ANHARMONIC CONFORMATIONAL ANALYSIS OF BIOMOLECULAR SIMULATIONS

by

Akash Parvatikar

B.E. Electronics and Communication, R.V. College of Engineering,

2016

Submitted to the Graduate Faculty of

the School of Computing and Information in partial fulfillment

of the requirements for the degree of

Master of Sciences

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This thesis was presented

by

Akash Parvatikar

It was defended on

April 13th, 2018

and approved by

Dr. S. Chakra Chennubhotla, Associate Professor,

Department of Computational and Systems Biology

Dr. Vladimir I. Zadorozny, Associate Professor,

Department of Informatics and Networked Systems, School of Computing and Information

Dr. Shikhar Uttam, Research Assistant Professor,

Department of Computational and Systems Biology

Thesis Advisor: Dr. S. Chakra Chennubhotla, Associate Professor,

Department of Computational and Systems Biology

ANHARMONIC CONFORMATIONAL ANALYSIS OF BIOMOLECULAR SIMULATIONS

Akash Parvatikar, M.S.

University of Pittsburgh, 2018

Anharmonicity in time-dependent conformational fluctuations is noted to be a key feature of functional dynamics of biomolecules. While anharmonic events are rare, long timescale ($\mu s - ms$ and beyond) simulations facilitate probing of such events. However, automated analysis and visualization of anharmonic events from these long timescale simulations is proving to be a significant bottleneck. Traditional analysis tools for biomolecular simulations have focused on spatial second order statistics. Previous work involved resolving *higher order spatial correlations* through quasi-anharmonic analysis (QAA). In this thesis, we extend this analysis to spatio-temporal domain in the form of anharmonic conformational analysis (ANCA).

We demonstrate ANCA on a publicly available millisecond long trajectory data of the protein Bovine pancreatic trypsin inhibitor (BPTI) using cartesian coordinates of the individual atoms selected for analysis. To overcome the limitation of finding a good reference structure through trajectory alignment, we propose ANCA in the dihedral space to make use of the internal angles derived from the backbone of a fluctuating biomolecule. We test this dihedral angle extension of ANCA on a microsecond long simulation of Drew-Dickerson Dodecamer B-DNA data. Our results indicate that ANCA provides a biophysically meaningful organizational framework for long timescale biomolecular simulations.

We have additionally built a scalable Python package for ANCA, namely pyANCA, with modules that can: (1) measure for anharmonicity in the form of higher order statistics and show its variation as a function of time, (2) output a story board representation of the simula-

tions to identify key anharmonic conformational events, and (3) identify putative anharmonic conformational substates and visualize transitions between these substates. ANCA is available as an open-source Python package under the BSD 3-Clause license. Python tutorial notebooks, documentation and examples can be downloaded from http://csb.pitt.edu/anca.

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of my thesis. Without the support, patience and guidance of the following people, this study would not have been completed. It is to them that I owe my deepest gratitude.

Firstly, I would like to thank my research adviser, Dr. Chakra S. Chennubhotla for inspiring me to understand the science that governs several physical phenomena. He has been my mentor, guide and a dear friend. Secondly, I express sincere thanks to my academic adviser, Dr. Vladimir I. Zadorozny for giving an opportunity to carry out research for Master's thesis. My special thanks to Dr. Shikhar Uttam for encouragement, support and timely guidance during the entire course of research work.

I am also grateful to my colleagues in Computational and Systems Biology department at University of Pittsburgh, Dr. Luong T.H. Nguyen, Dr. Filippo Pullara, Gengkon Lum, Dr. Akif Burak Tosun and Dr. Om Choudhary for their valuable and constructive suggestions during the planning and development of my research.

Dr. Arvind Ramanathan, staff scientist at Oak Ridge National Labs (ORNL) has been an excellent mentor who taught me to be "humble" towards the data. Everyday discussions with him during summer internship at ORNL helped me to develop an intuition to reveal the complexity of protein structures and make biophysically relevant remarks on the simulated data.

I would also like to thank our collaborator, Dr. Pratul K. Agarwal for providing DNA trajectory dataset to carry out alternate analysis. He has been helpful to make me understand the structural aspects of biomolecules and actively contributed towards processing the raw data to be made ready for analysis.

Finally, I would like to thank my family and friends who have willingly helped me in this endeavor. *Amma* for always energizing me with her words. *Appa* for believing in me and encouraging in whatever I would like to pursue. Mayur, my little brother, for all the affection. Priyanka deserves a special thanks. She has always been behind the scenes supporting my passion for science and provided healthy criticism just so that I can do things much better. Jennie, for her love and support.

1.0 INTRODUCTION

1.1 PROTEIN MOTIONS

Proteins are complex macromolecules consisting of single and/or multiple chains of amino acids. They are also called the building blocks of the body and are the second most abundant molecules behind water. Proteins play a critical role in several vital processes that span from tissue repair and maintenance to energy production, and from hormone creation to antibody formation for fighting diseases. The process by which a linear sequence of amino acids fold into a functional three-dimensional structure or conformation is a big open problem in molecular biophysics. A well-known paradigm is that the sequence determines structure and that structure determines function.

Proteins undergo large scale conformational fluctuations due to inter and intra-molecular interactions. Different conformational activities of proteins can be identified at different spatial and temporal scales as illustrated in Figure 1.1. Richard Feynman's statement on the science of 'understanding' being highly correlated to *jiggling* and *wiggling* of atoms has been widely accepted by the scientific community [1]. Due to the recent advancements in experimental techniques and computational modeling, we are beginning to infer the functionality of biomolecules based on the nature of their motions. An emerging paradigm is that the structure leads to dynamics and dynamics determines function.

Let us now investigate the internal proteins motions at different timescales. As shown in Figure 1.1, bond vibrations occur on a femtosecond to picosecond timescale [2]. These vibrations involve a small number of atoms in a spatially narrow region of the protein. These rapid motions involve bond stretching, angle bending, and twisting motions between planes formed by adjacent groups of atoms. Furthermore, at the nanosecond timescale we

can experimentally observe flexible vibrations of loops and rotations of side chains. Only when we go beyond microsecond timescale, the motions involve fluctuations in secondary structure elements comprising of α-helices, β-sheets and flexible loops or changes in the entire shape of a protein. These rather slow motions have attracted significant interest due to their possible linkage to biological function. A collective ensemble of conformational fluctuations involving multiple regions of a protein is more commonly referred to as *breathing motions*. In the following section, we will talk about different experimental and state-of-the-art computational methods that have evolved to measure and interpret protein dynamics.

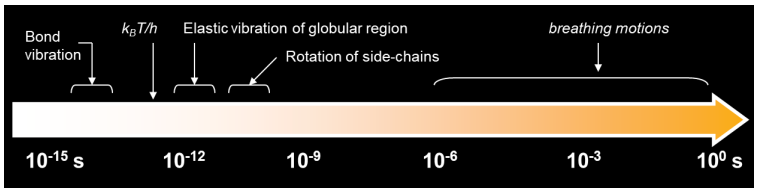


Figure 1.1: **Temporal scales for internal protein motions.** The internal motions of a protein range from femtoseconds ($10^{-15}s$) to microseconds and beyond ($> 10^{-6}s$).

1.2 METHODS TO STUDY PROTEIN MOTIONS

The following sections give an overview of the experimental techniques and molecular dynamics (MD) simulations that can provide insights into the dynamic fluctuations of proteins.

1.2.1 Experimental Techniques

Several experimental techniques have been useful in generating snapshots of protein trajectories and improving our understanding of the protein motions relevant to function. One such technique is to perform neutron-scattering experiments to provide measures of thermal vibrations and their associated frequencies [3]. With this technique fast thermal motions spanning from 1 - 100 picoseconds can be monitored.

Nuclear Magnetic Resonance (NMR) technique has certain merits which allows molecular level analysis of proteins. It has the ability to compute quantitative dynamics at a greater detail and is used to sample protein conformations at nanoseconds and longer timescales [4]. Spin-echo neutron scattering is another technique to probe the motions of macromolecules [5]. It can monitor motions from microsecond to millisecond timescale.

Another addition to the set of experimental techniques is exploring conformational landscape of proteins through hydrogen-deuterium exchange (HDX) monitored by NMR spectroscopy or mass spectrometry [6]. It can measure slow conformational changes in the order of milliseconds. X-ray crystallography has been beneficial to understand the characteristics of protein conformational substates that can relate to the biological function [7].

A common problem with experimental techniques is the inability to effectively analyze protein motions at an atomic scale. One of the reasons being short lifetime and low probability of the energetic conformers. Thus, molecular dynamics (MD) simulations that bridge theory and experiments can provide a fresh path to understand microscopic interactions between sub-units of biomolecules under physiologically relevant environmental conditions.

1.2.2 Molecular Dynamics (MD) simulations

Molecular Dynamics (MD) simulation refers to the science of simulating atomistic scale motions of biomolecules which can provide granular details about the nature of each particle in motion while the molecule is sampling a complex energy landscape [8]. They can be considered as a remarkable tool to explore the underlying geometry and functional roles of the macromolecules. However, the starting three-dimensional structure used by MD is generated from previously mentioned experimental techniques such as X-ray crystallography, NMR spectroscopy or formulated theoretical models. Further, an appropriate molecular mechanics (MM) forcefield is chosen to set the structure in action by modeling a suitable environment and ambient temperature [9]. In order to determine the behavior of this system temporally, we solve Newton's equations of motion at each time step as the molecule traces a rugged potential energy surface while assuming different conformations. This complex energy landscape can be thoroughly analysed to infer biologically relevant motion.

Force fields in MD describe atomic interactions that govern the overall fluctuations of biomolecule. They are commonly known as empirical force fields which are derived from ab-initio quantum mechanical calculations and fit into thermodynamic observations obtained from experiments to describe the dynamics of atoms of interest. Given a force field for a conformation ($\vec{r}$), the potential energy $V_{MM}(\vec{r})$ is given by the equation:

$$V_{\mathrm{MM}}(\vec{r}) = V_b + V_{nb}, \tag{1.1}$$

where, V_b gives bonded interactions and V_{nb} provides non-bonded interactions. The bonded interactions includes sum of three interactions namely V_{stretch}, V_{angle}, V_{dihedral} as reflected by the first three terms of the following equation respectively:

$$V_b = \sum_{i,j} K_b(b - b_o)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_o)^2 + \sum_{\text{dihedrals}} K_\phi[1 - cos(n\phi)]. \tag{1.2}$$

V_{stretch} indicates the harmonic potential of covalently bonded atomic pairs. V_{angle} provides a measurement for a shift of angle from ideal angle θ_o. The last term V_{dihedral} tries to replicate the steric barrier between four atoms separated by three covalent bonds. The corresponding motion associated with the previous term is the rotation along dihedral angle.

The non-bonded interactions given by V_{nb} comprises of two terms defined by the equation:

$$V_{nb} = \sum_{\text{non-bonded}} \left[\frac{A_{ik}}{r_{ik}^{12}} + \frac{C_{ik}}{r_{ik}^{6}}\right] + \sum_{\text{non-bonded}} \left[\frac{q_i q_j}{D r_{ik}}\right]. \tag{1.3}$$

The first term from the above equation describes Lennard-Jones potential giving a sense of van der Waals interactions. A and C are influenced by atom type and estimated through scattering experiments. r_{ik} is a measure of distance between atoms. This is important because van der Waals is heavily involved in attaining conformer's stability [9]. The second term in the equation represents Coulombic forces in which, q_i gives the charge of the atom and D is the dielectric function of the surrounding medium.

The total potential energy function from Eq. 1.2 and Eq. 1.3 is a differentiable equation which gives force acting on atoms which can be further used to integrate Newton's laws of motion. This is the underlying theory behind simulating motions of biomolecules through *molecular dynamics*. It extracts set of conformers/ ensemble for a given time frame. For

example, if a protein is simulated for 1 ms and conformers are sampled every 1 ns, then we can have a dataset with 1,000,000 conformers.

Although MD simulations make atomistic level analysis feasible, it has certain limitations. One such drawback being the effect of insufficient sampling due to all-atom simulations of biomolecules. Structural changes in protein ranges from nanoseconds to milleseconds and even longer. In order to achieve quantitative stability, simulations require time-steps of the range femto-pico seconds ($10^{-15} - 10^{-12}$). Only recently, due to the advancements in parallel computing and custom made hardware types such as Field Programmable Gate Arrays (FPGAs) [10] and Application-Specific Integrated Circuits (ASICS) [11] have tremendously increased simulation speed. Active research for algorithmic improvements is also being pursued by scientists world-wide to enable longer timescale simulations.

1.3 CONFORMATIONAL SUBSTATES OF PROTEIN ENSEMBLE

Proteins are not static structures, but are complex systems that are kinetically active at various spatial and temporal scales [12]. They take the form of several conformations within a short timescale that is difficult to probe. However, in the previous section we discussed about few experimental techniques, X-ray crystallography and Nuclear Magnetic Resonance spectroscopy in particular, which provides us information about three-dimensional organization of atoms that make-up a protein macromolecule and possibly understand the *structure-function* relationship. Emerging evidence claims that, protein conformation in its native/folded state (functioning structure) is not singular, but assumes multiple conformations in relation to the environment encompassing the molecule.

Proteins while undergoing dynamic conformational fluctuations samples multiple minima free-energy landscape that can be imagined to consist of *hills* and *valleys* comprising of certain number of conformers. These undulations can be thought of energy *wells* separated by barriers. Each well consists of certain conformers that share similar properties like internal energies, structure, and others. Such protein structures that have similar properties or conformations within a well is referred to as a *substate*. Transitions between these substates

can be coupled to biological function of the macromolecule as illustrated in Figure 1.2 [13]. Obtaining insights into how a native structure of protein samples free-energy landscape is necessary to provide in-depth description of the functioning mechanism. In order to achieve this, we should address the challenge of precise characterization of substates and extract conformers within these substates and later derive knowledge about pathway of protein function.

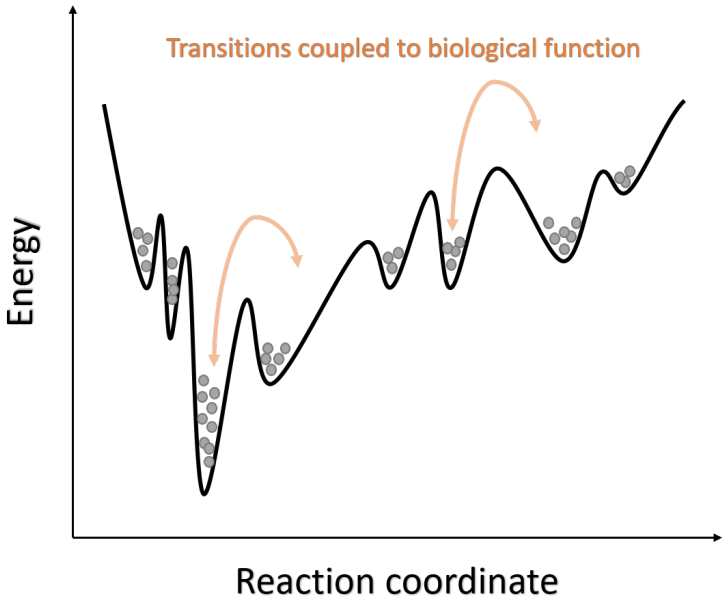


Figure 1.2: **Energy landscape of protein conformations.**

Experimental data obtained from X-ray crystallography and NMR spectroscopy doesn't reveal details that is necessary to relate transition between substates to the function. While X-ray crystallography performed at ambient temperature characterizes conformational ensembles, it fails to give quantitative information about infrequent states and transition between them [14]. Alternately, NMR spectroscopy requires specific inputs such as number of substates. Due to these limitations, and growing computation power, atomistic scale MD simulations has been instrumental in efficient characterization of substates [15]. Collection

of conformations from MD simulation is referred to as an *ensemble*.

