ANALYSIS OF MOLECULAR DYNAMICS SIMULATIONS OF PROTEIN FOLDING

BY

ARUNA RAJAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Doctoral Committee:

Assistant Professor Yann Chemla, Chair
Professor Yoshitsugu Oono, Director of Research
Professor Klaus Schulten
Assistant Professor Smitha Vishveshwara

Abstract

Microsecond long Molecular Dynamics (MD) trajectories of biomolecular processes are now possible due to advances in computer technology. Soon, trajectories long enough to probe dynamics over many milliseconds will become available. Since these timescales match the physiological timescales over which many small proteins fold, all atom MD simulations of protein folding are now becoming popular. To distill features of such large folding trajectories, we must develop methods that can both compress trajectory data to enable visualization, and that can yield themselves to further analysis, such as the finding of collective coordinates and reduction of the dynamics. Conventionally, clustering has been the most popular MD trajectory analysis technique, followed by principal component analysis (PCA). Simple clustering used in MD trajectory analysis suffers from various serious drawbacks, namely, (i) it is not data driven, (ii) it is unstable to noise and change in cutoff parameters, and (iii) since it does not take into account interrelationships amongst data points, the separation of data into clusters can often be artificial. Usually, partitions generated by clustering techniques are validated visually, but such validation is not possible for MD trajectories of protein folding, as the underlying structural transitions are not well understood. Rigorous cluster validation techniques may be adapted, but it is more crucial to reduce the dimensions in which MD trajectories reside, while still preserving their salient features. PCA has often been used for dimension reduction and while it is computationally inexpensive, being a linear method, it does not achieve good data compression. In this thesis, I propose a different method, a nonmetric multidimensional scaling (nMDS) technique, which achieves superior data compression by virtue of being nonlinear, and also provides a clear insight into

the structural processes underlying MD trajectories. I illustrate the capabilities of nMDS by analyzing three complete villin headpiece folding and six norleucine mutant (NLE) folding trajectories simulated by Freddolino and Schulten [1]. Using these trajectories, I make comparisons between nMDS, PCA and clustering to demonstrate the superiority of nMDS.

The three villin headpiece trajectories showed great structural heterogeneity. Apart from a few trivial features like early formation of secondary structure, no commonalities between trajectories were found. There were no units of residues or atoms found moving in concert across the trajectories. A flipping transition, corresponding to the flipping of helix 1 relative to the plane formed by helices 2 and 3 was observed towards the end of the folding process in all trajectories, when nearly all native contacts had been formed. However, the transition occurred through a different series of steps in all trajectories, indicating that it may not be a common transition in villin folding. The trajectories showed competition between local structure formation/hydrophobic collapse and global structure formation in all trajectories. Our analysis on the NLE trajectories confirms the notion that a tight hydrophobic core inhibits correct 3-D rearrangement. Only one of the six NLE trajectories folded, and it showed no flipping transition. All the other trajectories get trapped in hydrophobically collapsed states. The NLE residues were found to be buried deeply into the core, compared to the corresponding lysines in the villin headpiece, thereby making the core tighter and harder to undo for 3-D rearrangement. Our results suggest that the NLE may not be a fast folder as experiments suggest. The tightness of the hydrophobic core may be a very important factor in the folding of larger proteins. It is likely that chaperones like GroEL act to undo the tight hydrophobic core of proteins, after most secondary structure elements have been formed, so that global rearrangement is easier. I conclude by presenting facts about chaperone-protein complexes and propose further directions for the study of protein folding.

To Appa, Amma and Archana.

Acknowledgments

Many words of thanks are due to those,
Who've taught, explained and clarified;
Whose resolute patience I have tried,
As far as resolute patience goes!
To dearest ones, who've stood by me,
Through harsh times and misery;
To friends, for the love they have shown,
And for all the kindness I've ever known.
To those of you still plodding through,
This very painfully conceived device,
If fuzzy punctuation doesn't get to you,
It could be the addled passive voice.
To music,
That has always filled this world of mine,
With pleasant thoughts, and helped forget.
Its been terrible, but it will be fine,
In nostalgic retrospect.

I would like to thank my advisor, Yoshitsugu Oono for all the help, guidance and the immense patience he has shown me. From wading through my confused punctuation marks, color scheme of figures to clarifying every concept addressed within and without the scope of my thesis and advising me on every issue I have been faced with in graduate school, there is

much I owe him deep gratitude for. But for his support and intelligent insights into science, this thesis could not have taken shape. I would also like to thank Klaus Schulten for having first pointed me to my thesis problem, for valuable input in my research, computational support, research grants and various other help. This thesis would not have been possible without Peter Freddolino and Satwik Rajaram. I am grateful to Peter for spending many hours running long MD simulations that I have analyzed for this thesis and to Satwik for providing much code and clarification of every detail for the methods used in this thesis. I am deeply indebted to Peter and Satwik for providing me with figures, help with writing and for painstakingly carefully answering every question I have bothered them with. I thank John Stack, who has been very instrumental in guiding me through the PhD process, been very kind to me with the teaching assignments and has always been available to talk to me about my concerns. I would also like to thank my committee members for their useful suggestions and comments relating to my research and thesis. I thank Eduard, Eric, JC, Ramya, Jeremy and all the members of Yoshi's research group and TCBG for various useful discussions.

Had it not been for my dearest friend, Prabhu, I would not have ventured to do a PhD in Physics. I am grateful to him for his constant encouragement throughout my undergraduate years which led me to choose physics; his constant presence as a pillar strength throughout my years at graduate school, for being there with me when I have fallen sick or broken bones (as I have done very often!), his infinite patience, friendship and love. To Rajesh and Pavan, I owe a debt past telling. Rajesh has always been available to listen to everything, from my research to the broken heat vent in my apartment to my happiness at learning a new piece of music to my lacklustre description of my daily life. Pavan has always had more faith in me than I have had in myself and always been a cherished and wonderful friend. It would be insufficient to say that without these people, I would have never made it through. My heartfelt gratitude to my cousin Vijay, my uncle and family, and several relatives in this lonely country. Through all my times of strife, I have relied on them for much support.

Any mention of graduate school would be incomplete without the mention of all the lovely people in Champaign that have made it worthwhile. My oldest friend in Champaign, Abhishek Roy, has always been helpful with his intelligence and useful insights into all I have done, stayed by me through difficult times, cooked and eaten many wonderful meals with me and kept my life full of chatter about music, books, movies and all those wonderful things that can make daily life endearing. His friendship and love have gone a long way in making this thesis possible. I must thank Shiying, my constant companion and alter ego. Her chatter and company have gotten me through many dreary nights of research work and thesis writing. A special thanks to my most wonderful friends from Champaign: Laura for her friendship, love, food, yummy desserts, for wonderful company on fun trips and most endearing chatter, her patience for my woes and sound advice always; Esi for being my most intelligent and resourceful homework partner, for his friendship and chatter, for being a fantastic swimming partner and for inspiring me with his laid back and easygoing attitude; Roberto for making life in Chambana more bearable with Salsa lessons, movie sessions, cooking sessions and chatter; Themis and Abhi for being my home away from home: their parties, picnics, counseling sessions and their large couch and TV, and kitchen with unlimited food supply; Rahul, Satwik and Nayana for parenting me and advising me through my worst dilemmas in graduate school; Vidisha for movie sessions, chai sessions, unending chatter and pleasant company; Parag, Radhika, Travis, Akbar and Serena for their unlimited patience, love and support; and many other wonderful friends from UIUC. The last year at Champaign has been my best one thanks to the newly found music group : Ram, Satya, Shiral, Mohit, Aditya, Subha and Bharat. But for our music sessions, Friday night Blind Pig visits, coffee, dinner and chatter, for all their sound advice on music and life, I don't know how I could have gone through these last few months of research/thesis writing. Kuanwu Lin, for helping me through my difficulties, every time I have approached him, and for always encouraging me to use my intuition and better sense. Thank you to all my friends, from all over the world, who encourage me to talk and write my woes away

in verse. The most memorable and important part of graduate school will always be the friendship and love I have found here. Without that and without facebook, the wonderful social utility for passing time more usefully (some people call it procrastinating), without all the joy that youtube and the internet have brought, I would have never found the sense of humour to sail me through.

The PhD has been a long and difficult run and I am grateful to my family, without whom I could not have survived to make this finish line. I cannot find enough words to thank my parents and sister for their unending love, support and affections, for continuously believing in me when I lost hope and for edging me on to complete this ordeal. Every time I have felt worn out, I owe my return to research and renewed energy to all their love and encouragement, to the wonderful place that is India, with all its warm people and million colours that can make anybody happy.

This work was supported by NIH grant P41-RR05969 and NSF grant PHY0822613, NSF funded Center for the Physics of Living Cells at UIUC, Keio University through Yoshi Oono and the Department of Physics at UIUC. Computer time was provided by the NationalCenter for Supercomputing Applications through grant MCA93S028.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this thesis, I discuss the adaptation of dimension reduction methods to analyze Molecular Dynamics (MD) trajectories and illustrate our methods by analyzing the folding trajectories simulated by Freddolino and Schulten [1]. I will first develop the background necessary to understand protein folding in our context. Next, I will briefly introduce the simulations performed by Freddolino and Schulten, and then explain why it is crucial to develop and adapt trajectory analysis methods such as discussed in this thesis.

1 Background

Proteins are molecular machines that carry out most chemical, mechanical and other important cellular functions in living organisms. A protein's function is determined by its three-dimensional structure, which is in turn determined by the corresponding amino acid sequence [2]. Proteins begin as long polypeptide chains synthesized by the ribosome and then fold to their three-dimensional structure. When they do not fold correctly, proteins aggregate to cause various diseases, such as Alzheimer's, Parkinson's and Huntington's, to name a few. This has made protein folding a hot topic of study for over 50 years now.

1.1 General facts about protein folding

I will now briefly summarize whatever is known so far about folding.

1. Proteins are classified into structured (ordered) and not completely structured (amorphous proteins discussed in [3]). Amorphous proteins do not have definite natural

conformations, they change their structure depending on cellular conditions, chaperones acting upon them and the functionality required.

2. Structured proteins fold into definite 3D structure in a given cellular environment according to their primary sequences. There are many important structured proteins that cannot spontaneously fold [4], but need the assistance of chaperones, molecules that bind to proteins to accelerate folding. There is no experimental evidence to indicate that all proteins fold to their equilibrium structures. We should not expect that the native (biologically functional) conformation for a (large) protein (even assisted by chaperones) is the equilibrium state (the lowest free energy). It is more natural to assume that proteins are generally in metastable states. For small ($\leq$100 residues) proteins however, it is likely that the native state is the free energy minimum.

3. The folding free energy is a few k_BT atmost. Perhaps to make flexible molecular machines that can adapt to their environment, the free energy difference between the unfolded and folded conformations are not large. However, folded conformers must be stable against mutation, often with the help of chaperones [5].

4. Although ΔG may be small, however, we should not forget that ΔS from the unfolded to folded state and ΔH are both larger negative quantities. It is thus, hard to drive the system to the folded state. Chaperones prevent chains from going astray into wrong conformations [6, 4]. Especially, under different conditions, the same protein may fold via different pathways. Multiple ways to fold make folding robust, and evolution is likely to have selected for this robustness to ensure that a protein will fold under the varying conditions prevalent in different cellular contexts [7, 8]. Given this, it is questionable if a pathway is a meaningful concept.

Traditionally, the study of protein folding problem has been two fold: 1) Can we predict the 3-D structure of a protein given its sequence? 2) Starting from a linear chain of amino acids, coded by its sequence, how does a protein reach its biologically correct 3-D folded

state? This thesis addresses the study of the second problem. More specifically, it provides methods to analyze the folding and other dynamical changes in proteins, studied by molecular dynamic simulations.

Levinthal's paradox

In 1969, Levinthal estimated that for a protein with 100 amino acids, the time taken to thermally sample all available configurations (which he calculated to be 10^{67}) would be 10^{30} times longer than the expected life time of the universe [9]. He observed however, that many small proteins (in biological systems) fold spontaneously in a few milliseconds. Thus, a paradox seems to result. Levinthal himself resolved this "paradox" by concuding that proteins do not fold by sampling all available conformations but by following specific pathways. Following this, many researchers proposed various specific folding pathways, some of which are described below:

1. *The framework model or collision-diffusion model* proposes a hierarchical assembly by which most elements of the native secondary structure are formed first according to the primary sequence, but independent of tertiary structure. The tertiary structure results from the collisions of these secondary structure elements amongst themselves [10].

2. *The nucleation model* suggests folding is initiated by the formation of a seed or a unit of native secondary structure by only a few residues (e.g. a beta-turn or a helical turn) [11].

3. *The hydrophobic collapse model* proposes that the driving force in protein folding is hydrophobic collapse to form a molten globule. The native state then forms by the rearrangement of the collapsed molten globule structure [12, 13].

Each of these models has had a few successes in explaining experimental data of certain proteins [14, 15, 16].

We believe that the more fruitful way to think about the Levinthal's paradox is that biological proteins are not random sequences of amino acids, but have been evolutionarily selected because they fold fast. For example, for an N residue protein, there are 20^N possible

amino acid sequences. However, only a small fraction of these sequences are found in biology. This indicates that the notion of a pathway was ill founded, even conceptually, not just empirically. Biology does not solve the folding problem by remembering specific pathways of folding, instead, it selects special sequences that fold in multiple ways (not to be confused with pathways) over physiologically relevant timescales, spontaneously or with the help of chaperones.

1.2 MD studies of protein folding

Computational studies of folding of small proteins have always been going hand in hand with experiments. Such studies aim to find folding pathways, important intermediates and transition states. However, it should be clearly recognized that coarse-grained dynamics cannot be relied on to understand kinetics. Coarse graining implictly assumes that the potential energy can be re-parameterized at different length scales, however there is no proof of such re-parameterizations being reliable for proteins over folding timescales. Protein kinetics is very sensitive to small structural changes and thus, Monte Carlo simulation is not reliable; coarse-grained molecular dynamics (MD) is perhaps better than the former, but its reliability has not been characterized yet. Complete atomistic simulations are the only way we can hope to study protein folding honestly. We must note, however, that perhaps only a small class of proteins may be simulated in detail by MD. Many proteins have folded states that are very sensitive to the conditions inside the cell [5, 17] and MD simulations cannot simulate all the conditions of the cell accurately enough to study such proteins.

The biggest challenge that MD simulations of protein folding face are that of timescale and accuracy. The currently known fastest folding proteins fold over 0.7 - 1.0 μs at room temperature [18]. There is a hypothesized limit of around $(N/100)$ μs for the folding timescale of an N residue protein [19]. Only recently has technological progress enabled atomistic MD simulations to probe microsecond timescale [20, 21, 22].

One of the most commonly studied fast folding proteins is the villin headpiece, a 35

residue actin-binding domain, which folds into a three helix bundle with a hydrophobic core in about 5.5 μs [1] at room temperature. In 1998, Duan and Kollman simulated the villin headpiece in what had been the longest simulation (of 1 μs) until then [23]. Recently, a fast folding mutant (that folds in $\approx$ 1 μs) for villin has been found by experiments [18]. Kubelka et al., found that by substituting two of the buried lysines (residues 24 and 29) in villin with norleucine, a fast folding mutant can be engineered. Freddolino and Schulten have simulated first complete folding trajectories for the villin headpiece and its norleucine mutant in explicit solvent [1].

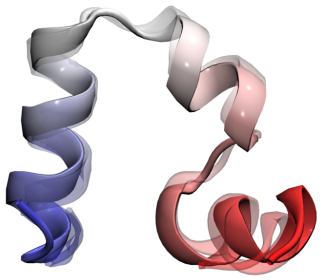

Figure 1.1: **Villin headpiece**: The final folded villin headpiece structure obtained from Freddolino and Schulten trajectories superimposed over the crystal structure of villin headpiece [1].

Simulations by Freddolino and Schulten

Three simulations of the villin-headpiece and six simulations of the norleucine (NLE) mutant were performed in NAMD 2.5/2.6 [24] by Freddolino and Schulten [1], with starting conformations prepared from a fully extended structure and heat denatured at 450 K. In all three villin headpiece simulations, the protein folded over $\sim$6 μs and stayed in the folded state for 1-2 μs after folding, when the trajectories were terminated. Only one of the NLE simulations reached completion. One more NLE trajectory explored a near native state, while the other four were trapped in non native hydrophobically collapsed states. The NLE simulations were terminated 1 μs after the protein folded or after 8 μs for the incomplete simulations.

The CHARMM22 force field with CMAP corrections [25] was used for the simulations. Short range nonbonded interactions were cut off at 8.0 Å with switching beginning at 7.0 Å; long range electrostatics were treated using the particle mesh Ewald method. All bonds involving hydrogens in the protein were constrained using the RATTLE algorithm [26] with water geometry maintained using SETTLE [27]. An integration timestep of 2.0 fs was used, with bonded and short range interactions evaluated every timestep and long range electrostatics once every three timesteps. A temperature of 337 K was maintained using a Langevin thermostat with a damping constant of 0.1 ps^{-1} to mimic experimental conditions [19, 18]. The crystal structure of the protein was simulated for 200 ns and found to be stable in the force field [1]. Below, I show the time trace of the local secondary structure of each residue, the Q value (defined as the fraction of native contacts present in a structure; this value increases from 0 for the completely unfolded state to 1 for the folded state), and the C_α RMSD (the root mean square distance of the amino acid C_α atoms in each frame from those of the crystal structure) for each villin headpiece trajectory [1].

