the upper 50 kDa domain including the cardiomyopathy loop and C-terminus of the HO-helix (Figure 5.2B). There is also a dramatic increase in the mobility of helix-loop-helix region of the lower 50kDa domain. Previous work has suggested that F441 may be a key residue that couples the switch II and the lower 50 kDa subdomain of the actin binding region (106). Our simulation results show that F441 is highly dynamic in G440A in the ATP state (Figure 5.3). Disruption of switch II rotation by the G440A mutation alters the mobility of F441, which may prevent its interaction with the surrounding hydrophobic cluster of the lower 50kDa sub-domain. These two structural uncouplings engender disruption of communication pathways between the active site and the upper and lower 50 kDa regions required for ATP-induced dissociation.

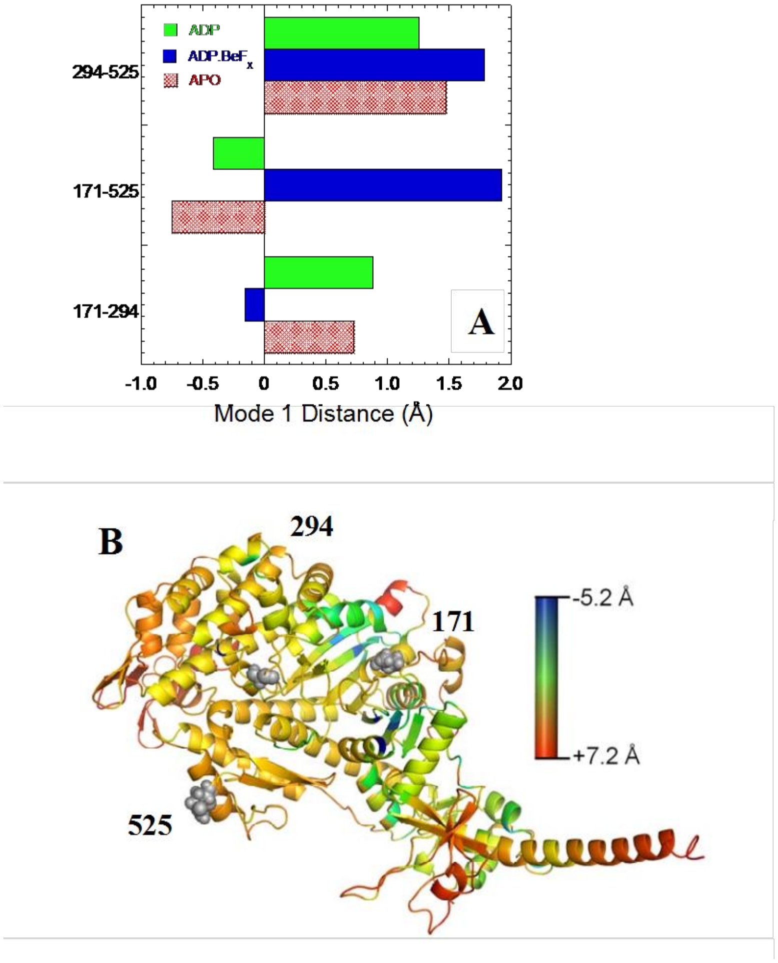


Figure 5.1 Intrinsic flexibility and dynamics of the MV crystal structures. A) We performed PCA analysis using FRODA to examine the relative motions of three residue pairs (294-171, 171-525, and 294-525) in all three crystal structures. The correlated motions were similar in the MV.ADP and MV.APO states although different in magnitude, while they were quite different in magnitude and direction in the MV.ADP.BeFX state. B) The residue root mean square deviation (RMSD) was examined in the MV.APO and MV.ADP crystal structures. The MV.APO residue RMSD was subtracted from the MV.ADP residue RMSD and the relative flexibility change is shown in a ribbon diagram The color scale is shown with red representing the most positive change in RMSD (MV.ADP more flexible than MV.APO) and blue representing the most negative change in RMSD (MV.APO more flexible than MV.ADP).

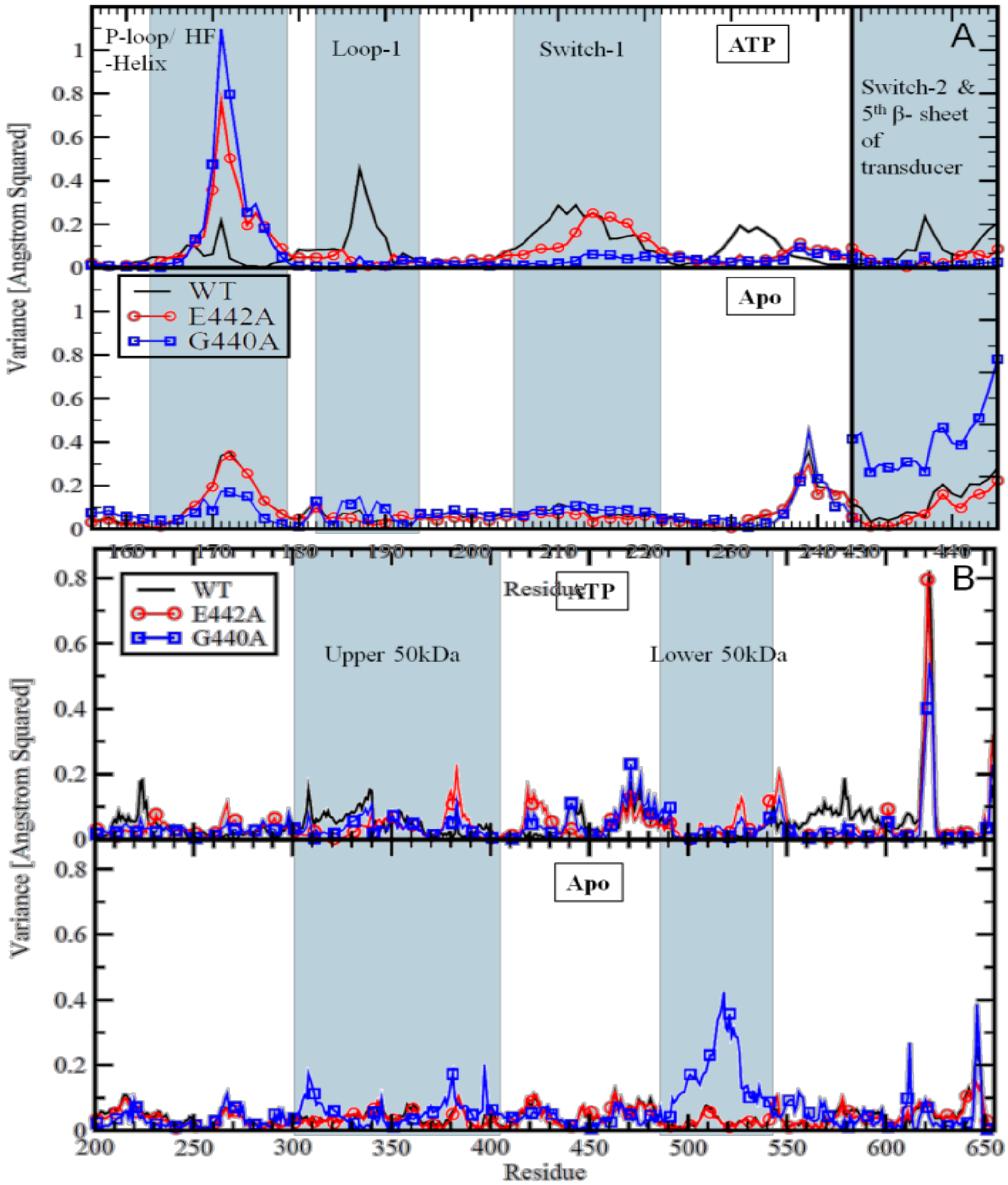


Figure 5.2 The first PC mode of the nucleotide and actin binding subsets of mutant and wild-type myosin V. (A) Analysis of the NBP subset defined by 106 residues (156-244, 429-445) is shown in the nucleotide-free and ATP bound state. (B) The ABR subset defined by 455 residues (201-655) nucleotide-free and ATP bound state. The shaded areas highlight the structural elements discussed in the text, which have significant changes in dynamics within the designated subsets.

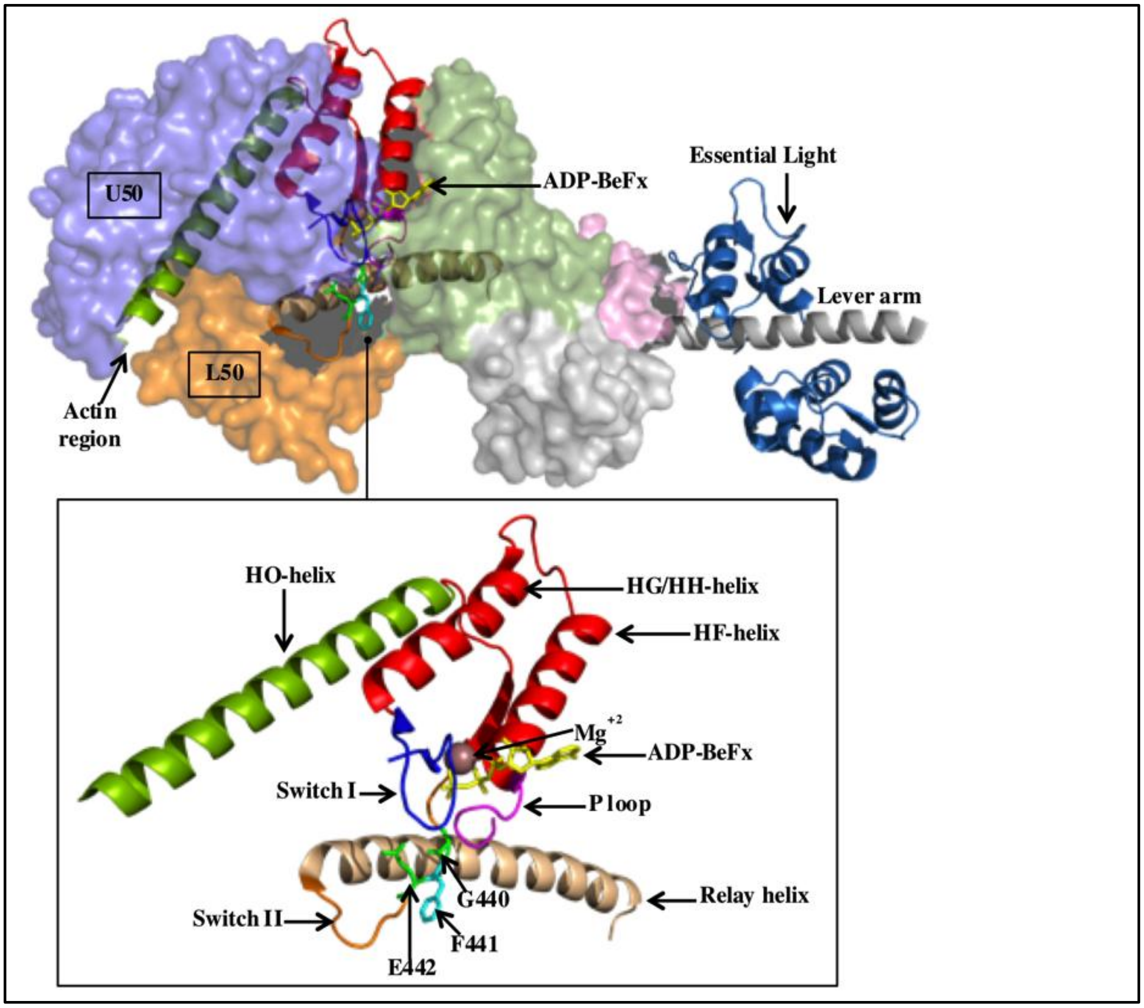


Figure 5.3 Crystal structure of myosin V in the ADP-BeFx state (PDB, 1w7j) showing the Upper 50kDa (U50) and Lower 50kDa (L50) sub-domains along with the actin and nucleotide binding regions (top). Elements involved in coordination of the nucleotide and communication between the actin and nucleotide binding regions are magnified (bottom).

CHAPTER 6: DISCUSSION

6.1 Summary

The work outlined in this dissertation has established the GSM as a viable alternative and/or co-model to be used with, or in place of, either an ENM or MD. The GSM is both efficient and effective at determining the native state dynamics of a protein. The model is also extensible in that it has shown efficiency and effectiveness when applied to very large proteins like MV, which has multiple domains. Specifically, the model not only has a strong intra-consistency, but it is also consistent in large part with other models as well as experiment. The application of ED to trajectories has been shown to be an effective technique to tease out the biological motions of a subset of residues from a protein. The metrics that we have developed clearly show that the model is very effective at elucidating the dynamics of a protein's native state. The question of statistical sampling of biologically relevant motions can be addressed by using subspace analysis, as we have demonstrated here. The GSM with ED and subspace analysis has been applied to a wide range of proteins with success including small, single domain proteins and large multi-domain, multi-state proteins like MV. While the method is able to capture the essence of state transits, the current limitations of the model preclude the elaboration of a biologically sound pathway.

6.2 Significance

We have clearly shown through our analysis that the GSM is a solid model that may be used reliably for the determination of native state dynamics of proteins. Our work indicates that the results obtained from the GSM are qualitatively and quantitatively similar to those from MD and ANM. While the GSM is not entirely new, it has not had a rigorous testing prior to this work. In addition, we have shown that depending on the application, results may be obtained thousands of times faster than from an MD simulation. Within the framework of biophysical simulations, we have established a new method for rapidly assessing the native state dynamics of a protein as well as the dynamics of multi-state proteins that have about 1000 residues. Specifically, our work on MV combining experimental and computational methods has helped to clarify biologically important mechanisms in that protein.

We have implemented a set of mathematical and statistical techniques to quantify the similarity of native state dynamics as captured in a trajectory from either an MD or GSM simulation. While ED has a long history, the application of subspace analysis using both RMSIP and PA is new and has been shown to be very effective for measuring the similarity of protein dynamics. The work we have done here not only extends the GSM paradigm, but also places it on solid ground for other scientists to use as an analysis tool for their research. Finally, our analysis tools can be applied to other proteins and molecules to help push forward the frontiers of biological science.

6.3 Publications

Characterizing Protein Motions from Structure. (2011)
Charles C. David and Donald J. Jacobs,
Journal of Molecular Graphics and Modelling 31, 41–56

Kinetics and Thermodynamics of the Rate Limiting Conformational Change in the myosin V Mechanochemical Cycle. (2011)
Donald J. Jacobs, Darshan Trivedi, Charles C. David, and Christopher M. Yengo
Journal of Molecular Biology, Apr 15;407(5):716-30

Switch II Mutants Reveal Coupling between the Nucleotide- and Actin-Binding Regions in Myosin V. (2012)
Darshan Trivedi, Charles C. David, Donald J. Jacobs, and Christopher M. Yengo
Biophysical Journal Volume 102 Issue 11 pages 2545-2555

Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins.
Charles C. David and Donald J. Jacobs
Submitted as book chapter on Protein Dynamics in Methods of Molecular Biology
Editor: Dennis Livesay (to appear 2013)

Essential Dynamics of Proteins with Subspace Analysis in Java.
Charles C. David and Donald J. Jacobs
In Preparation to be submitted to BMC Bioinformatics

REFERENCES

1. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer Jr., E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977). The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures. J. Mol. Biol., 112, 535.

2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Weissig, TNBH, Shindyalov, I.N., Bourne, P.E. (2000). The protein data bank. Nucl. Acids Res 28, 235–242. [PubMed: 10592235]

3. Balsera, M.A., Wriggers, W., Oono, Y., and Schulten, K. (1996). Principal Component Analysis and Long Time Protein Dynamics. Journal of Physical Chemistry, Vol. 100, pp. 2567-2572.