1.4 OUTLINE OF THE THESIS

In this thesis, we lay emphasis on the statistical analysis of conformational fluctuations that are important for biological function. As a means to achieve this goal, we have developed a toolbox to address the following two broad questions:

- How to characterize protein dynamics by considering spatio-temporal information?
- How to identify putative conformational substates and visualize the transition between them?

The first chapter of this thesis motivates the readers to understand the relationship between *structure-motion-function* of biomolecules. Second chapter illustrates the analysis of protein motions using a publicly available MD simulation data of *Bovine pancreatic trypsin inhibitor* (BPTI). We account for the non-Gaussian behavior of trajectory fluctuations and motivate the idea of chasing high-order statistics to resolve spatio-temporal correlations. In the third chapter, we provide an alternate analysis of macro-molecules by considering dihedral angles and elucidate its advantage over using Cartesian coordinate system which is considered in the previous chapter. In the final chapter, we describe the open-source Python package pyANCA that serves as a toolkit to quantitatively analyze biomolecular simulations.

2.0 SPATIAL AND TEMPORAL CORRELATION ANALYSIS TOOLS

Molecular dynamics (MD) simulation generates data which are highly correlated in spatial and temporal domain. Analysis of protein motions obtained from such simulations is observed to exhibit anharmonicity. This non-Gaussian behavior can be attributed to the dense conformational fluctuations as a result of constant interplay with surrounding environment (solvent, ions and other proteins) and the inter-residue force fields within the globule. Among protein fluctuations observed at different temporal resolution, the motions happening beyond micro-second (10^{-6}) time-scales is significant due to it's possible role in protein function. Such movements are often referred to as *breathing motions*. In this chapter, we discuss several tools to resolve spatial (SD2 and SD4) and temporal dependencies (TD2 and TD4) using second and fourth-order statistics to enable characterizing long time-scale protein fluctuations.

2.1 INTRODUCTION

Anharmonicity in time-dependent conformational fluctuations is noted to be a key feature of functional dynamics of biomolecules [16]. In this chapter, the primary focus is to resolve spatial and temporal dependencies observed from the analysis of molecular data. Previously, much of the work has been focussed on applying spatial statistics. Traditional tools such as Principal component analysis (PCA) [17] and Quasi-anharmonic analysis (QAA) [18] has been successfully tested to spatially resolve data in second and fourth-order. However, the correlations in temporal domain also needs to be resolved to improve probing of anharmonic behavior. In this chapter, we discuss several tools to decorrelate MD simulated milli-second

long trajectory data of the small globule Bovine pancreatic trypsin inhibitor (BPTI).

A data matrix (X_{orig}) is constructed with 412497 observations (conformers) spanned across 1.031 milli-seconds for the 58 residues of BPTI protein which was generated on the Anton supercomputer by D.E. Shaw Research group. Data is organized as $3N \times t$ matrix, where $3N$ represents *(x,y,z)* coordinates from individual atom selections and t represents conformations. The large matrix is passed through decorrelation modules (SD2, SD4, TD2 and TD4) for resolving anharmonic dependencies. Before supplying trajectory coordinates to these modules, the data is structurally aligned.

Understanding the flexible nature of proteins is essential to obtain insights about its activity. Different experimental conditions can generate conformations which are not identical. Thus, protein structures needs to be superimposed for identifying functionally relevant protein domains [19]. In order to achieve this, firstly rigid and flexible residues are identified through application of Gaussian-weighted RMSD superpositions. It translates and rotates the structures to minimize the arithmetic mean of positions of atoms in subsequent structures, that is, reducing root mean square deviations (RMSD) of ensembles. Rigid residues are used as an underlying structure to iteratively align the MD ensemble. After alignment, the data is processed through different modules to obtain uncorrelated components to gain an understanding of protein dynamics.

2.2 BOVINE PANCREATIC TRYPSIN INHIBITOR (BPTI): STRUCTURE AND STATISTICAL INSIGHTS INTO THE NATURE OF BIO-MOLECULE

Bovine pancreatic trypsin inhibitor is a three-dimensional small globular protein which forms the drug *Aprotinin* that suppresses the action of protein digestion. It is one of the most extensively studied protein whose structure was determined with a high resolution of 1.9 $\mathring{A}$ [20]. BPTI is also the first protein to be analysed experimentally through NMR spectroscopy in the late 20th century in Kurt Wuthrich's lab [21]. Due to its relatively stable structure comprising of 58 amino acids, BPTI was the first macromolecule to be simulated using

Molecular Dynamics (MD) simulation by Karplus and group [22].

BPTI performs its function of suppressing protein digestion by breaking down the macromolecules into their respective peptide blocks through restraining the action of trypsin enzyme produced in the bovine pancreas. The enzyme was first isolated independently in 1936 from active pancreatic extract [23]. Apart from studying the protein's structure, both experimentally and computationally, extensive research has been carried out to understand the dynamics and develop insights about its folding pathway. Due to its innate nature of forming complexes with several other enzymes, it has been chosen as an example study for investigating protein-protein interactions. In the following section, we will try to visualize BPTI's structure and dynamics by employing statistical measures.

As illustrated in Figure 2.1, the BPTI protein comprises of 58 amino acid residues. Amino acids are the monomers which are bonded together to form multiple chains of polypeptides collectively termed as proteins. BPTI takes the form of a tertiary fold comprising of two anti-parallel β sheets and short segments of two α helices.

2.2.1 Statistical insights of trajectory fluctuations

In our studies, we used the one millisecond trajectory of BPTI generated by D.E. Shaw Research group on the Anton supercomputer [24]. For simplicity and efficiency, we consider only the backbone atoms (C^{α}) over 1.1 ms trajectory generating about 412497 conformers. To complement insights from harmonic measures of conformational changes, such as the root-mean squared deviation (RMSD), we have used higher-order anharmonic measures, namely kurtosis (κ) [15]. κ measures the *peakiness* of the probability distribution of a random variable. κ is calculated from either the Cartesian coordinates or dihedral angle selections specified by the user. For unimodal distributions, κ quantifies the proportion of weights on the tails. A distribution with $\kappa = 3$ is called mesokurtotic or mesokurtic. A Gaussian distribution with zero mean, unit variance is mesokurtotic. A value of $\kappa > 3$ indicates a super-Gaussian distribution that is more peaked and heavier tailed than the baseline Gaussian. Such a distribution which has positive excess kurtosis is called leptokurtic

Figure 2.1: **Three-dimensional conformation of BPTI.** The cartoon representation of BPTI has been created using VMD software.

or leptokurtotic. Conversely, a distribution that is less peaked than the baseline Gaussian has kurtosis $\kappa < 3$ which is generally referred to as sub-Gaussian. Such a distribution with negative excess kurtosis is termed as platykurtic or platykurtotic. The statistical significance of κ is assessed through the kurtosis test, which rejects the hyposthesis of normality when the p-value < 0.005. Using κ, we quantify which parts of the protein exhibit anharmonic motions as illustrated in Figure 2.2 and for how long as seen in Figure 2.3. In the case of BPTI, we can observe that a majority of the C^{α} atoms spend at least 5% of their time exhibiting anharmonic motions. However, the interface formed by helices 1 and 2 is mostly

harmonic, because of the strong hydrophobic interactions and Cys-disulfide bonds.

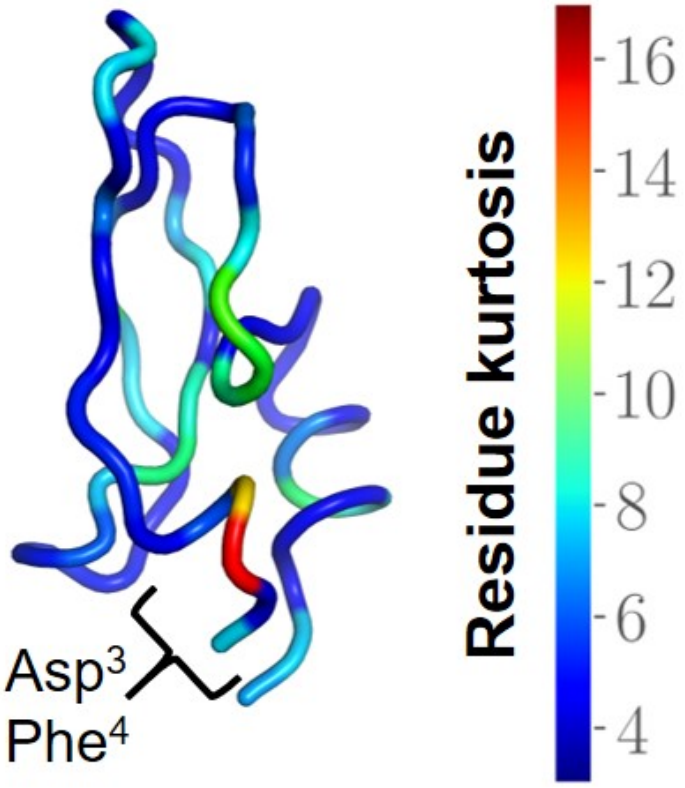


Figure 2.2: **Residues painted by individual kurtosis (κ) values.** Two residues $Asp^3 - Phe^4$, show the largest κ values while sampling anharmonic motions infrequently.

Further, we analyzed the variation of κ at each C^{α} coordinate (x,y,z), using an exponential sliding window with a half-life of $1\mu s$ from the trajectory [18]. Almost all the individual residues exhibit some degree of anharmonicity, while κ is more pronounced along individual coordinate directions. These conformational changes constitute events within the trajectory that may be of interest to the user for further analysis. Using κ, the user can identify conformational events that occur at distinct timescales and organize a conformational storyboard for the entire simulation(s). Figure 2.4 shows the variation of kurtosis over time using an

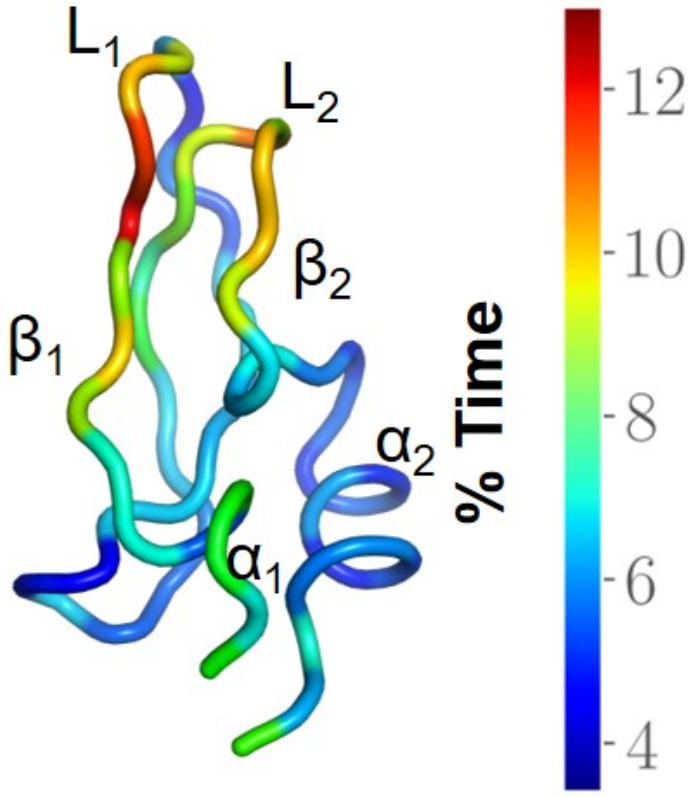


Figure 2.3: **Residues are colored by the time spent sampling anharmonic fluctuations.**

exponential window with a half-life of $1\mu s$. The filtering procedure is described in detail in [18]. Using a user-defined threshold (green line in Figure 2.4), a total of 17 conformational events are detected (labeled $E_1 - E_{17}$). Select events from this are organized as a story board in Figure 2.5.

These events summarize the time-points at which the BPTI loops L_1 and L_2 open/close.

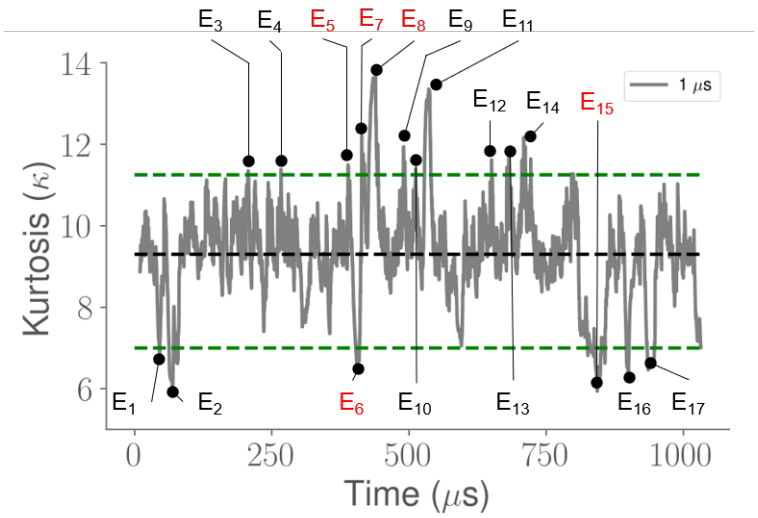


Figure 2.4: **Time evolution of kurtosis (κ) values seen through an exponential sliding window of 1 μs half-life.** Using a threshold of four standard deviations (green dotted lines) above and below the mean κ (black dotted line) identifies a total of 17 conformational events labeled $E_1 - E_{17}$.

The storyboard provides a means to quickly summarize large MD trajectories, while allowing the user to visually interact with events of interest and simultaneously track other quantities of interest (e.g. $RMSD, R_g$, etc) over the course of long simulations. In addition to using κ, conformational events can be detected with information theoretic measures such as mutual information [25]. However, these techniques are compuutationally expensive. Trajectory segments from the story board can be further analyzed to identify putative conformational substates.

As an additional feature, we also provide storyboards to analyze the fluctuations of individual residues as shown in Figure 2.6.