In all three trajectories, helix 1 and helix 3 (labelled in red and blue respecitively in Fig. 1.1) are formed within the first 1-2 μs. In trajectory 1, helix 2 (labelled in white in Fig. 1.1) forms only in the last microsecond [1], where as in the other two trajectories, helix 2 forms early on and associates with helix 1 to form two-helix states. In Fig. 1.2, we see that the Q value increases in punctuated steps from 0 to 1 in all trajectories indicating that there might be minimal frustration in the protein sequence. Frustration in protein folding refers to the formation of non native contacts which compete with the formation of native contacts.

To obtain the number of (pairs of) residues in contact, we counted all pairs of hydrophobic residues within 4 Å of each other and all pairs of polar residues/hydrogen bonding partners within 3.5 Å of each other in all trajectories. In all three trajectories, the number of non-native contacts formed was $\leq 10\%$ of the number of native contacts formed. Most non-native contacts were formed by the residues not involved in the formation of native secondary

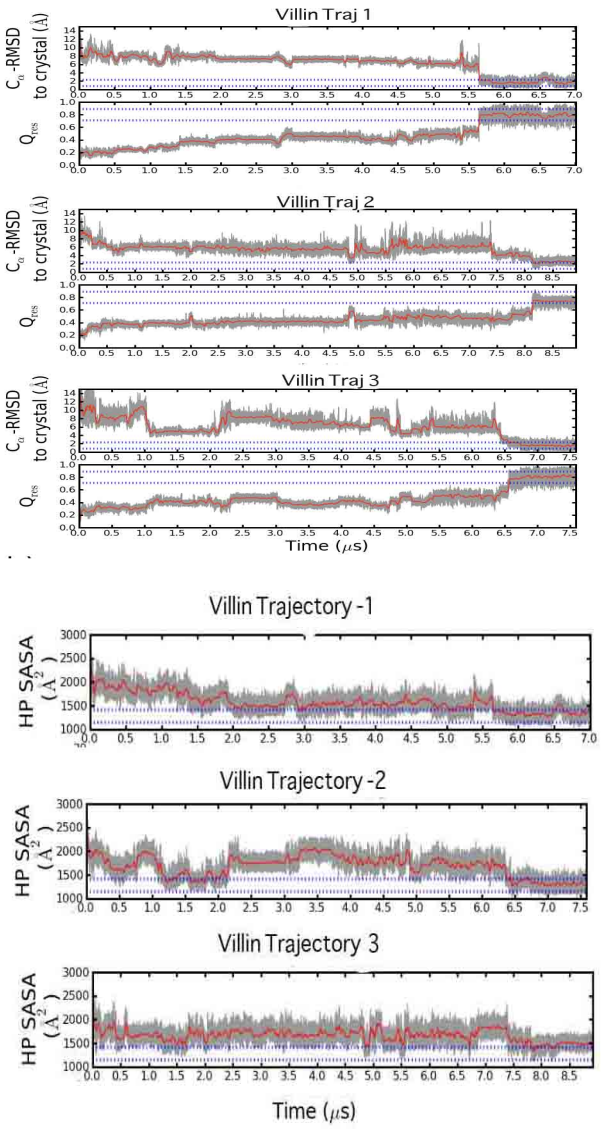

Figure 1.2: **Salient features of Villin headpiece folding trajectories**: Running averages over 30 ns are shown in red, and the range defined by the mean and two standard deviations from simulation of the native state as blue bars. HP SASA refers to solvent accessible surface area (SASA) of hydrophobic groups.(Adapted Fig. Courtesy: Peter Freddolino, TCBG, UIUC [1])

structure elements and floppy end terminal. Subsequent to formation, such contacts were quickly broken. However in the second trajectory, the protein is trapped in a state with non-native contacts for 1 μs after a rapid hydrophobic collapse (Fig. 1.2).

In case of the NLE mutant, while one trajectory (NLE-FOLD1) folded to its native state in roughly 2.5 μs, only one other trajectory, NLE-FOLD3, even reached a near native state.

NLE-FOLD3 showed near native states after about 1 μs and then again after 7 μs, but the protein is not considered folded as the distribution of folding observables does not match that from the crystal structure simulation [1].

2 Trajectory Analysis

Folding trajectories presented above [1] contain millions of frames (each frame being one snapshot in time of all of the protein's atomic coordinates) and in order to obtain a qualitative picture of the folding process or perform free energy calculations of relevant structural transitions, it is important to obtain reduced representations of these trajectories. Soon, longer folding trajectories may become available. It will then be crucial to develop methods that will distill salient features and enable quantitative analysis.

Conventionally, clustering algorithms have been the most popular method of choice for MD trajectory analysis [28]. Simple clustering algorithms used to analyze MD trajectories [28, 29, 30] require specification of the number of clusters or a cluster radius, making the clustering artificial and not data driven, that is, (i) the clusters are unstable against small changes in cutoff parameters and noise in the data and (ii) inter relationships between cluster members are not taken into account when clustering, leading to do artificial clusters. When simple cutoff based clustering was applied to villin folding trajectories, we found that the cluster centers produced were unstable to changes in coordinates included, binning time and cutoff parameters (to be discussed in Chapter 2). An additional goal of MD simulations of folding processes is to find collective coordinates. Clustering does not yield itself to such analysis. There is clearly a need to go beyond clustering to analyze MD folding trajectories. Dimension reduction methods are better choices to reduce long MD trajectories and this thesis discusses the use of a popular methods, principal component analysis (PCA) to reduce trajectories, as well as introduces a more stable, nonlinear dimension reduction method, nMDS and shows that it is a robust and dependable method to analyze MD trajectories.

3 Our Approach/Roadmap

In Chapter 2, I discuss conventional clustering methods used to analyze MD trajectories and some of their drawbacks. In Chapter 3, I discuss how dimension reduction methods can be applied to MD trajectories to represent them in a reduced space, introducing PCA. I discuss the usefulness of PCA for trajectory analysis and also enumerate its drawbacks. In Chapter 4, I present a robust nonlinear dimension reduction method, nMDS, that we have successfully adapted for the first time to analyze MD trajectories. I discuss the results obtained from nMDS and the insights obtained into villin headpiece and the norleucine mutant folding. I make comparisons between nMDS, PCA and clustering and show how nMDS performance is superior (except in terms of computational requirements). I also suggest various improvements to nMDS and ways to extract collective coordinates of more folding MD trajectories, which may become available in the near future. In Chapter 5, I conclude by summarizing and analyzing existing viewpoints on protein folding and propose directions to study protein folding.

Chapter 2

Cluster Analysis of MD Trajectories

1 Introduction

The protein folding trajectories described in Chapter 2 have more than 10^7 frames containing all the atomic coordinates of the protein. It is, hence, very important to find a reduced representation for these trajectories so that we may obtain a picture of the underlying biological processes. Conventionally, clustering algorithms have been the methods of choice to distill the salient features of MD trajectories [28, 29, 30].

Clustering algorithms are a class of unsupervised data-reduction methods that classify patterns into groups known as clusters, such that the patterns within a cluster are related in some way [31]. Cluster analysis has become one of the most popular and widely used data-reduction methods. It continues to be used in a diverse range of fields and applications including social network analysis [32], market research [33], search result grouping [34] and image segmentation [35]. For MD trajectory analysis, cluster analysis holds a dominant position, being used to classify the frames in a trajectory according to their similarities in structure, measured with a Euclidean metric in configuration space.

Undoubtedly, cluster analysis is the best choice for dimensional reduction in many contexts. However, the simplicity of cluster analysis results belies the many assumptions that are implicitly made (depending on the specific algorithm chosen), which in turn limit its use as a data driven method. Although many cluster analysis algorithms are currently used to analyze MD trajectories [28, 29, 36, 30], I will discuss a popular clustering technique used by Freddolino and Schulten in [1]. This technique is implemented in a program called

GROMACS [37] for use with MD trajectories. I will list the drawbacks of this method and also illustrate its unreliability when applied to MD trajectories.

As far as application to MD trajectories is concerned, we will restrict our discussion to the set of algorithms that conform to the following definition: Given a set of N patterns $S = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}\}$, that either a) exist in space with a well defined distance measure or b) have a known set of pairwise distances d_{ij}, a cluster analysis algorithm is one that generates a partition $P = \{P_i\}$ of S in a way that the patterns in each P_i are maximally alike, while those in different ones differ.

Clustering methods used to analyze MD trajectories usually specify a cutoff on the cluster radius or total number of clusters desired [30, 36]. In this thesis, we use the GROMACS program for comparative studies. In the GROMACS algorithm used with a gromos field [30] in GROMACS, the RMSD of all atomic coordinates (can restrict it to C_α coordinates or any heavy atom coordinates also) between all pairs of structures are determined. For each structure, the number of structures for which the ℓ_2 or root mean square distance (RMSD) was less than the desired cutoff is calculated. The structure with the highest number of neighbours is then taken as the center of a cluster, and formed together with all its neighbours a (first) cluster. This process is repeated until all the structures (frames) in the trajectory are successfully classified into clusters. Thus, the output of the GROMACS program is a set of frame numbers designated as cluster centers and the corresponding members in each cluster.

There are severals problems with clustering and though they may not affect qualitative visualization of trajectories, they can be disastrous if clustering is used for quantitative analysis. Some serious problems of clustering methods are listed below :

a) They are not data driven. They implicitly assume that there exists a certain cut-off parameter which can be used to bin the data.

b) They are often unstable to changes in cutoff parameters and noise.

c) They do not provide any information about the inter-relationships between clusters.

d) Partitions generated by clustering are generally validated by visual inspection of the structures returned as cluster centers. Since little is known about protein dynamics en-route to folding, visual inspection is not a reliable way of validating clustering techniques applied to MD simulations of protein folding.

e) They do not easily yield themselves to analysis involving finding of collective coordinates.

2 Unreliability of clustering results

The instabilities of conventional clustering is presented with quantitative analysis in this section.

2.1 Change in cutoff parameters

First, all the trajectories were binned at 3 ns and each frame was read in as a vector containing all the C_α coordinates of the protein. We now calculate pairwise RMSDs between frames and construct an internal coordinate system as follows. Suppose that there are N frames in the trajectory (or that the trajectory is divided into N equally spaced snapshots). Then, let us construct a symmetric matrix M, defined by $M(i, j) = \ell_2$ distance (RMSD) between frame i and frame j. The matrix was constructed by computing the RMSD between all heavy atoms across frames after discarding some of the initial unfolded state frames and aligning all the frames (by appropriate rotations and translations). Each trajectory, thus, consisted of about 2000 vectors of 2000 dimensions each. Clustering was done using the Euclidean ℓ_2 (dRMSD) metric and a cutoff of 1.5 Å . We obtained about 100 cluster centers. When the cutoff was changed to 3 Å , we obtained only 73 cluster centers. Some of the clusters that were obtained when binned at 1.5 Å merged into larger clusters. It is impossible to tell whether 3 Å achieves better clustering than 1.5 Å without visually inspecting the data. Visual examination of partitions generated by clustering is not only cumbersome, but also

unreliable for protein folding trajectories as explained before. There is, hence, no way to judge which clusters may be representative of the pathway for quantitative calculations such as described in [38].

In order to demonstrate the unreliability of cluster centers, we chose the cluster centers (returned by cluster analysis) with at least 20 members and use nMDS to project them onto a two dimensional euclidean space. Figures **??** show clearly that the cluster centers shift around in the projected space when cutoff parameters are changed. Some clusters are merged, while others disappear. Although I emphasise that visual inspection is not a reliable way to validate clustering data for folding trajectories, to get a qualitative picture of which cluster centers were affected by varying cut offs. The clustering was tested using a range of cutoffs from 1.5 to 5 Å. The figure shows some conformations that appeared in two different clusters when binned at 1.5 Å but looked very similar visually.

2.2 Change in binning time

We binned the trajectories more coarsely in time by a factor of 10^3, 10^4 and 10^5 to see how the cluster centers shift when analyzed in the dihedral angle space (each trajectory frame read in as a 70 dimensional vector consisting of backbone angles of all residues). The number of cluster centers obtained went down from 100 for the 6 ps sampled trajectory to 88 to 72 to 30 (for each 10 fold increase in binning time). The drop in the number of cluster centers was not linear with increase in binning time. While this suggests that there might be a loss of features when the data are binned more coarsely in time, it is not possible to visually examine all the cluster centers produced to ascertain an ideal binning time.

What relationships do the cluster centers bear with each other? We attempt to examine this in a 2-D space (for visualization purposes). In Fig. 2.3, we show by projecting on to a 2-D euclidean space, how cluster centers were affected when the data from villin trajectory 2 in the dihedral angle space was binned in 1, 6 and 30 ns time steps. We show clusters with at least 20 members.

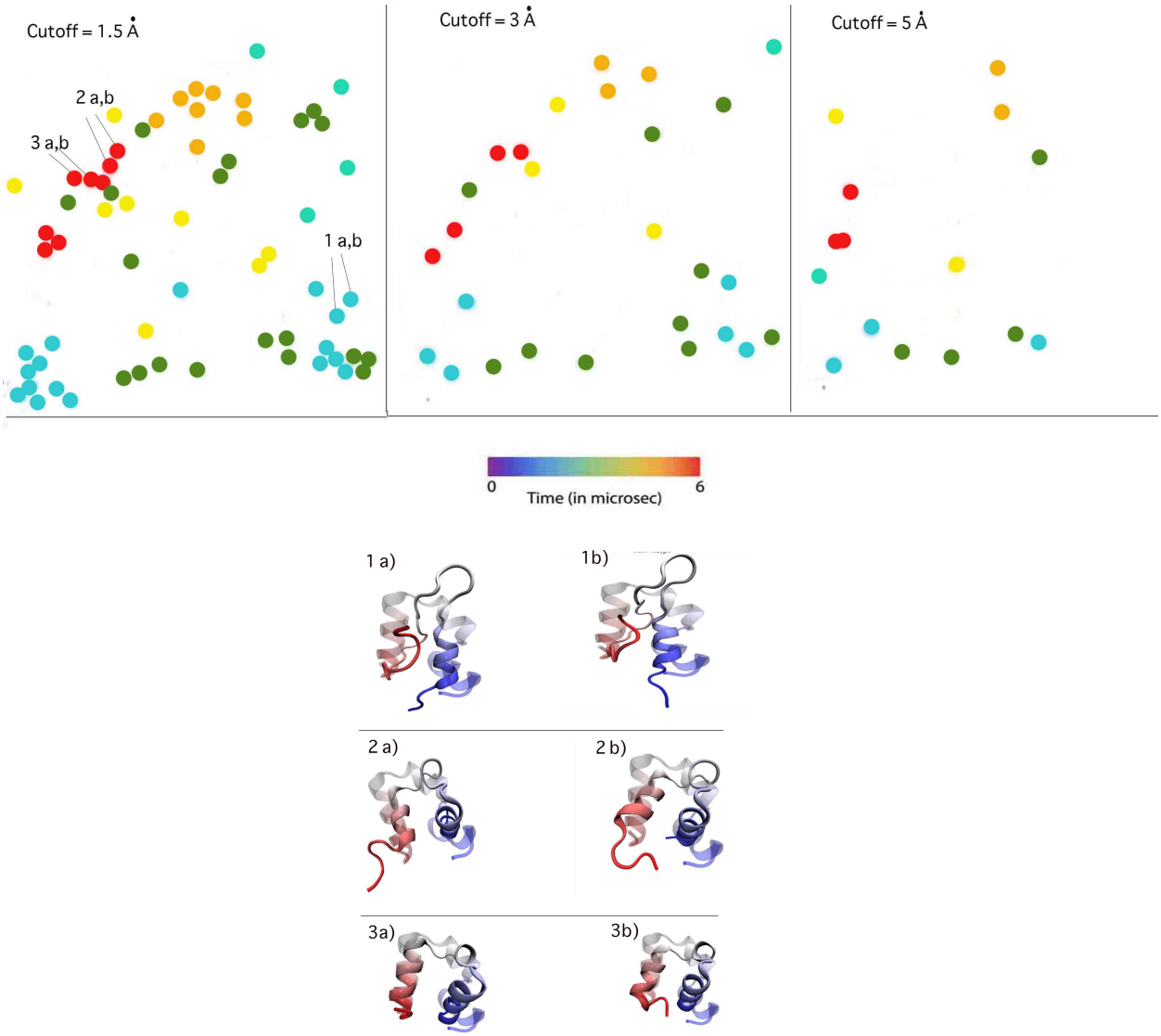

Figure 2.1: **Cluster centers with varying cutoff parameters**: The figure shows how cluster centers (with at least 20 members in their cluster) of villin trajectory 2, binned every 3 ns shifted, when varying clustering cutoffs between 1.5 Å and 5 Å, by projecting them onto a two dimensional Euclidean space using nMDS. An algorithm similar to the GROMACS with the gromos method, was used for the clustering. The trajectory was binned at 3 ns and pairwise RMSDs between frames were calculated to apply clustering using a dRMSD metric. We see that the cluster centers shift around in the projected space. Some clusters were merged and some disappeared. The centers marked as 1 a,b; 2 a,b; and 3 a,b are visualized in the figure to give qualitative examples of similarities of cluster centers.

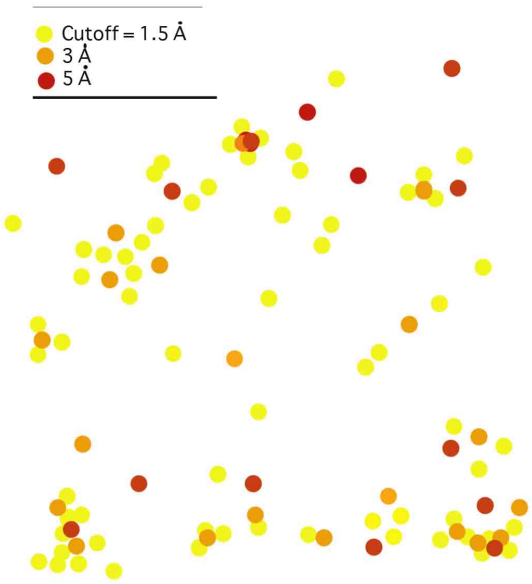

Figure 2.2: **Cluster centers with varying cutoff parameters superimposed**: The figure superimposes cluster centers obtained by varying cutoffs while analyzing villin trajectory 2, binned every 3 ns, using the gromos field in GROMACS.