4. Bahar, I., Atilgan, A.R., Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. Folding Des 2, 173–181.

5. Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J 80, 505–515. [PubMed: 11159421]

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

7. Krebs, W.G., Alexandrov, V., Wilson, C.A., Echols, N., Yu, H., Gerstein, M. (2002). Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins 48, 682–695. [PubMed: 12211036]

8. McCammon, J.A., Gelin, B.R., Karplus, M. (1997). Dynamics of folded proteins. Nature 267, 585–590. [PubMed: 301613]

9. Smith, L., Daura, X., van Gunsteren, W. (2002). Assessing Equilibration and Convergence in Biomolecular Simulations. PROTEINS: Structure, Function, and Genetics 48:487–496.

10. Wells, S.A., Menor, S., Hespenheide, B.M., and Thorpe, M.F. (2005). Constrained geometric simulation of diffusive motion in proteins. Phys. Biol., 2, S127–S136.

11. Farrell, D.W., Kirill, S., Thorpe, M.F. (2010). Generating stereochemically acceptable protein pathways. Proteins, 78, 2908-2921.

12. Jacobs, D.J., Rader, A.J., Kuhn, L.A., and Thorpe, M.F. (2001). Protein flexibility predictions using graph theory. PROTEINS: Struct., Func. and Gen., 44, 150–165.

13. Wells, S. A., Jimenez-Roldan, J. E., and Römer, R. A. (2009). Comparative analysis of rigidity across protein families. Phys. Biol., 6, 046005, doi: 10.1088/1478-3975/6/4/046005.

14. Berendsen, H.J., Hayward S. (2000). Collective protein dynamics in relation to function. Curr Opin Struct Biol 10: 165-169.

15. Amadei, A., Linssen, A.B., de Groot, B.L., van Aalten, D.M., Berendsen, H.J. (1996). An efficient method for sampling the essential subspace of proteins. J. Biomol. Struct. Dyn 13, 615–625. [PubMed: 8906882]

16. Amadei, A., Linssen, A.B., Berendsen, H.J. (1993). Essential dynamics of proteins. Proteins 17, 412–425. [PubMed: 8108382]

17. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science 2, 572.

18. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. J. Educational Psychol 24, 441.

19. Manly, B. (1986). Multivariate statistics - A primer. Boca Raton: Chapman & Hall/CRC

20. Abdi. H., & Williams, L.J. (2010). "Principal component analysis.". Wiley Interdisciplinary Reviews: Computational Statistics, 2: 433–459.

21. Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4

22. Brüschweiler, R. (1995) Collective protein dynamics and nuclear spin relaxation. J. Chem. Phys., 102(8), 3396-3403.

23. Sanejouand, T.F. (2001). Conformational change of proteins arising from normal mode calculations. Protein Eng 14, 1–6. [PubMed: 11287673]

24. Yang, L., Song, G., Carriquiry A., Jernigan R.L. (2008). Close Correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. Structure 16, 321-330.

25. David, C.C., Jacobs, D.J. (2011). Characterizing Protein Motions from Structure. Journal of Molecular Graphics and Modelling 31, 41–56.

26. Van Aalten, D.M.F., De Groot, B.L., Findlay, J.B.C., Berendsen, H.J.C., Amadei, A. (1997). A Comparison of Techniques for Calculating Protein Essential Dynamics. Journal of Computational Chemistry, Vol. 18, No. 2, 169-181.

27. Rueda, M., Chaco ́, P., Orozco, M. (2007). Thorough Validation of Protein Normal Mode Analysis: A Comparative Study with Essential Dynamics. Structure 15, 565–575.

28. Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems. Edited by Qiang Cui, Ivet Bahar. Published December 12th 2005 by Chapman and Hall/CRC – 432 pages

29. Kitao, A., Go, N. (1999). Investigating protein dynamics in collective coordinate space. Current Opinion in Structural Biology, 9:164-l 69.

30. Ma, J. (2005). Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. Structure, Vol. 13, 373–380. DOI 10.1016/j.str.2005.02.002

31. Hayward, S., Kitao, A., Go, N. (1995). Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. Proteins, 23(2):177-86. PMID: 8592699

32. Hayward, S., Kitao, A., Go, N. (1994). Harmonic and anharmonic aspects in the dynamics of BPTI: a normal mode analysis and principal component analysis. Protein Science, 3(6):936-43. PMID: 7520795

33. Scholkopf, B., Smola, A., Muller, K-R. (1999). Kernel Principal component Analysis in B. Scholkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in Kernel Methods–Support Vector Learning, pages 327-352. MIT Press, Cambridge, MA.

34. Sapra, S. (2010). Robust vs. classical principal component analysis in the presence of outliers. Applied Economics Letters, 17, 519–523.

35. Storer, M., Peter, M., Roth, P.M., Urschler, M., and Bischof, H. Fast-Robust PCA. Institute for Computer Graphics and Vision Graz University of Technology Inffeldgasse 16/II, 8010 Graz, Austria

36. Gnanadesikan, R., Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28:81–124.

37. Huber, P. (1981). Robust Statistics. Wiley and Sons

38. De La Torre, F., Black, M. (2003). A framework for robust subspace learning. International Journal on Computer Vision, 54:117–142.

39. Handling of data containing outliers. Wolfram Stacklies and Henning Redestig CAS-MPG Partner Institute for Computational Biology (PICB) Shanghai, P.R. China and Max Planck Institute for Molecular Plant Physiology Potsdam, Germany

40. Joint Outliers and Principal Component Analysis. Georgy Gimel'farb, Alexander Shorin, and Patrice Delmas. Dept. of Computer Science, University of Auckland, P.B. 92019, Auckland, New Zealand.

41. Kriegel, H. P.; Kröger, P.; Schubert, E.; Zimek, A. (2008). A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms. Scientific and Statistical Database Management. Lecture Notes in Computer Science 5069: 418. DOI:10.1007/978-3-540-69497-7_27. ISBN 978-3-540-69476-2.

42. Cattell, R. B. (1966). The Scree Test for the Number of Factors. Multivariate Behavioral Research, 1(2), 245-276.

43. Cattell, R. B. & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. Multivariate Behavioral Research, 12, 289-325.

44. Jacobs, D.J., Trivedi, D., David, C.C., Yengo, C.M. (2011). Kinetics and Thermodynamics of the Rate Limiting Conformational Change in the myosin V Mechanochemical Cycle. Journal of Molecular Biology, Apr 15;407(5):716-30.

45. Trivedi, D., David, C.C., Jacobs, D.J., Yengo, C.M. (2012). Switch II Mutants Reveal Coupling between the Nucleotide- and Actin-Binding Regions in Myosin V. Biophysical Journal Volume 102 Issue 11 pp.2545-2555. doi:10.1016/j.bpj.2012.04.025

46. Amadei, A., Ceruso, M.A., Di Nola, A. (1999). On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. Proteins 36, 419–424. [PubMed: 10450083]

47. Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., Ortiz, A.R. (2005). An analysis of core deformations in protein superfamilies. Biophys. J 2005;88:1291–1299. [PubMed: 15542556]

48. Miao, J. and Ben-Israel, A. (1992). On Principal Angles between Subspaces. Lin. Algeb. and its Appl. 171, 81-98.

49. Gunawan, H., Neswan, O., Setya-Budhi, W. (2005). A Formula for Angles between Subspaces of Inner Product Spaces. Contributions to Algebra and Geometry, Volume 46, No. 2, 311-320.

50. Absil, P.A., Edelman, A., and Koev, P. (2006). On the largest principal angle between random subspaces, Linear Algebra and its Applications, Volume 414, Issue 1, Pages 288-294.

51. Hess, B. (2002). Convergence of sampling in protein simulations, Phys. Rev. E 65, 031910

52. Cerny, C.A., Kaiser, H.F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivariate Behavioral Research, 12(1), 43-47.

53. Hyvärinen, A., Oja, E. (2000). Independent Component Analysis: Algorithms and Applications. Neural Networks, 13(4-5):411-430.

54. Hyvärinen, A. (1999). Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Trans. on Neural Networks, 10(3):626-634.

55. Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, Volume 15, Number 2, Pages 265–286.

56. Yao, F., Coquery, J., Lê Cao, K. (2012). Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinformatics, 13:24 doi:10.1186/1471-2105-13-24

57. Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

58. Vojtechovsky, J., Chu, K., Berendzen, J., Sweet, R.M., Schlichting, I. (1999). Crystal structures of myoglobin-ligand complexes at near-atomic resolution. Biophys.J., 77, 2153-2174.

59. Fong, S., Hamill, S.J., Proctor, M., Freund, S.M., Benian, G.M., Chothia, C., Bycroft, M., Clarke, J. (1996). Structure and stability of an immunoglobulin superfamily domain from twitchin, a muscle protein of the nematode Caenorhabditis elegans. J.Mol.Biol., 264, 624-639 .

60. Vijay-Kumar, S., Bugg, C.E., Cook, W.J. (1987). Structure of ubiquitin refined at 1.8 A resolution. J.Mol.Biol, 194, 531-544.

61. Lolis, E., Alber, T., Davenport, R.C., Rose, D., Hartman, F.C., Petsko, G.A. (1990). Structure of yeast triosephosphate isomerase at 1.9-A resolution. Biochemistry, 29, 6609-6618.

62. Radestock, S., Gohlke, H. (2008). Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng. Life Science 8, 507-522.

63. Ahmed, A., Villinger, S., Gohlke, H. (2010). Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. Proteins, 78, 3341–3352.

64. Livesay, D.R., Dallakyan, S., Wood, G.G., and Jacobs, D.J. (2004). A flexible approach for understanding protein stability. FEBS Letters, 576, 468-76.

65. Jacobs, D.J., Livesay, D.R., Hules, J., and Tasayco, M.L. (2006). Elucidating quantitative stability-flexibility relationships within thioredoxin and its fragments using a distance constraint model. Journal of Molecular Biology, 358, 882-904

66. Livesay, D.R., Huynh, D.H., Dallakyan, S., and Jacobs, D.J. (2008). Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family. Chemistry Central Journal 2,17, 1-20.

67. Sweeney, H.L. & Houdusse, A. (2010). Structural and functional insights into the Myosin motor mechanism. Annu Rev Biophys 39, 539-557.

68. Malnasi-Csizmadia, A. & Kovacs, M. (2010). Emerging complex pathways of the actomyosin powerstroke. Trends Biochem. Sci. 35, 684-690.

69. Coureux, P.D., Wells, A.L., Menetrey, J., Yengo, C.M., Morris, C.A., Sweeney, H.L. & Houdusse, A. (2003). A structural state of the myosin V motor without bound nucleotide. Nature 425, 419-423.

70. Sun, M., Rose, M.B., Ananthanarayanan, S.K., Jacobs, D.J. & Yengo, C.M. (2008). Characterization of the pre-force-generation state in the actomyosin cross-bridge cycle. Proc. Natl. Acad. Sci. USA 105, 8631-8636.

71. Conibear, P.B., Bagshaw, C.R., Fajer, P.G., Kovacs, M. & Malnasi-Csizmadia, A. (2003). Myosin cleft movement and its coupling to actomyosin dissociation. Nat. Struct. Biol. 10, 831-835.

72. Yengo, C.M., De La Cruz, E.M., Chrin, L.R., Gaffney, D.P., 2nd & Berger, C.L. (2002). Actin-induced closure of the actin-binding cleft of smooth muscle myosin. J. Biol. Chem. 277, 24114-24119.

73. Irving, M., St Claire Allen, T., Sabido-David, C., Craik, J.S., Brandmeier, B., Kendrick- Jones, J., Corrie, J.E., Trentham, D.R. & Goldman, Y.E. (1995). Tilting of the light-chain region of myosin during step length changes and active force generation in skeletal muscle. Nature 375, 688-691.

74. Goldman, Y.E. (1998). Wag the tail: structural dynamics of actomyosin. Cell 93, 1-4. 9.

75. Dunn, A.R. & Spudich, J.A. (2007). Dynamics of the unbound head during myosin V processive translocation. Nat. Struct. Mol. Biol. 14, 246-248.

76. Forkey, J.N., Quinlan, M.E., Shaw, M.A., 48. Corrie, J.E. & Goldman, Y.E. (2003). Three-dimensional structural dynamics of myosin V by single-molecule fluorescence

polarization. Nature 422, 399-404.

77. Shiroguchi, K. & Kinosita, K., Jr. (2007). Myosin V walks by lever action and Brownian motion. Science 316, 1208-1212.

78. Toprak, E., Enderlein, J., Syed, S., McKinney, S.A., Petschek, R.G., Ha, T., Goldman, Y.E. & Selvin, P.R. (2006). Defocused orientation and position imaging (DOPI) of myosin V. Proc. Natl. Acad. Sci. USA 103, 6495-6499.

79. Whittaker, M., Wilson-Kubalek, E.M., Smith, J.E., Faust, L., Milligan, R.A. & Sweeney, H.L. (1995). A 35-Å movement of smooth muscle myosin on ADP release. Nature 378,748-751.

80. Jontes, J.D., Ostap, E.M., Pollard, T.D. & Milligan, R.A. (1998). Three-dimensional structure of Acanthamoeba castellanii myosin-IB (MIB) determined by cryoelectron microscopy of decorated actin filaments. J. Cell Biol. 141, 155-162.

81. Volkmann, N., Liu, H., Hazelwood, L., Krementsova, E.B., Lowey, S., Trybus, K.M. & Hanein, D. (2005). The structural basis of myosin V processive movement as revealed by electron cryomicroscopy. Mol. Cell. 19, 595-605.

82. Nyitrai, M. & Geeves, M.A. (2004). Adenosine diphosphate and strain sensitivity in myosin motors. Philos Trans R Soc Lond B Biol Sci 359, 1867-1877.

83. Huxley, A.F. (1957). Muscle structure and theories of contraction. Prog. Biophys. Biophys. Chem. 7, 255-318.

84. Siemankowski, R.F., Wiseman, M.O. & White, H.D. (1985). ADP dissociation from actomyosin subfragment 1 is sufficiently slow to limit the unloaded shortening velocity in vertebrate muscle. Proc. Natl. Acad. Sci. USA 82, 658-662.

85. Oguchi, Y., Mikhailenko, S.V., Ohki, T., Olivares, A.O., De La Cruz, E.M. & Ishiwata, S. Robust processivity of myosin V under off-axis loads. Nat Chem Biol 6, 300-305.

86. Oguchi, Y., Mikhailenko, S.V., Ohki, T., Olivares, A.O., De La Cruz, E.M. & Ishiwata, S. (2008). Load-dependent ADP binding to myosins V and VI: implications for subunit coordination and function. Proc. Natl. Acad. Sci. USA 105, 7714-7719.

87. Uemura, S., Higuchi, H., Olivares, A.O., De La Cruz, E.M. & Ishiwata, S. (2004). Mechanochemical coupling of two substeps in a single myosin V motor. Nat. Struct. Mol. Biol. 11, 877-883.

88. Sellers, J.R. & Veigel, C. Direct observation of the myosin-Va power stroke and its reversal. Nat. Struct. Mol. Biol. 17, 590-595.

89. Sun, M., Oakes, J.L., Ananthanarayanan, S.K., Hawley, K.H., Tsien, R.Y., Adams, S.R. & Yengo, C.M. (2006). Dynamics of the upper 50-kDa domain of myosin V examined with fluorescence resonance energy transfer. J. Biol. Chem. 281, 5711-5717.

90. Bauer, C.B., Kuhlman, P.A., Bagshaw, C.R. & Rayment, I. (1997). X-ray crystal structure and solution fluorescence characterization of Mg.2'(3')-O-(N-methylanthraniloyl) nucleotides bound to the Dictyostelium discoideum myosin motor domain. J. Mol. Biol. 274, 394-407.

91. Hannemann, D.E., Cao, W., Olivares, A.O., Robblee, J.P. & De La Cruz, E.M. (2005). Magnesium, ADP, and actin binding linkage of myosin V: evidence for multiple myosin V-ADP and actomyosin V-ADP states. Biochemistry 44, 8826-8840.

92. Rosenfeld, S.S., Houdusse, A. & Sweeney, H.L. (2005). Magnesium regulates ADP dissociation from myosin V. J. Biol. Chem. 280, 6072-6079.