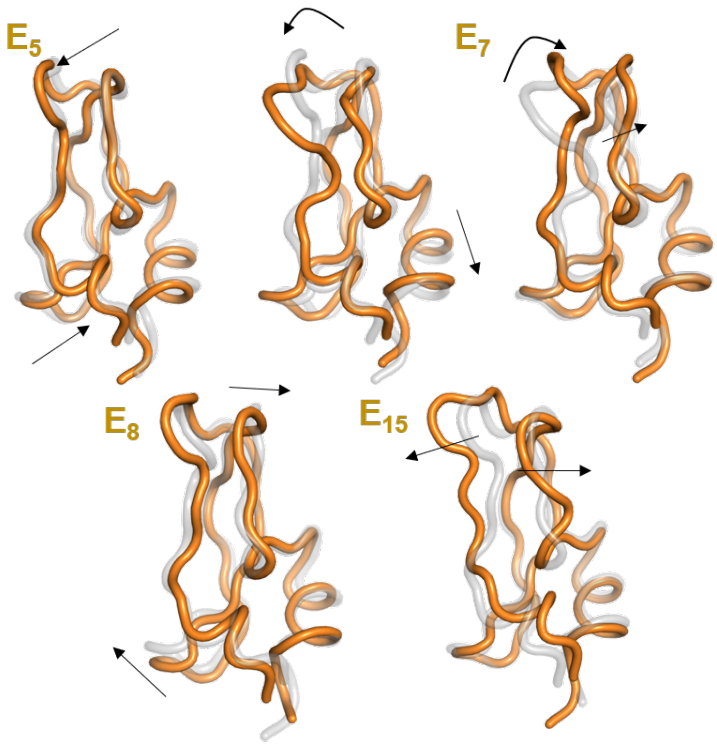
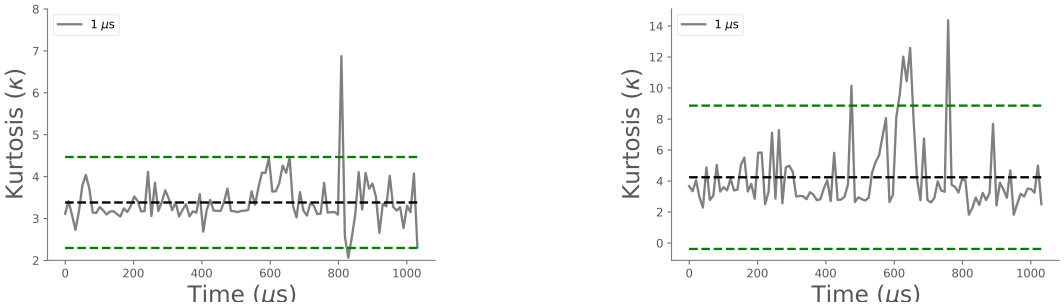


Figure 2.5: **Five selected events $E_5, E_6, E_7, E_8 and E_{15}$ as ensembles, with gray cartoon representing the previous event and the orange cartoon representing the current event. Arrows are used to highlight the opening/ closing of the flap regions of BPTI in each event.**

2.3 SPATIAL DECORRELATION IN SECOND-ORDER (SD2)

SD2 module exploits the established methods of PCA to remove dominant second-order spatial correlations. The practical application for smart computing through PCA was first proposed by H. Hotelling [26]. PCA tries to reduce a larger dimensional dataset of correlated variables into small number of transformed uncorrelated variables. The underlying assumption in applying PCA lies in the assumption that described observations can be explained by variances and computed covariances of data matrix. While considering the analysis of MD simulation data, PCA provides visualization of residues that are harmonically resolved.



(a) Asp-3 residue fluctuations

(b) Phe-4 residue fluctuations

Figure 2.6: **Residue-based event detection for Asp-3 and Phe-4 in BPTI using kurtosis measured with an exponential sliding window of 1 μs half-life, captures significant conformational changes.**

2.3.1 Performing second-order decorrelation in space

Prior to resolving spatial dependencies through SD2 module, data needs to be centered by subtracting mean of positional observations for each residue. The mean removed data matrix (X_{orig}) undergoes transformation to compute a spatial covariance matrix followed by principal component analysis. The covariance matrix is defined by:

$$C_y = \langle X_{orig}(t) X_{orig}(t)^T \rangle \tag{2.1}$$

Principal component analysis is then performed by doing an eigenvalue decomposition of the covariance matrix:

$$C_y = U \; \Sigma \; U^T, \tag{2.2}$$

where, Σ and U indicate eigenvalues and eigenvectors respectively. Eigenvectors are an indication of the direction of dominant atomic displacements, whereas eigenvalues provide a numerical measure to assess frequency of corresponding fluctuations. The principal components (PCs) provides a measure to describe protein conformational distribution along the dominant eigenvectors. The eigenvalues are typically arranged in descending order such that i^{th} value corresponds to i^{th} eigenvector also referred to as *loadings* of principal components.

Low eigenvalues correspond to large-scale global motions (slow modes) whereas large values reflect frequently occurring localized protein motions (fast modes) [27]. Significant contribution of slow modes in relevance to protein function has been a motivation factor to choose a lower dimensional subspace. The subspace number during protein analysis indicates a number which explains for the most variance in the dataset. We observe from Figure 2.7 that, each eigenvalue index doesn't contribute equally to the amount of variance in the data. This lower value of subspace m is determined by the inflection point observed in the scree plot as illustrated in Figure 2.7.

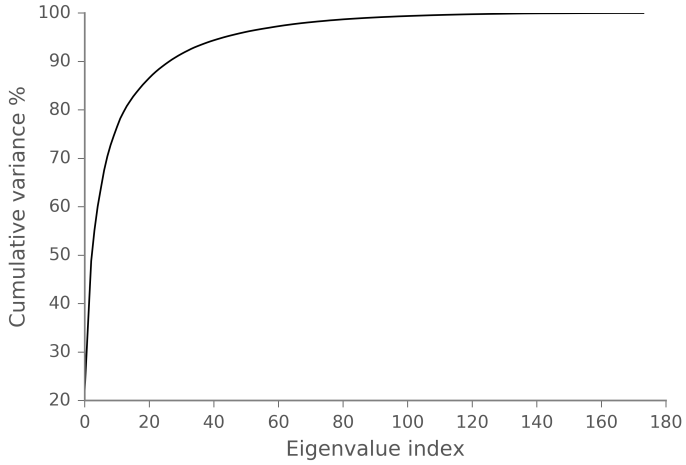


Figure 2.7: **Scree plot of BPTI data.** An illustration of the trend in cumulative variance of the motions with respect to the eigen value index. 28 PCA modes is sufficient to explain 90.38% of the total variance. This helps us to choose the value of subspace m that quantifies the amount of variance preserved.

SD2 module removes dominant second-order spatial correlations. The function diagonalizes covariance matrix R_y to obtain the projection matrix $Y = B^T X_{orig}$ $(3N \times t)$, where m is subspace dimensionality and B $(3N \times 3N)$ are the dominant eigenvectors. PCA although holds merit in finding projection vectors with maximal variance, it fails to differentiate between slow vibrational modes since large variance need not account for the temporal cor-

relations that exists in the molecular data. Hence, we test the 2^{nd} order spatially resolved BPTI data matrix Y to spatially resolve in fourth-order (SD4) to find components which are energetically separate.

In this section, we tried to understand about doing PCA on high-dimensional molecular data, its implications and drawbacks. In the following section, we will examine projections of the modes obtained from SD2 and test if that alone is sufficient to characterize the conformational landscape of a millisecond long MD simulated molecular data.

2.3.2 SD2 modes of motion in BPTI

The BPTI protein data used for analysis has 58 residues (N) which is sampled over 1.1 ms generating 412497 conformers considering only the backbone atoms. The 178 observations ($3 \times$ N) is projected onto the rotation matrix B to obtain a second-order spatially resolved data Y. Further, for the sake of visualization we have considered top 3 dominant SD2 modes as seen in Figure 2.8.

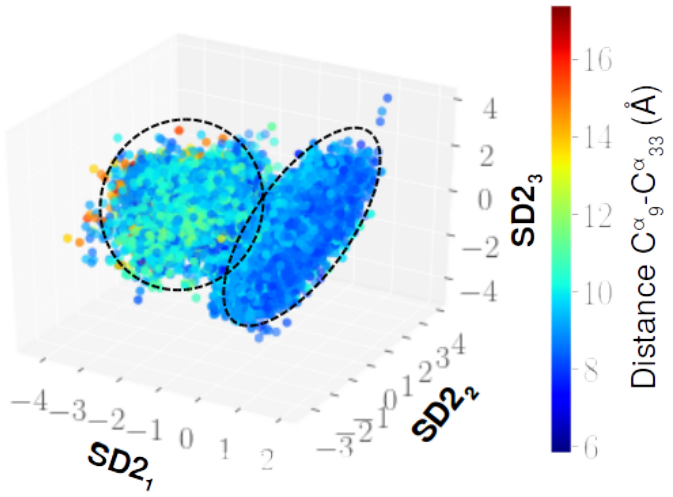


Figure 2.8: **SD2: removing second-order spatial correlations using principal component analysis (PCA).** Interestingly, the top three modes of SD2 capture the open and closed state of BPTI, it is not able to identify the separation between the two flap regions formed by L1 and L2 which is represented in Figure 2.3. The movie like representations in Figure 2.9. and the supporting movies highlight the differences in the motions captured by SD2.

SD2 reveals two clusters after projecting the BPTI data onto top 3 dominant modes to collectively characterize the atomistic fluctuations as a result of removing any second-order spatial correlations that might exist in the conformational ensemble. Each conformation is represented by a dot in the 3-dimensional scatter plot as seen in Figure 2.8. To quantify the motions, we use a reaction coordinate based on the distances between residues Proline-9 (Pro^{9}) and Phenylalanine-33 (Phe^{33}). Further, movie like representation can be seen from Figure 2.9 in order to gain biophysical insights about the data after performing SD2.

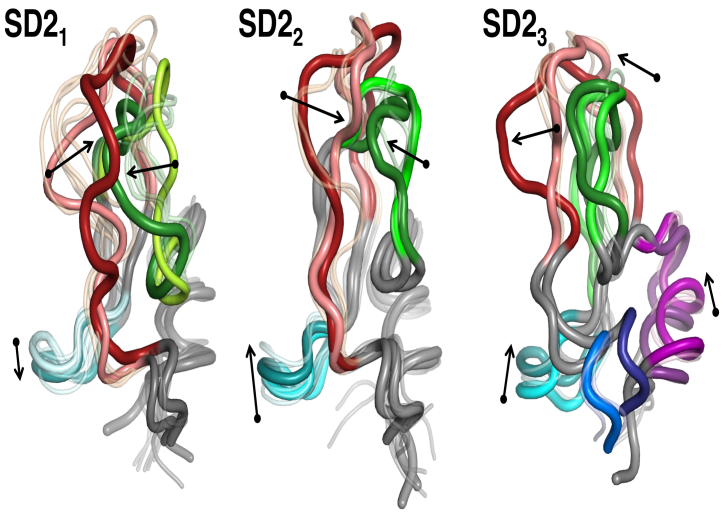


Figure 2.9: **The three principal modes of motion determined from the SD2 analysis of the BPTI simulations** ($SD2_1 - SD2_3$) **shown in a movie like representation.** The light to dark transitions indicate direction of motion for the region in BPTI that are highlighted in different colors along with arrows to depict the start and end-state of the protein. Notably, in all the three modes, the flap regions (L_1 shown in salmon-red transition and L_2 shown in light-green to green transition) move in a concerted fashion towards each other ($SD2_1$ and $SD2_2$) or move together in an outward manner ($SD2_3$). Additional fluctuations are visible between the $\beta_1 - \beta_2$ loop (shown in cyan). In $SD2_3$, the displacements are larger, depicting displacements in both α_1 (blue) and α_2 (purple).

Notably, SD2 doesn't characterize anharmonic fluctuations in conformational levels and fails to separate out the conformational landscape based on its energetic homogeneity. Due to the high intrinsic dimensionality of the data obtained from MD simulations and its anharmonic behavior, we progress towards higher order statistics and also deal with temporal correlations later in this chapter.

2.4 SPATIAL DECORRELATION IN FOURTH-ORDER (SD4)

While observing the probability distribution of positional deviations of atomistic fluctuations of BPTI data, it is evident that it exhibits a non-Gaussian behavior as seen from Figure 2.10. This anharmonic (non-Gaussian) behavior observed from the long-tails cannot be resolved by SD2. Thus, in order to characterize such motions, we make use of the fourth-order moment, kurtosis (κ), that measures the peakiness of the distribution. For a Gaussian distribution, the value of kurtosis (κ) is equal to 3.0.

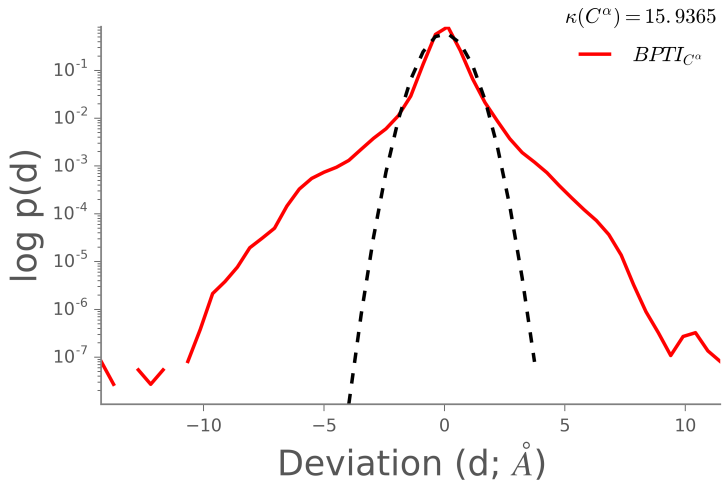


Figure 2.10: **Positional deviation histogram reveals statistical diversity.** Non- Gaussian (anharmonic) behavior is observed from the 1.1 millisecond MD simulated BPTI data. The overall fourth-order moment kurtosis (κ) is equal to 15.94. The value of κ for a Gaussian distribution is equal to 3.0.

2.4.1 Performing fourth-order decorrelation in space

In order to characterize atomic fluctuations in terms of its anharmonic behavior, we have developed SD4 module. This function attempts to resolve the intrinsic non-orthogonal dependencies in atomistic fluctuations by reducing the fourth moment. The second-order projections from SD2 are used to build a fourth-order spatially correlated cumulant tensor. SD4 approximately diagonalizes this tensor to return an anharmonic mode matrix.

The covariance matrix R_y obtained from SD2 module captures correlations in atomic fluctuations derived from second-order with an assumption that the basis vectors be orthogonal. However, since we are dealing with trajectories with non-orthogonal basis vectors, SD4 uses independent component analysis [28] as a technique to extract anharmonic sources $\vec{x}$, such that:

$$\vec{s} = W\vec{x}. \tag{2.3}$$

In this equation, W represents separating matrix which describes fourth-order correlations between different residues of BPTI protein. The anharmonic modes can be quantified through:

$$\vec{x} = W^{-1}\vec{s}. \tag{2.4}$$

In order to derive the basis matrix W which can comprise of non-orthogonality, we compute fourth-order cumulant tensor given by the equation:

$$Q_{ij} = E\left\{Y^4\right\} - 3E^2\left\{Y^2\right\}, \tag{2.5}$$

where, Y is the second-order spatially resolved matrix obtained from SD2 module and Q_{ij} $\in \mathbb{R}^{m\times(m\times k)}$, where $k = [m \times (m+1)]/2$ is the generalized cumulant matrix. This higher order matrix is efficiently diagonalized using methods of Jacobian rotations [29] to compute a rotation matrix G. The rotation matrix G is further transformed to calculate the separating matrix W that resolves data in fourth-order. The spatially decorrelated matrix of fourth-order is computed by obtaining:

$$Z_{SD4} = WX_{orig}, \tag{2.6}$$

where W attempts to separate sources from signal mixture X_{orig} by finding directions, such that projections onto these directions have maximum statistical independence. The computed parameter Z_{SD4} is fourth-order spatially resolved matrix.

To summarize, we describe anharmonic modes of motion by performing matrix operations considering fourth-order statistics. In the following section, emphasis is laid on analyzing

fourth-order spatially resolved data through projections and if the spatial resolution of data is sufficient to group the conformations painted with *reaction coordinates* do exhibit similar properties.