2.3 Effect of noise

To illustrate the effect of noise on our clusters, we calculated RMSDs for all pairs of frames of a trajectory binned at 6 ps, taking into consideration the coordinates of all heavy atoms. We obtained about 200 clusters using the gromos method in the GROMACS program with a cutoff of 2 Å in this way. Fig. 2.4 shows how cluster centers (with at least 80 members) shifted in the projected 2-D space (obtained by nMDS) when contributions from end residues were removed. For visualization purposes, Figure 2.5 shows examples of some of the cluster centers that were mostly similar, but showed some variation in the coordinates of the end residues.

When we removed the contributions of the end residues, some cluster centers merged and reduced the total number of clusters (with more than 80 members) by more than 50 percent. In the case of villin, we know that residues 1-3 and 32-35 are not involved in secondary structure formation and contribute to noise in the data as their coordinates fluctuate a lot. However, in a system where such information is not available, it is not possible to use a

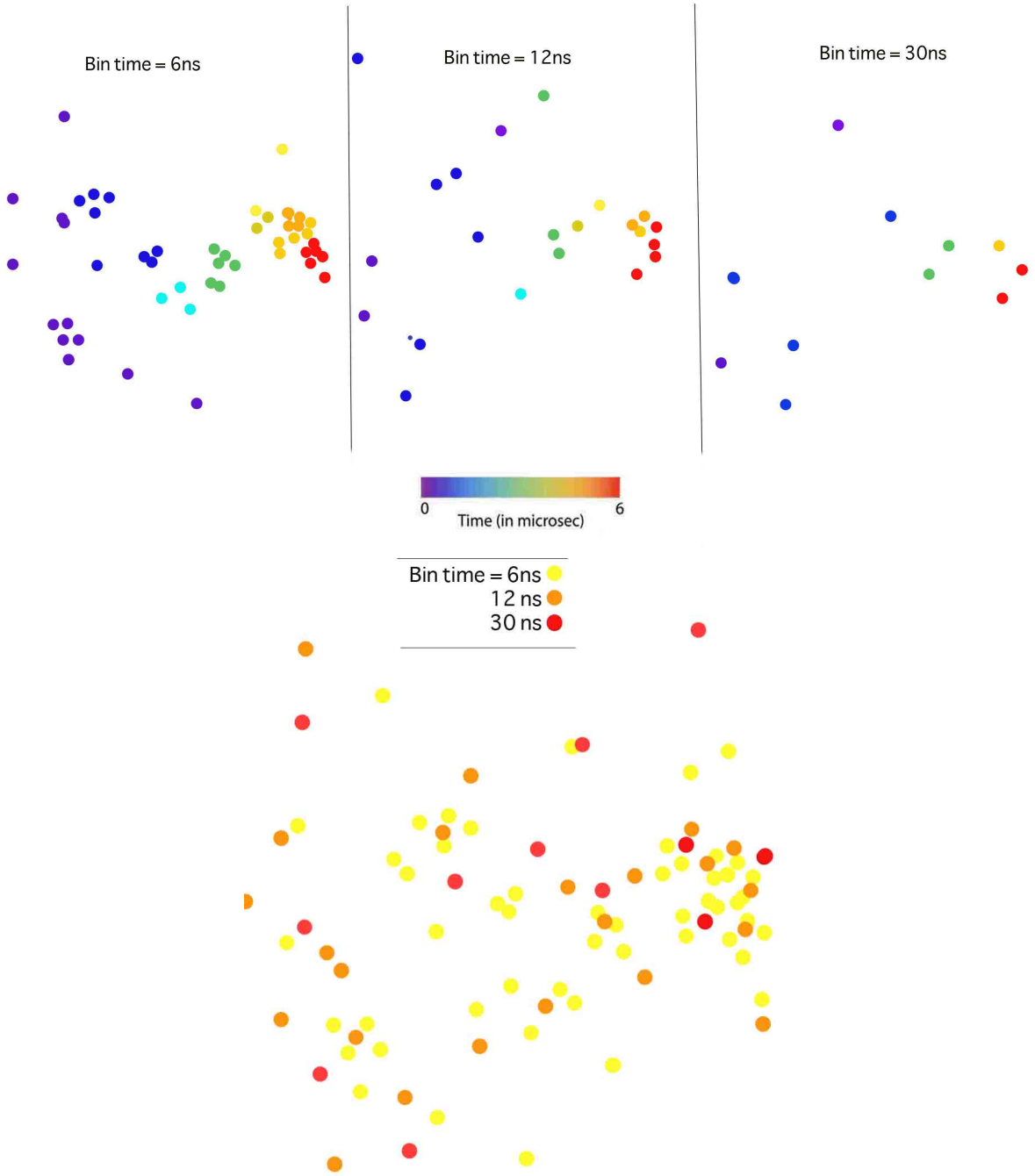

Figure 2.3: **Cluster centers by varying binning time**: The figure shows cluster centers (with at least 20 members in their cluster) of villin trajectory 1 in the dihedral angle space, by projecting them onto a two dimensional Euclidean space using nMDS. Centers were obtained by varying binning time between 1 ns and 30 ns. An algorithm similar to the GROMACS program was used for the clustering. We see that the cluster centers shift around in the projected space. Some clusters were merged and some disappeared. A superimposition of all three binning times is shown (inset) in the figure to help visualize the shifting of centers.

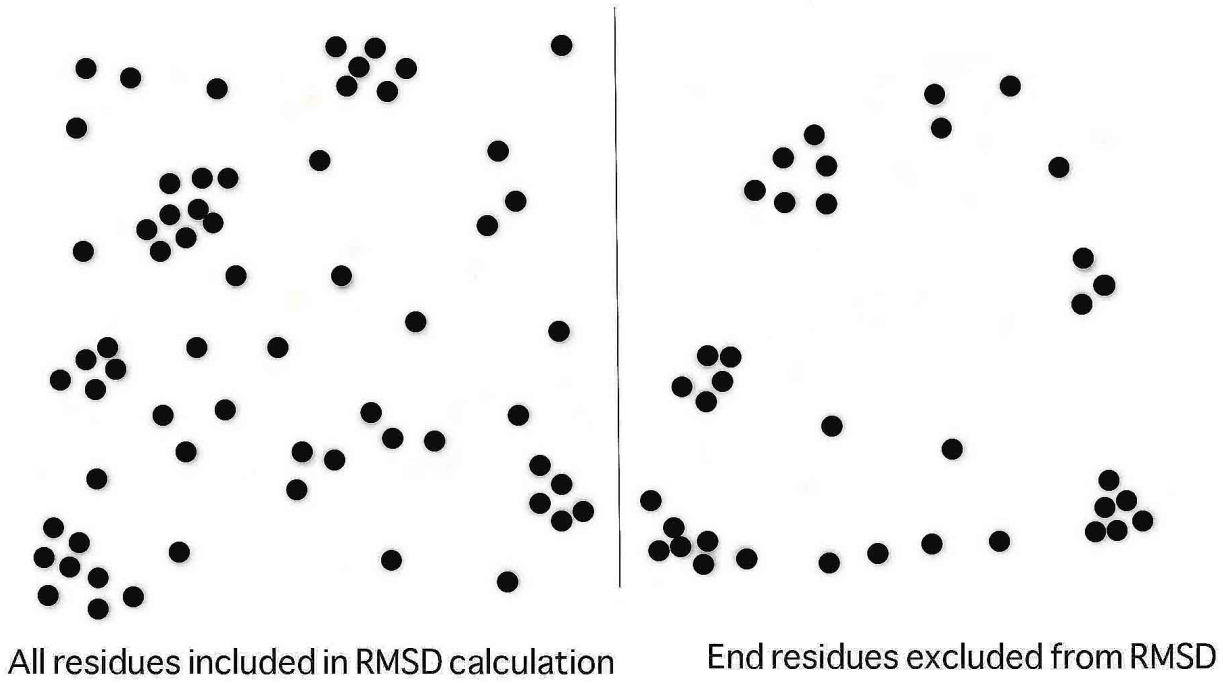

Figure 2.4: **Cluster centers with varying cutoff parameters**: The figure shows how cluster centers (with at least 80 members in their cluster) of villin trajectory 2 shifted when removing noisy coordinates, by projecting them onto a two dimensional Euclidean space using nMDS. An algorithm similar to the GROMACS program with the gromos method was used for the clustering, with trajectory binned at 6 ps and cutoff of 2 Å. The left panel shows the cluster centers obtained when coordinates of all the residues were used and the right panel shows the cluster centers obtained after removing contributions of end residues. We see that the number of clusters with more than 30 members go down by more than 50 percent when noise was removed. This shows that simple clustering is unstable to noise.

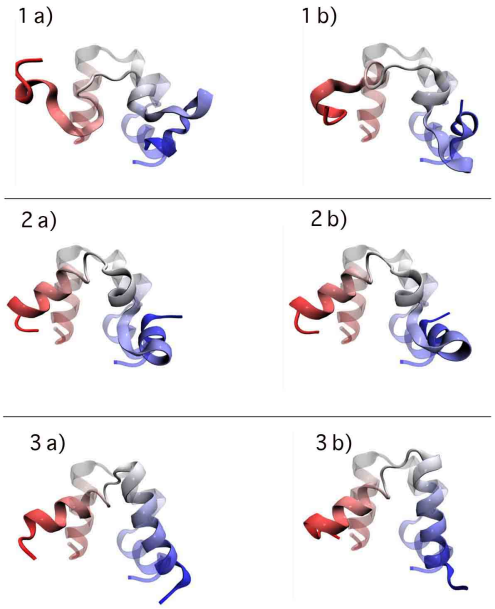

Figure 2.5: **Cluster instability to noise**: The left column and right columns show two different cluster centers obtained by clustering villin headpiece trajectory 2, binned at 6 ps in the internal frame wise RMSD space using the coordinates of all heavy atoms and a cutoff of 2 Å. These are examples of cases were visually similar structures were separated into distinct clusters. The differences arose due to coordinates of the floppy end residues being taken into consideration. These cluster centers merged when coordinates of the end residues were left out.

simple clustering method to tell which coordinates to filter out as noise and which ones to retain for clustering.

2.4 Inter-relationships between cluster centers

Let us assume that an ideal cutoff parameter is known or is otherwise obtainable. Let us also assume that the data is noise free and all the relevant coordinates have been included before. Even then, one of the most important drawbacks is that we do not have a way to tell how dissimilar cluster centers are after performing clustering. If a simple technique like GROMACS is applied to uniformly distributed data in any space, cluster centers are still returned as the clustering method does not consider the inter-relationships of data points. In the case of protein folding, if clusters are used as distinct states in a Markov chain or a transition matrix is built between the clusters [36], then misleading information about the underlying dynamics can be obtained. More specifically, we do not know if the cluster centers are distinct regions of the phase space that the protein hops between.

We can see from Fig. 2.4 that inter relationships between cluster centers are not taken into account by a simple program like GROMACS. Various rigorous cluster validation methods, which take into account inter-cluster relationships have been developed in the field of bioinformatics [39]. It can nevertheless be quite difficult to choose the necessary and sufficient set of validation techniques for MD trajectories without prior knowledge of the structural processes underlying folding.

3 Collective coordinates of folding

Cluster analysis may be acceptable for qualitatively visualizing MD trajectories, but their use to study the number of structural transitions present in the trajectories and perform free energy calculations such as in [38], may lead to serious artifacts. If we use clustering to analyze our trajectories, and then construct free energy plots based on the clusters obtained,

it is necessary to impose further assumptions, such as, considering Q values or radius of gyration to be important quantities. However, these quantifiers need not be meaningful for folding. Clustering does not give us any information about which coordinates to choose to follow the folding process.

4 Correct Usage of Cluster Analysis

Although I have pointed out various problems faced when using cluster analysis, these are largely due to the use of cluster analysis to perform tasks it was not intended to, rather than due to inherent problems with cluster analysis itself. The goal of cluster analysis is to assign patterns to group. It is therefore implicitly assumed that these groups are meaningful, and sufficiently different from each other. Cluster analysis was not intended to provide a compact depiction of the relations between the various patterns, and it is ill-suited for this task. However, for analysis of biological data, such pattern relations are precisely what we need. This makes cluster analysis a poor choice for such tasks.

In order to find patterns in MD trajectories, it is necessary to look beyond clustering to distill the underlying information. Our trajectories reside in a high dimensional space as every snapshot has information about all atomic coordinates. However, not all coordinates are important to folding; many coordinates are likely to be correlated and, thus, if viewed in the correct reduced coordinate space, the folding trajectories might lie in some lower dimensional space. The extraction of a correct reduced coordinate space has been the goal of a variety of dimensional reduction methods. We explain the use of dimension reduction methods such as PCA and nMDS for MD trajectory analysis in subsequent chapters. We also illiustrate in the next chapter, how dimension reduction methods overcome the drawbacks of clustering so far discussed.

Chapter 3

Principal Component Analysis

1 Introduction

As I have shown in the previous chapter, cluster analysis has several serious drawbacks and does not help find collective coordinates in MD trajectories. Dimension reduction methods can be very useful for reducing large trajectories as they are not cutoff dependent, and take inter-relationships between data points into consideration, making the analysis data driven. I will first introduce a simple dimension reduction method used very widely to analyze biological data, principal component analysis (PCA) [40]. PCA has been very popular to reduce MD data [41], but it suffers from certain limitations that become important in the context of protein folding. I present our PCA results below and show that while PCA achieves reasonable data compression in some configuration spaces, we still need better methods for trajectory analysis.

Principal component analysis (PCA) is used to retrieve dominant patterns and representative distributions from noisy data. The idea is to map the system in question (in our case, MD trajectories) from a multidimensional space to a reduced space spanned by a few principal components (PCs), thus elucidating the principal/dominant features underlying the observed data. Often, a small number of components (compared to the dimension of the input space) are sufficient to describe the structure in the data.

Given a set of N centered attributes $S = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}\}$, such that $\sum_{k=1}^{N} \boldsymbol{x_k} = 0$, PCA

diagonalises the covariance matrix:

$$C = \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{x_j} \boldsymbol{x_j}^T.$$

To do this, one has to solve the eigenvalue equation:

$$C\boldsymbol{v} = \lambda \boldsymbol{v},$$

where $\lambda \geq 0$ and $\boldsymbol{v}$ subsequently span the same space as $\boldsymbol{x_k}$'s.

The eigenvectors corresponding to the largest eigenvalues are, thus, the axes along which the data shows maximal variation. Given an eigenvalue λ_k, the percentage of the total amplitude variation captured by the corresponding eigenmode is given by $\lambda_k / \sum_j \lambda_j$. Usually, the first few (5 or 6) modes capture most of the variation in the data. To visualize the data, it is projected on to a lower dimensional space spanned by the first few principal axes. By doing this, we can also deduce coordinates along which the data shows interesting patterns in the projected space. Note that the algorithm described above captures only linear relationships between data points when finding eigenmodes and implicity uses the Euclidean metric to do so. There are versions of PCA which attempt to capture nonlinear relationships amongst data points, such as kernel PCA [42], but they are not used in our study. We have developed and used a robust nonlinear dimension reduction method (nMDS), which will be explained in the next chapter.

2 PCA applied to villin headpiece trajectories

We now examine the results obtained by applying linear PCA to the villin headpiece trajectories.

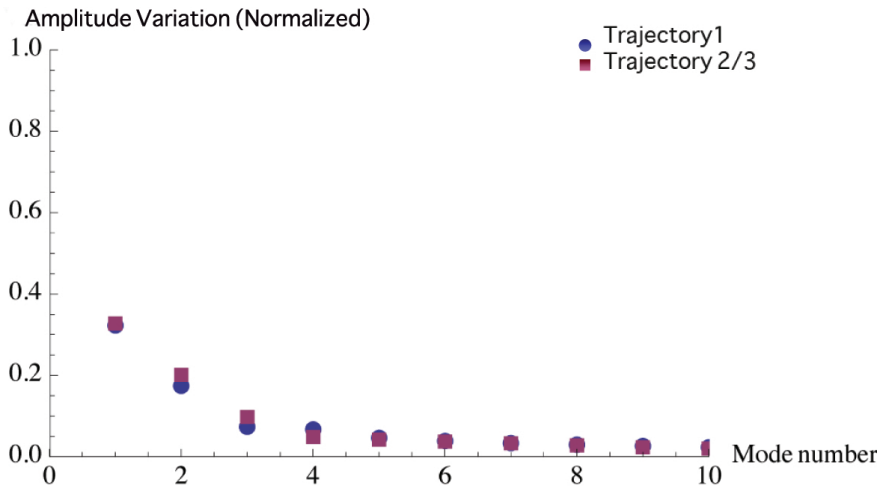

Figure 3.1: **PCA: variation vs principal modes in dihedral angle space**: The graph shows eigenvalues as a function of mode number on applying PCA to two of the villin headpiece trajectories binned at 6 ns, in the dihedral angle space. The third trajectory showed a very similar graph and overlapped with trajectory 2, it has been left out for easy reading of the graph. The first six modes capture about 90 % of the amplitude variation in the data.

2.1 Dihedral angle space

The trajectories were binned at every 6 ns and the resultant snapshots/frames were read in as 70-dimensional vectors (ϕ/ψ angles for the 35 residues) to obtain about 1000 vectors for each trajectory. On applying linear PCA, we found that in all three trajectories, 90 percent of the total amplitude of variations was captured by the 6 largest amplitude modes. Figure 3.1 shows this for the first two trajectories, trajectory 3 followed a trend identical to trajectory 2. Hence, we find that six dimensions should be enough to represent the data using PCA.