93. Sellers, J.R. (2000). Myosins: a diverse superfamily. Biochim Biophys Acta, 1496(1): p. 3-22.

94. De La Cruz, E.M. and E.M. Ostap. (2004). Relating biochemistry and function in the myosin superfamily. Curr Opin Cell Biol, 16(1): p. 61-7.

95. Kull, F.J., Vale, R.D., Fletterick, R.J. (1998). The case for a common ancestor: kinesin and myosin motor proteins and G proteins. J Muscle Res Cell Motil, 19(8): p. 877-86.

96. Kintses, B., et al. (2007). Reversible movement of switch 1 loop of myosin determines actin interaction. EMBO J, 26(1): p. 265-74.

97. Coureux, P.D., Sweeney, H.L., Houdusse, A. (2004). Three myosin V structures delineate essential features of chemo-mechanical transduction. EMBO J, 23(23): p. 4527-37.

98. Holmes, K.C., Geeves, M.A. (2000). The structural basis of muscle contraction. Philos Trans R Soc Lond B Biol Sci, 355(1396): p. 419-31.

99. Geeves, M.A., Holmes, K.C. (1999). Structural mechanism of muscle contraction. Annu Rev Biochem, 68: p. 687-728.

100. Kuhner, S., Fischer, S. (2011). Structural mechanism of the ATP-induced dissociation of rigor myosin from actin. Proc Natl Acad Sci U S A. 108(19): p. 7793-8.

101. Forgacs, E., et al. (2009). Switch 1 mutation S217A converts myosin V into a low duty ratio motor. J Biol Chem, 284(4): p. 2138-49.

102. Sasaki, N., Shimada, T., Sutoh, K. (1998). Mutational analysis of the switch II loop of Dictyostelium myosin II. J Biol Chem, 273(32): p. 20334-40.

103. Shimada, T., et al. (1997). Alanine scanning mutagenesis of the switch I region in the ATPase site of Dictyostelium discoideum myosin II. Biochemistry, 36(46): p. 14037-43.

104. Holmes, K.C., et al. (2004). The structure of the rigor complex and its implications for the power stroke. Philos Trans R Soc Lond B Biol Sci, 359(1452): p. 1819-28.

105. Mehta, A.D., et al. (1999). Myosin-V is a processive actin-based motor. Nature, 400(6744): p. 590-3.

106. Yildiz, A., et al. (2003). Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. Science, 300(5628): p. 2061-5.

107. Yengo, C.M., et al. (2002). Kinetic characterization of the weak binding states of myosin V. Biochemistry, 41(26): p. 8508-17.

108. Hannemann, D.E., et al. (2005). Magnesium, ADP, and actin binding linkage of myosin V: evidence for multiple myosin V-ADP and actomyosin V-ADP states. Biochemistry, 44(24): p. 8826-40.

109. De La Cruz, E.M., et al. (1999). The kinetic mechanism of myosin V. Proc Natl Acad Sci U S A, 96(24): p. 13726-31.

110. Watanabe, T.M., et al. (2010). Mechanical characterization of one-headed myosin-V using optical tweezers. PLoS One. 5(8): p. e12224.

111. Jacobs, D.J., et al. (2011). Kinetics and thermodynamics of the rate-limiting conformational change in the actomyosin V mechanochemical cycle. J Mol Biol. 407(5): p. 716-30.

112. Sun, M., et al. (2006). Dynamics of the upper 50-kDa domain of myosin V examined with fluorescence resonance energy transfer. J Biol Chem, 281(9): p. 5711-7.

113. Sun, M., et al. (2008). Characterization of the pre-force-generation state in the actomyosin cross-bridge cycle. Proc Natl Acad Sci U S A, 105(25): p. 8631-6.

114. Okimoto, N., et al. (2001). Theoretical studies of the ATP hydrolysis mechanism of myosin. Biophys J, 81(5): p. 2786-94.

115. Fisher, A.J., et al. (1995). X-ray structures of the myosin motor domain of Dictyostelium discoideum complexed with MgADP.BeFx and MgADP.AlF4. Biochemistry, 34(28): p. 8960-72.

116. Ovchinnikov, V., Trout, B.L., Karplus, M. (2010). Mechanical coupling in myosin V: a simulation study. J Mol Biol, 395(4): p. 815-33.

117. Conibear, P.B., et al. (2003). Myosin cleft movement and its coupling to actomyosin dissociation. Nat Struct Biol, 10(10): p. 831-5.

118. Cecchini, M., Houdusse, A., Karplus, M. (2008). Allosteric communication in myosin V: from small conformational changes to large directed movements. PLoS Comput Biol, 4(8): p. e1000129.

119. Kintses, B., Yang, Z., Malnasi-Csizmadia, A. (2008). Experimental investigation of the seesaw mechanism of the relay region that moves the myosin lever arm. J Biol Chem, 283(49): p. 34121-8.

120. Decarreau, J.A., et al. (2011). Switch I closure simultaneously promotes strong binding to actin and ADP in smooth muscle myosin. J Biol Chem, 286(25): p. 22300-7.

121. Nagy, N.T., et al. (2010). Functional adaptation of the switch-2 nucleotide sensor enables rapid processive translocation by myosin-5. FASEB J. 24(11): p. 4480-90.

APPENDIX A: RECIPE FOR PCA

Essential Dynamics using Coordinate PCA applied to myoglobin PDB ID: 1a6n

1. Obtain trajectories (1 or more) from dynamic simulation. For illustrative examples, one MD and three FRODA trajectories for myoglobin (PDB ID 1a6n) are considered to explain aspects of PCA. For this purpose, details about the setup of the various simulations are ignored, except when it pertains to methodology. The MD trajectory consists of 2,000 frames after equilibration. One FRODA trajectory has 2,000 frames (100,000 explored conformations), and the other two FRODA trajectories have 10,000 frames. The sampling rate of FRODA is normally set at 1 out of 50 conformations generated. Here, one long trajectory is obtained from sampling every conformation (10,000 explored conformations), meaning it is 10% as long as the 2,000 frame FRODA trajectory in terms of MC-steps, while the other is obtained from sampling every tenth conformation (100,000 explored conformations), is of equal length.
2. Remove overall translations and rotations by aligning each frame to a reference structure.
   - We use the starting (crystal) structure as our reference, and our quaternion alignment program to optimally align each structure to the reference structure. Only the alpha carbon atoms were included in the alignment process.
3. Choose the set of atoms for the analysis: This forms the data matrix A_{Aligned}.

- Protein conformations (observations or frames) define columns, and rows describe the x, y and z coordinates of the alpha carbon atoms. In this example, all 151 of the alpha carbons are used, giving 453 total DOF (variables).

4. Examine the descriptive statistics for the variables.
   - Table 1 shows some statistics for three selected coordinates (variables) to highlight the non-uniformity of the standard deviations.
5. Examine the Kaiser-Meyer-Olkin sampling adequacy scores for each coordinate. The MD and FRODA trajectories each with 2,000 samples are compared in Figure A1. Most coordinates from (MD, FRODA) simulation (do not, do) meet the recommended KMO cutoff criterion of 0.50. We assess how the KMO statistic changes when the number of FRODA samples is increased from 2,000 to 10,000, and investigate how the sampling frequency effects the sampling adequacy in Figure A1-B. The overall KMO statistic remains about the same, and the individual coordinates that had a low KMO statistic did not improve by increasing the number of samples. Even more surprising, a sample rate of 1 leads to a slight improvement of the KMO. Thus, there exists a trade-off between the amount of conformational space that a simulation explores and the statistical sampling adequacy of those states. How to improve sampling adequacy in locations with low KMO statistics is not clear, since more sampling in the same way has diminishing returns.
6. Center the variables of A_{Aligned} (row centering).
   - This forms the centered data matrix A'.
7. Construct the covariance matrix, Q, of {x,y,z} positions for the atoms using A'.
   - For comparisons, construct the correlation matrix R.

8. Diagonalize Q or R using an eigenvalue decomposition.
9. Examine the eigenvalue scree plot to determine the number of eigenvectors to include in the reduced vector space that describes the most relevant features. Figure A2 shows these plots in Panel A along with the conformational and residue RMSDs in Panels B & C.
   - It is not advisable to include all modes up to a preset percent of variance cutoff. Note that the characteristics of the scree plot depend heavily on whether one is analyzing fluctuations within a single native basin or is analyzing combined trajectories of multiple states. For a single native basin of random motions, many modes will be required to achieve 50% of the variance. For multiple states/configurations, the first two modes may subsume more than 50% of the variance. Our example MD plot shows that most of the variance is captured by one mode, because its CE clusters into two conformational states. In contrast, the FRODA plot does not have a dominant mode, but rather shows a monotonically decreasing trend indicative of random fluctuations about the native state of the protein (the input structure).
10. Select the top set of eigenvectors for forming the principal components (PCs) (Usually 2-20). In our MD example, the top two modes reveal how two distinct states of the protein were sampled. However, at least 10 modes are required to define the essential subspaces for a comparison between MD and FRODA CEs (See the RMSIP plots below).
11. Examine the component loadings, which are the product of the square root of the eigenvalue with the eigenvector. When the correlation matrix is used, they are also

the correlation coefficients (cosines) between the variables and factors (PCs). Analogous to Pearson's r, the squared component loading (squared cosine) is the percent of variance in that variable explained by the PC. In Table 2, PC1 is clearly capturing the behavior of the first three variables.

- Scatterplots of the component loadings for the top 2 factors should be examined. In Figure A3 the first ten variables (Var 1 to Var 10) are seen to cluster. The angle between the variables on this scatterplot indicates the level of correlation, with (0, 90, 180) degrees indicating a correlation of (1, 0, -1).

12. Examine the squared cosines of the variables. These values indicate whether a correlation is worthy of interpretation or likely an artifact of projection into a low dimensional subspace. Only the first 3 are shown in the Table 3, and they strongly support the correlations shown in Figure A3.
13. Examine the contribution of the variables. Here we show only the first 3 in Table 4, but even from this truncated list, it is clear that the N-terminus residues have a large contribution to the first mode.
14. Examine the eigenvector collectivity, Figure A4. The top modes tend to be more collective than lower modes indicating that many residues are participating in collective motions. For our example, the FRODA eigenvector collectivity drops off rather steeply suggesting that the top forty or so modes capture most of the collective motions occurring in the native state. This trend of having a set of highly collective modes highlights the fact that real protein motions tend to be captured by a superposition of PC modes, not a single mode. In contrast, the MD collectivity does not drop off rapidly suggesting that many more modes may be required to properly

capture the dynamics that the MD simulation produced. These results also clearly demonstrate that while PC modes in totality always form a complete basis set, they are derived from statistics, and will be dependent on the sampling. The top PC modes describe the sampling, not necessarily anything of biological importance. It is therefore important to carefully choose what and how to sample so that biological interpretations can be made.

15. Construct the weighted RMSD modes: Here we map the 3m components of the eigenvectors to m new variables that capture the squared displacements of each residue to visualize which residues contribute most to the fluctuations of each PCA mode. For each eigenvector i, the new mode N_i has m components, with each component defined by the square root of the sum of the squares of the 3 variables that contribute to the associated residue, scaled by the square root of the corresponding eigenvalue. These results are shown in Figure A5. The mapping equation is given by:

$$N_i = \sqrt{\lambda_i \begin{pmatrix} x_1^2 + y_1^2 + z_1^2 \\ \vdots \\ x_m^2 + y_m^2 + z_m^2 \end{pmatrix}} \qquad \text{Eq. (A1)}$$

- Weighting is done by multiplying by the square root of the eigenvalue for the mode, λ_i.
- It is often useful to compare the RMSD modes to the overall residue RMSD plot from the entire trajectory. Also, one may use the un-weighted RMSD modes to see relative displacements that are hard to see in the weighted plots due to the typical rapid decrease in the eigenvalues with mode index.

16. Construct the displacement vectors (DV) for the trajectory, given by $DV_i = X_i - X_{ref}$ and construct the Principal Components (PCs) (Factor Scores)
    - PC_i is formed by taking the inner product between eigenvector i and each DV (Observation). Projections can be made on singe modes to view as line graphs. Projections on sets of 2 PCA modes create scatter plots that show how the simulation explored the configuration space defined by the selected set of modes. In Figure A6, it is evident that the MD trajectory sampled two states of the protein as seen by the two clusters in the scatterplot of PC1 versus PC2. In contrast, the projection of the FRODA trajectory onto the top two modes shows a uniform distribution.
17. Check the contribution of the observations to the PCs to see if there are particular ones that unduly influence the analysis. Here we show only the first 3 observations in Table 5 and the values are percentages.
18. We also examine the squared cosines of the observations when determining if an observation belongs to a particular cluster or not. In Table 6, we show values for the first 3 observations. Values in bold are significant at the 0.01 level.
19. Since the sampling in the MD simulation was poor for many variables, we check the cosine content of the top two principle components. Comparing PC1 to a half-period cosine, we find a 0.63 correlation and in comparing PC2 to a full period cosine, we find a 0.16 correlation. The high cosine content in mode one suggests that the MD simulation should be run longer.
20. When examining two or more sets of PCA modes, determination of how similar the trajectories are to each other may be assessed using the CO, RMSIP or PA metrics.

- In Figure A7, we compare the vector space of the top modes from the MD trajectory to that of the FRODA trajectory, each with 2000 frames. Note that the various metrics for SS comparisons depend on the size of the VS and SS. As the SS DIM increases, the ability of that SS to capture a given eigenvector increases. Because all the metrics have dependencies on dimensionality, it is best to have a baseline score for random comparisons as a function of the dim(VS) and dim(SS).

Table 1: Descriptive statistics for 3 variables in the MD simulation data.

Variable	Minimum	Maximum	Mean	Std. deviation
Var 1	3.456	11.489	7.085	1.610
Var 10	9.568	12.980	11.530	0.707
Var 20	8.390	10.467	9.423	0.301

Table 2: Component Loadings for the first 3 variables in the MD trajectory.