2.4.2 SD4 modes of motion in BPTI

SD4 modes are statistically ordered based on the kurtosis of the projected coordinates. This strategy used for constructing the projected conformational space may not always reveal biophysically meaningful information. However, in pursuit of establishing frameworks to identify certain critical events that occur sparsely during the motion of proteins, we have built models that support to discover rarely occuring important events. As a method to build associations from SD4 modes, user can choose desired physical observables such as radius of gyration (R_g), scaled internal energies, or pairwise distance between residues. We continue to choose the residue distances between Pro^{9} and Phe^{33} to build the SD4 conformational landscape as seen in Figure 2.11.

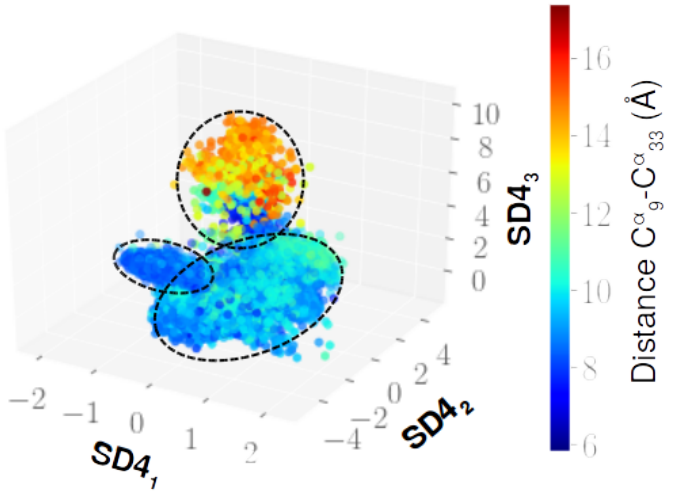


Figure 2.11: **SD4: resolving fourth-order spatial correlations by minimizing kurtosis.** The top three modes from SD4 identify the separation of the two loops clearly identifying a putative conformational substate showing an open state of BPTI.

SD4 performs better to group the conformations corresponding to the value of residual distance that can be visualized from three clusters as seen in Figure 2.11. This helps to identify putative conformational substates that could be biophysically relevant. The movie like representations help to visualize the spatial anharmonic modes through cartoon like representation of BPTI. By looking at the motions captured by $SD4_3$, we can clearly see an increase in the distance between flap regions of BPTI which is illustrated in Figure 2.12.

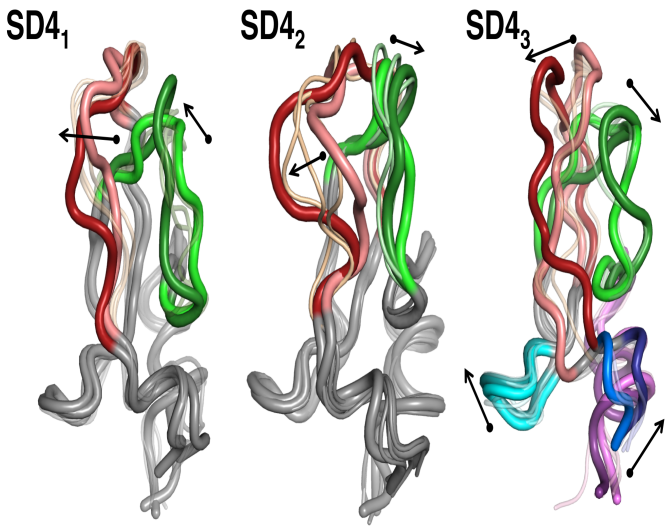


Figure 2.12: **The three principal modes of motion determined from the SD4 analysis of the BPTI simulations ($SD4_1$ - $SD4_3$) shown in a movie like representation.** The light to dark transitions are representative of the conformations within BPTI similar to Figure 2.3. Arrows indicate the predominant directions of the motions. $SD4_1$ identifies motions that enable the two flap regions (L_1 and L_2) to move together in a concerted fashion leading to a closure of this region (i.e., both flaps come together closer). $SD4_2$ describes motions that enable the flap regions to move apart from each other, however capturing only an intermediate separation between the flap regions where by L_2 stays relatively stable compared to L_1, which moves further apart. $SD4_3$ describes a motion which allows both the flap regions L_1 and L_2 to separate, while simultaneously displacing the α_1(blue) and α_2(purple) helices.

2.5 TEMPORAL DECORRELATION IN SECOND-ORDER (TD2)

The trajectories extracted as a result of performing MD simulations are correlated at various spatial and temporal scales [30]. In order to characterize the conformational fluctuations, it's necessary to be able to collectively summarize the spatio-temporal patterns with which the protein residues are interactively moving. While SD2 module is beneficial to obtain directions of dominant variance, no information can be derived about the temporality of the simulation.

2.5.1 Performing second-order decorrelation in time

Temporal decorrelation (TD2) is performed on a second-order spatially resolved data Y to find maximal autocorrelation for a given parameter *lag time*. This function removes dominant second-order temporal correlations by computing a time-delayed (specified by a lag time τ) covariance matrix and performing PCA. The time-lagged covariance matrix C_z is given by the equation:

$$C_z(\tau) = \langle Y(t)Y(t-\tau)^T \rangle. \tag{2.7}$$

The possibility of computational errors such as round-off errors can destroy the symmetricity of the covariance matrix. In order to make sure that the matrix is symmetric, a mathematical computation is done:

$$C_z(\tau) = \frac{1}{2}\left[C_z(\tau) + C_z(\tau)^T\right]. \tag{2.8}$$

An eigenvalue decomposition is performed over this time-lagged symmetrized covariance matrix,

$$C_z(\tau) = U_{td2}\ \Sigma_{lag}\ U_{td2}^T, \tag{2.9}$$

to obtain spatially whitened and temporally decorrelated data. A matrix Z is obtained by projecting the spatially resolved data matrix Y onto the dominant eigenvectors B_{TD2}. The eigendecomposition yields values for autocorrelations Σ_{lag} and temporal principal components given by the eigenvectors U_{td2}.

2.5.2 Harmonic double-well experiment

In order to estimate slow order parameters which might be of interest to analyze fluctuations from biomolecular simulations, SD2 yields misleading results since it only finds directions along which we can observe high variation. However, we are interested to find correlations between same variables at a different time instance that occurs due to dependencies within the data. TD2 precisely performs the operation of removing autocorrelations of the time series data to obtain statistically independent components that can be classified to have homogenous properties. To demonstrate the use of two approaches, i.e. passing the data into SD2 and TD2 module, we generated a Hidden-Markov model (HMM) to obtain a two-dimensional Gaussian data $X_{experimental}$ (x) with zero mean (μ) and unit variance (σ). A Gaussian function is given by:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-} \frac{(x-\mu)^2}{2\sigma^2}. \tag{2.10}$$

In our experiment, we generated data from 50000 samples having two degrees of freedom which can be seen in Figure 2.13. This is a representation of double harmonic well potential. We then performed SD2 and TD2 onto this data containing two wells.

The direction of arrow marked by SD2 doesn't reveal any important information about what event might be occuring. However, TD2 arrow deals with resolving the data by finding directions where the autocorrelation is maximum. This might hint us about the rare event that might be taking place in between two harmonic wells. Applying this strategy to biomolecular data, in the next section we project spatially resolved data Y onto the rotation matrix obtained from TD2 and visualize the temporal harmonic modes through cartoon representations of BPTI.

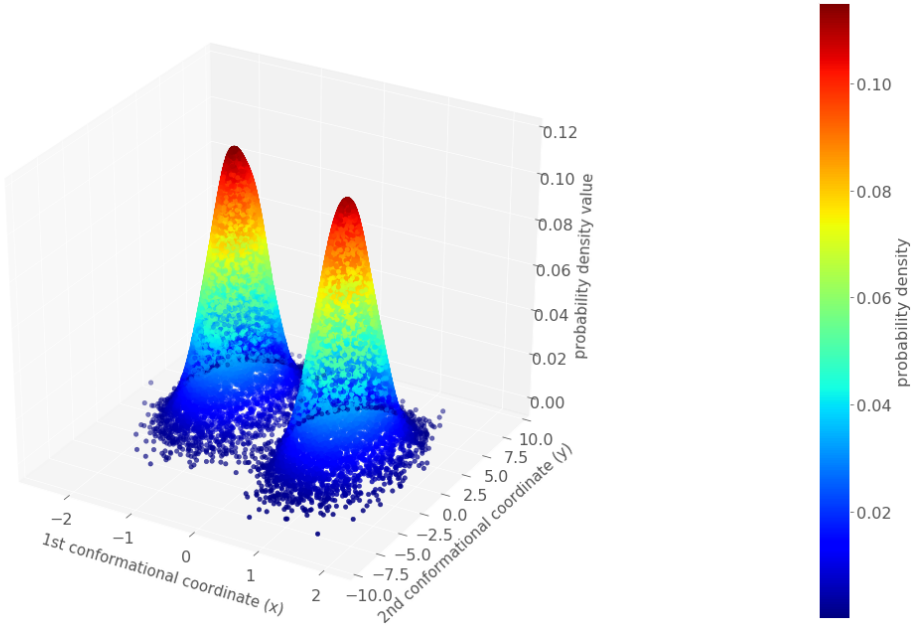


Figure 2.13: **Two-dimensional Gaussian double-well.**

2.5.3 TD2 modes of motion in BPTI

TD2 modes are arranged according to the temporal harmonic coordinates that tries to find principal components having exclusive energetic properties. The conformational space is built by projecting the spatially decorrelated data onto the top three dominant conformational coordinates that are temporally resolved. Each conformation is marked as a dot in the three-dimensional scatter plot and is colored according to the positional displacements observed between Pro^{9} and Phe^{33} as shown in Figure 2.14.

From the movie like visualization of BPTI protein by projecting onto the temporally resolved conformational coordinates, we observe that in comparison to $TD2_1$ the conformational transitions are unable to pick the transition that involves a larger separation between the flap regions since the projections from the simulations indicate the presence of both closed and open states in this transition. $TD2_2$ describes the motion involving the movement of the flap regions in a concerted fashion. This can be seen from the cartoon version in Figure 2.15.

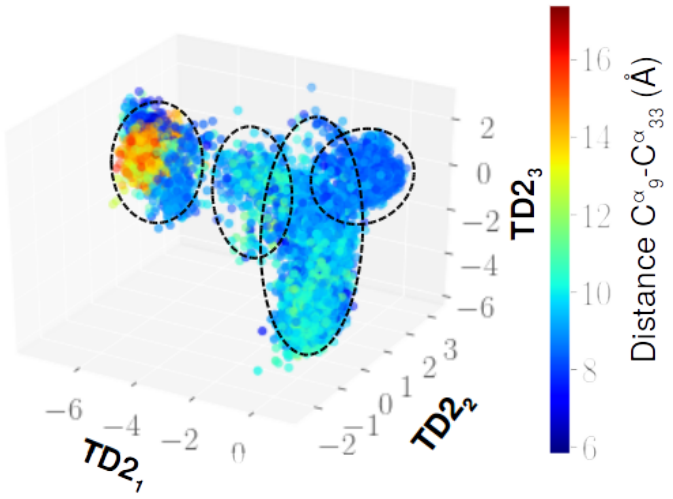


Figure 2.14: **TD2: removing dominant second-order temporal correlations using time-delayed principal component analysis.** We see the separation of open and closed states of BPTI, however, the open state still consists of conformations that have closed state as seen in Figure 2.9.

2.6 TEMPORAL DECORRELATION IN FOURTH-ORDER (TD4)

In the interest of resolving spatial and temporal anharmonic dependencies in the molecular simulation trajectories, we have designed the TD4 module which performs joint diagonalization of time-delayed cumulant matrices (a tensor of fourth-order time-delayed statistics signifying kurtosis). TD4 is the counterpart of SD4, where fourth-order spatial correlations are minimized, implying zero time lag.

2.6.1 Performing fourth-order decorrelation in time

Conceptually, the assumption we make is that a molecular simulation trajectory is a linear combination of independent, anharmonically fluctuating protein motions. To discover these anharmonic motions, we borrow a technique from signal processing literature, called Blind Source Separation (BSS) [31], which attempts to extract or unmix independent non-Gaussian sources from signal mixtures with Gaussian noise. To facilitate the extraction of anharmonic

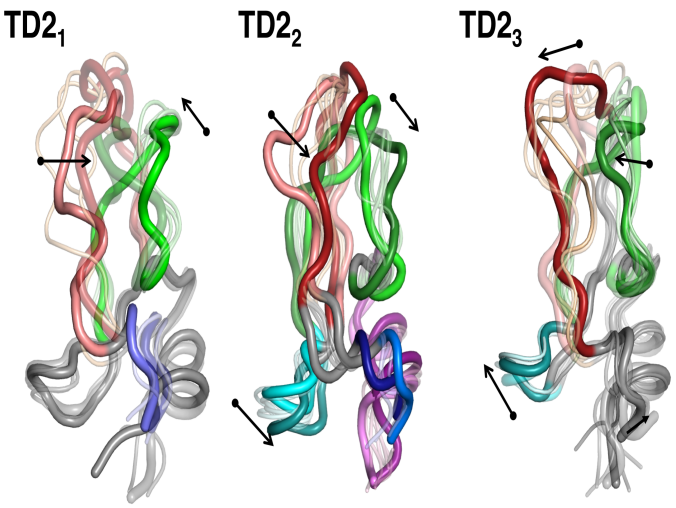


Figure 2.15: **The three principal modes of motion described by TD2 analysis of the BPTI simulations ($TD2_1 - TD2_3$) shown in a movie like representation.** The light to dark transitions are representative of the conformational within BPTI similar to Figure 2.3. Arrows indicate the predominant directions of the motions. $TD2_1$ describes the direction of motion that brings together the flap regions ($L_1 and L_2$) involving a partial motion of α_1 region that is displaced.

modes of motion of the fourth-order, the trajectory data $X_{orig} \in \mathbb{R}^{3N \times t}$, where $3N$ represents *(x,y,z)* coordinates from individual atom selections and t represents conformations is decorrelated for second-order dependencies both spatially and temporally by transforming it through the modules of SD2 and TD2.

Algorithmically, the method of unmixing temporally correlated signals of fourth-order can be viewed as a symmetric eigenvalue problem of a generalized cumulant matrix Q_{ij}. As a measure of statistical independence, we will consider the 'diagonality' of a set of cumulant matrices. The cumulant matrices are generated in a low-dimensional subspace denoted by m, which is the best guess for the most compact summary of the fourth-order statistics. The subspace dimensionality can be adjusted by examining the inflection points in the cumulative variance plots generated from SD2 module as seen from Figure 2.7. In order to generate

the cumulant matrices, a time-lagged covariance matrix is defined by:

$$R_z(\tau) = E\left\{ZZ_\tau^T\right\}, \tag{2.11}$$

where $Z \in \mathbb{R}^{m\times t}$ is second-order spatially and temporally resolved molecular simulation data, τ is time delay and $Z_\tau = Z(t-\tau)$ is the time-lagged version of Z. A fourth-order cumulant matrix Q_{ij} of this data matrix Z is defined by:

$$Q_{ij} = E\left\{ZZ^TZ_\tau^TZ_\tau\right\} - E\left\{ZZ^T\right\}\operatorname{tr}E\left\{Z_\tau Z_\tau^T\right\} - 2E\left\{ZZ_\tau^T\right\}E\left\{Z_\tau Z^T\right\}, \tag{2.12}$$

where $Q_{ij} \in \mathbb{R}^{m\times m}$ computes a time-lagged cumulant matrix. The possibility of computational errors, such as round-off errors, can destroy the symmetricity of the cumulant matrix which is restored by performing:

$$Q_{ij} = \frac{1}{2}\left[Q_{ij} + Q_{ij}^T\right]. \tag{2.13}$$

A time-lagged cumulant tensor $\mathbb{Q} \in \mathbb{R}^{m\times(m\times k)}$, where $k = [m\times(m+1)]/2$ is defined for the storage of cumulant matrices computed by the symmetric Q_{ij} matrix. Joint diagonalization of these time-lagged cumulant matrices reduces fourth-order temporal dependencies leading to anharmonic modes of motion of the trajectory data. This is done through Jacobi's iterative method of finding solution to a system of linear equations. In particular, the method uses successive transformations to calculate diagonal elements of the cumulant tensor by decimating off-diagonal elements with each iteration. The spatio-temporally decorrelated matrix of fourth-order is computed by obtaining:

$$Z_{TD4} = WX_{orig}, \tag{2.14}$$

where W attempts to separate sources from signal mixture X_{orig} by finding directions, such that projections onto these directions have maximum statistical independence. The computed parameter Z_{TD4} is fourth-order spatially and temporally resolved matrix.