In Fig. 3.2, we show two of the villin headpiece trajectories embedded in the space spanned by the first two principal components. The first two axes capture only about 50 % of the variation in the data. The compression achieved seems very good and on visual inspection, it was found that structurally disparate frames were well separated in the projected space.

When the trajectories were binned more coarsely or finely by five times, the structure of the trajectories in the projected space remained unchanged, that is, the densely populated regions of the projected space shown in Fig. 3.2 remained so even on changing the binning

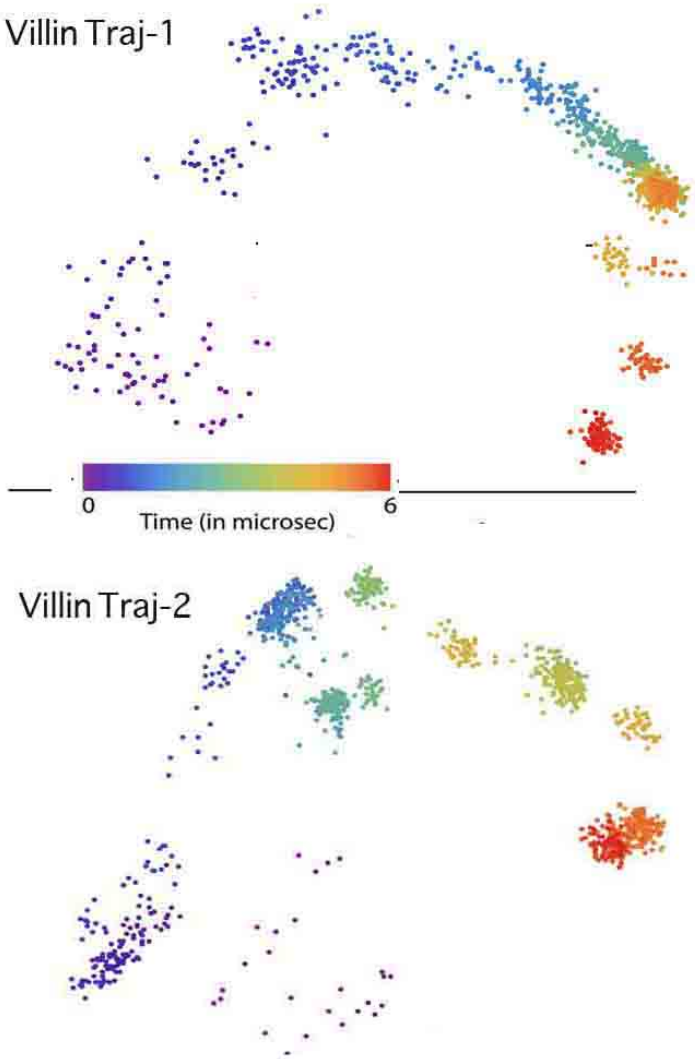

Figure 3.2: **PCA results in dihedral angle space**: The graph shows villin trajectories 1 and 2 (binned at 6 ns) projected along the first two principal components. These axes only capture 50 % of the variation in the data. The compression achieved seems very good and on visual inspection it was found that structurally disparate frames were well separated in the projected space. Both trajectories look qualitatively similar in the projected space, however when nMDS was used to find similarities, apart from helix 1 and 3 formation, no commonalities were found between the trajectories.

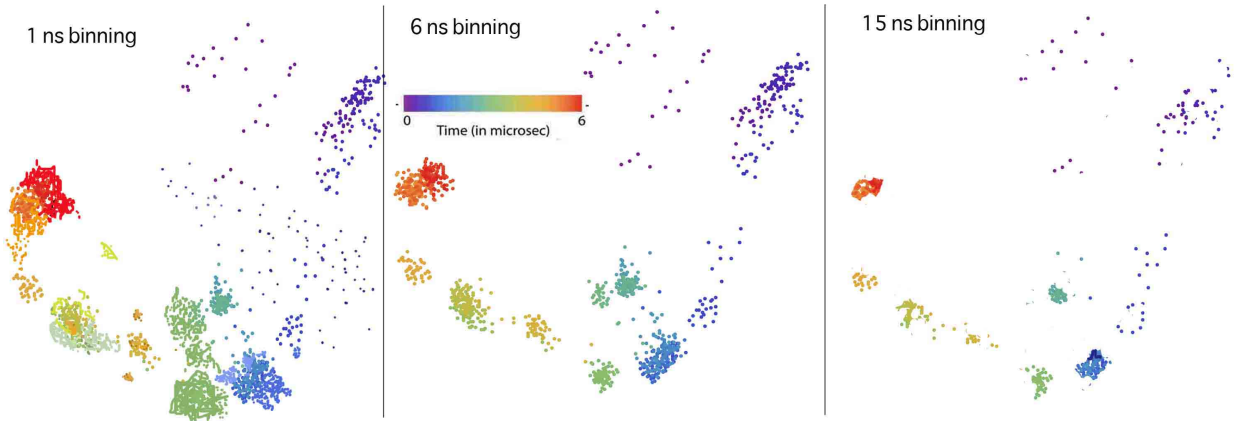

Figure 3.3: **Stability of PCA to change in binning time**: The graph shows villin trajectory 2 (input space: dihedral angles) projected along the first two principal components when the binning time was changed from 1 through 30 ns. The basic structure of the trajectory along the projected space remained stable.

time (Fig. 3.3). Also, when the when the dihedral angles of floppy residues were removed, the PCA projection did not change significantly.

PCA is linear compression method, as we discussed before. To see how good the compression achieved by PCA is, we must compare PCA results to a nonlinear dimension reduction method. Figure 3.4 shows the embedded space representation for trajectory 1 obtained by applying PCA and nMDS, respectively. Note that if some structures lie closer to each other than the other structures in the projected space (there is a peak in the local density), we call them a "cluster" for the purpose of qualitative analysis. nMDS proved to be robust in preserving inter-relationships between structures, although it is computationally expensive when the data size was increased. PCA still proves to be a computationally inexpensive first look at the trajectories and nMDS can be used to find further structure in the data set if needed. PCA results may be used to construct linear maps from PCA axes to nMDS axes. However, we do not have sufficient data to do this.

We now look at how PCA does is in the frame to frame RMSD space which has been briefly described below before illustrating PCA results.

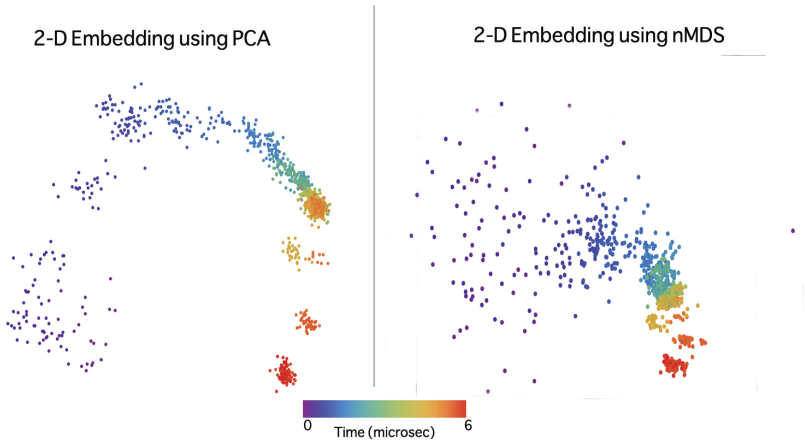

Figure 3.4: **PCA and nMDS embedded representation of trajectory 1 applied to dihedral angle space**: Two panels showing the embedding of trajectory 1 from the dihedral angle space, binned at 6 ns, to a 2-D space obtained by PCA (left) and nMDS (right). The first two axes suffice to embed all the data in nMDS, where as with PCA, they capture only 50 % of the total amplitude variation in the data. While PCA may show a clearer separation of data points along the two axes, we must remember the low percentage of fluctuations captured by the first two axes in PCA and not over interpret the results. PCA results may be used to construct linear maps from PCA axes to nMDS axes. However, we do not have sufficient data to do this.

2.2 Cartesian coordinate space

We chose an internal coordinate system (described below) for each trajectory (similar to that used in [1] with the gromos [30] method in GROMACS program [37] for clustering) to apply dimension reduction. Each trajectory (binned at 6 ns) hence consisted of about 1000 vectors of 1000 dimensions each. PCA applied to these vectors again showed that the largest 6 modes captured 90 percent of the total amplitude of variations and hence, dimension reduction could be applied in this coordinate space (Fig. 3.5).

In this coordinate space, PCA did not do well as well as nMDS. It is likely that the correlation between atomic coordinates is nonlinear, whereas the backbone dihedral angles are probably not correlated in any significantly non linear way for nMDS to have a clear advantage. PCA embedding in 2-D obtained for trajectory 1 and 2, when applied to the internalised coordinate system, is shown in Fig. 3.6. The first two axes only capture 40 % of the variation in the data. PCA was unable to separate the trajectories into clearly separated regions of the phase space. As explained before, the PCA we used captures only

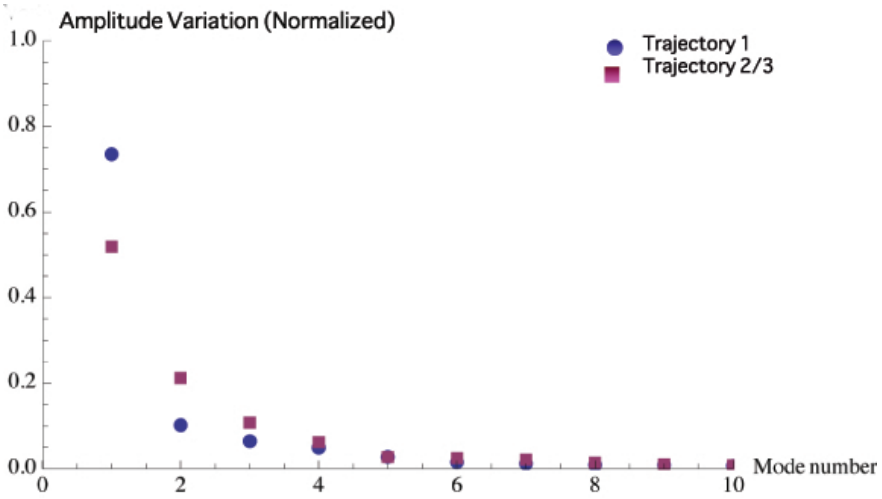

Figure 3.5: **PCA: variation vs mode number in internal coordinate space**: The graph shows eigenvalues as a function of mode number on apply PCA to two of the trajectories. The trajectories were binned at 6 ns and the input space was the internal coordinate space. The third trajectory showed a very similar graph and overlapped with trajectory 2 mostly, it has been left out for easy reading of the graph. We see that the first six modes capture about 90 % of the amplitude variation in the data.

linear relationships between data points. It is likely that many coordinates are nonlinearly correlated and PCA was unable to capture this when projecting onto a lower dimensional space. Unless we apply a nonlinear compression method, it is not possible to tell if the compression was poor due to PCA's failing or if the trajectories lack any structure when projected from the internal coordinate space. In Fig. 3.7, I compare PCA to nMDS (which we will be introduced in the next chapter) in the internal coordinate space, to illustrate the poor compression achieved by PCA.

We see that while in the dihedral angle space, nMDS and PCA results could be mapped to each other (by appropriate translations and rotations), this was not true in the cartesian coordinate space. This indicates that correlations in dihedral angles are not significantly nonlinear for the folding trajectories, but correlations in the cartesian space are likely nonlinear.

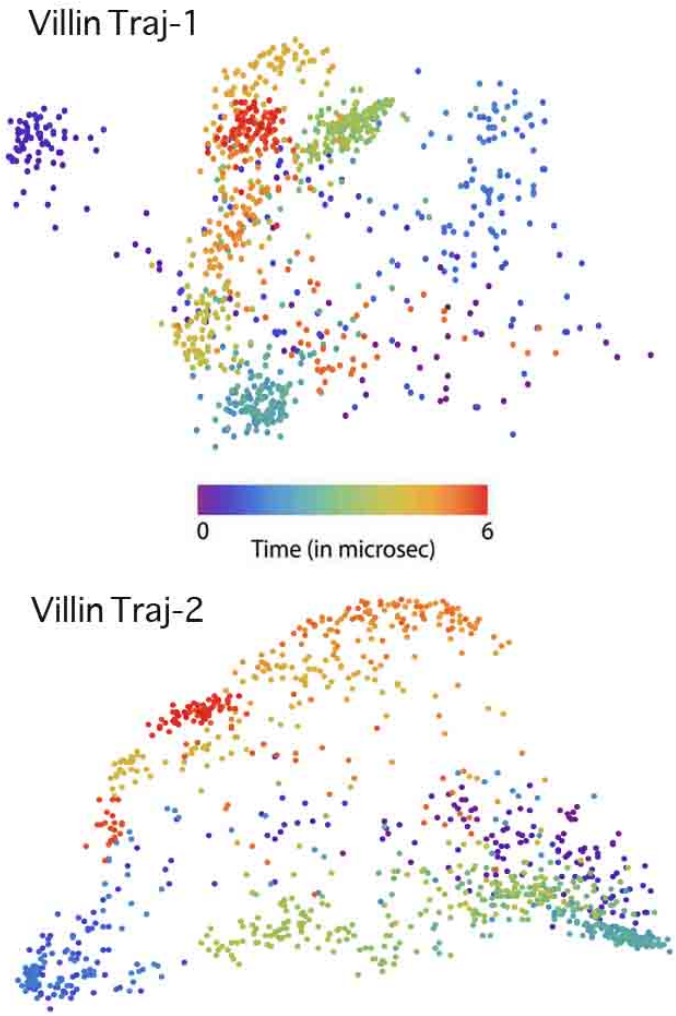

Figure 3.6: **PCA results in internal coordinate space**: The graph shows villin trajectories 1 and 2 projected (binned at 3 ns) along the first two principal components. These axes only capture 40 % of the variation in the data. PCA was unable to separate the trajectories into clearly separated regions of the phase space. Unless we apply a nonlinear compression method, it is not possible to tell if the compression was poor due to PCA's failing or if the trajectories lack any structure when projected from the internal coordinate space.

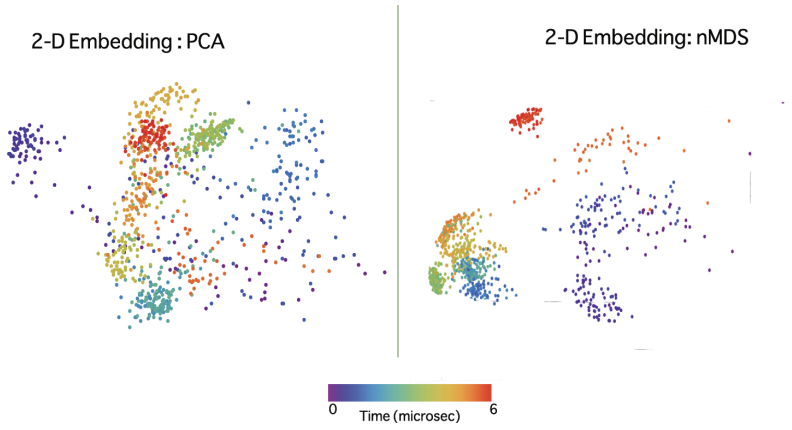

Figure 3.7: **PCA and nMDS embedded representation of trajectory 1 applied to internal coordinate space**: Two panels showing the embedding of trajectory 1, binned at 3 ns, from the internal coordinate space (described in the paper) to a 2-D space obtained by PCA (left) and nMDS (right). PCA does not do as well to separate out and correctly order the clusters present in the data, nMDS does a better job of preserving the interrelationships between data points while embedding them onto a lower dimensional space. Additionally, the first two PCA axes capture only 40 % of the total amplitude fluctuation in the data.

3 Advantages and drawbacks of PCA

To summarise, I enumerate the advantages and shortcomings of PCA:

1. PCA is a powerful dimension reduction method that is not cutoff dependent like clustering, and also is amenable to finding coordinate coordinates. Compared to clustering, it is more stable against noise and addition of coordinates.

2. PCA is computationally inexpensive and easy to implement, as it only involves diagonalization of the covariance matrix. It is a very useful first look at large trajectories to check for any structure that might be present.

3. PCA results can be affected by the sampling time chosen and introduce artifacts in the data as shown in [43].

4. PCA is however a linear analysis method and only captures linear relationships between data points, when projecting onto a lower dimensional space. Kernel PCA would be a natural next extension to PCA, but it is difficult to tune the kernel to achieve

sufficient compression. In the next chapter, we elaborate on nMDS, and show how it achieves superior data compression and provides key insights into villin headpiece and the mutant folding trajectories.

Chapter 4

Non-metric multidimensional scaling method: Application to villin trajectories

1 Introduction

To achieve maximal compression of folding/other MD trajectories while preserving salient features, it is important to design nonlinear dimension reduction schemes as we have shown in the previous chapters. We have adapted such a dimension reduction scheme, non-metric multidimensional scaling (nMDS) method to analyze villin headpiece and norleucine trajectories simulated by Freddolino and Schulten [1]. I will first introduce nMDS [44, 45, 46, 47] and then discuss its application to the analysis of villin trajectories.

2 The concept of nMDS

Given a set of points lying in some high-dimensional space, the basic goal of dimensional reduction is to find a lower dimensional representation capturing the relative relations. Based on the system in question, there are many nonlinear dimension reduction schemes (reviewed in [48]) available. nMDS is a completely data driven scheme, and in our experience its performance is superior to other methods of its class (except in terms of computational requirements) [49, 50].