Variable	PC1	PC2	PC3
Var 1	0.807	-0.218	-0.056
Var 2	0.890	-0.223	-0.095
Var 3	0.867	-0.254	-0.111

Table 3: Squared Cosines of the Variables

Variable	PC1	PC2	PC3
Var 1	0.651	0.048	0.003
Var 2	0.791	0.050	0.009
Var 3	0.752	0.065	0.012

Table 4: Contribution of the Variables to the PCs as Percent

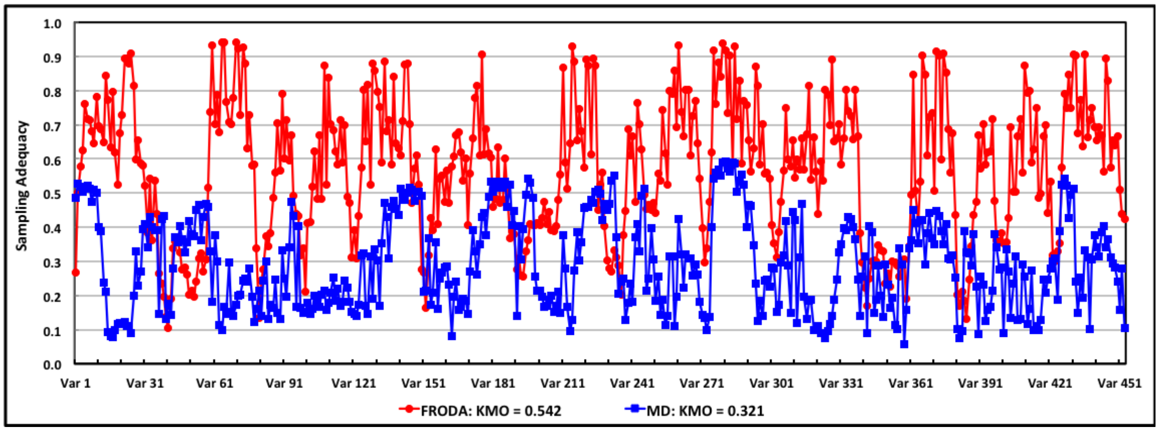
Variable	PC1	PC2	PC3
Var 1	0.570	0.137	0.014
Var 2	0.693	0.143	0.039
Var 3	0.659	0.186	0.053

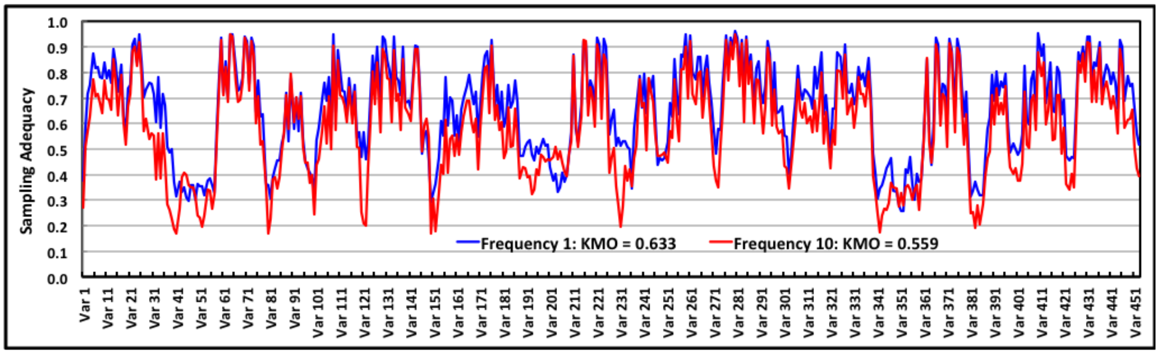
Table 5: Contribution of the Observations to the PCs as Percent

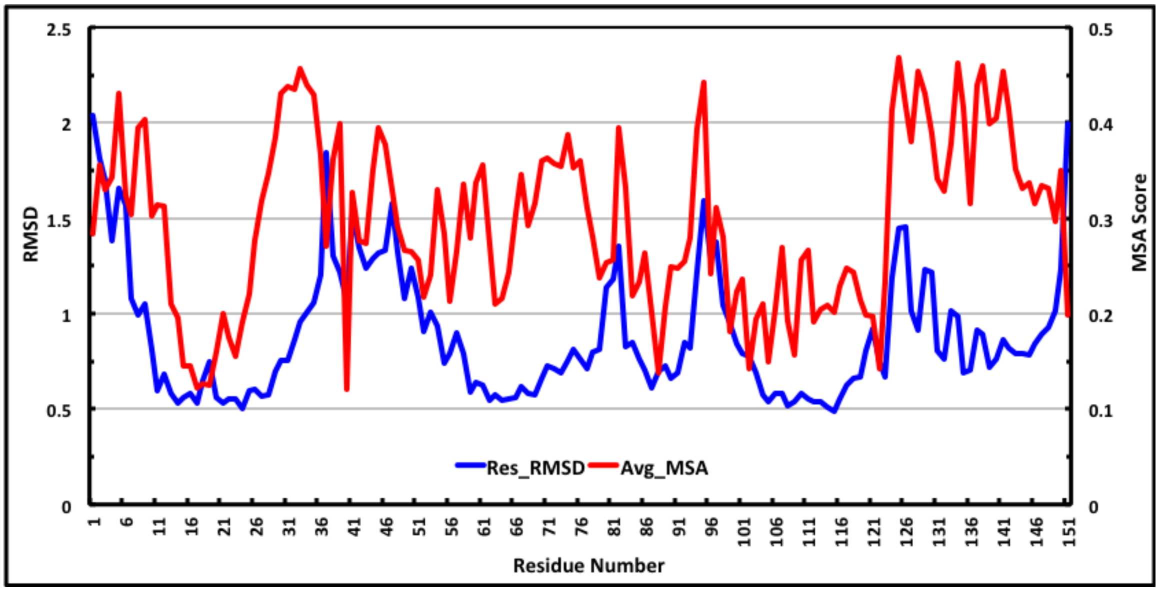
Observation	PC1	PC2	PC2
Obs 1	0.015	0.529	0.147
Obs 2	0.002	0.329	0.121
Obs 3	0.003	0.485	0.033

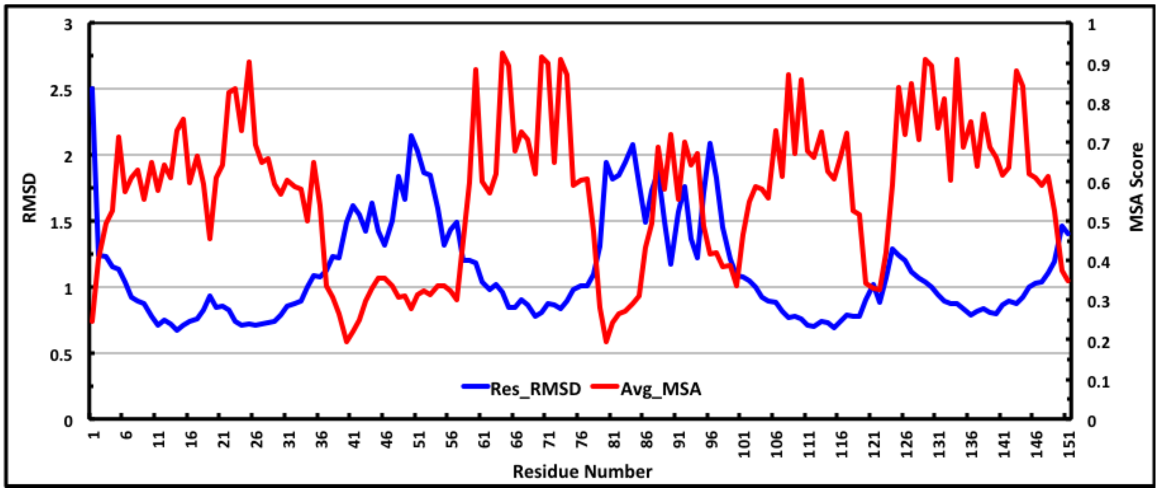
Table 6: Squared Cosines of the Observations

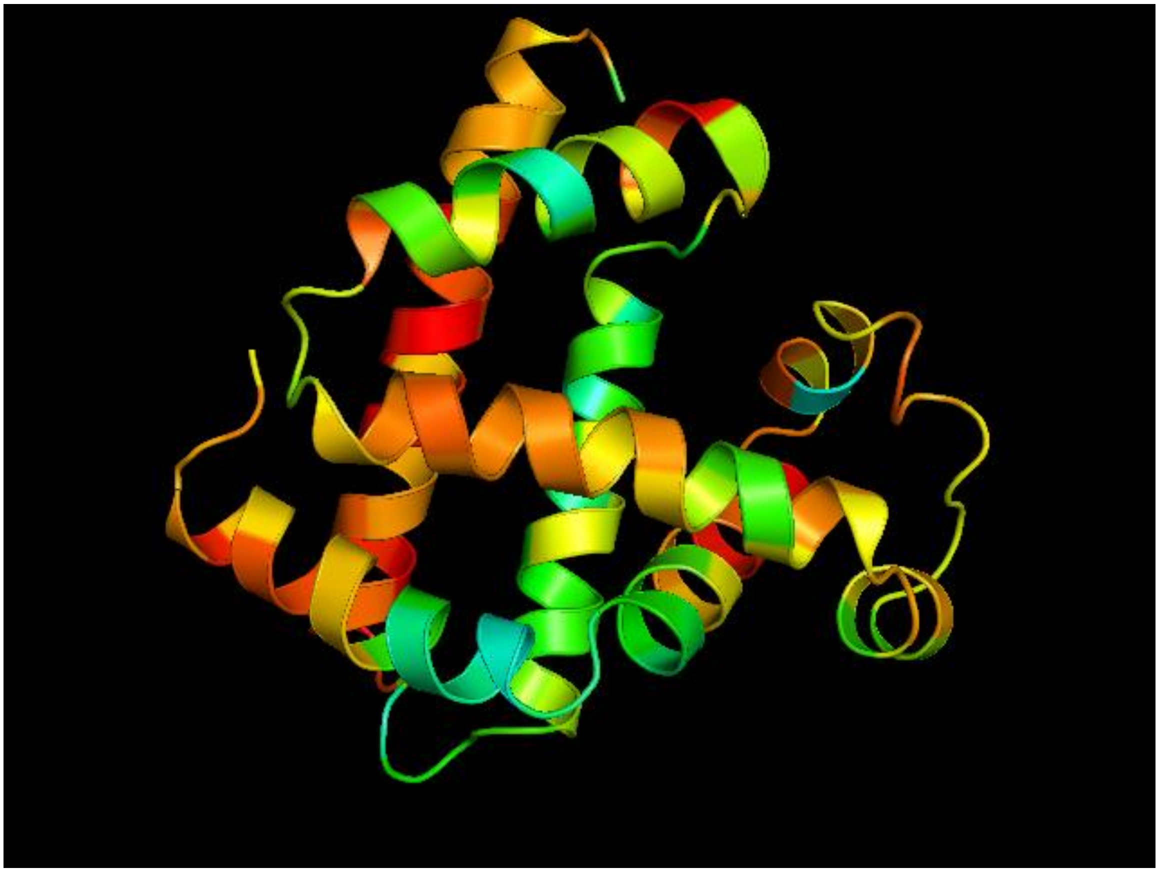
Observation	PC1	PC2	PC3
Obs 1	0.026	0.285	0.052
Obs 2	0.005	0.222	0.054
Obs 3	0.007	0.351	0.016




Sampling Adequacy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency 1: KMO = 0.633
Frequency 10: KMO = 0.559
Var 1
Var 11
Var 21
Var 31
Var 41
Var 51
Var 61
Var 71
Var 81
Var 91
Var 101
Var 111
Var 121
Var 131
Var 141
Var 151
Var 161
Var 171
Var 181
Var 191
Var 201
Var 211
Var 221
Var 231
Var 241
Var 251
Var 261
Var 271
Var 281
Var 291
Var 301
Var 311
Var 321
Var 331
Var 341
Var 351
Var 361
Var 371
Var 381
Var 391
Var 401
Var 411
Var 421
Var 431
Var 441
Var 451


RMSD
2.5
2
1.5
1
0.5
0
MSA Score
0.5
0.4
0.3
0.2
0.1
0
Res_RMSD
Avg_MSA
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
81
86
91
96
101
106
111
116
121
126
131
136
141
146
151
Residue Number


RMSD
3
2.5
2
1.5
1
0.5
0
MSA Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Res_RMSD
Avg_MSA
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
81
86
91
96
101
106
111
116
121
126
131
136
141
146
151
Residue Number



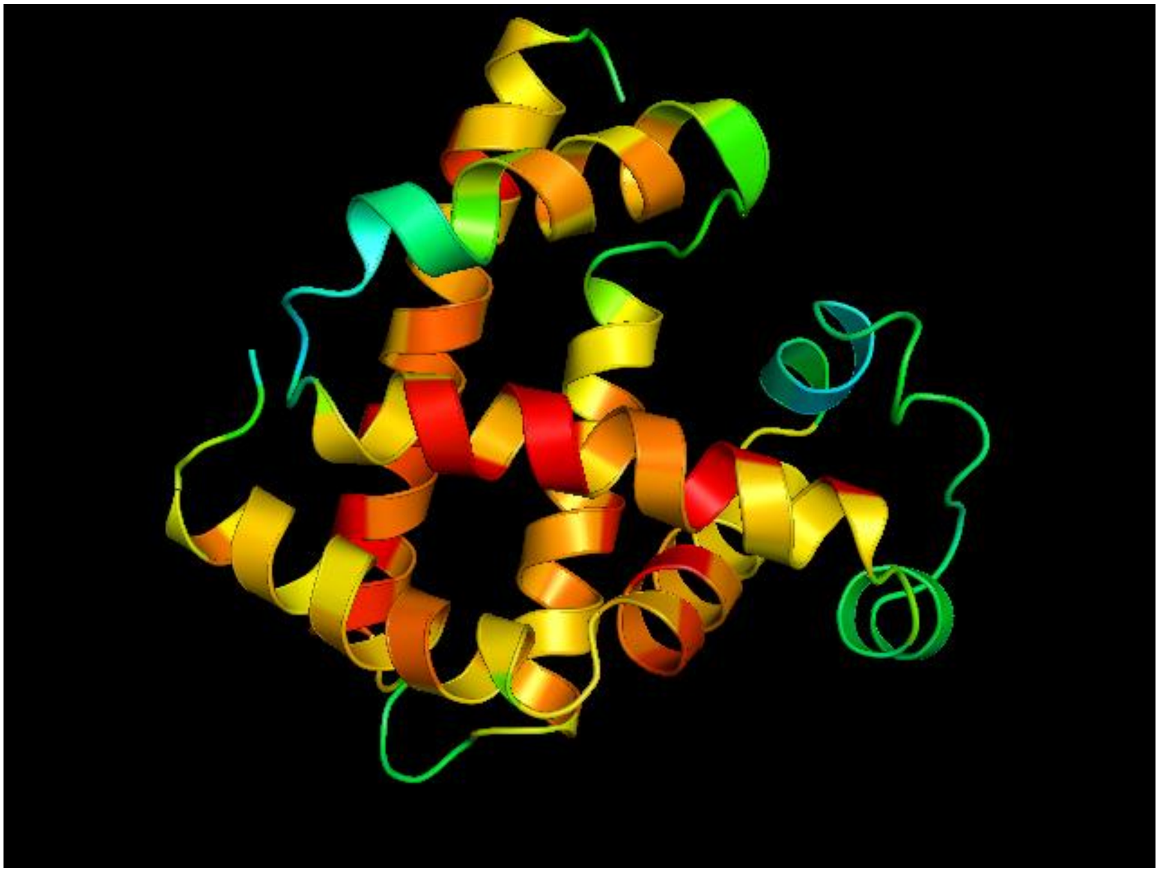
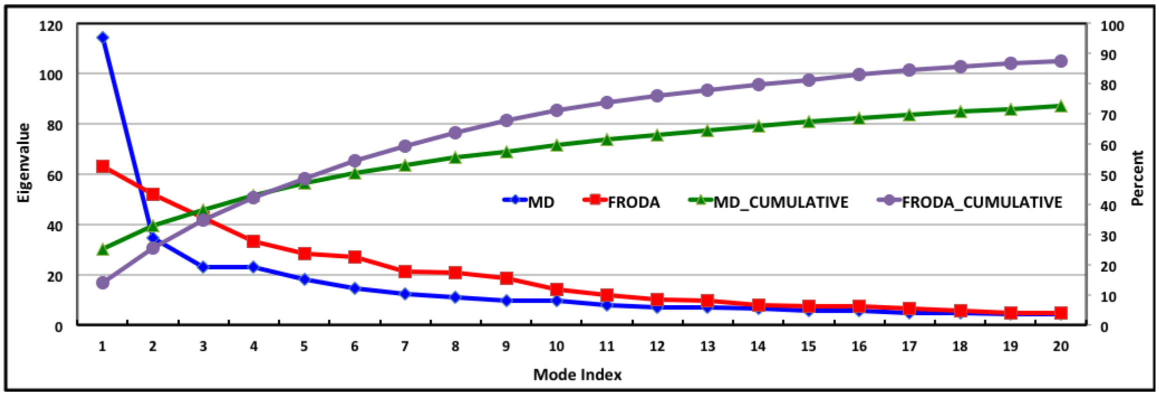


Figure A1: The Kaiser-Meyer-Olkin MSA for (A) the FRODA and MD CEs each with 2000 frames, and (B) for the FRODA CEs each with 10,000 frames. The overall KMO score is shown parenthetically in the legend. (C) Relationship between residue RMSD and MSA for MD. (D) Relationship between residue RMSD and MSA for FRODA. (E) Ribbon diagram colored by the MSA scores for MD. (F) Ribbon diagram colored by the MSA scores for FRODA.