2.6.2 Harmonic triple-well experiment

While analyzing data from MD simulations, it's essential to identify metastable states [15] that can explain the behavior of protein's motion and hence its function. Due to the complexity of large datasets, it is challenging to be able to identify such states and second-order statistics fails to resolve them correctly. Hence, we pursue fourth-order statistics as a method to characterize anharmonic protein motions. In order to illustrate the application of TD4, we generated a three well harmonic potential having 100000 samples as described in Figure 2.16.

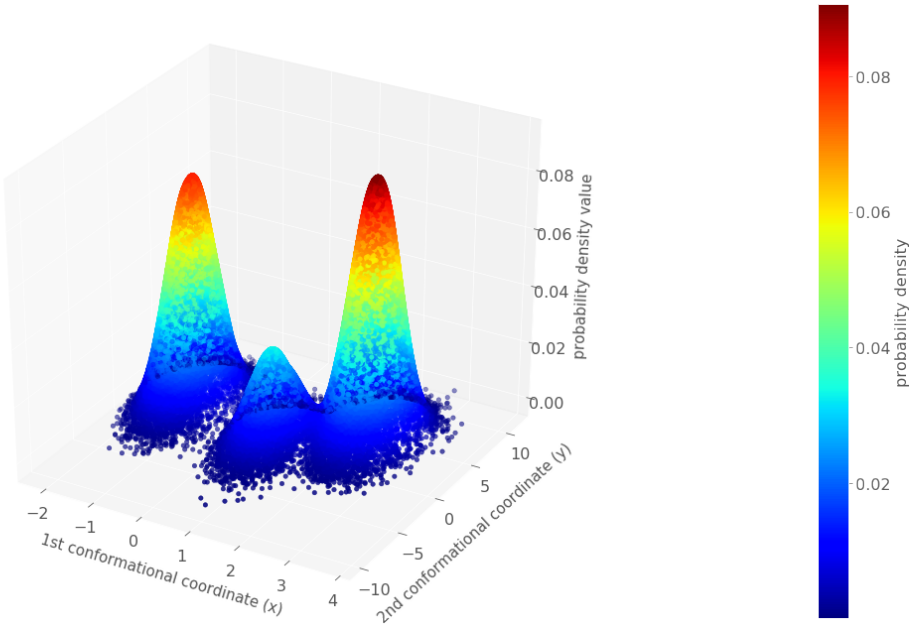


Figure 2.16: **Two-dimensional Gaussian triple-well.**

The data from triple well was passed through modules SD2, TD2 and TD4 successively to spatially and temporally resolve the dependencies in fourth-order. After obtaining harmonic and anharmonic modes from decorrelation functions, data was projected onto the second principal/independent component. From Figure 2.17 it is evident that SD2 and

TD2 doesn't differentiate the three states while TD4 helps to achieve this. As a method to improve identification of such metastable states, we project BPTI data onto three dominant anharmonic modes in the following section.

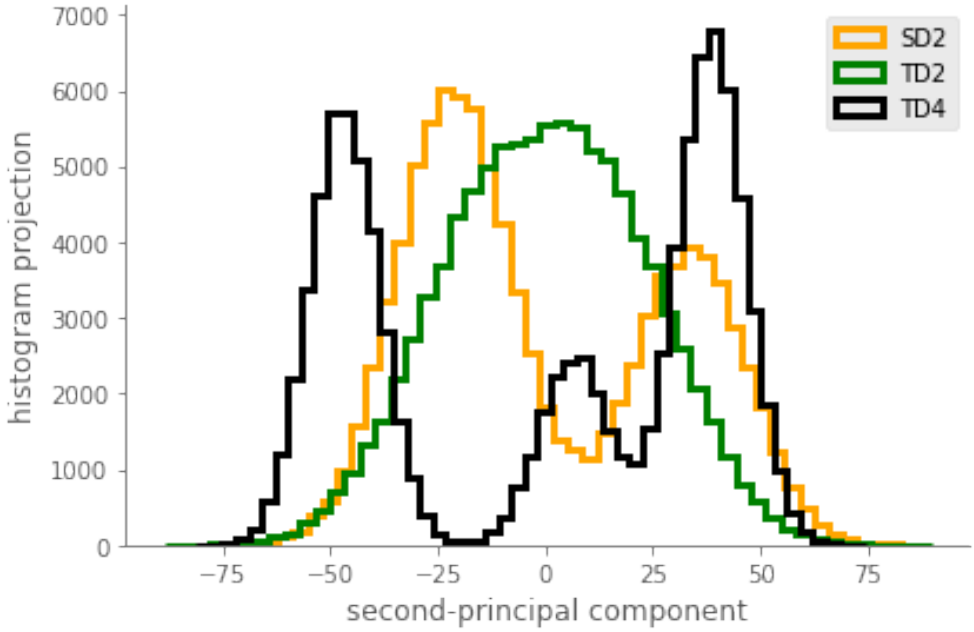


Figure 2.17: **Histogram projections of spatially and temporally resolved data.**

2.6.3 TD4 modes of motion in BPTI

TD4 module constructs a time-delayed fourth-order kurtosis tensor, which is then approximately diagonalized to obtain anharmonic modes of motions once the second-order spatial and temporal correlations are resolved. TD4 module is the temporal analog of the spatial SD4 module. For BPTI, the projections from three TD4 modes ($TD4_1 - TD4_3$) as depicted in Figure 2.18 describe essential motions of the flap region along two distinct directions. To quantify these motions, we use a reaction coordinate bases on the distances between Pro^9 and Phe^{33}.

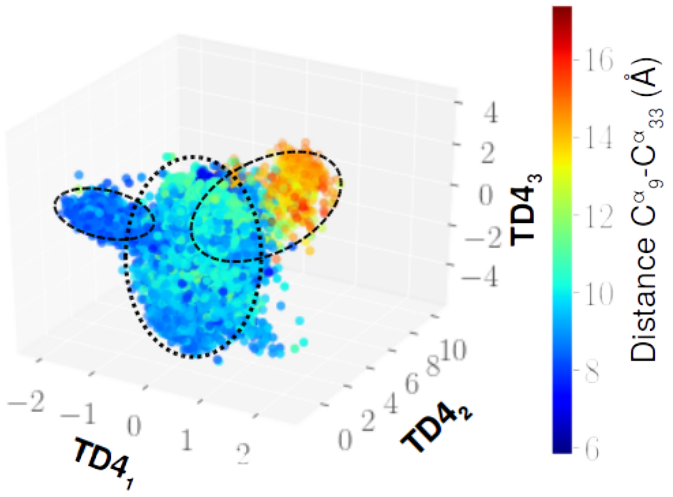


Figure 2.18: **Multi-dimensional description of the simulation data using the top three time-delayed anharmonic modes.** Each conformation, represented by a dot, is colored by the distance between the centers of mass of the flap regions. Three putative conformational substates are demarcated by dotted ellipses depicting the closed (I) and open (III) states that pass through an intermediate state (II), as seen by the colored distance distribution. Arrows indicate how to reach the closed and open states by walking along anharmonic modes $TD4_1$ and $TD4_2$ from the intermediate state.

In order to understand the motions seen from Figure: 2.18, we depict movies that capture the conformational transitions in BPTI(see Figure: 2.19). In each case, the flaps open/close albeit in distinct directions and in some cases even capturing rare transitions involved in the exchange of the flaps. The ANCA modes enable us to quantitatively understand the extent to which the relative motions between the flaps expose opening/closing of this region.

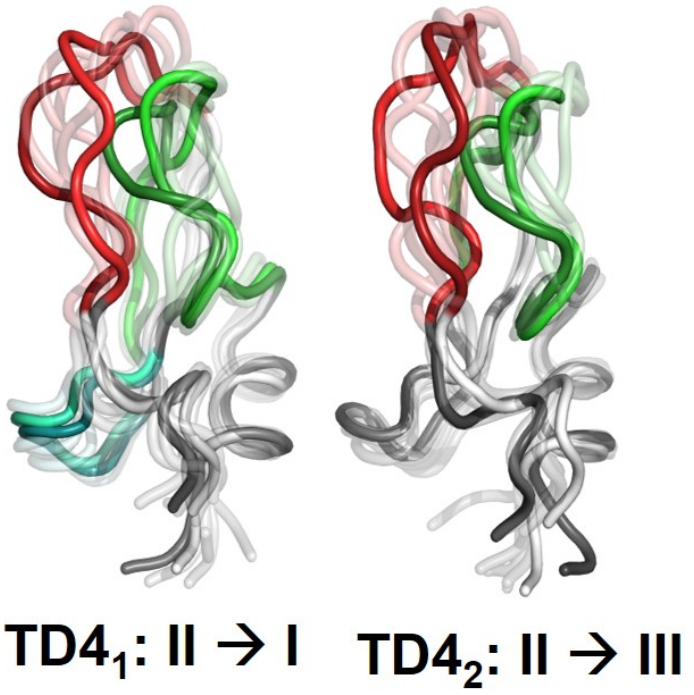


Figure 2.19: **The three principal modes of motion determined from the TD4 analysis of the BPTI simulations ($TD4_1$ - $TD4_3$) shown in a movie like representation.** These motions are shown in an ensemble form, where the time evolution is highlighted with loops L_1 (red), L_2 (green) and $\beta_1\beta_2$ (cyan) and the rest of the protein (gray) depicted from light to dark colors, denoting start-to-end progression..

2.7 CONCLUSION

To summarize, we observe anharmonicity in the collective fluctuations of BPTI from the 1 milli-second long MD simulation. In order to provide a compact summary of its motions and establish a grouping of conformations which are energetically similar, second-order decorrelations doesn't provide insightful details. Thus, higher-order statistics is implemented for discovering sub-states that can be clustered based on it's characteristics. Even while doing fourth-order computations, SD4 (previosuly QAA [15]) doesn't provide details about temporal correlations which needs to be resolved to be able to extract components that are independent with respect to a specified biophysically relevant property. With the help of

TD4 we observe, large inter-residue distances cluster together and the ones which tend be farther away are clustered separately.

3.0 DIHEDRAL ANHARMONIC CONFORMATIONAL ANALYSIS

For biomolecular simulations that can include both proteins and nucleotides, it is more convenient to eliminate the sensitivity to Cartesian alignment to a reference structure and perform anharmonic analysis in the dihedral space directly. We explore this new approach on the dodecamer B-DNA (*Deoxyribonucleic acid*) molecule.

3.1 STRUCTURE OF DODECAMER B-DNA

The basic structural unit of a DNA is a nucleotide comprising of a nucleoside attached to atleast one phosphate group. The DNA molecule is achiral and asymmetric. The nucleoside unit includes nitrogenous base and five-carbon sugar. Nitrogenous bases (molecule containing nitrogen) can be further classified into four types: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The shape of the DNA is largely influenced by hydrophobic interactions between bases and the allowed bond angles in the sugar-phosphate backbone. They can assume the structure of either A, B or Z-DNA. Among the three forms, analyzing conformations of B-DNA is popular due to its predominant occurence in the cells. B-DNA is a three-dimensional right-handed helix structure with equal spacing of each base pair (20 base-pairs in total). For our studies, we have considered synthetic dodecamer DNA d(C-G-C-G-A-A-T-T-C-G-C-G) whose individual crystal X-ray structure was confirmed by Drew *et al.* [32]. Due to the limitations of experimental techniques to provide information about the flexibility and complex movements of coupled bonds between individual units forming strands of DNA, atomistic scale MD simulations have made it possible to perform conformational analysis on this complex system. The structure of DNA derived from microsecond

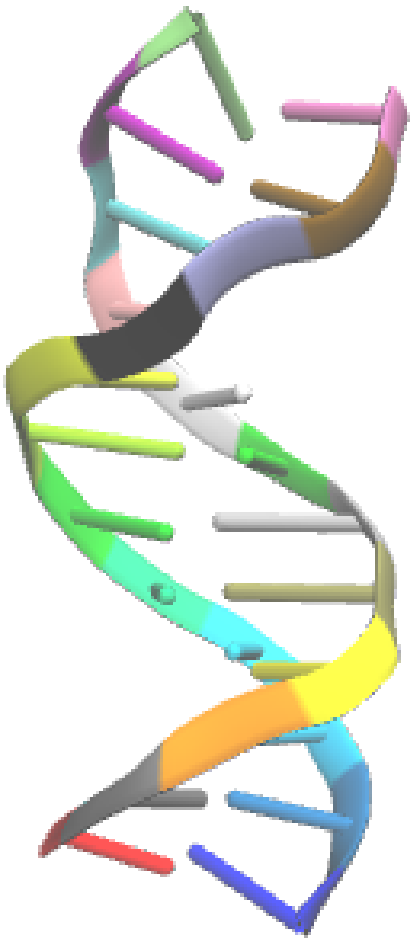


Figure 3.1: **Structure of Dickerson-Drew dodecamer B-DNA.** The molecule assumes the shape of a right-handed double stranded B helix. It consists of 24 base-pairs. Each strand S1 and S2 has 12 base pairs. Following the rule of base pairing, purine adenine always pairs with pyrimidine thymine and pyrimidine cytosine always pairs with purine guanine. Purines and Pyrimidines refer to the number of carbon nitrogen ring bases. The former has two-carbon nitrogen ring base, whereas the latter has a single carbon nitrogen ring base.

long simulation is visualized in Figure 3.1. The 1 μs simulation was performed by our collaborators (agarwal-lab.org).

3.1.1 Torsion angles of nucleic acid conformers

Classical molecular dynamics (MD) simulation has been an all-powerful tool to describe three-dimensional structure, motion and function of macro-molecules such as proteins and nucleic acids [33]. The data obtained from MD as an arrangement of $3N \times t$ matrix has attracted significant attention for analyzing functional dynamics. Dimensionality reduction of the large input data matrix to fewer coordinates describing a majority of molecular dynamics has been an area of active research. Recent studies investigating reversible folding and unfolding in water for a penta-alanine simulation [35] suggested that torsional angle analysis might be more insightful than Cartesian coordinate analysis [34] in differentiating the internal motions based on torsional angles from overall motions. An additional advantage of using angular representation for analysis is that it does not require rigid-body alignment of the conformer data with anyreference structure.