Before laying out the algorithm for nMDS, I will illustrate the method with the following example. 1000 major cities were considered all around the earth, and the distance between them was calculated as

$$\delta_{ij} = -\cos\theta_{ij},$$

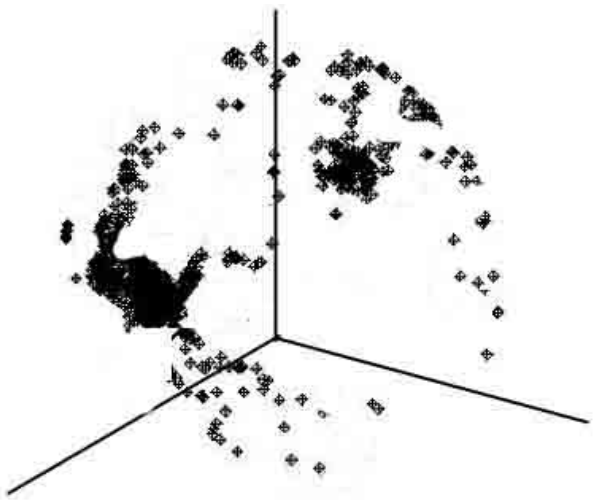

Figure 4.1: **nMDS embedding of cities on the surface of a globe** Fig. modeled after [51, 48]

where θ_{ij} is the angle between cities i and j measured with the origin as the center of the globe. Then the distances between the different cities were compared, and this information (i.e., the inequalities and not the actual distances) was passed to nMDS. nMDS constructed a representation of these 600 points by embedding them in a 3 dimensional Euclidean space, as in [48, 51]. The result is shown in Fig. 4.1.

Not only is the spherical structure of the earth automatically generated, without any metric or geometric information. Thus, nMDS generates a representation of data points in a low dimensional space based just on the relations between the relative pairwise distances. In practice, for computational reasons, rather than passing inequalities, the actual data from which the distances can be calculated are usually passed to nMDS. The distances may be computed using any reasonable metric (that does not change the global ordering of distances).

3 Impelementation of nMDS

In this section, I will explain the implementation of nMDS more rigorously. nMDS is an unsupervised data geometrization method [52] placing N points representing the objects

under study (in our case, the N frames of an MD trajectory), in a certain metric space E (to be discussed for our particular case later), such that the pairwise distances $d(i,j)$ of the points in E have consistency with the pairwise dissimilarities $\delta(i,j)$ of the corresponding objects in the input data. More precisely, nMDS tries to ensure that if $\delta(i,j) > \delta(k,l)$, then $d(i,j) > d(k,l)$ for all i, j, k and l denoting objects being analyzed. It is considered non-metric because, strictly speaking, the $\delta(i,j)$ values need not be known; only their order relationships need to be known, i.e., whether $\delta(i,j) > \delta(k,l)$ holds or not. If we have a reasonable number ($N > 30$) of points, this condition is typically strong enough to ensure a unique geometrical pattern for good data [49]. There are many possible implementations of an nMDS algorithm outlined above [44, 45, 46, 47]. We use an algorithm that was designed by Taguchi and Oono and successfully adapted to large sets of gene expression time series data to unravel relational patterns among genes [51, 48, 50, 53]. A flowchart explaining the application of nMDS to MD trajectory data is shown in Fig. 4.2.

If the pairwise dissimilarity $\delta(i,j)$ has ranking R_{ij} in the set of all the available dissimilarities, and $d(i,j)$ has ranking r_{ij} in the set of all the pairwise distances of the points in E, the points in E are positioned to minimize

$$\Delta \equiv \sum_{i \neq j} (R_{ij} - r_{ij})^2.$$

The minimum of this is clearly when $R_{ij} = r_{ij}$ for all i and j, *i.e.*, when the pairwise rankings in the original and embedded space exactly match. This is achieved through an over-damped dynamics driven by the ranking mismatch. The updating scheme used is:

$$\boldsymbol{x}_i \equiv \boldsymbol{x}_i + \alpha \sum_{j \neq i} [R_{ij} - r_{ij}] \frac{\boldsymbol{x}_i - \boldsymbol{x}_j}{|\boldsymbol{x}_i - \boldsymbol{x}_j|},$$

where the positions of the points in E are given by $\boldsymbol{x}_i$ and α is an appropriately small number to make the relaxation dynamics stable. The $\boldsymbol{x}_i$ are initially chosen randomly, and

the positions are updated using the rule above until a fixed point is reached. It should be clear that the desired minimum of Δ does correspond to fixed point of the dynamics. Like all nonlinear-optimization methods, there is a risk of getting trapped in a fixed point that is a local minimum, although in practice the dependence on initial condition seems to be weak compared to the dependence found for other methods. Comparison of the embedding to that produced by PCA is done in our case to check if nMDS achieves a reasonable embedding. Figure 4.3 demonstrates the weak dependence on intial condition, as nMDS projections remain stable over different binning times for the MD trajectories analyzed in this thesis.

4 Application to villin headpiece trajectories

4.1 Villin headpiece folding - Dihedral angle space

The trajectories were binned at every 6 ns and resultant snapshots/frames were read in as a 70-dimensional vectors (ϕ/ψ angles for the 35 residues) to obtain about 1000 vectors for each trajectory. Since PCA showed that 6 modes were enough to capture all fluctuations assuming linear relationships, a nonlinear dimension reduction method is sure to yield good compression for up to 3 or 4 dimensions. We applied nMDS to all trajectory data using a Euclidean distance ($d\sin\theta$) metric in the 70-dimensional input space as a metric to assign dissimilarities. It was found that two dimensions were enough to capture the variations in the data (this was checked by applying PCA to the embedded results obtained from nMDS reduction to 2, 3 , 4 and 5 dimensions as described in [48]). The idea is as follows: Suppose we use nMDS to embed identical data into n-dimensional and $(n+1)$-dimensional spaces. Using the embedded results, we can construct principal axes with the aid of PCA. Then, we study the correlation coefficients of the coordinates of the points. Usually, the first n principal axes of the $(n+1)$ dimensional embedding result have high correlation coefficients with the n principal axes of the n-dimensional embedding result. If the correlation between the $(n+1)$th axes of an $(n+1)$-dimensional embedding with that of the n axes of an n-

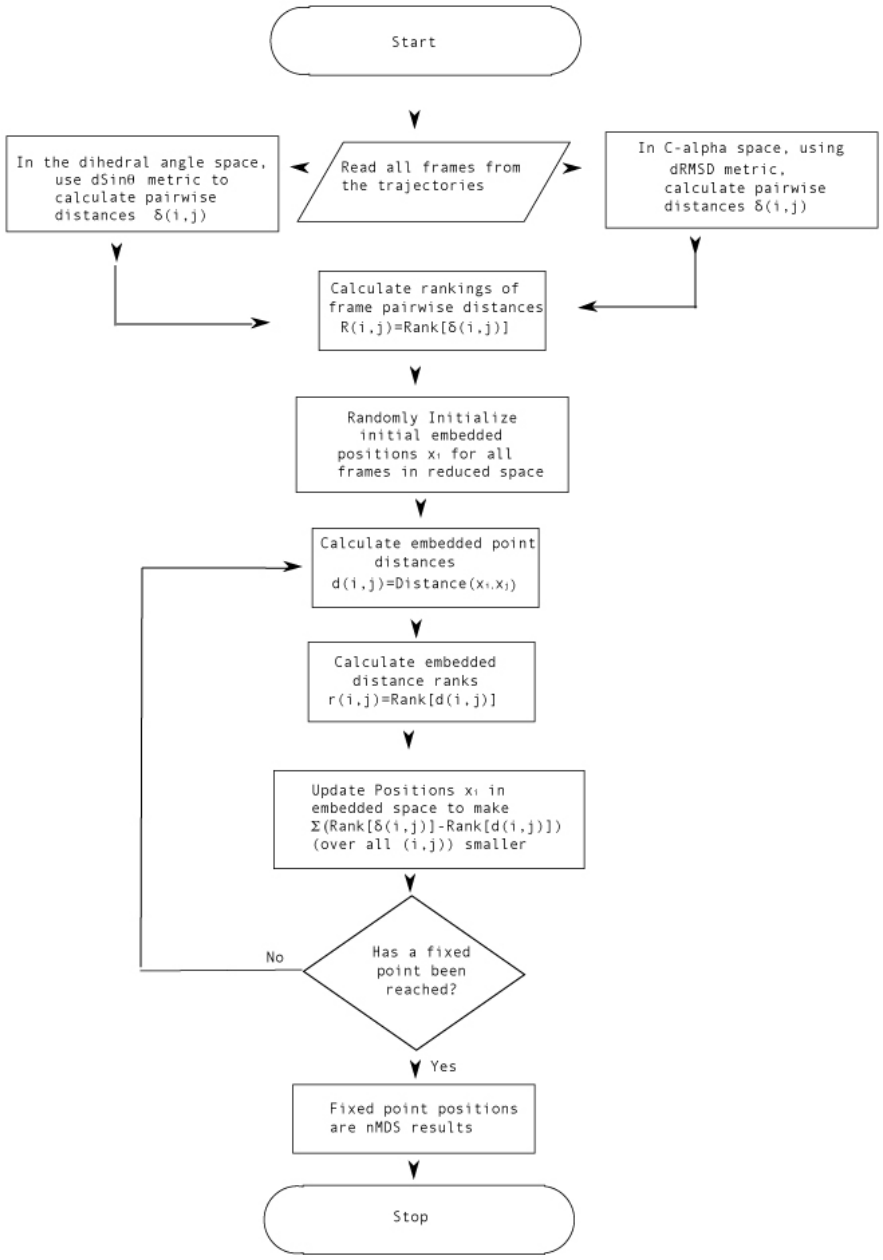

Figure 4.2: **nMDS flowchart describing data embedding.**

dimensional embedding, then we may say that the n-dimensional reduced space captures the main features in the data. In Table 4.1, we show the application of this method to $n = 2$ in the dihedral angle space for all trajectories. In Table 4.2, we show that 1D is not sufficient

for embedding the data. From these tables, we may conclude that a 2D Euclidean space is necessary and sufficient to capture the main features in the data. We could even devise a statistical test based on the correlation coefficients, but we do not dwell on this in our thesis, as our purposes are at present limited to distilling qualitative features of the underlying trajectories.

Table 4.1: Correlation coefficients between 2D and 3D axes obtained by applying PCA to nMDS results on all villin trajectories in the dihedral angle space.

Traj 1	3DI	3DII	3DIII
2DI	0.982	0.012	0.002
2DII	0.012	0.975	0.005
Traj 2	3DI	3DII	3DIII
2DI	0.992	0.009	0.001
2DII	0.011	0.989	0.001
Traj 3	3DI	3DII	3DIII
2DI	0.965	0.018	0.003
2DII	0.03	0.934	0.01

Table 4.2: Correlation coefficients between 2D and 1D axes obtained by applying PCA to nMDS results on all villin trajectories in the dihedral angle space.

Traj 1	2DI	2DII
1D	0.632	0.294
Traj 2	2DI	2DII
1D	0.812	0.178
Traj 3	2DI	2DII
1D	0.56	0.32

We also checked that the reduced representation in two dimensions remained unaffected by binning our trajectories by up to five times more coarsely or finely in time. Hence, nMDS proves to be stable when used to view trajectories at different time scales.

We picked representative structures from the densely occupied portions of the reduced conformational space (in 2-D) to obtain a reduced representation for the trajectories (see

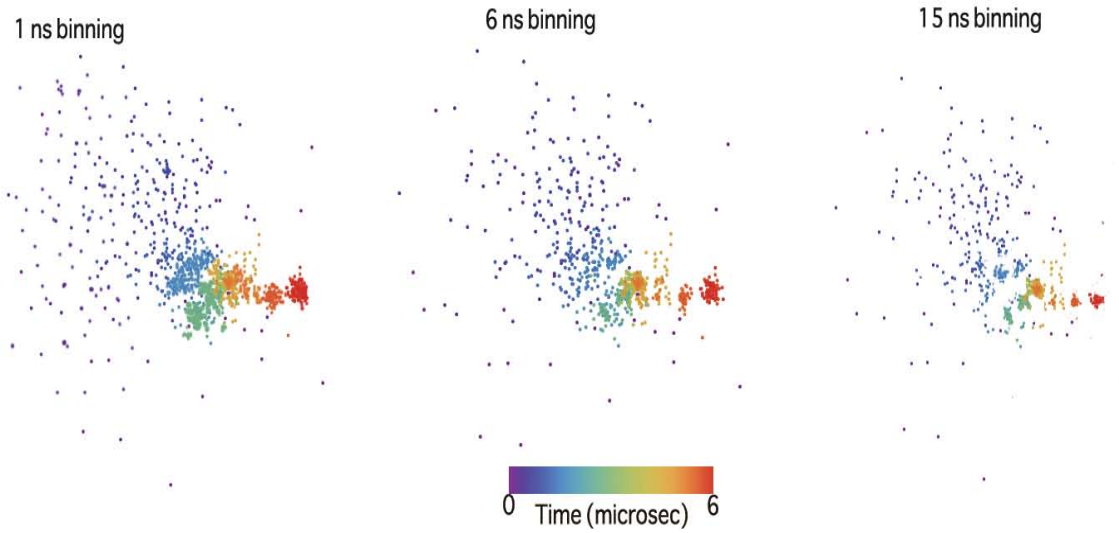

Figure 4.3: **Stability of nMDS to bin size**: nMDS emedding (in 2-D) of villin trajectory 1 data in the dihedral angle space is shown when the binning time was varied between 1 ns and 30 ns. The patterns in the projected space remained stable with change in binning time.

Figs. 4.4, 4.5 and 4.6). nMDS results show that the conformational space explored by the protein narrowed with time (in the projected 2-D space) as expected. In all three trajectories, the protein initially explores conformational space in what seems like random motion after which the secondary structure elements begin to form. In trajectory 2 and 3, all three helices form within the first 400 ns. In Trajectory 1, it takes up to 1 μs for helix 1 and helix 3 to form and helix 2 forms only in the last microsecond. We can see natural clustering in all cases, which implies that there are many fairly well defined metastable states. When simple clustering was used, a large number of clusters ($\sim$100, see Chapter 2) were found, but nMDS and PCA show clearly that there are not more than 5 or 6 distinctly densely populated regions in the explored phase space. This shows again that simple clustering can create more clusters than actually exist.

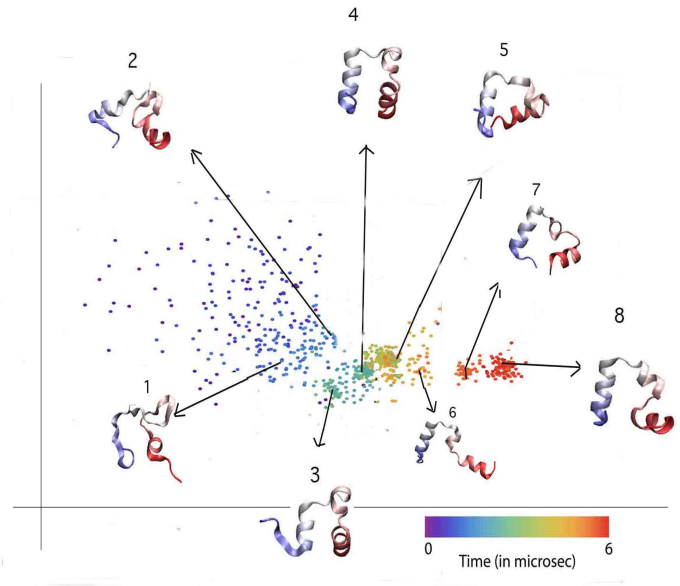

Figure 4.4: **Reduced representation of trajectory 1 binned at 6 ns in embedded 2-D space (structures numbered chronologically)**: Helix 1 and 3 (in red and blue resp.) form very quickly, but helix 2 (in white) forms only towards the end when helix 1 adopts the right orientation with respect to the rest of the structure. Each representative structure is superimposed over the native state to show folding.

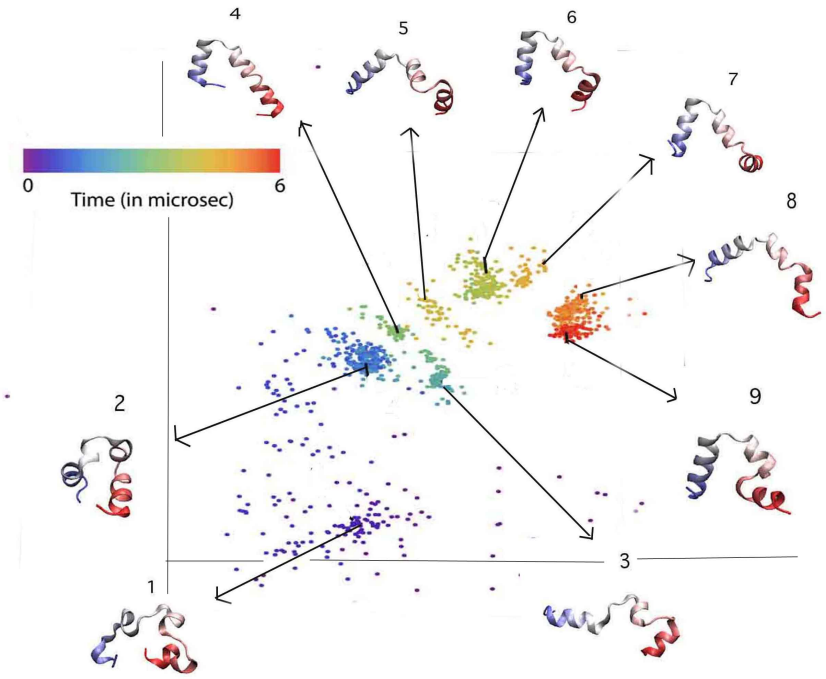

Figure 4.5: **Reduced representation of trajectory 2 in embedded 2-D space (structures numbered chronologically)**: All three helices form very quickly but their relative orientations are incorrect. Parts of these helices then dissociate, form non native contacts and finally rearrange to reach the correct structure.