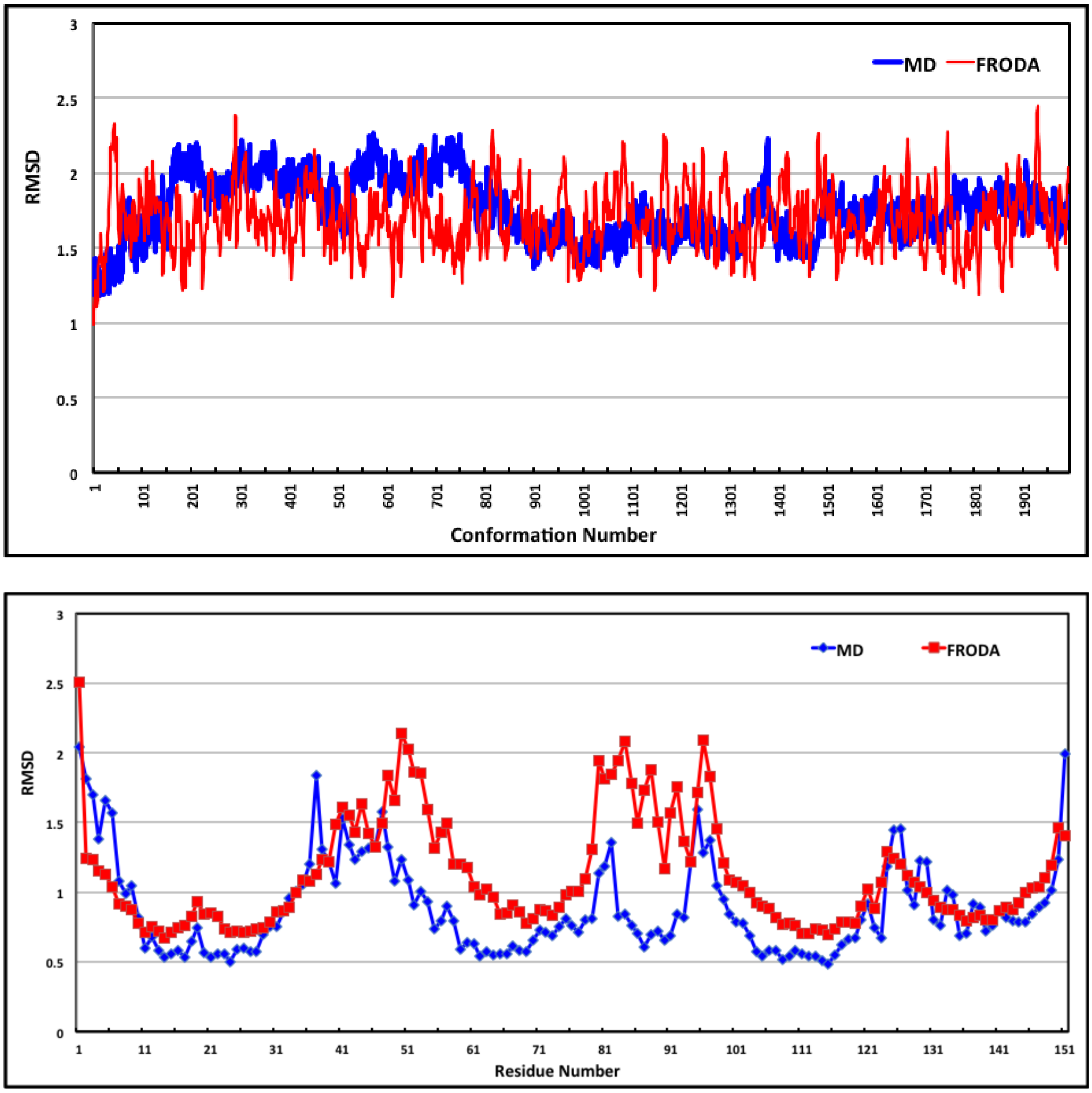


Figure A2: (A) Eigenvalue scree plots for the FRODA and MD CEs showing both the correlation explained in each mode and the cumulative correlations (Since the PCA was based on the correlation matrix). (B) The conformation RMSD of the MD and FRODA trajectories. Each value is with respect to the starting structure (crystal structure). (C) The residue RMSD for the MD and FRODA trajectories.

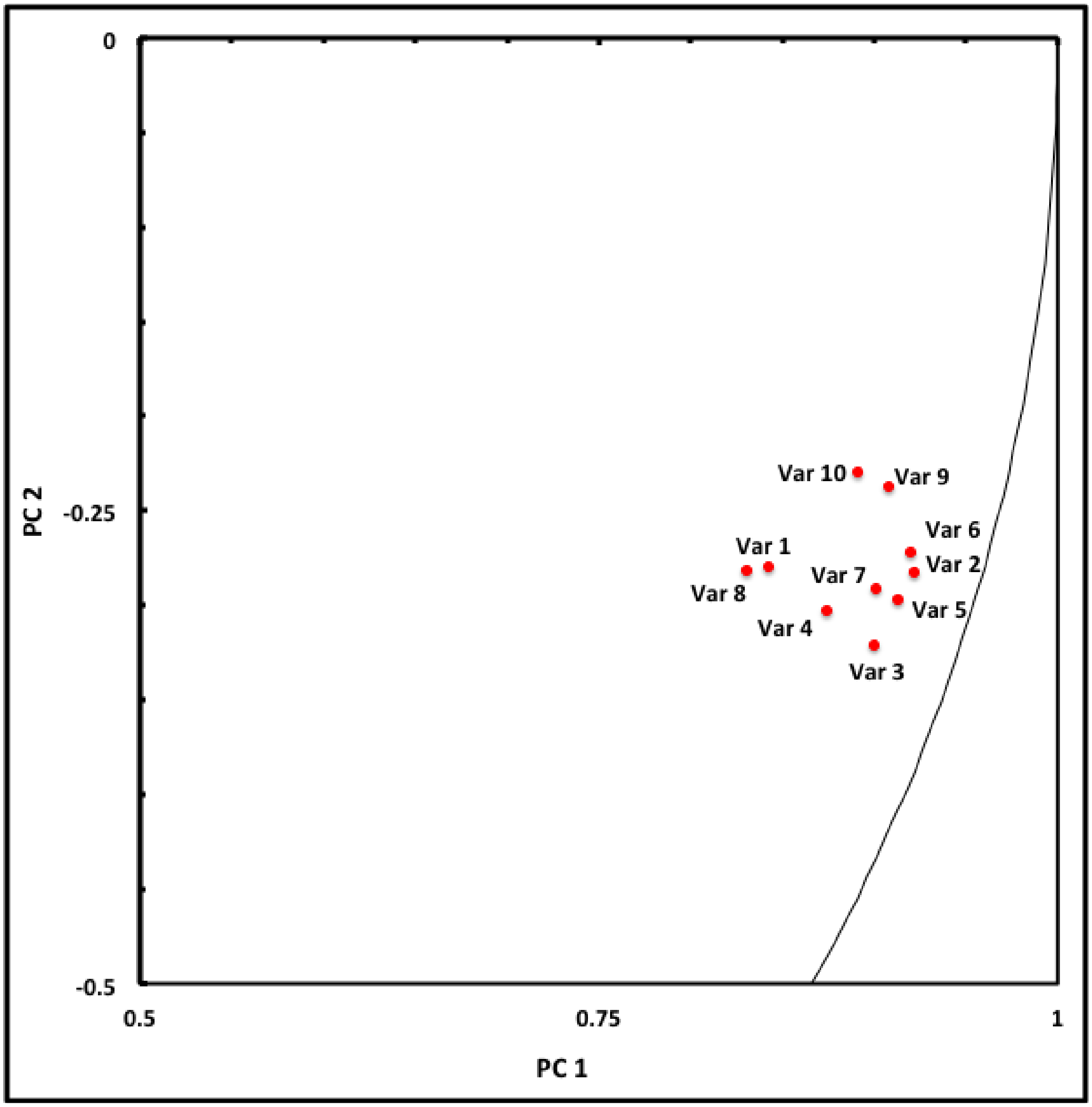


Figure A3: The correlations between the first 10 variables and the top 2 PCs. Notice how these variables form a tight cluster with small angles between each, indicating that they are correlated on these PCs. The boundary line on right is an arc of the unit circle to indicate how close the values are to 1.

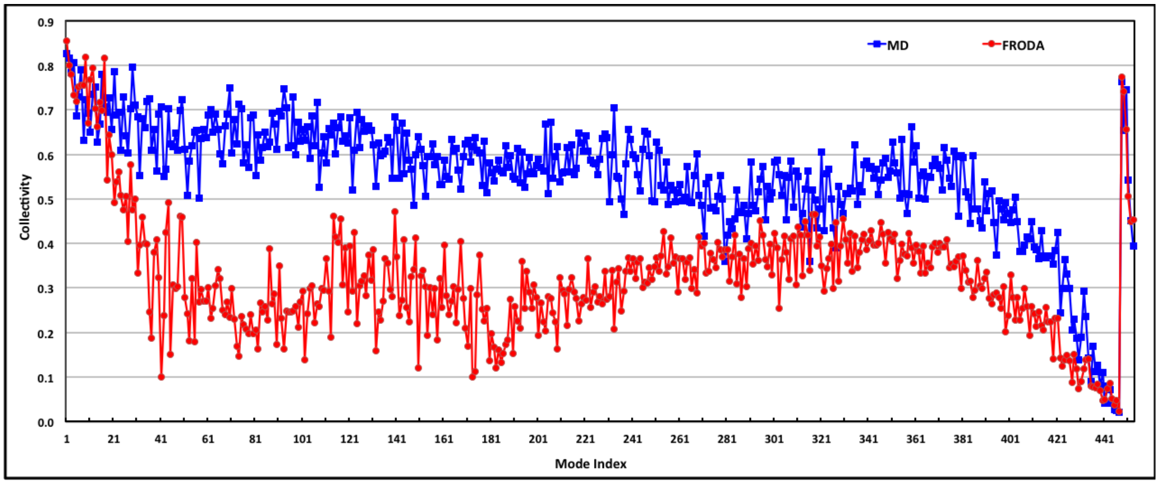
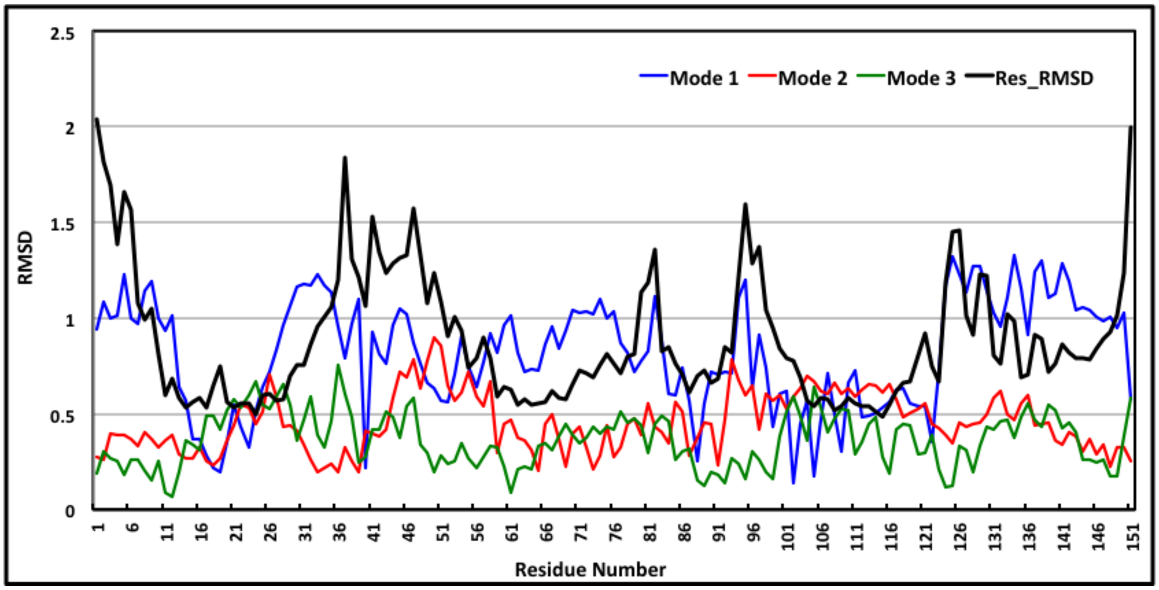


Figure A4: The eigenvector collectivity for the entire set of eigenvectors from both the MD and FRODA PCA. Note that the mode index is plotted with decreasing size of the eigenvalue, so mode index 1 is the top mode. This plot indicates that the collectivity measure should not be of primary concern.