Due to the benefits of investigating in the dihedral angle space, several studies have been reported analyzing the distribution of angles and the flexibility shown by the molecule as it bends and twists during the simulation [36, 37, 38]. In this section, we describe six torsional angles ($\alpha - \zeta$) present in the backbone of a nucleic acid that specify the conformation of a nucleotide and an additional torsional angle (χ) which gives the orientation of an N-terminal base with respect to the glycosidic bond. For ease of understanding, we will consider a single nucleotide chain of B-DNA and illustrate how the angles are computed (Figure 3.2).

The nucleic acid backbone parameters involving seven torsion angles is computed by considering the angle between two adjacent planes formed by the x, y, z coordinates of successive four atoms described below [32].

- α : O3'(i-1) - P - O5' - C5'
- β : P - O5' - C5' - C4'
- γ : O5' - C5' - C4' - C3'
- δ : C5' - C4' - C3' - O3'
- ϵ : C4' - C3' - O3' - P(i+1)
- ζ : C3' - O3' - P(i+1) - O5' (i+1)
- $\chi - purines$: O4' - C1' - N1 - C2

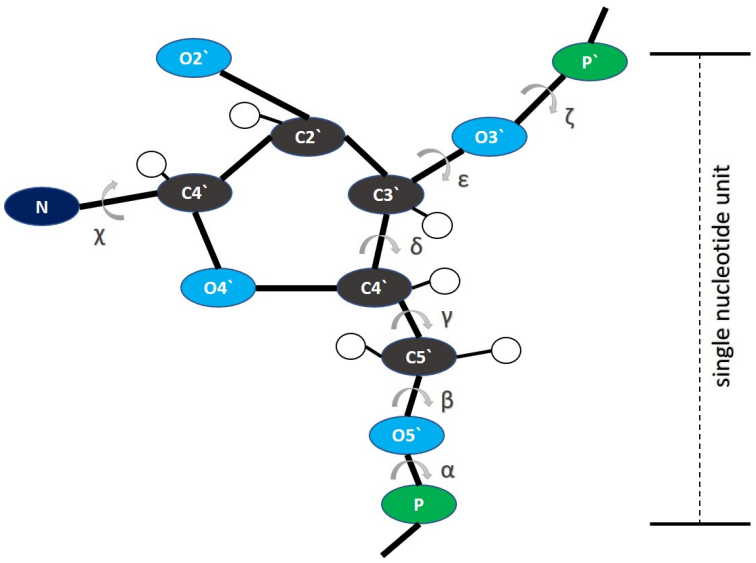


Figure 3.2: **Single nucleotide of B-DNA.** A nucleic acid conformer is defined from one phosphate group to the next. Conformation is described through six torsional angles namely $\alpha, \beta, \gamma, \delta, \epsilon, \zeta$ and the glycosidic torsion angle by χ.

- $\chi - pyrimidines$: O4' - C1' - N9 - C4

We provide visualization tools to depict the distribution of torsional angles for both strands (Figure 3.3). Before starting to analyze these angles, we need to understand the circularity of data and modifications necessary to compute statistical parameters [38].

3.1.2 Circular Statistics

Unlike Cartesian coordinates, representing circular data is not straightforward. By observing two angles $\phi_1 = 40^o$ and $\phi_2 = 340^o$, the mean of the angles $\langle \phi_{mean} \rangle = 190^o$. However, by considering the boundaries of angles which are defined to be in the range -180^o to 180^o, we have, $\phi_{1'} = 40^o$ and $\phi_{2'} = -20^o$. Considering the new values for angles, we obtain $\langle \phi_{mean'} \rangle = 10^o$. In order to avoid this problem, we convert each angle, say ϕ, into its Euclidean

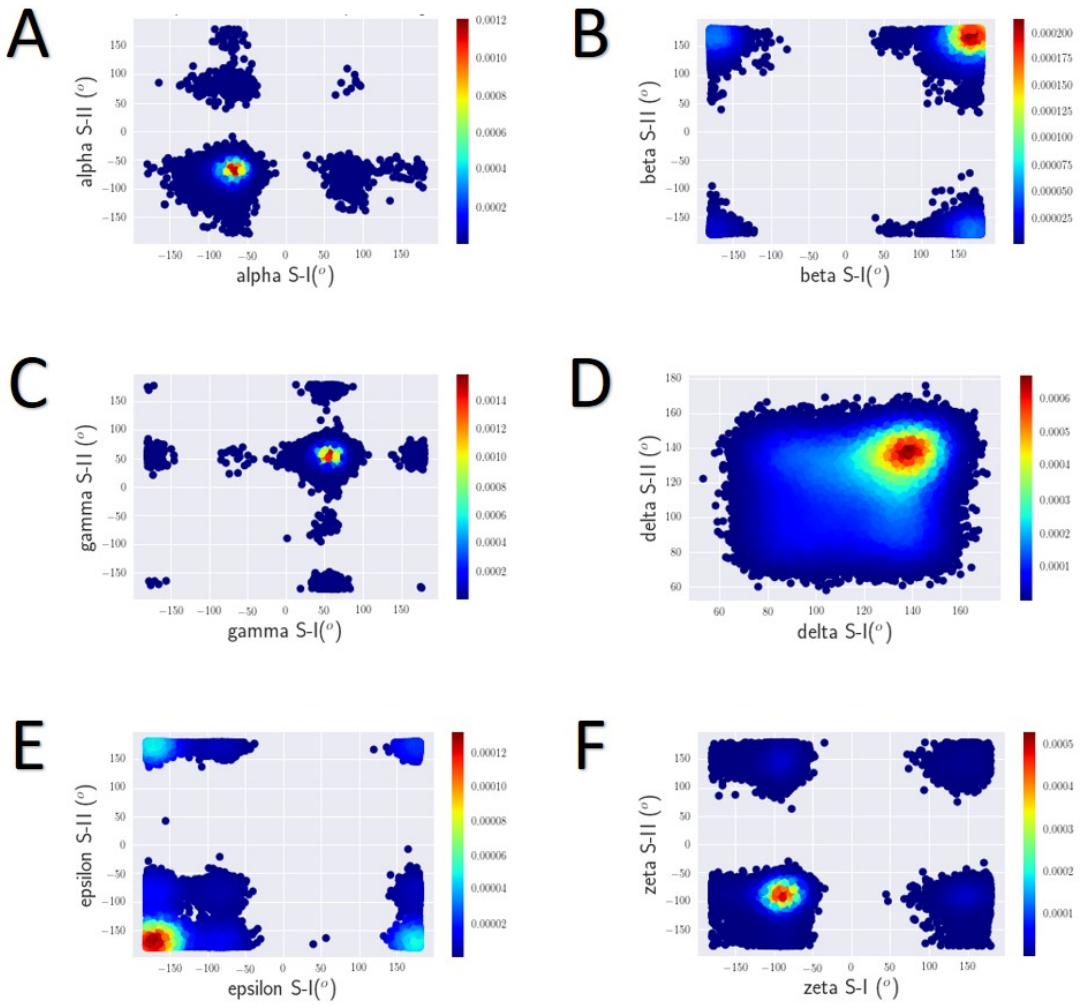


Figure 3.3: **Density plots of the torsion angles considering strand-1 and strand-2.**

representation with the following transformation:

$$x = \cos(\phi), \tag{3.1}$$
$$y = \sin(\phi).$$

This linear transformation of angle ϕ into (x, y) makes it possible to establish the Euclidean metric which is given by the equation:

$$\Delta = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{3.2}$$

Thus, the problem associated with circularity is avoided. We make use of *pycircstat* package [39] to compute circular statistical values such as circular mean, circular variance, circular kurtosis, and others which will be discussed later in this chapter.

3.2 DATA EXTRACTION

For our analysis, we have obtained a 1 μs long simulation data of Drew-Dickerson dodecamer B-DNA having the sequence $[CGCGCAATTCGCG]_2$. Each frame is 100 ps apart, providing 10,000 conformers in the microsecond timescale. It is extracted using AMBER force fields which gives a *topology* DNA sequence file and *mdcrd* trajectory file of Binary NetCDF format. We use *mdanalysis* libraries [40] to capture and process trajectories. MDAnalysis recognizes binary trajectories through *.ncdf* suffix and is read by the package function *NCDFReader*. Since we are interested in using torsion angles, we transform the AMBER force fields generated data using CPPTRAJ [41] commands as given below for first base-pair of strand-1 for all the 10,000 conformations:

- α - dihedral strand1-alpha :1@O3' :2@P :2@O5' :2@C5' out alpha.dat
- β - dihedral strand1-beta :2@P :2@O5' :2@C5' :2@C4' out beta.dat
- γ - dihedral strand1-gamma :2@O5' :2@C5' :2@C4' :2@C3' out gamma.dat
- δ - dihedral strand1-delta :2@C5' :2@C4' :2@C3' :2@O3' out delta.dat
- ϵ - dihedral strand1-epsilon :2@C4' :2@C3' :2@O3' :3@P out epsilon.dat
- ζ - dihedral strand1-zeta :2@C3' :2@O3' :3@P :3@O5' out zeta.dat

The keyword *dihedral* used above directs CPPTRAJ software to output dihedral angle and it expects the user to input the symbolic names of four atoms preceeded by ':' for each torsion angle. *out* and *filename.dat* at the end specifies the type and name of the file respectively. *strand1-alpha* that follows *dihedral* command is a reference name and can be ommitted. By repeating these commands for rest of the base-pairs of strand-1 and strand-2, we obtain 120 torsion angles (20 angles of each type) which can be broken down into [6 (torsion-angle types) $\times$ 10 (base-pairs in each strand) $\times$ 2 (number of strands)]. These 120

torsion angles are converted to Euclidean representation as described in Eq. 3.1 thereby giving two equivalent vectors for each angle. A single conformation at a particular time-step i is represented by a vector:

$$\phi_i = (\alpha_{S1}, \alpha_{S2}, \beta_{S1}, \beta_{S2}, \ldots \ldots, \zeta_{S1}, \zeta_{S2})_i \tag{3.3}$$

The data matrix (240×10000) is provided as an input to the decorrelation modules. It is constructed as follows:

$$\begin{pmatrix}
\alpha_{S1}^{1} & \alpha_{S1}^{2} & \alpha_{S1}^{3} & \ldots & \ldots & \alpha_{S1}^{10000} \\
\alpha_{S2}^{1} & \alpha_{S2}^{2} & \alpha_{S2}^{3} & \ldots & \ldots & \alpha_{S2}^{10000} \\
\beta_{S1}^{1} & \beta_{S1}^{2} & \beta_{S1}^{3} & \ldots & \ldots & \beta_{S1}^{10000} \\
\beta_{S2}^{1} & \beta_{S2}^{2} & \beta_{S2}^{3} & \ldots & \ldots & \beta_{S2}^{10000} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\epsilon_{S1}^{1} & \epsilon_{S1}^{2} & \epsilon_{S1}^{3} & \ldots & \ldots & \epsilon_{S1}^{10000} \\
\epsilon_{S2}^{1} & \epsilon_{S2}^{2} & \epsilon_{S2}^{3} & \ldots & \ldots & \epsilon_{S2}^{10000} \\
\zeta_{S1}^{1} & \zeta_{S1}^{2} & \zeta_{S1}^{3} & \ldots & \ldots & \zeta_{S1}^{10000} \\
\zeta_{S2}^{1} & \zeta_{S2}^{2} & \zeta_{S2}^{3} & \ldots & \ldots & \zeta_{S2}^{10000}
\end{pmatrix} \tag{3.4}$$

In the following section, we will discuss resolving the data spatio-temporally in second and fourth-order.

3.3 RESOLVING SPATIAL AND TEMPORAL CORRELATIONS THROUGH TORSION ANGLE ANALYSIS

We perform spatial and temporal decorrelations on a data that has been transformed into its Euclidean representation from Cartesian coordinates. The theory behind resolving data in space and time using second and fourth-order statistics has been described in 2.3, 2.4, 2.5 and 2.6.

3.3.1 Dihedral Spatial Decorrelation

Dihedral spatial decorrelation (dSD2) is performed via PCA to represent correlated internal motions. PCA is carried out on the *sin* and *cos* transformed torsion angles through eigenanalysis of the covariance matrix as proposed by Mu *et. al* [35]. The subspace dimensionality m is adjusted by examining the inflection points that dSD2 module returns as shown in Figure 3.4. A subspace $m = 24$ is chosen which explains 70% of the cumulative variance.

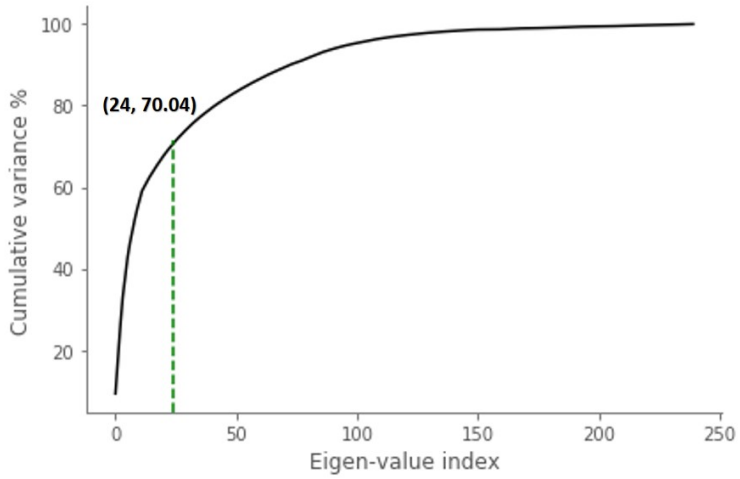


Figure 3.4: **Scree plot of B-DNA data.** 24 PCA modes explains 70.04% of the total variance.

dSD2 removes dominant second-order spatial correlations. In addition to the simulation data, it requires an input m. dSD2 module diagonalizes the covariance matrix and returns the eigenvalues S (size $m \times 1$), eigenvectors B and the projection matrix Y. Eigenvectors B indicate the direction of dominant motions. Figure 3.5 reveals that the top three dominant eigenvectors obtained from dSD2 captures correlated motions of ϵ and ζ in strands 1,2,3 and 9. The top three modes from this module is shown in Figure 3.6A and the corresponding movies are shown in Figure 3.6B. From the movie like representation for dSD2, we observe

that the first (top) and last three base-pairs (bottom) of two strands transits from a slightly bent structure (I) to a vertical shape (II) which is approximately parallel to the surface.