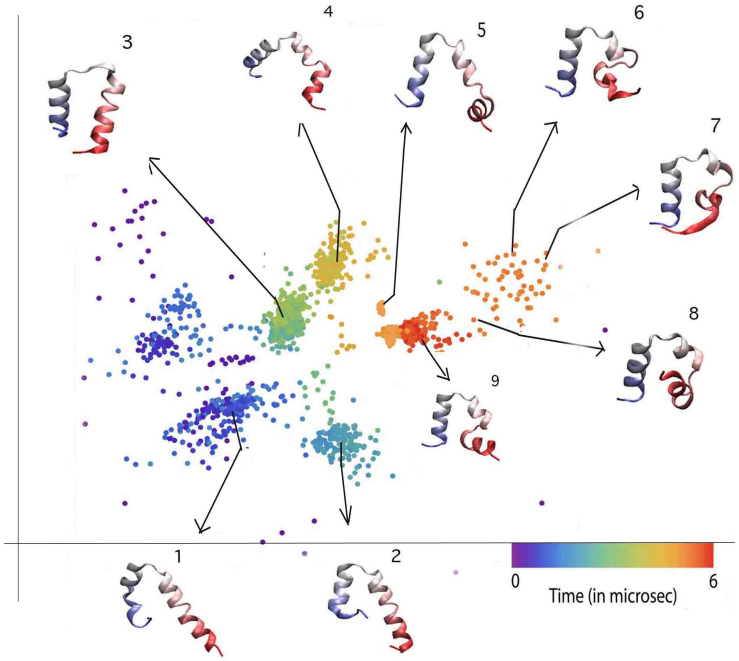

Figure 4.6: **Reduced representation of trajectory 3 in embedded 2-D space (structures numbered chronologically)**: A two-helix conformation with helix 2 and helix 1 joined is very stable for the first 3 μs; the protein then dissociates these helices and adopts the correct tertiary structure.

4.2 Using nMDS to filter noise in the trajectories

We have already shown that nMDS has been able to distill qualitative features of the folding trajectories, while reducing their dimensions and proved to be stable when the binning time was changed. nMDS is also stable to noise (addition of extra coodinates) and furthermore, the method itself can be used to extract coordinates of interest and throw out noisy coordinates. To find the dihedral angle coordinates that show a definite correlation and filter out noisy dihedral angle coordinates, we apply nMDS to the coordinate vectors across all trajectories. All the values of a dihedral angle coordinate across the simulation time make up one coordinate vector. Hence, for each trajectory sampled at 6 ns, we obtained 70 coordinate vectors of about 1000 dimensions each. When we applied nMDS to these vectors, we found that in each trajectory, both the backbone dihedral angles of the residues forming helices 1 and 3 fell into clusters and all the other dihedral angles were scattered around the projected space as shown in Fig. 4.7. This indicates that the non helix residues fluctuate apparently randomly throughout the trajectory and do not show any interesting pattern. In order to find collective coordinates, we must remove the noisy coordinates and then apply nMDS to our data. nMDS can thus automatically show us which coordinates might be of interest and which to discard, where as clustering or PCA will be unable to provide such information.

Another way to remove noisy coordinates is also outlined below.

In practice, it is found that certain data points (frames) are consistently embedded much more poorly than others. To identify such points, a local quantity called the (rank) mismatch is used. The mismatch $\Delta(i)$ for a point i is defined by

$$\Delta(i) = \sum_{j \neq i} [R_i(j) - r_i(j)]^2,$$

where $R_i(j)$, as defined in the nMDS algorithm, is the ranking of $\delta(i, j)$ (i.e., the dissimilarity between the i^{th} and j^{th} points) among the dissimilarities of all the points to the i^{th} point. $r_i(j)$ is the corresponding ranking in the embedded result. We can remove these less

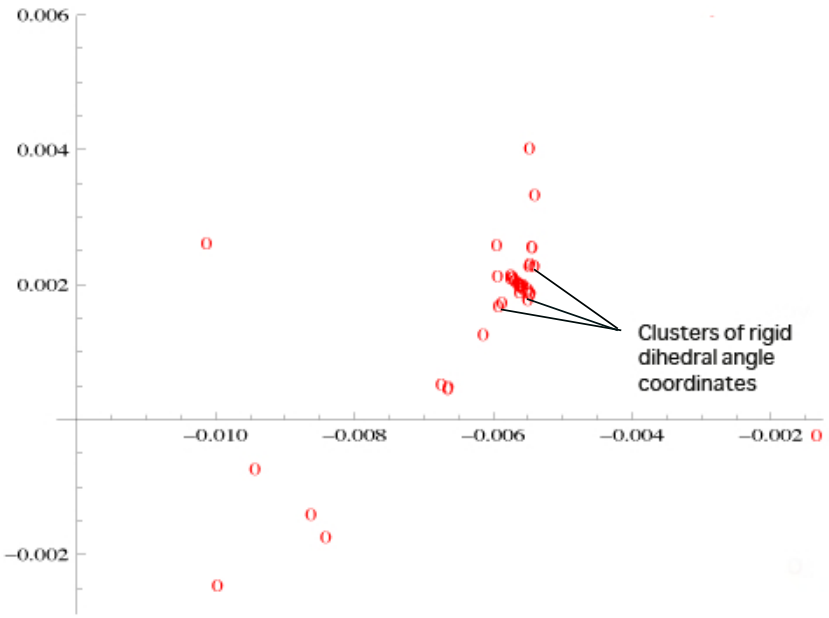

Figure 4.7: **nMDS applied on dihedral angle coordinate vectors of trajectory 1**: When nMDS was applied to the 70 dihedral angle vectors of trajectory 1 binned at 6 ns, we found that both dihedral angles of residues forming helices 1, 2 and 3 fell into clusters as indicated in the figure. The non helix residues formed scatter. This indicates that contributions from the non helix residues can be excluded from trajectory analysis. Trajectory 2 and Trajectory 3 dihedral angle coordinates showed a similar pattern when nMDS was applied.

consistent data points to hone the nMDS projection onto the reduced space. We again found that the non helical residues contributed to noise and if contributions from these residues was removed, the projection was cleaner. Such a honing scheme was implemented for all the trajectories before finding similarities.

4.3 nMDS to find similarities between trajectories

In order to find collective coordinates for villin folding, we must ask how similar the three trajectories are. Are there some (clusters of) structures that occur in all three trajectories? To answer this, we must study how close to each other the data points across three trajectories lie in the reduced (projected) space. We applied nMDS to data from all three trajectories together after removing noisy coordinates (dihedral angles of floppy residues: residue numbers 1-3, 11-12 and 32-35) and found that the structures from different trajec-

tories clustered very differently in the 2-D projected space, except for a few similarities. We found that along one of the axes in the 2-D projection, the trajectories met at a few points, which on visual examination showed that helix 1 and helix 3 were completely formed for those data points in all trajectories. However, along the second dimension, the trajectories were still slightly separated (the separation could be due to the difference in conformation of the residues forming helix 2) except for around the native state where they met again (Fig 4.8). Trajectories 2 and 3 had more similarities with each other than with trajectory 1 in the second projected axis. It was found that both trajectories 2 and 3 had two-helix structures similar to that shown in Fig. 4.8. Although in trajectory 2, these structures occurred transiently, in trajectory 3 they seemed to be very stable and lasted for up to 3 μs. Some of the qualitative features discussed above, like the early formation of helix 1 and 3, were obtained from careful visual inspection by Freddolino and Schulten [1]. However, nMDS is able to glean this information easily and is more reliable than visible inspection.

nMDS results thus show that the path to the native state along the dihedral angle coordinates differs qualitatively for all trajectories, although a few trivial similarities like the rapid formation of helix 1 and 3 exist. One axis may be interpreted as pertaining to formation of helices 1 and 3 and the other pertaining to local structure of the residues forming helix 2 and the coil regions. Since PCA results and nMDS results bore a qualitative resemblance in the dihedral angle space, if enough trajectories become available in the future, it may be possible to construct a map between the PCA axes and the nMDS axes. The axes obtained by PCA are known linear combinations of the input dimensions. By constructing a map (for e.g., quadratic or some power series) from the coordinates of the projected data in the PCA-reduced space to the corresponding coordinates in nMDS-reduced space, we can attempt to reconstruct the nMDS coordinates. However, when we attempted to map PCA axes to the corresponding nMDS axes using three trajectories, the data was insufficient for a clear interpretation to emerge. This is perhaps because we have only three trajectories showing large heterogeneity. Empirical evidence from the use of nMDS in bioinformatics suggests

that If we had about 30 trajectories, it is likely that such mappings may become statistically possible [50].

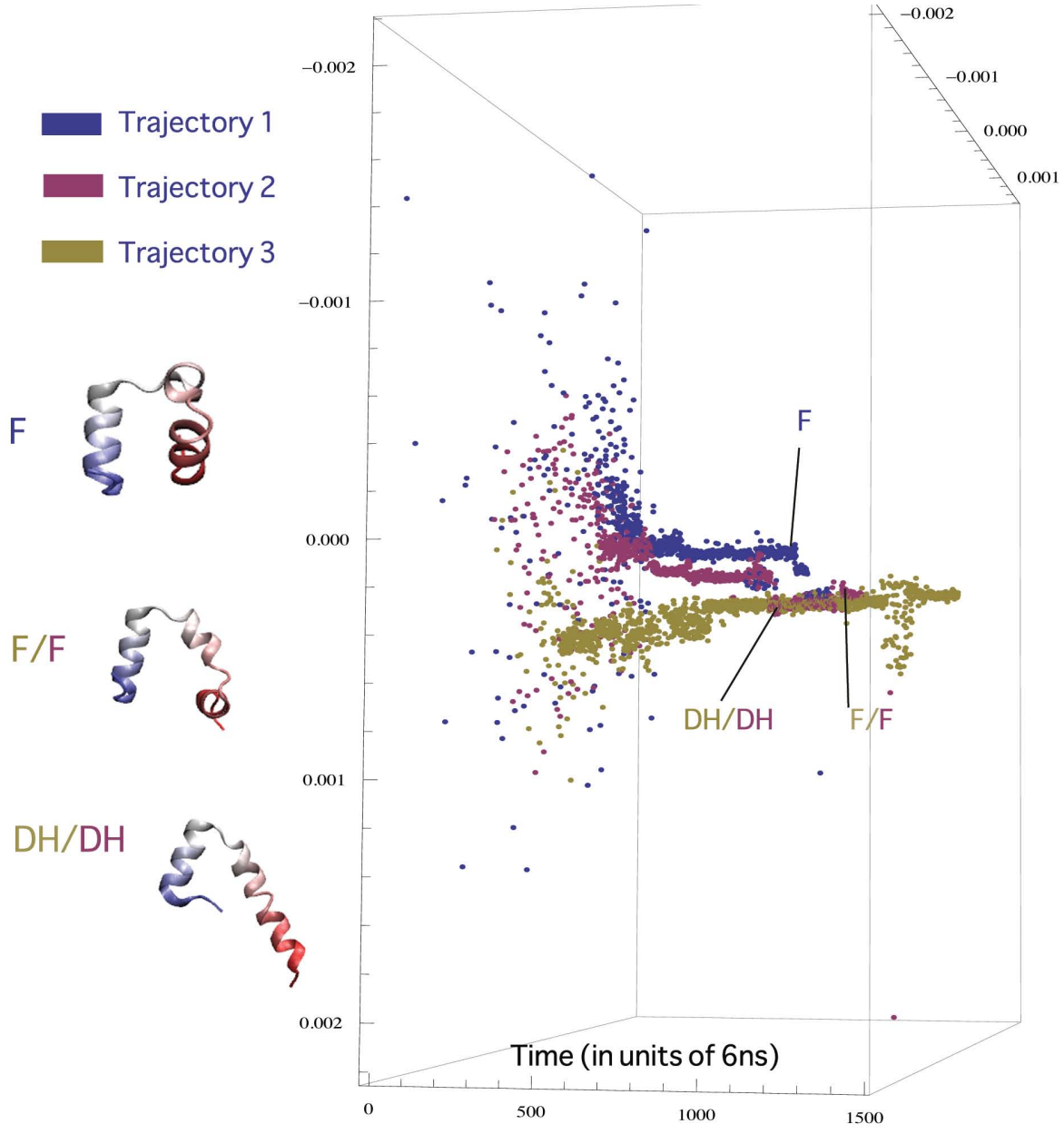

Figure 4.8: **Separation of the trajectories in reduced dihedral angle space**: nMDS was applied to dihedral angle data, binned at 6 ns, from all three trajectories. Along one of the axes, there are many crossing points beteween the trajectories. The crossing points were found to correspond to similar secondary structure elements forming, i.e. formation of helix 1 and 3. Along the other axis however, trajectory 1 is separated until it reaches the cluster containing the native state. Trajectories 2 and 3 meet at the two-helix states. The double helix (DH) and flipped (F) states are marked for each trajectory in the figure. Note that the flipped state of trajectory-1 is different from that of trajectories 2 and 3.

In order to understand the tertiary rearrangement in folding, we need to look at the trajectories in cartesian coordinate space. After that, we attempt to combine both dihedral angle and cartesian coordinates before applying PCA/nMDS to obtain a better picture of the folding process.

4.4 nMDS in Cartesian coordinate space

We chose an internal coordinate system (described in Chapter 3) for each trajectory (similar to that used in [1] with the gromos [30] method in GROMACS program for clustering) to apply dimension reduction. As PCA applied to these vectors had showed that the largest 6 modes captured 90 percent of the total amplitude of all the modes, dimension reduction could be applied in this coordinate space. A comparison of PCA and nMDS 2-D embedding obtained for trajectory 1, when applied to the internalised coordinate system was shown in Figure 3.7 and we clearly saw that nMDS was better than PCA in compressing the information in this space. It is likely that the correlation between cartesian coordinates of heavy atoms across frames is nonlinear, whereas the backbone dihedral angles are probably not correlated in any significantly non linear way for nMDS to have a clear advantage.

We found that three dimensions were enough to represent the data after applying nMDS. All three trajectories again (after removing noisy coordinates using nMDS as described before) showed completely different structures in the reduced 3-D space. While in trajectory 2, rapid hydrophobic collapse led to structures similar to Structure 3 in Fig. 4.4 to be stable over 1-2 μs, in trajectory 3, a two-helix structure as shown in Fig. 4.8 was the most stable. In all three trajectories, a similar transition referred to as a "flipping transition" in [1] occurred towards the last 500 ns prior to folding. The flipping transition involved the reversal of helix 1 (flipping from pointing into the page to out of the page, with the page aligned along the plane formed by helix 2 and helix 3). The structure before helix 1 flipped into the correct native conformation will be called the flipped state in our discussion.

Does the flipping transition occur similarly in all three trajectories? In order to answer

this question, we chose C_α coordinates of only five of the residues (residues: 5, 8, 15, 23 and 27) representing the locations and orientations of the three helices and used the contact distances between them in all three trajectories to apply nMDS. Two axes were found to be sufficient to represent the data and we found that the trajectories explored different portions of the projected space and met only close to the native state (Fig. 4.9). On visual inspection, no clear interpretation of the projected axes emerged but the points of meeting for trajectory 2 and 3 showed a series of two-helix conformations and a common flipped conformation (marked on Fig. 4.9). Trajectory 1 only explored transiently some of the structures that were common to Trajectory 2, and a two-helix state never occured. The flipped state was found to be different in Trajectory 1 as compared to that of Trajectories 2 and 3 (Fig. 4.9) in that the second helix was formed after the flipping happened in Trajectory 1. In Trajectories 2 and 3, helix 2 and 3 dissociated from the two-helix state described before and the protein quickly locked itself in the correct tertiary state after helix 1 flopped around exploring various non-native conformations. Although the flipped structure occured in all three trajectories (and was slightly different in trajectory 1 structurally compared to trajectories 2 and 3 as explained above), the flipping transition was observed to occur through a different series of steps in all three trajectories. The flipping transition in the three villin headpiece trajectories is shown in Fig. 4.10. The flipped state in trajectory 1 lacked a well formed helix 2, while the flipped states of trajectories 2 and 3 had all three helices and closely resembled each other. Note that despite starting out at the same flipped conformation, trajectories 2 and 3 flip into the native conformation in a different series of steps.

When both C_α coordinates and dihedral angle coordinates were used to apply nMDS reduction, the resultant representation was dominated by the C_α coordinate values and no new similarities between the trajectories emerged. This shows that the similarities in local structure formation are trivial and the global folding path is very different for all trajectories. Dihedral angles may hence not be good candidates for collective coordinates, at least for small proteins such as villin headpiece.