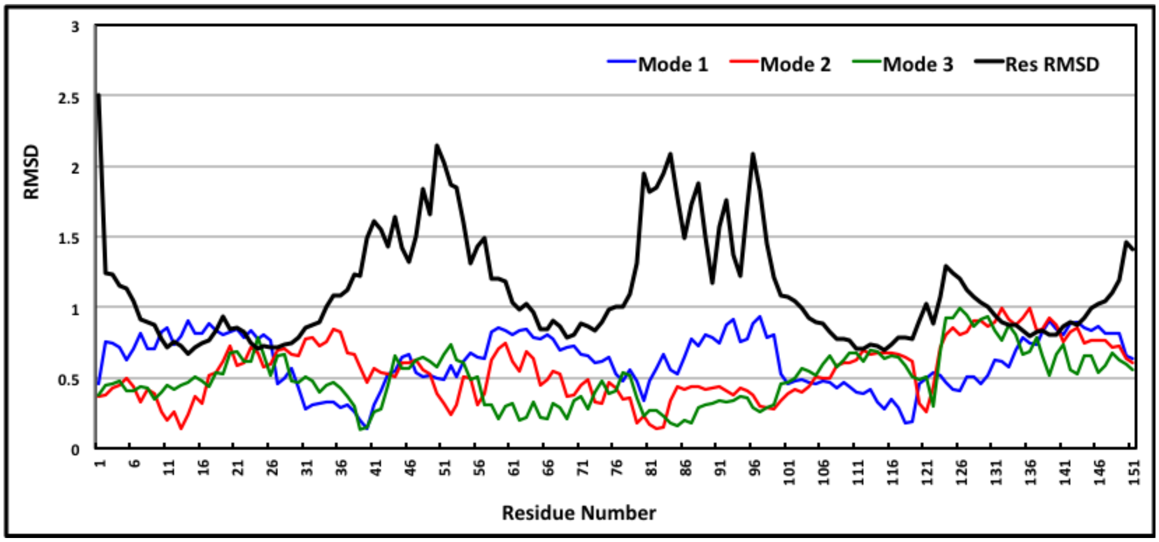
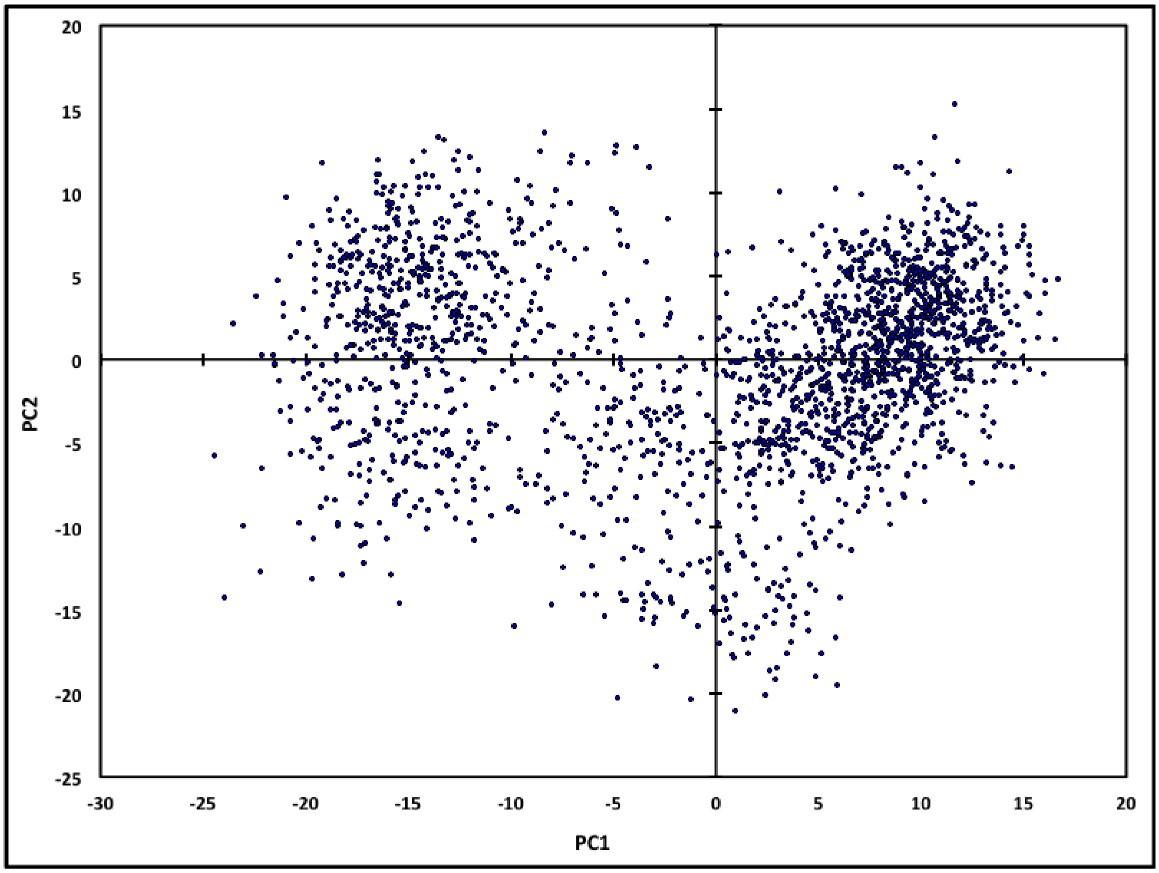
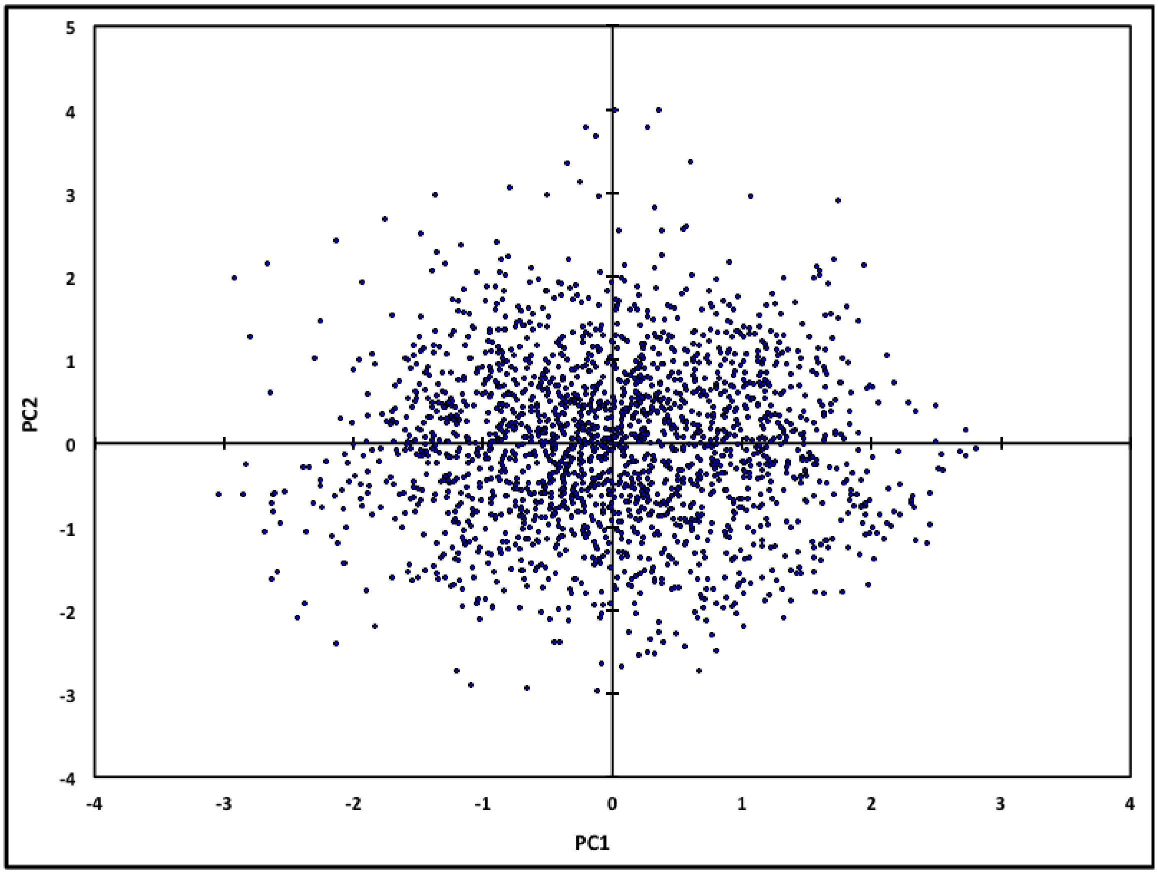
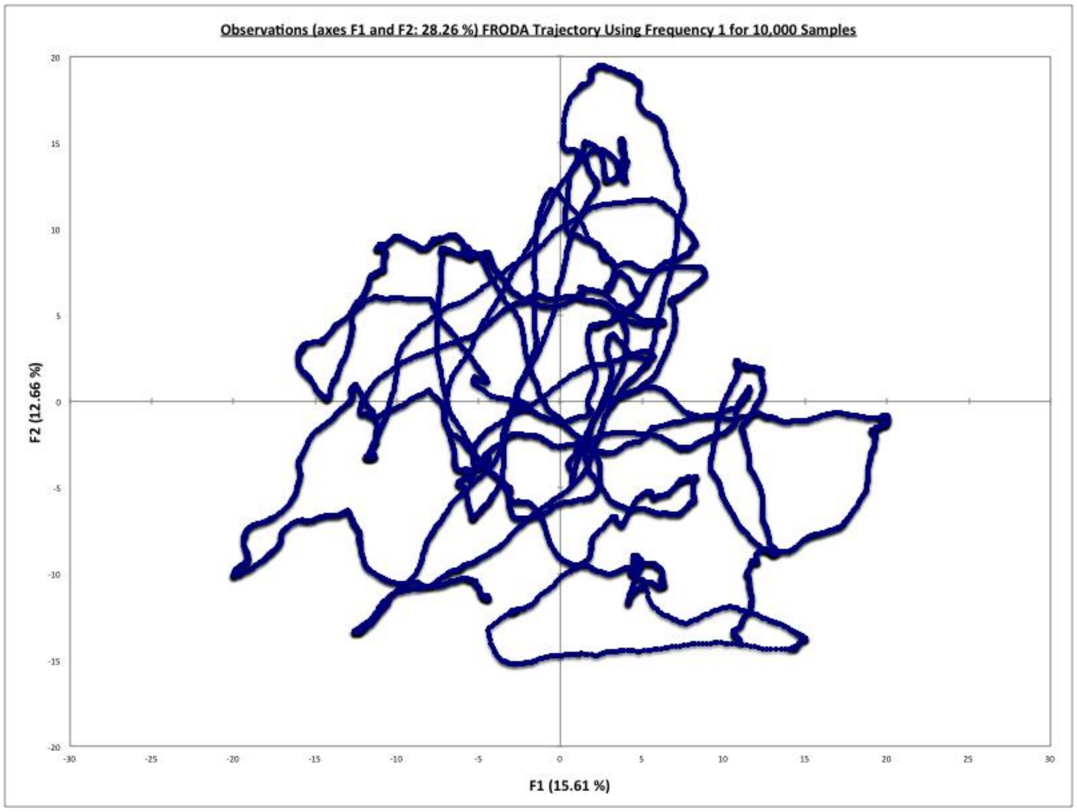


Figure A5: The RMSD and the top 3 RMSD modes are compared from (A) MD and (B) FRODA PCA.




5
4
3
2
1
0
-1
-2
-3
-4
PC2
-4
-3
-2
-1
0
1
2
3
4
PC1


Observations (axes F1 and F2: 28.26 %) FRODA Trajectory Using Frequency 1 for 10,000 Samples
20
15
10
5
0
-5
-10
-15
-20
F2 (12.66 %)
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
30
F1 (15.61 %)

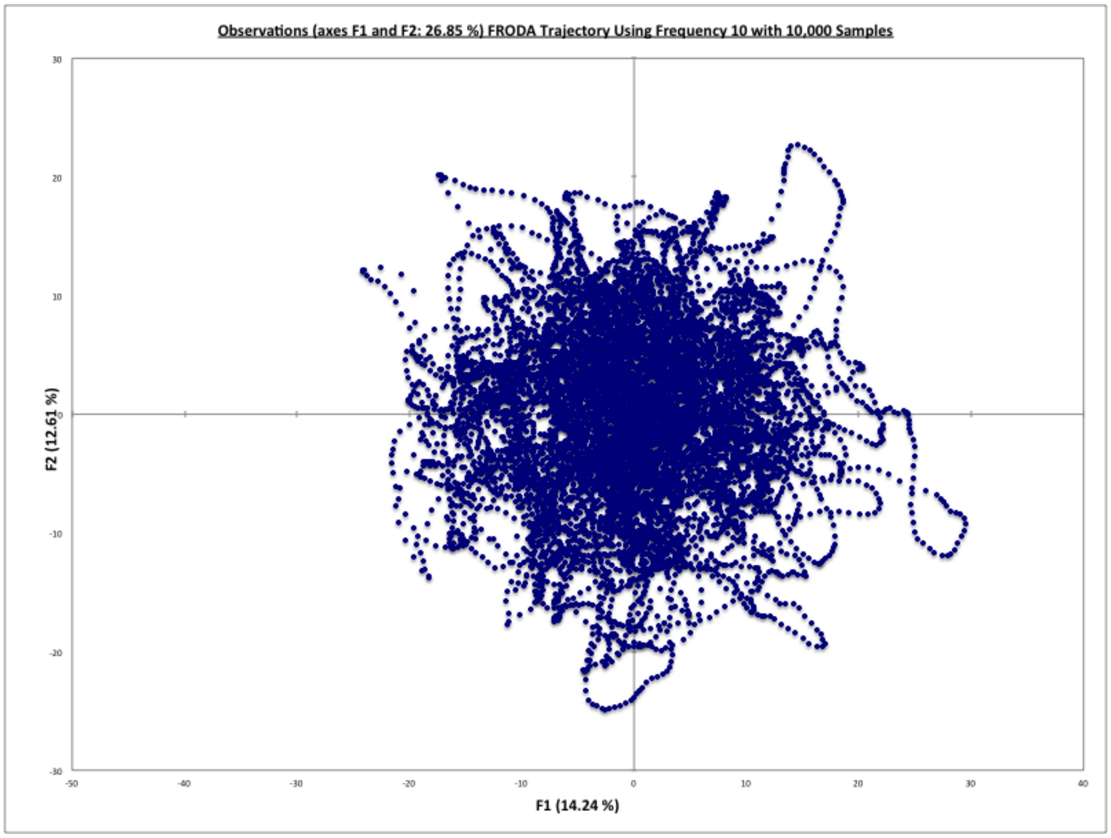
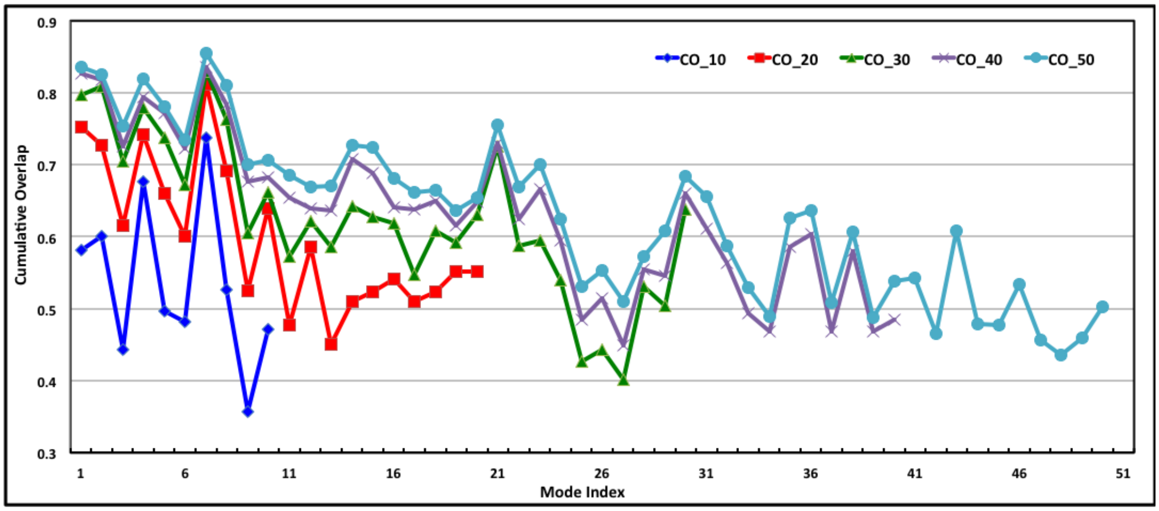


Figure A6: (A) MD and (B) FRODA displacement vectors are projected onto their respective top 2 PCs. (C) Freq 1 and (D) Freq 10 displacement vectors are projected onto their respective top 2 PCs.



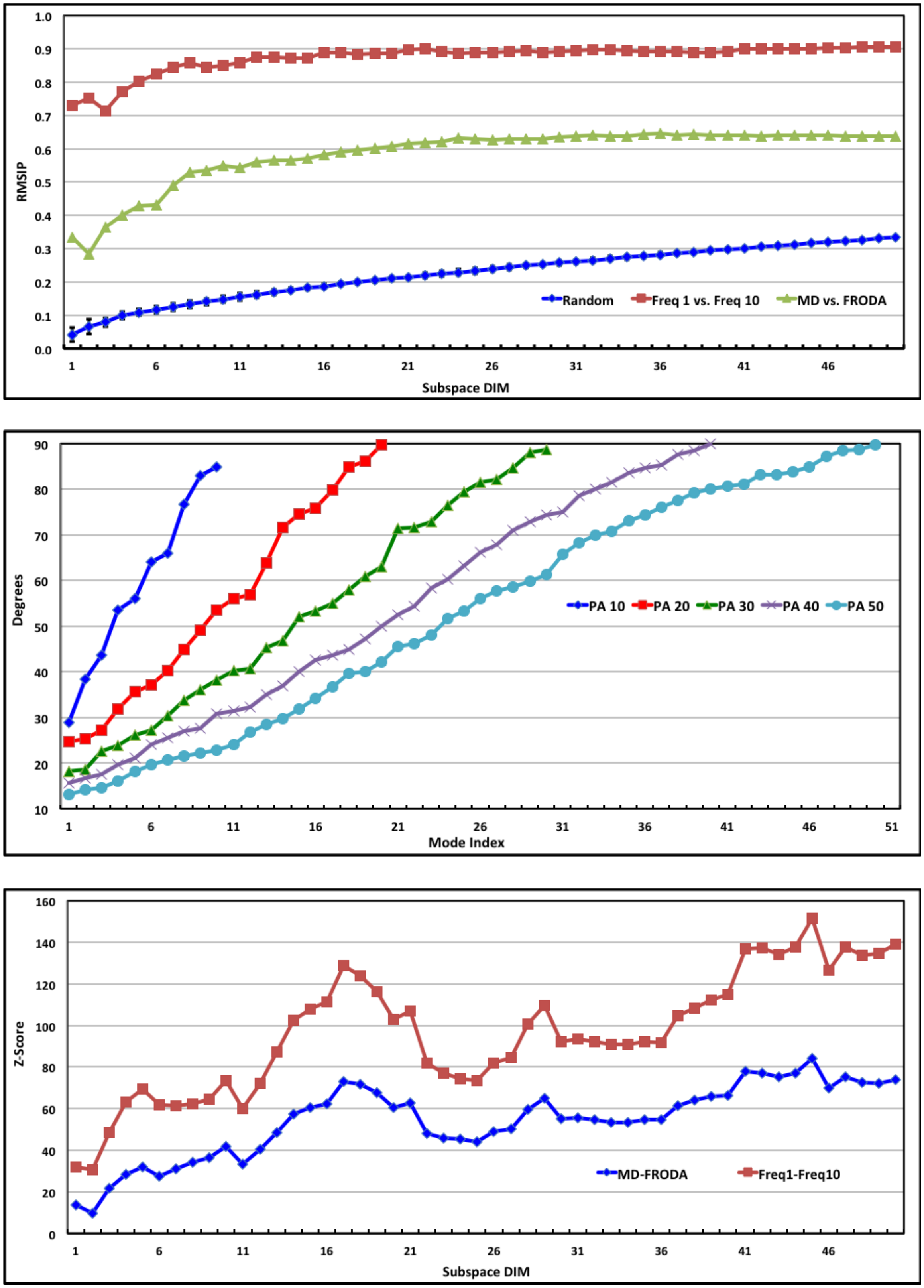


Figure A7: (A) The cumulative overlap of each MD eigenvector with the entire set of FRODA eigenvectors defining the subspace of indicated size. We do not show the

reverse metric, which is not symmetric, but yields similar values. (B) The RMSIP scores for the comparisons of random processes with 453 DOF, two FRODA simulations using the same conditions, and the MD and FRODA simulations. Error bars on the random process scores indicate plus and minus one standard deviation for 50 iterations. (C) The PA spectra for the comparisons of the MD and FRODA simulations using the indicated SS DIM. (D) The Z-scores for the RMSIP scores.