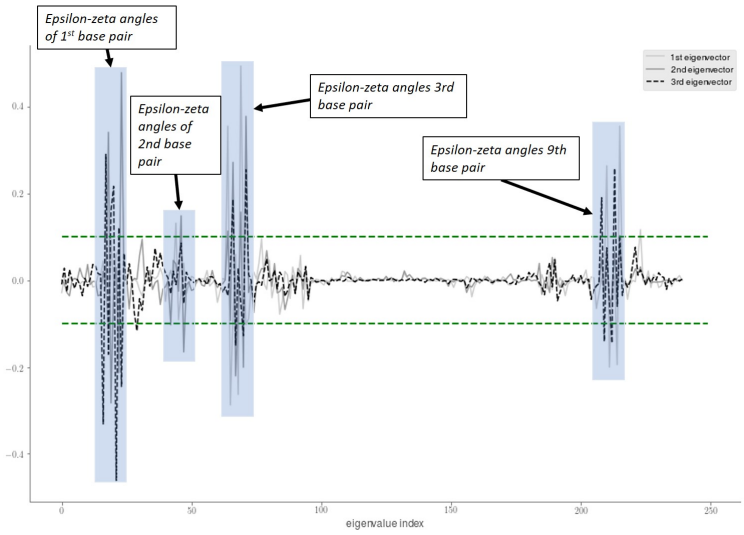


Figure 3.5: **Visualization of top three eigenvectors.**

Dihedral spatial decorrelation in fourth-order (dSD4) attempts to resolve the intrinsic non-orthogonal spatial dependencies in atomistic fluctuations. The second order projections, Y from dSD2 are used to build a fourth-order spatially correlated cumulant tensor. dSD4 approximately diagonalizes this tensor to return an anharmonic mode matrix W ($3N$ or $D \times m$). To build associations between dSD4 modes and biophysically meaningful reaction coordinates, we have used internal energy values from the simulations and have visualized how the physical observable is mapped onto each of the dSD4 modes. For B-DNA simulations, the top three modes from dSD4 are shown in Figure 3.7A. We capture the conformational transition from state I to state II from the corresponding movies as shown in Figure 3.6B, where state II undergoes an anticlockwise rotation. Here, we consider the cartoon representation of state I as the reference structure. From both the movies of dSD2 and dSD4 modes, we observe that the motions are not pronounced.

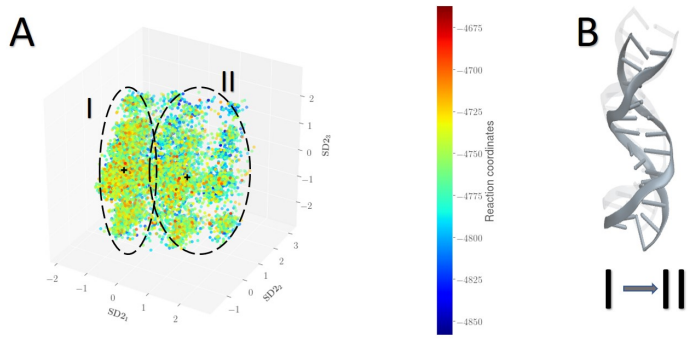


Figure 3.6: **Dihedral spatial decorrelation in the second-order of a microsecond long simulation of B-DNA.** (*A*) Multi-dimensional description of the simulation data using top three dSD2 modes and colored by the internal energy values of each conformer. (*B*) Motions are shown in ensemble form, where the light to dark transition indicated by two superimposing structures can be reached by walking along the cluster center from I to II.

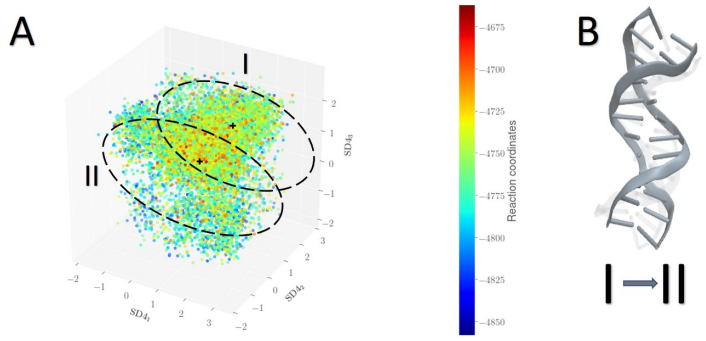


Figure 3.7: **Dihedral spatial decorrelation in the fourth-order of a microsecond long simulation of B-DNA.** (*A*) Multi-dimensional description of the simulation data using top three dSD4 modes and colored by the internal energy values of each conformer. (*B*) Motions are shown in ensemble form, where the light to dark transition indicated by two superimposing structures can be reached by walking along the cluster center from I to II.

3.3.2 Dihedral Spatio-temporal Decorrelation

Dihedral temporal decorrelation in second-order (dTD2) removes dominant second-order temporal correlations by computing a time-delayed covariance matrix and performing PCA. The inputs to this module are similar to dSD2, with an additional user specified parameter, τ, that denotes the lag time over which the temporal correlations are to be resolved. In order to choose a specific lag time, we computed the cosine similarity by comparing the absolute value of the elements in square matrix of top six eigenvectors from dTD2 with lag = 1 and lag = 1 to 100. From Figure 3.8, we can observe that at a lag time of 25 and above, the cosine similarity value stabilizes at around 0.94 before it degrades during larger lags. Thus, to perform temporal decorrelation, we choose a lag time of 25.

The outputs of this module include Z, a matrix obtained by projecting the simulation data on the dominant time-delayed eigenvectors and the corresponding eigenvalues. The top three modes from dTD2 module for B-DNA are shown in Figure 3.9A and the corresponding movies are depicted in Figure 3.9B. Clearly, the plots obtained after resolving temporal correlations is interesting. We observe from Figure 3.9B that the conformers are clustered based on their internal energy values. Lowest energy conformers are clustered as state I and high energy conformers as state II. The intermediate energetic conformers occupy clusters III and IV. The structural deformation is more severe while observing the change from I to IV. However, not much structural change is seen while transiting from I to II which is more diffcult since transition happens from lowest energy to highest energy state. dTD2 fails to capture these motions.

Dihedral temporal decorrelation in fourth-order (dTD4) constructs a time-delayed fourth-order kurtosis tensor, which is then approximately diagonalized to obtain anharmonic modes of motions once the second-order spatial and temporal correlations are resolved. dTD4 module is the temporal analog of the spatial dSD4 module. The input parameters to this module include the matrix Z (from the dTD2 module), a user specified subspace m denoting the number of desired anharmonic modes of motion, the lag time τ and the matrix V. The output from the module includes the separating matrix W. For B-DNA, the projections from

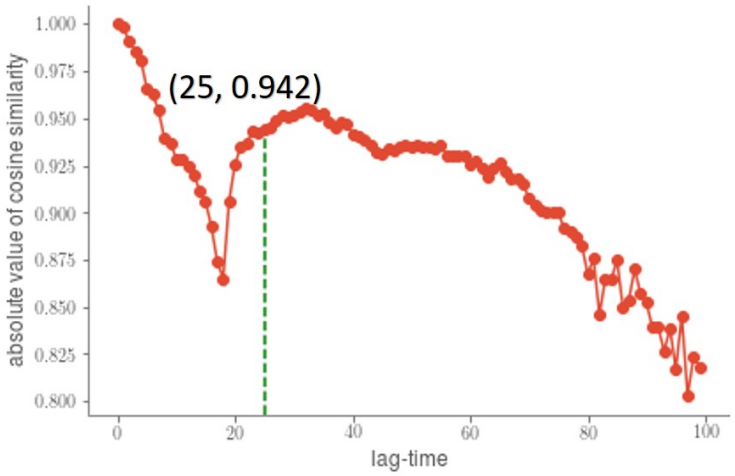


Figure 3.8: **Choosing a lag time by computing cosine similarity between eigenvectors obtained from dTD2 module with lag=1 and lag ranging from 1 to 100.**

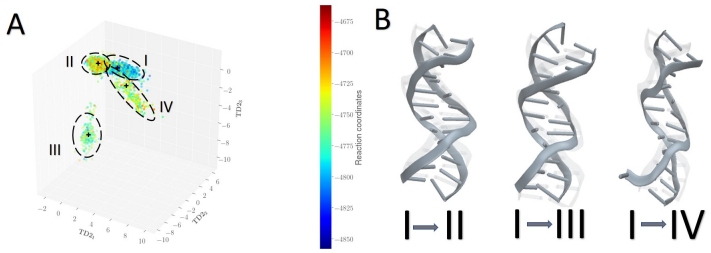


Figure 3.9: **Dihedral temporal decorrelation in the second-order of a microsecond long simulation of B-DNA.** (*A*) Multi-dimensional description of the simulation data using top three dTD2 modes and colored by the internal energy values of each conformer. (*B*) Motions are shown in ensemble form, where the light to dark transition indicated by two superimposing structures can be reached by walking along the cluster centers from I to II, I to III and I to IV.

three principal modes ($TD4_1 - TD4_3$) is depicted in Figure 3.10. In order to quantify the motions, we use a reaction coordinate based on the internal energy values of conformations which was previously used in dSD2, dSD4 and dTD2. To understand these motions further, we depict movies that capture the conformational transitions in B-DNA. From the movie

representation, we see that structural modification is more prominent while walking along the centers of clusters I to III. This is convincing by taking into account the fact that transition from low to high energy state is more difficult than vice-versa.

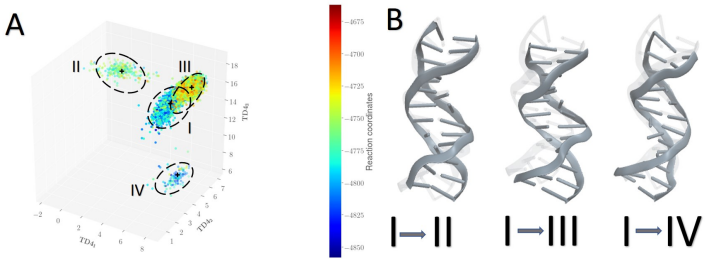


Figure 3.10: **Dihedral temporal decorrelation in the fourth-order of a microsecond long simulation of B-DNA.** (*A*) Multi-dimensional description of the simulation data using top three dTD4 modes and colored by the internal energy values of each conformer. (*B*) Motions are shown in ensemble form, where the light to dark transition indicated by two superimposing structures can be reached by walking along the cluster centers from I to II, I to III and I to IV.

3.4 CONCLUSION

In this chapter, we studied theory behind performing analysis on circular data and necessary transformations. We discuss the advantages that dihedral spatial and temporal decorrelations have over Cartesian coordinate analysis. An example of microsecond long simulation of Drew-Dickerson Dodecamer B-DNA is considered to test the dihedral extension. Through three-dimensional projections of the modes and movie like representation, we make an attempt to derive biophysically meaningful insights from the motions of conformers and their transitions.

4.0 PYANCA: A SCALABLE TOOLKIT TO ANALYZE HIGHER-ORDER ANHARMONIC MOTION SIGNATURES FROM MOLECULAR DYNAMICS SIMULATIONS

pyANCA is a python library to measure anharmonicity in Molecular Dynamics (MD) simulation data. The idea is to use higher order motion signatures in the simulation data for organizing the conformational landscape into putative conformational substates. pyANCA has modules which can: (1) measure for anharmonicity in the form of higher order statistics and show its variation as a function of time, (2) output a story board representation of the simulations to identify key anharmonic conformational events, and (3) identify putative anharmonic conformational substates and visualize transitions between these substates.

4.1 INTRODUCTION

Traditional analysis tools for biomolecular simulations have focussed on second-order statistics [42, 43, 38]. Anharmonicity in time-dependent conformational fluctuations is noted to be a key feature of functional dynamics of biomolecules [44, 16, 45]. Although anharmonic events are rare, long timescale ($\mu s - ms$ and beyond) simulations facilitate probing their behavior. However, automated analyses and visualization of anharmonic events from these long timescale simulations is proving to be a significant bottleneck.

Anharmonicity as an organizing principle for conformational landscape of proteins and other biomolecules is proposed in this thesis [46]. Previously, quasi-anharmonic analysis (QAA) was built to resolve *higher order spatial correlations* [15, 47, 48, 18]. pyANCA was built as an extension to the QAA toolbox to resolve *higher order temporal correlations* from

long timescale simulations.

4.2 METHODS

pyANCA can process trajectories in many formats commonly used by the biophysics community, including Protein Data Bank (PDB), CHARMM DCD, AMBER coordinates, Gromacs xtc files. pyANCA uses *mdanalysis* [40, 49] and *mdtraj* [50] to capture and process coordinate (or other feature) information from MD trajectory files. Further, user can specify which features to select and process using an extensive set of coordinate and feature selection commands within two packages. Using Python's inbuilt capabilities to process memory-mapped arrays, we can process large trajectories up to several terabytes.

4.2.1 Data Extraction

pyANCA makes use of the powerful *MDAnalysis* libraries to extract coordinates or angles from molecular dynamics trajectories. The function used for data extraction is a generic driver for obtaining coordinates of interest. We can use a variety of atom selections and the results are returned as numpy arrays.

getCoordinates: Method used to extract MD trajectories when provided with the path for *topology* and *trajectory* files. While performing analysis in the dihedral space, the (x,y,z) coordinates can be transformed to compute the torsion angles.

4.2.2 Alignment

Alignment of the selected coordinates is done through this function. This is to ensure that we have removed translations and rotations before analysing. pyANCA offers two rigid-body alignment algorithms:

- Iterative alignment
- Standard Kabsch alignment

IterativeMeansAlign: It takes the extracted coordinates as input and provides an aligned trajectory free from rotational and translational degrees of freedom. While performing alignment, we assume that protein chains or independent molecules are put together. Different softwares for performing simulation gives different ways for putting individual chains/ molecules. Input arrays are expected to be passed in the form of $Ns \times 3 \times Na$, where Ns denotes number of snapshots, Na represents number of atoms and 3 indicates the x, y, z directions.

4.2.3 Resolving spatial and temporal correlations

pyANCA provides four core modules for analyzing MD trajectories. These modules take as input X, either Cartesian coordinates of dimensions $3N \times t$, where $3N$ represents the 3D (x, y, z) coordinates of the individual atoms selcted for analysis, or cosine/sine transformed dihedral angles, namely (ϕ, ψ, χ) resulting in a $D \times t$, where D represents the total number of transformed dihedral angle selections. In both cases t represents the conformations from the simulations.

SD2: This module removes dominant second-order spatial correlations by computing a spatial covariance matrix and performing principal component analysis (PCA). In addition to the simulation data, SD2 requires as input m, the subspace dimensionality. m can be adjusted by examining the inflection points in the cumulative variance plots that this module returns. SD2 diagonalizes the covariance matrix nd returns the eigenvalues S (size $m \times 1$), eigenvectors B ($3N$ or $D \times m$) and the projection matrix $Y = B^T X(m \times t)$.

SD4: This module (previously QAA [15]) attempts to resolve the intrinsic non-orthogonal spatial dependencies in atomistic fluctuations. The second order projections, Y, from SD2 are used to build a fourth order spatially correlated cumulant tensor. SD4 approximately diagonalizes this tensor to return an anharmonic mode matrix W ($3N$ or $D \times m$). The default ordering of the ANCA modes is based on the kurtosis of the projected coordinates. However, this ordering may not always correspond to a biophysically relevant reaction coordinate [18]. This can be attributed to the fact that ANCA pursues rare conformational events and if the projected coordinates correlate with such rare events, then ANCA can indeed provide

biophysically meaningful projections.

TD2: TD2 module removes dominant second-order temporal correlations by computing a time-delayed covariance matrix and perfroming PCA. The inputs to this module are similar to the SD2 module, with one additional parameter, τ, that denotes the lag time over which the temporal correlations are to be resolved. The outputs of this module include Z, a matrix obtained by projecting the simulation data on the dominant time-delayed eigenvectors and the corresponding eigenvalues.

TD4: TD4 module constructs a time-delayed fourth-order kurtosis tensor which is then approximately diagonalized to obtain anharmonic modes of motions once the second-order spatial and temporal correlations are resolved [31]. TD4 module is the temporal analog of the spatial SD4 module. The input parameters to this module includes the matrix Z (from the TD2 module), a user specified subspace value m denoting the number of desired anharmonic modes of motion, the lag time τ and the matrix V. The outputs from the module includes the separating matrix W.