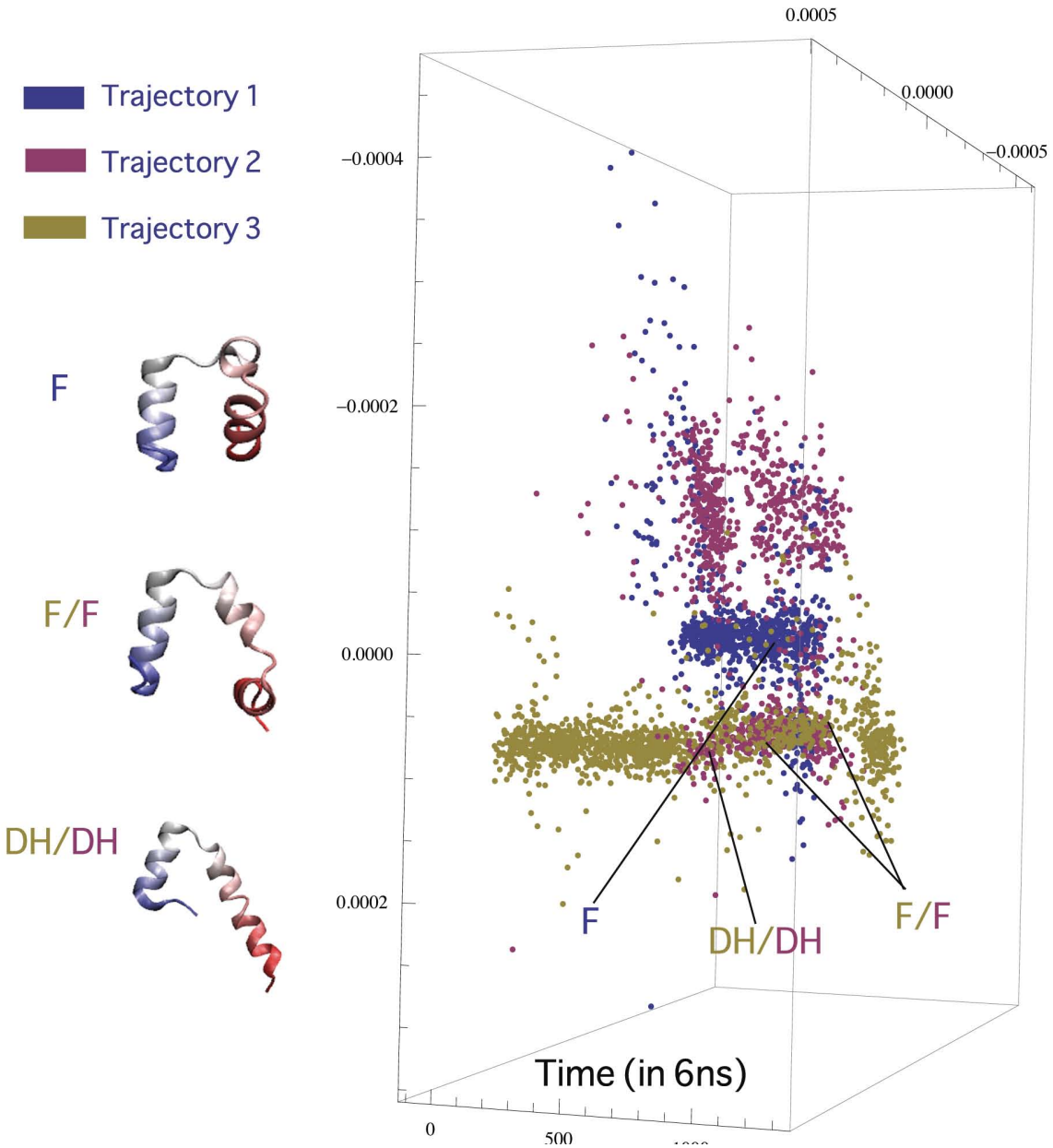

Figure 4.9: **Separation of the trajectories in reduced C_α contact distances space**: nMDS was applied to data from all the trajectories, binned at 6 ns, using only the C_α coordinates to calculate distances between residues 5, 8, 15, 23 and 27 while calculating RMSDs used to assign dissimilarity ranks for nMDS. The points of meeting for trajectory 2 and 3 are a series of two-helix conformations (labelled as DH) and the flipped state (F). Trajectory 1 does not meet the other two trajectories except towards the last 500 ns when the protein is nearly folded. No obvious interpretation of the axes emerged on visual inspection, the trajectories showed more marked difference in the parts of projected space they explored.

We also used a more rigorous method to find commonalities amongst trajectories, ICS Survery (described in detail in Appendix A), which shows that the global folding process is

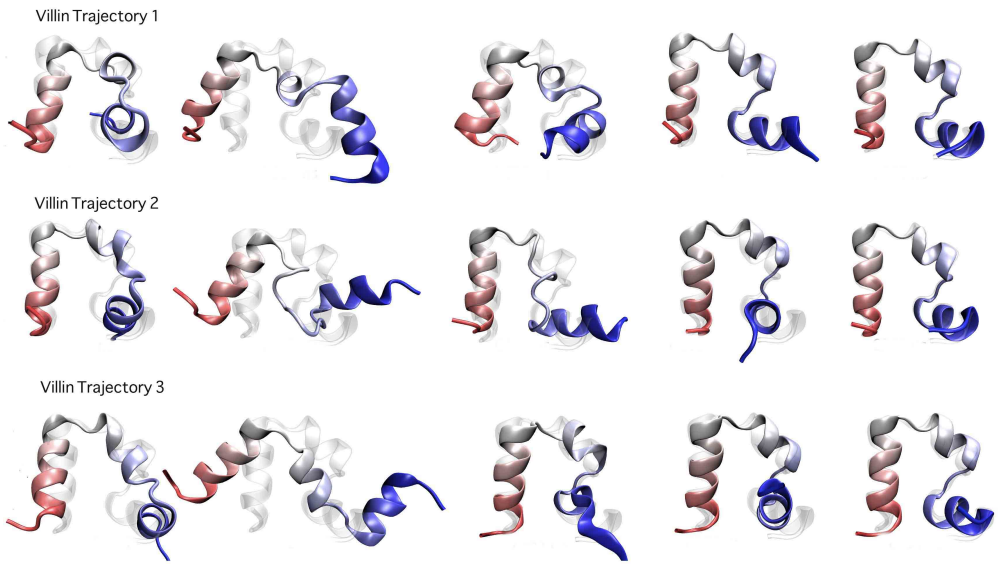

Figure 4.10: **Flipping transition in the three trajectories**: Representative structures from the transition between the flipped and native state conformations in all three villin headpiece trajectories. Protein coloring runs from blue to red from N terminus to C terminus. The crystal structure is superimposed in gray for comparison. The flipped state in trajectory 1 lacked a well formed helix 2, while the flipped states of trajectories 2 and 3 had all three helices and closely resembled each other. Note that despite starting out at the same flipped conformation, trajectories 2 and 3 flip into the native conformation in a different series of steps.

very different in all three trajectories and there do not seem to be special units of atoms that move similarly in all the trajectories. In each trajectory, a different set of coordinates seemed to move in tandem. We hence confirm the results obtained from nMDS analysis.

The most notable common feature found across trajectories using nMDS on all input spaces was the competition between local and global structure formation. If the protein formed all three helices very early like in trajectories 2 and 3, it spent a long time exploring non-native two-helix or collapsed conformations before dissociating and locking into the correct global structure. However, small changes in folding time such as this are not significant for small proteins. To understand the competition between global arrangement and local structure formation, it is important to study folding trajectories of larger proteins. To this end, we need at least 3-5 orders of magnitude faster computational speed.

4.5 What we can learn about villin headpiece folding

nMDS and ICS Survey confirm that all villin headpiece trajectories showed structural heterogeneity and the only trivial commonalities found were that some elements of secondary structure formed earlier on. This was to be expected as the protein shows secondary structure even in the denatured state in experiments [19]. The only other common feature found across trajectories using nMDS on all input spaces was the competition between local and global structure formation. If the protein formed all three helices very early like in Trajectories 2 and 3, it spent a long time exploring non-native two-helix or collapsed conformations before dissociating and locking into the correct global structure. However, small changes in folding time such as this are not significant for small proteins. To understand the competition between global arrangement and local structure formation, it is important to study folding trajectories of larger proteins. To this end, we need at least 3-5 orders of magnitude faster computational speed.

5 Norleucine trajectories

In the case of the NLE mutant, while one trajectory (NLE-FOLD1) folded to a native state in 2.5 μs, only one other trajectory, NLE-FOLD3 transiently explored a state with native-like interactions. On analyzing the NLE trajectories using nMDS, we did not find any similarities across the six trajectories whatsoever. This protein showed great structural heterogenity. The only striking features we found were that the double helix and various other hydrophobically collapsed conformations were very stable and lasted more than 2 to 3 μs.

We found that 3 dimensions (Tables **??**) were enough, but 2 dimensions were not enough to embed the trajectory data for all trajectories. Only in trajectory 1 (which folded), we found that 2-D was enough to embed the results, interestingly.

Table 4.3: Correlation coefficients between 3D and 4D axes or NLE trajectories, obtained by applying PCA to nMDS results

Traj 1	4DI	4DII	4DIII	4DIV
3DI	0.882	0.110	0.012	0.002
3DII	0.012	0.87	0.005	0.003
3DIII	0.02	0.021	0.925	0.001
Traj 2	4DI	4DII	4DIII	4DIV
3DI	0.92	0.10	0.002	0.002
3DII	0.01	0.92	0.11	0.001
3DIII	0.21	0.011	0.825	0.001
Traj 3	4DI	4DII	4DIII	4DIV
3DI	0.95	0.01	0.012	0.008
3DII	0.01	0.97	0.005	0.001
3DIII	0.19	0.001	0.825	0.001
Traj 4	4DI	4DII	4DIII	4DIV
3DI	0.682	0.110	0.22	0.012
3DII	0.11	0.77	0.13	0.002
3DIII	0.25	0.03	0.725	0.001
Traj 5	4DI	4DII	4DIII	4DIV
3DI	0.562	0.120	0.28	0.002
3DII	0.02	0.93	0.014	0.002
3DIII	0.15	0.026	0.842	0.001
Traj 6	4DI	4DII	4DIII	4DIV
3DI	0.909	0.092	0.012	0.007
3DII	0.012	0.79	0.15	0.003
3DIII	0.04	0.25	0.755	0.003

Table 4.4: Correlation coefficients between 2D and 3D axes or NLE trajectories, obtained by applying PCA to nMDS results

Traj 1	3DI	3DII	3DIII
2DI	0.89	0.10	0.001
2DII	0.09	0.85	0.007
Traj 2	3DI	3DII	3DIII
2DI	0.652	0.29	0.12
2DII	0.25	0.45	0.25
Traj 3	3DI	3DII	3DIII
2DI	0.78	0.13	0.005
2DII	0.28	0.47	0.30
Traj 4	3DI	3DII	3DIII
2DI	0.75	0.20	0.28
2DII	0.33	0.28	0.35
Traj 5	3DI	3DII	3DIII
2DI	0.58	0.22	0.28
2DII	0.45	0.42	0.12
Traj 6	3DI	3DII	3DIII
2DI	0.6	0.18	0.28
2DII	0.43	0.38	0.19

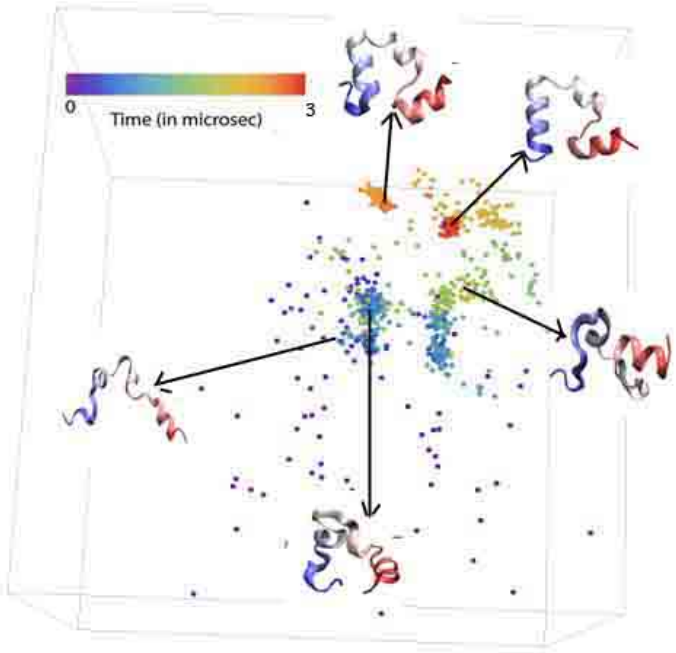

Figure 4.11: **Norleucine trajectory 1**: This trajectory reached the folded state in 2.5 μs.

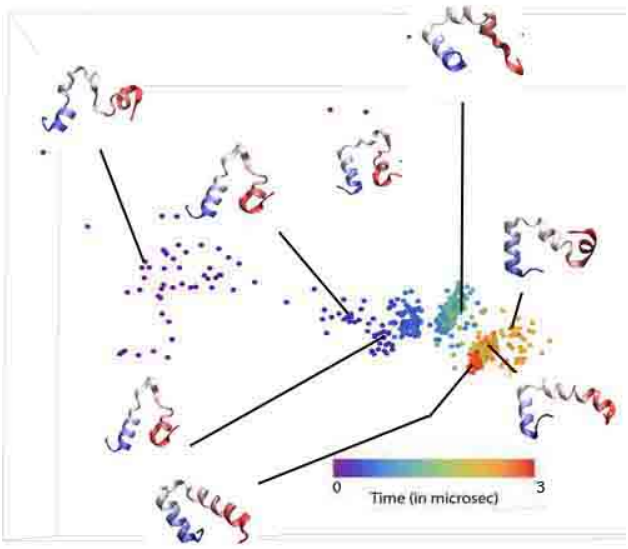

Figure 4.12: **Norleucine trajectory 2**: The trajectory gets stuck in a non native state and does not fold over the simulated timescale $\sim$8 μs.

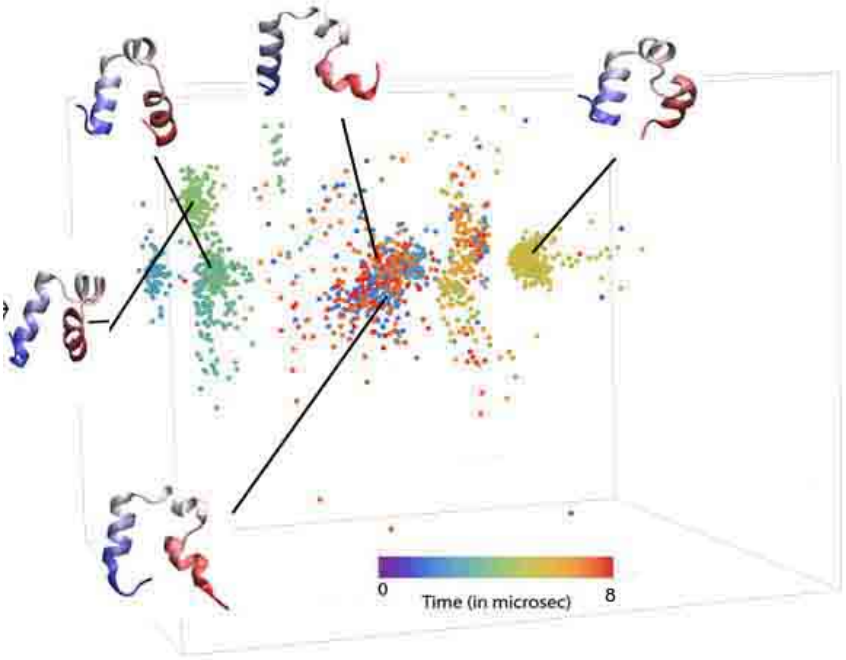

Figure 4.13: **Norleucine trajectory 3**: The trajectory transiently explores a near native state, but gets stuck in a non native state.

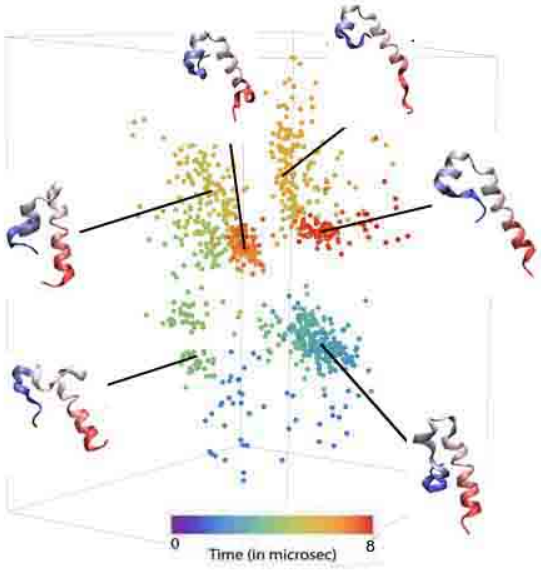

Figure 4.14: **Norleucine trajectory 4**: The protein does not fold over simulated time ∼8 μs.

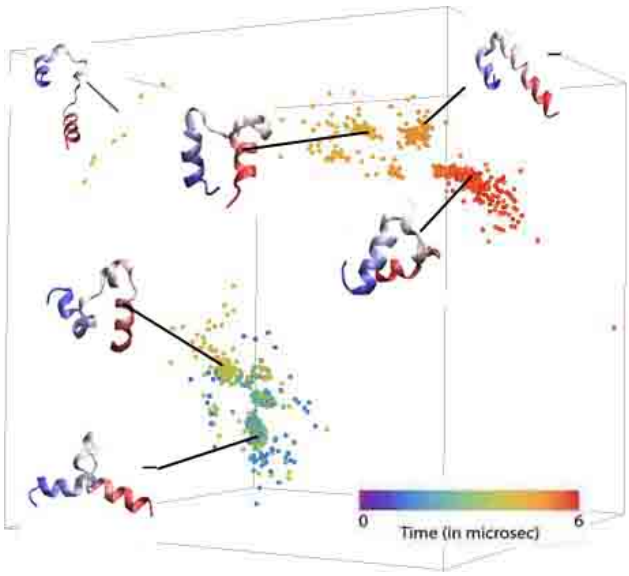

Figure 4.15: **Norleucine trajectory 5**: The protein does not fold over simulated time ∼8 μs.