APPENDIX B: RECIPE FOR DISTANCE BASED PCA

1. Obtain trajectories (1 or more) from dynamic simulation.
2. There is no need to remove overall translations and rotations as internal coordinates are being used.
3. Choose the set of atoms.
   - For a set of N atoms, there will be $N(N-1)/2$ modes. It is recommended that less than ten atoms be selected, because otherwise the interpretation of the resulting modes becomes increasingly difficult.
4. Construct an all-to-all distance matrix D for the residue set chosen for each trajectory.
5. Construct the centered data matrix D' by centering the variables (row center).
6. Construct the covariance (or correlation) matrix, Q_D (or R_D), from D'.
7. Diagonalize Q_D (or R_D) using an eigenvalue decomposition.
   - It is best to implement both methods.
8. Examine the eigenvalue scree plot.
9. Select the top set of modes, typically, this is one or two.
   - Each component of the distance PCA modes indicates how the relative distance between a pair of atoms change. There is no way to map the mode components to individual residues.
10. Construct the weighted distance modes.

- Weighting is done by multiplying by the square root of the eigenvalue for the mode, λ_i.

11. Construct the displacement vectors (DV) for the trajectory, given by $DV_i = X_i - X_{ref}$, and construct the Principal Components (PCs).

    - Although there is a physical difference between using internal and Cartesian coordinates, mathematically the same procedures described above in terms of taking inner products and forming projections are identical.

12. When examining two or more sets of PCA modes, determination of how similar the trajectories are to each other may be assessed using the CO, RMSIP or PA metrics.

APPENDIX C: RECIPE FOR KERNEL PCA

1. Obtain trajectories (1 or more) from dynamic simulation. For this example analysis, four proteins from different structural classes were shortened to 75 residues, and all included the N-terminal domain, as a "worst case scenario" of shared behaviors. These proteins were simulated under the same conditions and then subjected to kernel PCA as a combined set with 880 observations.
2. Remove overall translations and rotations by aligning each frame to a reference structure.
3. Select the set of atoms for the analysis to define the data matrix, A.
4. Center the variables of A and row center it to define the data matrix A'.
5. Construct the kernel matrix, K, of {x,y,z} positions for the atoms using A'.
   - The matrix K has dim (n x n) where n is the number of observations.
   - Each element (i, j) in the kernel is determined using a chosen kernel function, which has the general form as $K_{i,j} = K\left(k\left(x_i, x_j\right)\right)_{i,j}$. A linear kernel is given as $K(x,y) = (x \cdot y)$, and a homogeneous polynomial is given by $K(x,y) = (x \cdot y)^d = \left(C_d(x) \cdot C_d(y)\right)$ where C_d maps x to the vector $C_d(x)$ with entries that are all possible n-th degree ordered products of the entries of x. Another kernal type uses a Gaussian weighting function given by

$K(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ where the standard deviation, σ, is an adjustable parameter. A neural net kernel is given as $K(x,y)=\tanh\left(m(x\cdot y)+b\right)$, and a mutual information kernel is given as $K(x,y)=MI(x,y)$ where $MI(X,Y)=\sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$. These are commonly employed kernels in many fields, and are not necessarily particularly useful for protein dynamics. Nevertheless, because higher order correlations in large datasets can be filtered with these kernels, and as such, we have explored all of them.

6. Diagonalize K using an eigenvalue decomposition, and ignore the zero eigenvalues.
7. Examine the scree plot, and from where the kink is, select the top modes.
    - The characteristics of this plot depend heavily on whether one is analyzing fluctuations within a single native basin or is analyzing combined trajectories of multiple states. In kPCA, typically the first few eigenvalues are much larger than the remainder.
8. Determine the eigenvector collectivity. When using kPCA with properly tuned parameters, the top eigenvector often has a collectivity of 0.5 or higher.
9. Select the top set of eigenvectors for forming the kernel principal components (Usually 2-5).
10. Scale the top eigenvectors using the condition $1=\lambda_n(\alpha_n\cdot\alpha_n)$ where α_n is the n$^{\text{th}}$ eigenvector (a column vector) of K and λ_n is the corresponding n$^{\text{th}}$ eigenvalue of K.

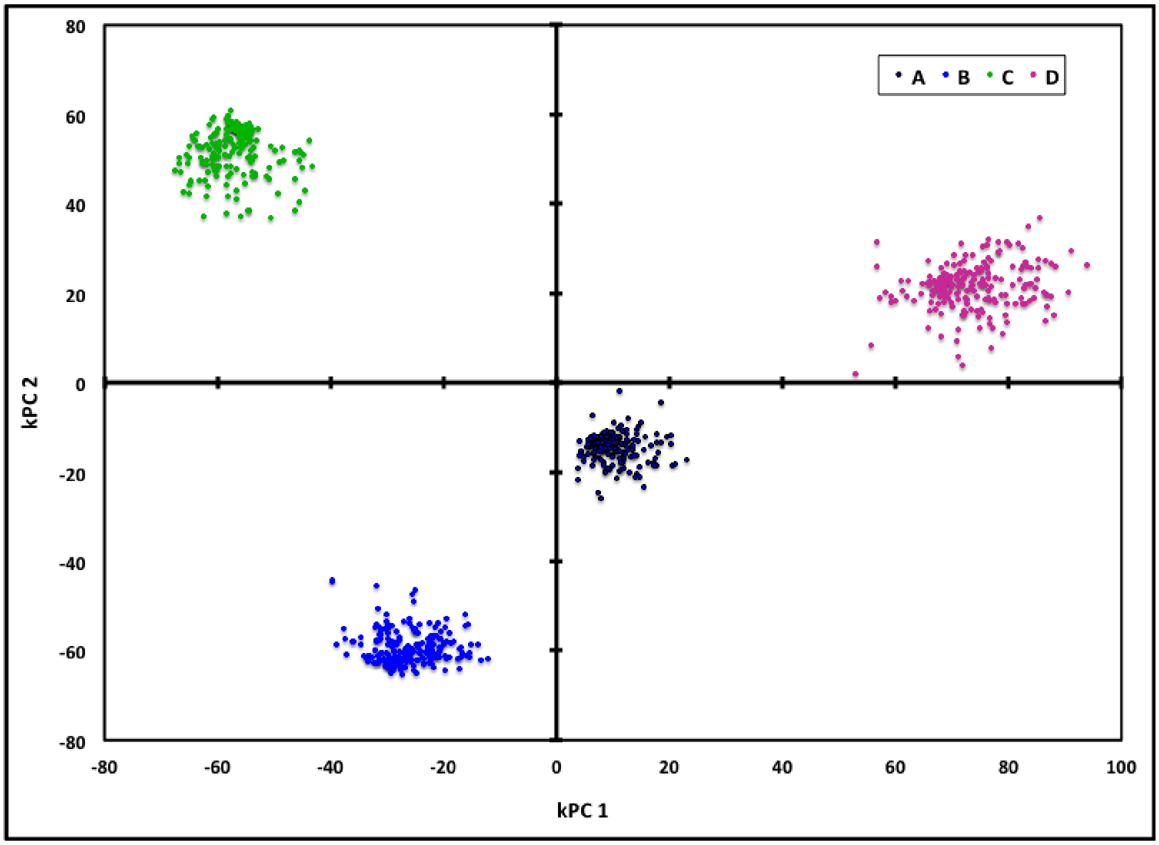
- The eigenvectors are derived from the feature space and usually do not have a meaningful interpretation in the sample space.

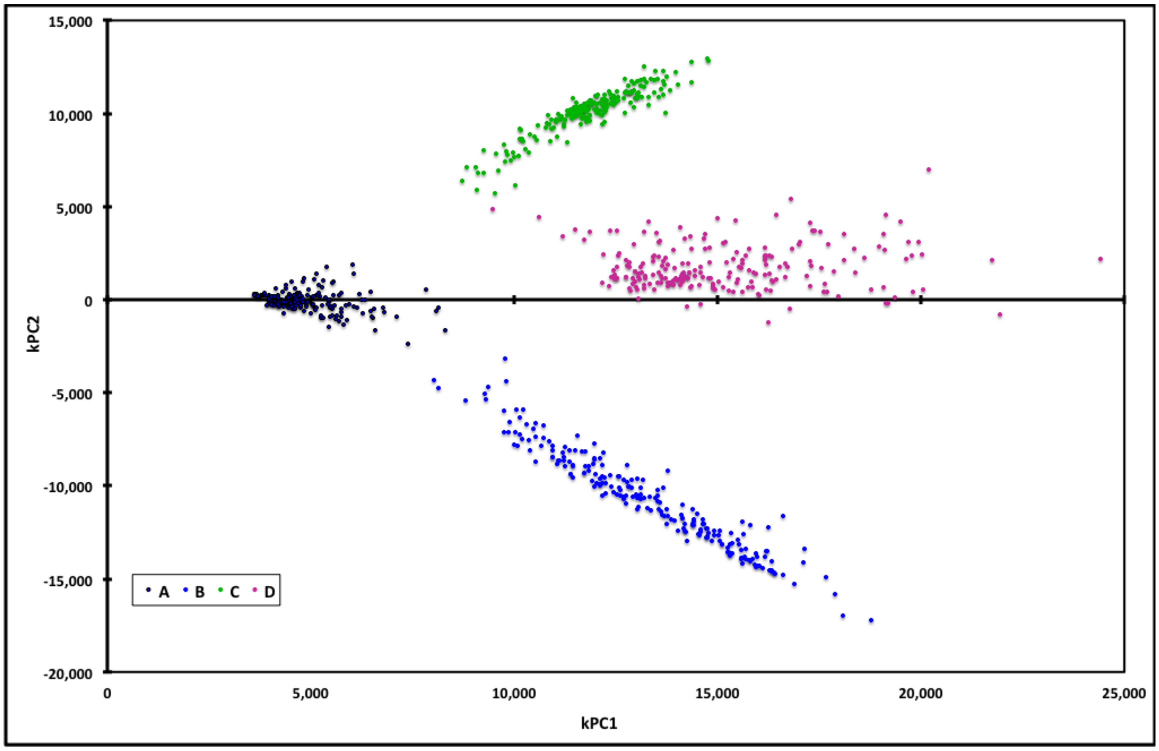
11. Construct the displacement vectors (DV) for the trajectory given by $DV_i = X_i - X_{ref}$, and then construct the Kernel Principal Components (kPCs).

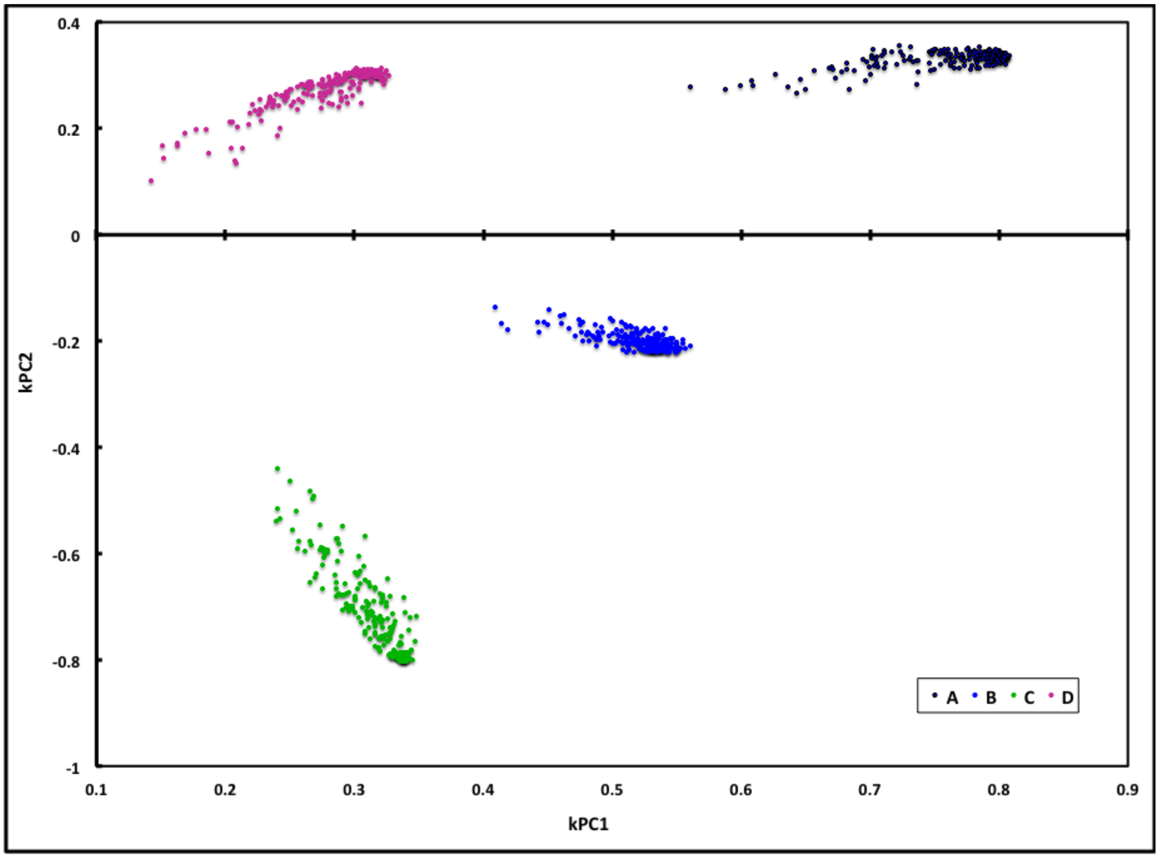
    - Calculate kPC$_n$ using $\left(kPC\right)_n\left(x\right) = \sum_{i=1}^{M} \alpha_i^n k\left(x_i, x\right)$. Note that x is a test vector, and not a training vector (a vector are used to create the kernel). If only the original centered data is to be used, i.e., the data used to construct K, then all the elements of K are already determined. Projections can be made on singe modes to view as line graphs or on 2 PCA modes create scatter plots that show how the simulation explored the configuration space defined by the selected set of modes.
    - We applied PCA and kPCA to the set of four 75 residue proteins to assess the ability of the methods to achieve cluster separation. The results are shown in Figure C1.