The theory and interpretation of the four modules has been described in detail in Chapter 2.

4.2.4 Visualization

We provide the user with example iPython notebooks to visualize the results from the analyses over a web-browser. In order to visualize structural data obtained from ANCA, we provide scripts for generating anharmonic modes using PyMOL or VMD. Individual regions in the protein can be colored using the output PyMOL files.

Following are some of the modules to obtain insights about the statistics of biomolecular fluctuations. An example of BPTI (as extensively detailed in Chapter 2) is used to illustrate the functions.

getLongTails: This module takes as input the aligned coordinates and plots fourth-order statistical profile for MD simulations. It gives insights about anharmonicity which is quantified by fourth-order moment, kurtosis (κ). Plot derived from this module gives a summary through histogram of C^{α} postional deviations of atoms. The tails observed from

Figure 2.10 provides strong evidence for the existence of anharmonicity.

perResidueRMSF: The root mean square fluctuation (RMSF) provides a measurement of deviation between positions of a residue and some reference position which is averaged over time. Through this measurement, it provides an indication about the backbone flexibility of biomolecules as illustrated in Figure 4.1. Since rmsf is computed per residue, it takes as input the aligned trajectories and number of residues.

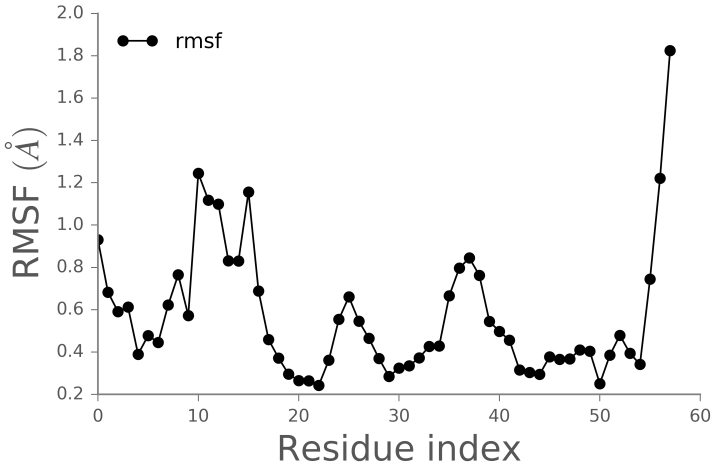


Figure 4.1: **RMSF of the backbone C^{α} atoms of 1.1 ms trajectory of BPTI comprising of 58 residues and 412497 conformers.**

perResidueAnharmonicityTime: Another approach towards quantifying anharmonicity in atomic fluctuations is to represent the time spent by each residue in all three directions in the tails of the distribution observed from Figure 2.10. Using this piece of information, we can color code each residue of the three-dimenstional protein structure based on the time spent sampling anharmonic fluctuations.

perResidueKurtosis: In order to understand the presence of higher-order moments in each residue over the entire trajectory, we compute kurtosis of each residue averaged over each of the (x, y, z) directions. This is shown in Figure 4.3

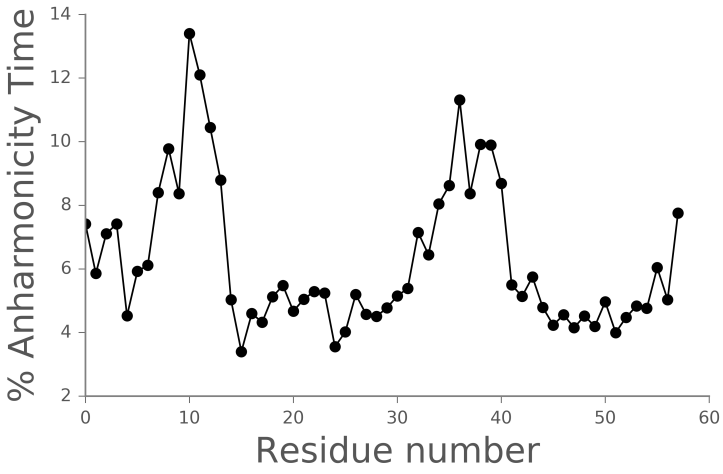


Figure 4.2: **Percentage time spent by residues sampling anharmonic conformational fluctuations.**

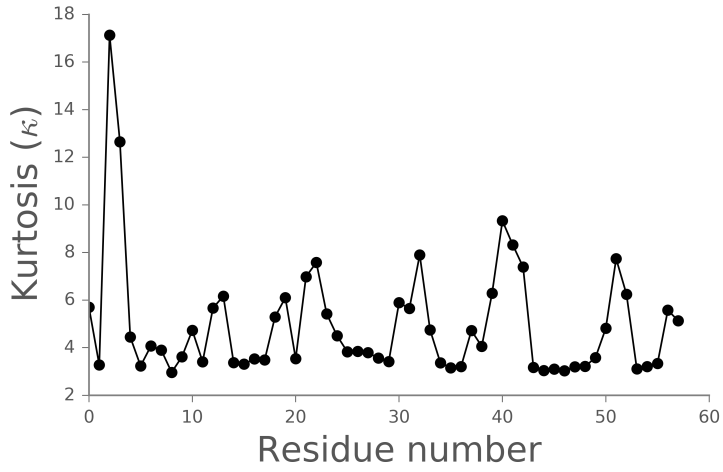


Figure 4.3: **Kurtosis values spread across different residues over the entire 1.1 ms trajectory frame.**

4.3 CONCLUSION

Several applications support analyses of MD trajectories based on second-order statistics, including MDAnalysis and mdtraj. To complement these tools, we have developed pyANCA as a package for analyzing higher-order anharmonic motion signatures from MD simulations. pyANCA provides a biophysically meaningful organizational principle for long timescale biomolecular simulations and can be integrated with software such as PyEMMA [51] to build Markov models of MD simulations, which we are pursuing as part of our future work. pyANCA is available as an open-source Python package under BSD 3-Clause license. Python tutorial notebooks, documentation and examples are available from http://csb.pitt.edu/anca for download.

BIBLIOGRAPHY

[1] Richard P Feynman, Robert B Leighton, and Matthew Sands. *Six not-so-easy pieces: Einstein's relativity, symmetry, and space-time*. Basic Books, 2011.

[2] Arvind Ramanathan and Pratul K Agarwal. Computational identification of slow conformational fluctuations in proteins. *The Journal of Physical Chemistry B*, 113(52):16669–16680, 2009.

[3] Giuseppe Zaccai. How soft is a protein? a protein dynamics force constant measured by neutron scattering. *Science*, 288(5471):1604–1607, 2000.

[4] M Kovermann, P Rogne, and M Wolf-Watz. Protein dynamics and function from solution state nmr spectroscopy. *Quarterly reviews of biophysics*, 49:e6, 2016.

[5] Bela Farago, Jianquan Li, Gabriel Cornilescu, David JE Callaway, and Zimei Bu. Activation of nanoscale allosteric protein domain motion revealed by neutron spin echo spectroscopy. *Biophysical Journal*, 99(10):3473–3482, 2010.

[6] Kasper D Rand, Martin Zehl, and Thomas JD Jørgensen. Measuring the hydrogen/deuterium exchange of proteins at high spatial resolution by mass spectrometry: overcoming gas-phase hydrogen/deuterium scrambling. *Accounts of chemical research*, 47(10):3018–3027, 2014.

[7] Gerhard Hummer, Friedrich Schotte, and Philip A Anfinrud. Unveiling functional protein motions with picosecond x-ray crystallography and molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15330–15334, 2004.

[8] Martin Karplus and Gregory A Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294):631, 1990.

[9] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[10] Sadaf R Alam, Pratul K Agarwal, Melissa C Smith, Jeffrey S Vetter, and David Caliga. Using fpga devices to accelerate biomolecular simulations. *Computer*, 40(3), 2007.

[11] Fernanda Duarte, Beat Anton Amrein, and Shina Caroline Lynn Kamerlin. Modeling catalytic promiscuity in the alkaline phosphatase superfamily. *Physical Chemistry Chemical Physics*, 15(27):11160–11177, 2013.

[12] Katherine A Henzler-Wildman, Ming Lei, Vu Thai, S Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913, 2007.

[13] David D Boehr, Dan McElheny, H Jane Dyson, and Peter E Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *science*, 313(5793):1638–1642, 2006.

[14] James S Fraser, Michael W Clarkson, Sheena C Degnan, Renske Erion, Dorothee Kern, and Tom Alber. Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, 462(7273):669, 2009.

[15] Arvind Ramanathan, Andrej J Savol, Christopher J Langmead, Pratul K Agarwal, and Chakra S Chennubhotla. Discovering conformational sub-states relevant to protein function. *PLoS One*, 6(1):e15827, 2011.

[16] Toshiko Ichiye and Martin Karplus. Anisotropy and anharmonicity of atomic fluctuations in proteins: analysis of a molecular dynamics simulation. *Proteins: Structure, Function, and Bioinformatics*, 2(3):236–259, 1987.

[17] Martin Karplus and Joseph N Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.

[18] Arvind Ramanathan, Andrej J Savol, Pratul K Agarwal, and Chakra S Chennubhotla. Event detection and sub-state discovery from biomolecular simulations using higher-order statistics: Application to enzyme adenylate kinase. *Proteins: Structure, Function, and Bioinformatics*, 80(11):2536–2551, 2012.

[19] Di Wu and Zhijun Wu. Superimposition of protein structures with dynamically weighted rmsd. *Journal of molecular modeling*, 16(2):211–222, 2010.

[20] Johann Deisenhofer and W Steigemann. Crystallographic refinement of the structure of bovine pancreatic trypsin inhibitor at l. 5 å resolution. *Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry*, 31(1):238–250, 1975.

[21] Gerhard Wider, Kong Hung Lee, and Kurt Wüthrich. Sequential resonance assignments in protein 1h nuclear magnetic resonance spectra: glucagon bound to perdeuterated dodecylphosphocholine micelles. *Journal of molecular biology*, 155(3):367–388, 1982.

[22] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585, 1977.

[23] MJ Kunitz and John H Northrop. Isolation from beef pancreas of crystalline trypsinogen, trypsin, a trypsin inhibitor, and an inhibitor-trypsin compound. *The Journal of general physiology*, 19(6):991–1007, 1936.

[24] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[25] Christopher L McClendon, Gregory Friedland, David L Mobley, Homeira Amirkhani, and Matthew P Jacobson. Quantifying correlations between allosteric sites in thermodynamic ensembles. *Journal of chemical theory and computation*, 5(9):2486–2502, 2009.

[26] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[27] Wenjun Zheng, Bernard R Brooks, and D Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proceedings of the National Academy of Sciences*, 103(20):7664–7669, 2006.

[28] Jean-François Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157–192, 1999.

[29] H GOLUB Gene and F Charles. Matrix computations. *Johns Hopkins Universtiy Press, 3rd edtion*, 1996.

[30] Hans Frauenfelder, Fritz Parak, and Robert D Young. Conformational substates in proteins. *Annual review of biophysics and biophysical chemistry*, 17(1):451–479, 1988.

[31] Pando Georgiev and Andrzej Cichocki. Robust independent component analysis via time-delayed cumulant functions. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 86(3):573–579, 2003.

[32] Horace R Drew, Richard M Wing, Tsunehiro Takano, Christopher Broka, Shoji Tanaka, Keiichi Itakura, and Richard E Dickerson. Structure of a b-dna dodecamer: conformation and dynamics. *Proceedings of the National Academy of Sciences*, 78(4):2179–2183, 1981.

[33] John D Chodera, William C Swope, Jed W Pitera, and Ken A Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Modeling & Simulation*, 5(4):1214–1226, 2006.

[34] TH Reijmers, R Wehrens, and LMC Buydens. Circular effects in representations of an rna nucleotides data set in relation with principal components analysis. *Chemometrics and Intelligent Laboratory Systems*, 56(2):61–71, 2001.

[35] Yuguang Mu, Phuong H Nguyen, and Gerhard Stock. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Structure, Function, and Bioinformatics*, 58(1):45–52, 2005.

[36] K Gunasekaran, C Ramakrishnan, and P Balaram. Disallowed ramachandran conformations of amino acid residues in protein structures. *Journal of molecular biology*, 264(1):191–198, 1996.

[37] Nicholas J West and Lorna J Smith. Side-chains in native and random coil protein conformations. analysis of nmr coupling constants and $\chi 1$ torsion angle preferences1. *Journal of molecular biology*, 280(5):867–877, 1998.

[38] Alexandros Altis, Phuong H Nguyen, Rainer Hegger, and Gerhard Stock. Dihedral angle principal component analysis of molecular dynamics simulations. *The Journal of chemical physics*, 126(24):244111, 2007.

[39] Philipp Berens et al. Circstat: a matlab toolbox for circular statistics. *J Stat Softw*, 31(10):1–21, 2009.

[40] Naveen Michaud-Agrawal, Elizabeth J Denning, Thomas B Woolf, and Oliver Beckstein. Mdanalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry*, 32(10):2319–2327, 2011.

[41] Daniel R Roe and Thomas E Cheatham III. Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, 9(7):3084–3095, 2013.

[42] Andrea Amadei, Antonius Linssen, and Herman JC Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425, 1993.

[43] Oliver F Lange and Helmut Grubmüller. Can principal components yield a dimension reduced description of protein dynamics on long time scales? *The Journal of Physical Chemistry B*, 110(45):22842–22852, 2006.

[44] Boryeu Mao, Michael R Pear, JA McCammon, and SH Northrup. Molecular dynamics of ferrocytochrome c: anharmonicity of atomic displacements. *Biopolymers*, 21(10):1979–1989, 1982.

[45] Toshiko Ichiye and Martin Karplus. Anisotropy and anharmonicity of atomic fluctuations in proteins: implications for x-ray analysis. *Biochemistry*, 27(9):3487–3497, 1988.

[46] Arvind Ramanathan, Andrej Savol, Virginia Burger, Chakra S Chennubhotla, and Pratul K Agarwal. Protein conformational populations and functionally relevant substates. *Accounts of chemical research*, 47(1):149–156, 2013.

[47] Andrej J Savol, Virginia M Burger, Pratul K Agarwal, Arvind Ramanathan, and Chakra S Chennubhotla. Qaarm: quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin. *Bioinformatics*, 27(13):i52–i60, 2011.

[48] Virginia M Burger, Arvind Ramanathan, Andrej J Savol, Christopher B Stanley, Pratul K Agarwal, and Chakra S Chennubhotla. Quasi-anharmonic analysis reveals intermediate states in the nuclear co-activator receptor binding domain ensemble. In *Biocomputing 2012*, pages 70–81. World Scientific, 2012.

[49] Richard J Gowers, Max Linke, Jonathan Barnoud, Tyler JE Reddy, Manuel N Melo, Sean L Seyler, David L Dotson, Jan Domanski, Sébastien Buchoux, Ian M Kenney, et al. Mdanalysis: a python package for the rapid analysis of molecular dynamics simulations. In *Proceedings of the 15th Python in Science Conference*, pages 98–105, 2016.

[50] Robert T McGibbon, Kyle A Beauchamp, Matthew P Harrigan, Christoph Klein, Jason M Swails, Carlos X Hernández, Christian R Schwantes, Lee-Ping Wang, Thomas J Lane, and Vijay S Pande. Mdtraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal*, 109(8):1528–1532, 2015.

[51] Martin K Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Perez-Hernandez, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noe. Pyemma 2: a software package for estimation, validation, and analysis of markov models. *Journal of chemical theory and computation*, 11(11):5525–5542, 2015.