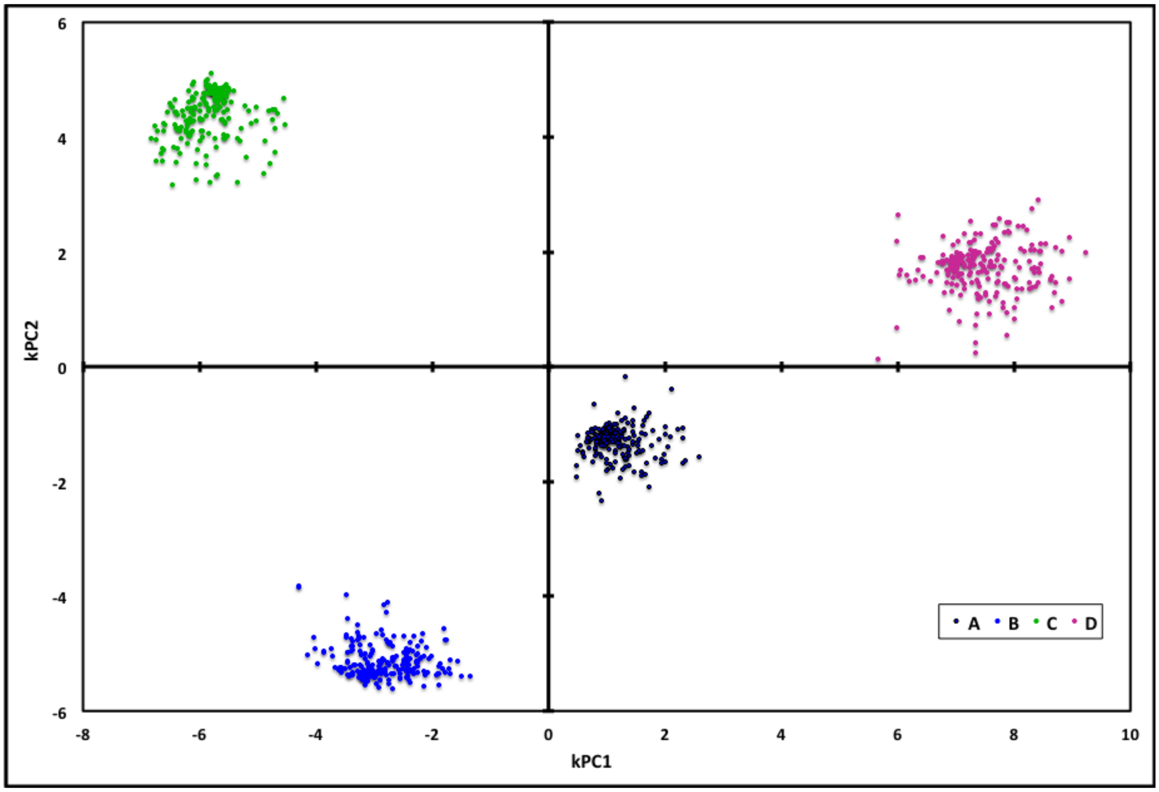
12. When examining two or more sets of kPCA modes, determination of how similar the trajectories are to each other may be assessed using the following metrics. We note that the essential subspaces in kPCA are quite small, comprised of usually 5 or so modes. This is especially true when standard PCA was used as a pre-processing dimensional reduction step. Additionally, subspace comparisons require that the parent vector spaces have the same dimensionality. Therefore, it is possible to compare the essential subspaces derived from different kernels only when the same number of samples are used in each case.

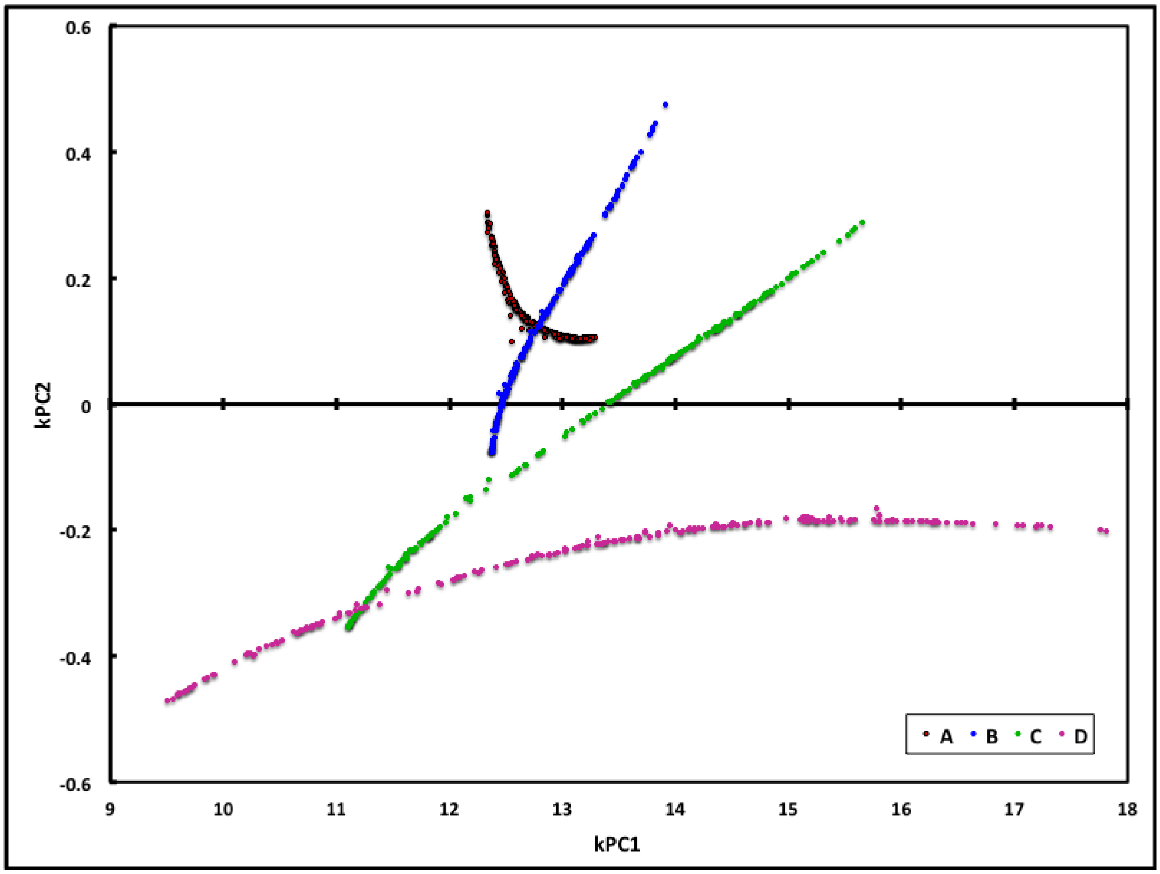
- In Figure C1-F, we show that the subspaces for the top modes generated from the different kPCA approaches are quite similar using the RMSIP scores and the first PA. The most dissimilar was the SS derived from the MI kernel.




15,000
10,000
5,000
0
-5,000
-10,000
-15,000
-20,000
kPC2
0
5,000
10,000
15,000
20,000
25,000
kPC1
• A • B • C • D


0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
kPC2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
kPC1
• A • B • C • D


kPC2
6
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
6
8
10
kPC1
● A ● B ● C ● D


kPC2
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
9
10
11
12
13
14
15
16
17
18
kPC1
● A ● B ● C ● D

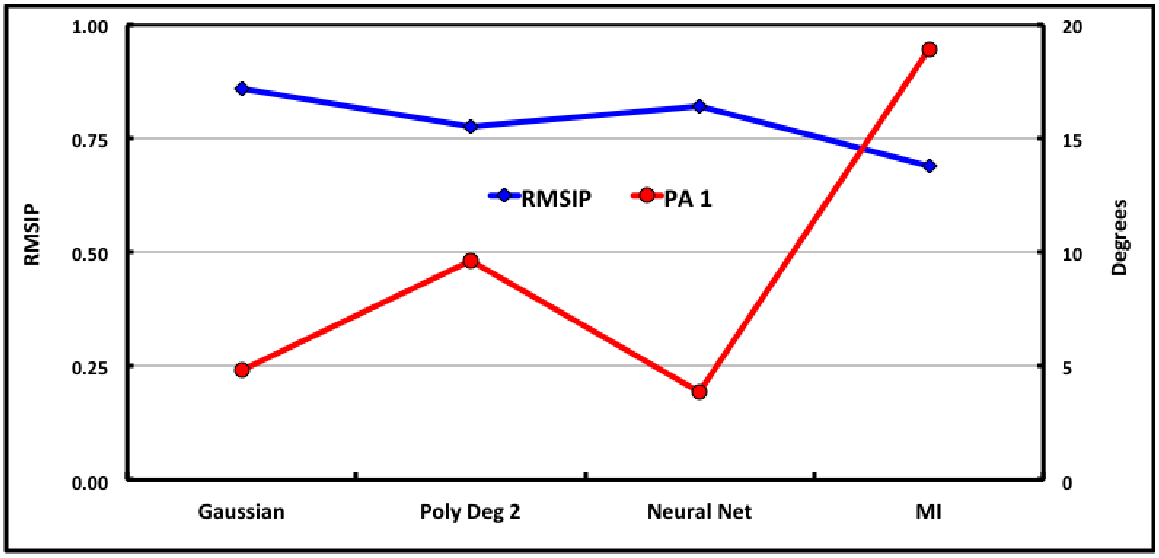


Figure C1: Cluster separation for the dynamics of four different proteins using different kernels, but all using the same CE containing trajectories involving 2000 FRODA frames for each of the four proteins. (A) Linear kernel equivalent to standard PCA. (B) Homogeneous polynomial kernel of degree two, which is sensitive to fourth order statistics. (C) Gaussian kernel with standard deviation set to 50. (D) Neural net kernel with no offset and a slope parameter set to 10^{-4}. (E) Mutual Information kernel. (F) Subspace comparisons of the 4 kernels in B-D using the linear kernel essential space as the reference. The SS DIM in all cases was five.

APPENDIX D: QUATERNION ROTATION ALGORITHM

The user must specify the subset of atoms to use for the analysis. The subset may be all atoms, all alpha carbons, or a subset of either of these. The alignment of a generated structure to a reference structure involves two sets of points, A and B for which we want to find an optimal alignment, defined as the alignment that minimizes the RMSD between each point in A and its corresponding point in B. In this situation, B is the reference structure and A is each successive conformation generated by FRODA. We know which point in A corresponds to which point in B, thus we have the correspondence set. The optimal translation involves superimposing the centroids of the two sets. The optimal rotation can be found as an optimization problem in which a quadratic is extremized. The form is that of the Raleigh Quotient, which tells us that the best RMSD alignment can be achieved by finding the largest eigenvalue of the matrix of the quadratic form.

Here we will use quaternions to carry out the rotation operation:

$$\begin{aligned} \vec{r} &= x\hat{i} + y\hat{j} + z\hat{k} \\ q &= a + b\hat{i} + c\hat{j} + d\hat{k} \\ q^* &= a - b\hat{i} - c\hat{j} - d\hat{k} \\ \vec{r}\,' &= qrq^* \end{aligned}$$

Where r is a vector, q is a quaternion, and q^* is the conjugate of q.

We want to find the rotation on set A that maximizes the sum of the dot products of the rotated vectors of A with the vectors of B, all expressed as offsets from the set of centroids:

$$\sum_{i=1}^{n} q a_i' q^* \circ b_i' = \sum_{i=1}^{n} \left(q a_i' \right) \circ \left(b_i' q \right)$$

$$q a_i' = \begin{pmatrix} 0 & -a_{i,x}' & -a_{i,y}' & -a_{i,z}' \\ a_{i,x}' & 0 & a_{i,z}' & -a_{i,y}' \\ a_{i,y}' & -a_{i,z}' & 0 & a_{i,x}' \\ a_{i,z}' & a_{i,y}' & -a_{i,x}' & 0 \end{pmatrix} q = A' q; \quad b_i' q = \begin{pmatrix} 0 & -b_{i,x}' & -b_{i,y}' & -b_{i,z}' \\ b_{i,x}' & 0 & -b_{i,z}' & b_{i,y}' \\ b_{i,y}' & b_{i,z}' & 0 & -b_{i,x}' \\ b_{i,z}' & -b_{i,y}' & b_{i,x}' & 0 \end{pmatrix} q = B' q$$

$$\sum_{i=1}^{n} \left(A_i' q \right) \circ \left(B_i' q \right) = \sum_{i=1}^{n} q^T A_i^{T'} B_i' \, q = q^T \left(\sum_{i=1}^{n} A_i^{T'} B_i' \right) q = q^T \left(\sum_{i=1}^{n} N_i \right) q = q^T N \, q$$

$$N_i = A_i^{T'} B_i'; \quad N = \sum_{i=1}^{n} N_i$$

$$\det\left(N - \lambda I_4 \right) = 0$$

$\left(N - \lambda_{\max} I_4 \right) \vec{v} = 0$ Where the optimal rotation is found by mapping the

$\vec{v} \to q_{optimal\ rotation}$

eigenvector v to q.

This eigenvalue equation is a quartic in lambda, and is solved for the largest such value.

The Algorithm:

1. Read in the x, y, z coordinates of the points in set A (this is one trajectory structure).
2. Calculate the centroid of set A: $a_c = \frac{1}{n} \sum_{i=1}^{n} a_i$
3. Generate the centered coordinates: $a_i' = a_i - a_c$

4. Repeat Steps 1 – 3 for set B (this is the reference structure and only needs to be done once).

    - The same subset of atoms must be used so as to maintain a proper correspondence set.

5. Calculate the matrices N$_i$ where: $N_i = A_i'^{T} B_i'$ and where we have:

$$A' = \begin{pmatrix} 0 & -a'_{i,x} & -a'_{i,y} & -a'_{i,z} \\ a'_{i,x} & 0 & a'_{i,z} & -a'_{i,y} \\ a'_{i,y} & -a'_{i,z} & 0 & a'_{i,x} \\ a'_{i,z} & a'_{i,y} & -a'_{i,x} & 0 \end{pmatrix}; \quad B' = \begin{pmatrix} 0 & -b'_{i,x} & -b'_{i,y} & -b'_{i,z} \\ b'_{i,x} & 0 & -b'_{i,z} & b'_{i,y} \\ b'_{i,y} & b'_{i,z} & 0 & -b'_{i,x} \\ b'_{i,z} & -b'_{i,y} & b'_{i,x} & 0 \end{pmatrix}$$

6. Calculate the matrix N where: $N = \sum_{i=1}^{n} N_i$

7. Solve: $\det\left(N - \lambda I\right) = 0$ to get the maximum eigenvalue, $\lambda_{\max}$

8. Determine the eigenvector corresponding to $\lambda_{\max}$ by solving: $\left(N - \lambda I\right)\vec{v} = \vec{0}$

9. Generate the quaternion $\bar{q}$ from the eigenvector $\vec{v}$: $\bar{q} = 0 + v_x\hat{i} + v_y\hat{k} + v_z\hat{k}$

10. Generate the set of rotated vectors R using: $r' = q\vec{r}q^*$

    - The set R is the set A after the optimal rotation has been applied.
    - The RMSD of the set R is the least RMSD.

11. Calculate the RMSD of the set R for each structure that is input:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|r'_{i\,ref} - r'_{i\,r}\right|^2}$$

    - This metric is called the conformational RMSD.
    - This is a good measure of how different the FRODA structure is from the original structure.

- With translation and rotation accounted for, the RMSD for the simulation should saturate when the trajectory equilibrates, with changes in RMSD showing as random fluctuations from the maximum value.
- This is especially relevant for simulations based on compositional invariant rigid clusters.

12. Calculate the RMSD for each atom in the transformed vectors of set R over all structures.
    - This metric is called the residue RMSD.
    - This will give a measure of how much each atom moves throughout the simulation.