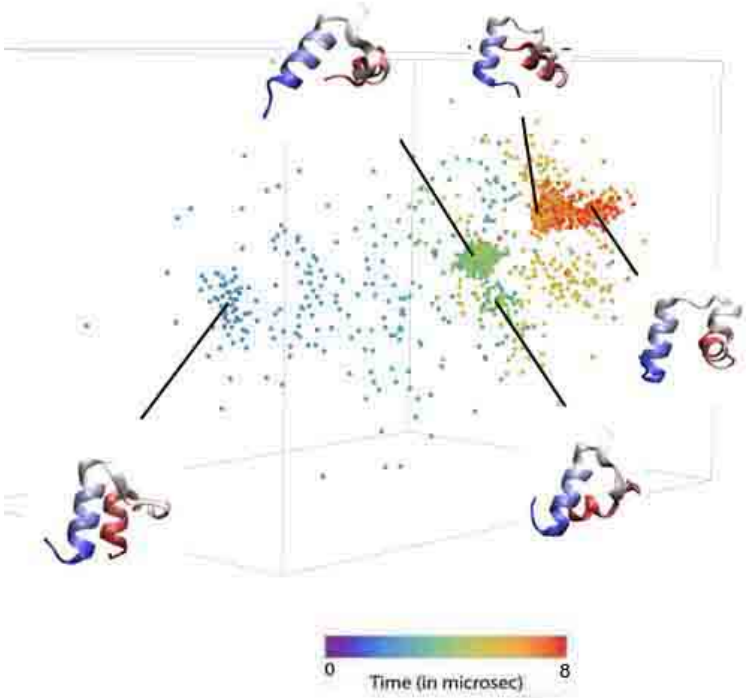

Figure 4.16: **Norleucine trajectory 6**: The protein does not fold over simulated time ∼8 μs.

In five of the six NLE trajectories, the protein was stuck in hydrophobically collapsed states with a tighter hydrophobic core than the native state. The norleucine residues are more buried in our trajectories compared to the crystal structure of the norleucine mutant [18]. It is possible that the CHARMM 22 force field does not fold the norleucine mutant correctly. However, experiments have only probed the formation of helix 3 in the NLE mutant and have not established the reported folding timescale of 1 μs unambiguously. A detailed discussion can be found in [1].

To find any commonalities between trajectories, we input all C_α coordinates and backbone dihedral angles of the protein across the six trajectories to nMDS and found that apart from a few common hydrophobically collapsed states (shown in Fig. 4.17), the trajectories showed large heterogeneity.

Why were most of the NLE trajectories unable to fold? We observe from the NLE trajectories that the formation of a very tight hydrophobic core at least delays (if not precludes) folding. The only trajectory which folds does not contain a flipped structure similar to the villin headpiece, and shows no competition between global and local structure formation. The other five trajectories were stuck in collapsed states as indicated in Fig. 4.17. We performed nMDS by considering the C_α coordinates of the core forming residues : 6, 10, 17 (phenylalanine) and 20 (leucine) and both the NLE residues. We found from nMDS analysis that the core was very tight in all trajectories, although the core was differently organized in each trajectory.

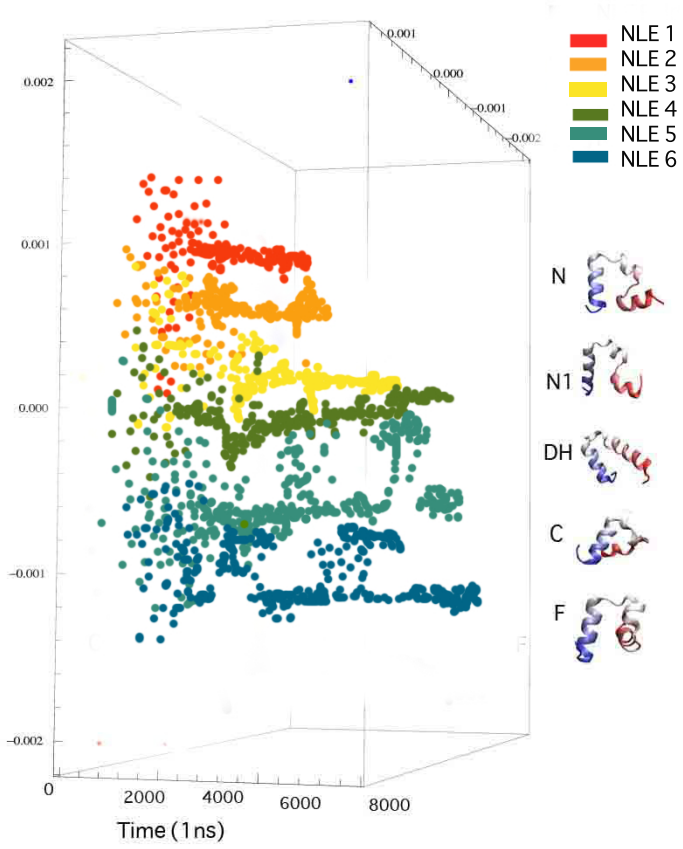

Figure 4.17: **Commonalities between NLE trajectories**: The protein shows large structural heterogeneity in the MD trajectories, and a few common hydrophobically collapsed states are labelled in the figure.

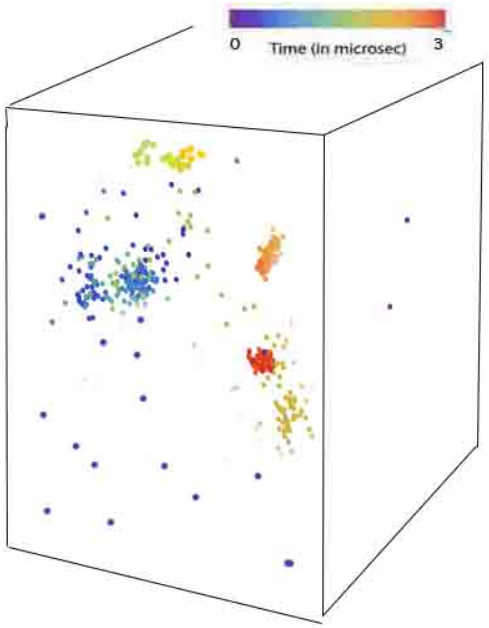

Figure 4.18: **Hydrophobic core in Trajectory 1**: We performed nMDS, binning the trajectories at 6 ns, by considering the C_α coordinates of the core forming residues: 6, 10, 17 (phenylalanine) and 20 (leucine) and both the NLE residues. We see that for trajectory 1, the core rearranges itself during folding.

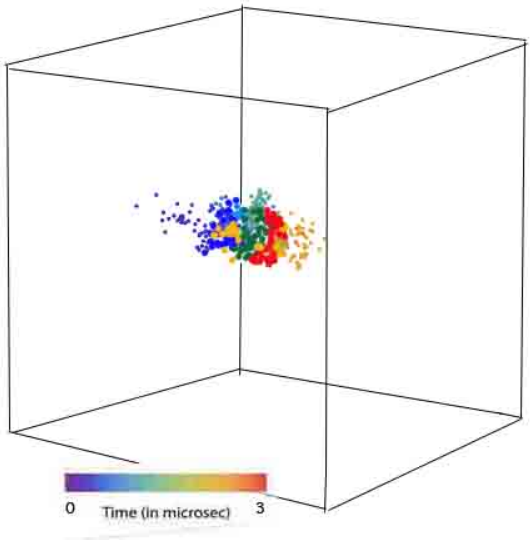

Figure 4.19: **Stable core shown for Trajectory 2**: We performed nMDS, binning the trajectories at 6 ns, by considering the C_α coordinates of the core forming residues: 6, 10, 17 (phenylalanine) and 20 (leucine) and both the NLE residues. We see that the core was very tight.

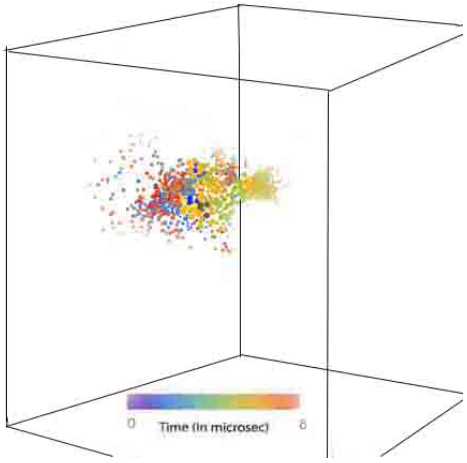

Figure 4.20: **Stable core shown for Trajectory 3**: We performed nMDS, binning the trajectories at 6 ns, by considering the C_α coordinates of the core forming residues: 6, 10, 17 (phenylalanine) and 20 (leucine) and both the NLE residues. We see that the core was very tight.

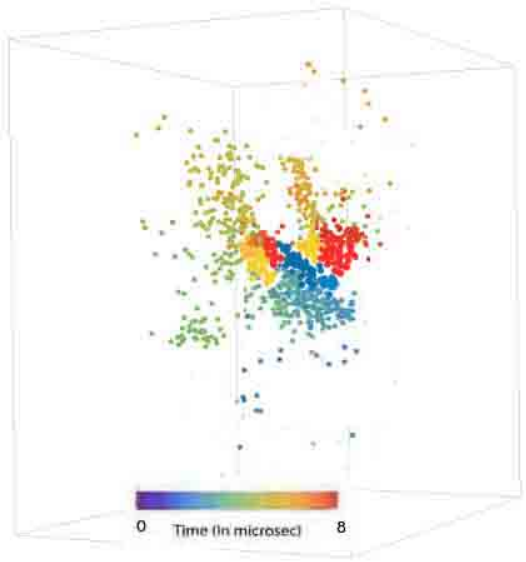

Figure 4.21: **Stable core shown for Trajectory 4**: We performed nMDS, binning the trajectories at 6 ns, by considering the C_α coordinates of the core forming residues: 6, 10, 17 (phenylalanine) and 20 (leucine) and both the NLE residues. We see that the core was very tight.

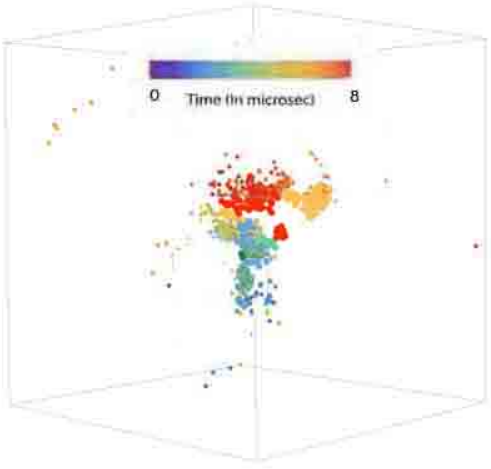

Figure 4.22: **Stable core shown for Trajectory 5**: We performed nMDS, binning the trajectories at 6 ns, by considering the C_α coordinates of the core forming residues: 6, 10, 17 (phenylalanine) and 20 (leucine) and both the NLE residues. We see that the core was very tight.

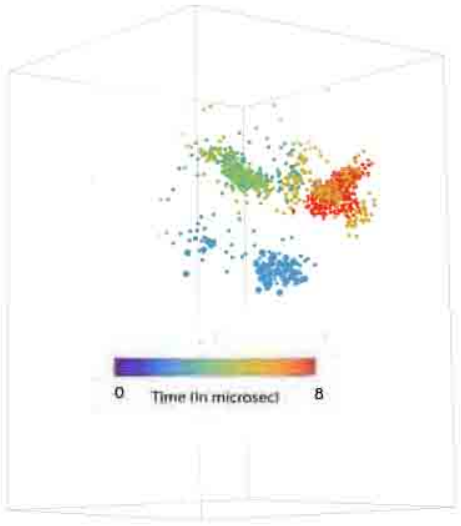

Figure 4.23: **Stable core shown for Trajectory 6** : We performed nMDS, binning the trajectories at 6 ns, by considering the C_α coordinates of the core forming residues: 6, 10, 17 (phenylalanine) and 20 (leucine) and both the NLE residues. We see that the core was very tight. There are two very close arrangements of the core found in this trajectory. The core slightly rearranges itself to proceed from one collapsed state to another.

The tightness of the core may prevent correct 3-D rearrangement of residues. In the villin trajectories, the core was rearranged during the flipping transition. In the NLE trajectory 1 the core seemed to form simultaneously with correct 3-D alignment, as illustrated in Fig. 4.18. Our results suggest that a tight hydrophobic core may be undesirable for biological proteins. While for small proteins, such effects may be small, for large proteins, delays due to getting trapped by tight hydrophobic cores cannot be ignored.

6 Summary: Why nMDS?

To summarize, nMDS results were stable to change in binning time and noise and hence, overcame the drawbacks of clustering. However, as we observed PCA was stable to binning time and achieved a reasonable compression for some configuration spaces. Why use nMDS over PCA? I wish to recall some of the issues with PCA we have already encountered in Chapter 3 and restate our preference for nMDS.

1. As we saw in Chapter 3, PCA did not do well to compress the data in the cartesian coordinate space. Cartesian coordinates are likely correlated nonlinearly in villin folding, and a linear method like PCA cannot pick out such correlations. Hence, it is important to devise/apply nonlinear methods for data reduction.

2. Note that in PCA, a Euclidean metric is implictly used to calcuate distances. In nMDS too, we used a Euclidean metric to assign dissimilarities/inequalities in ranking to the data points. How is nMDS better then? The answer is that in PCA, the exact distances between data points matters, whereas in nMDS, only the relative ordering of data points matters. This is a very important difference when analyzing structures in an MD trajectory. For example, if structure A and B are separated by 2 RMSD units and structure A and C by 4 units, when PCA projects the structures onto a lower dimensional subspace, it preserves the ratio of the distances between these structures. However, the RMSD distance does not mean anything significant, as we saw from our

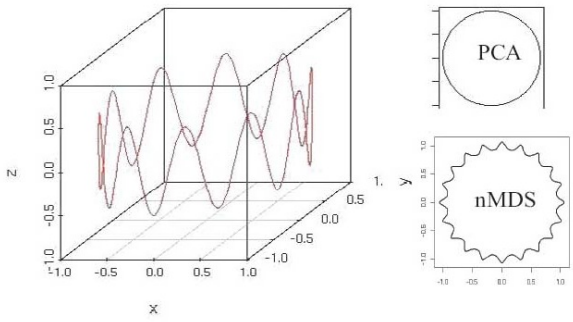

Figure 4.24: **nMDS over PCA**: Illustration to show that nMDS is superior to PCA in data compression. The left panel shows the input data in 3D and the right panel shows the corresponding projections obtained in 2D space by PCA and nMDS, respectively. Fig. Courtesy: Y. H. Taguchi and Y. Oono [51, 48].

clustering results. Two structures can have a significant RMSD difference, even if they are very similar, and vice versa. In nMDS, the exact RMSD values are immaterial, and only the fact that A is farther from C than it is from B is recorded. Dissimilarities in structures are more important than the exact distances computed by any metric between two structures. Hence, nMDS is a more powerful method for trajectory analysis than PCA or any metric dependent method.

7 Further improvement of methods

nMDS was always used with a Euclidean metric in our results reported in this thesis. However, the Euclidean metric may not be suitable to the study of all proteins/biological systems. In certain systems, it might be wiser to design more intuitive metrics that (at least) separate visually disparate structures even if their RMSDs are close. We also used a Hamming distance metric to rank the frames in our trajectories before applying nMDS, but this did not yield any new information compared to our results using the dRMSD or ℓ_2 metric. It is likely that the folding pathways are structurally heterogeneous and there are not many significant intermediates that new metrics may find. However, if larger proteins are studied, it would be desirable to design metrics that can distinguish topologically distinct structures that may lie close together if viewed in Euclidean RMSD space or Hamming distance space alone.

Much work still remains to be done with nMDS. While the compression achieved is certainly very high, nMDS is a computationally costly method and PCA may be a cheaper alternative if sufficient compression can be achieved using linear PCA. The interpretation of axes obtained after nMDS is still very dependent on visual inspection of embedded space data and perhaps some attempts can be made to construct linear maps between axes obtained by nMDS and those obtained by PCA, as described ahead. However for such mappings to be statistically meaningful, one will need to work with at least about 30 trajectories [50]. It can be expected that with the advance of technology, a large number of folding trajectories will soon become available and nMDS can prove to be a robust method to find collective coordinates for description of folding processes. Although we have illustrated our case for protein folding trajectories, nMDS should be able to reduce any MD trajectories effectively. Some methodological challenges for trajectory analysis will be: a) to devise ways to speed up nMDS to analyze large volumes of data and b) to devise better metrics to use nMDS with, depending on the protein in study.

In this Chapter, I have illustrated the use of nMDS to distill information from MD trajectories and suggested several ways to use nMDS for MD trajectory analysis applications in the future. In the next chapter, I will use the general observations made in this chapter to discuss the future of protein folding studies, after criticizing and analyzing current understanding of the field, in light of the work presented so far in this thesis.

Chapter 5

Conclusions and the view ahead for protein folding

1 Summary of work presented

In this thesis, I presented a non-metric multidimensional scaling (nMDS) method, that proved superior (except in terms of computational requirements) to conventional clustering techniques and PCA in distilling main features of folding trajectories. I illustrated how cluster analysis results were unstable against changes in cutoff parameters, changes in binning time and noise. I also showed that PCA, although computationally inexpensive, cannot achieve desired levels of compression for viewing MD trajectories in a reduced space.

I discussed results obtained from analyzing three complete villin headpiece folding trajectories and six norleucine mutant trajectories simulated by Freddolino and Schulten [1]. I will summarize below, the findings from this analysis.

1. **Structural Heterogeneity**: The folding trajectories of villin headpiece and its norleucine mutant showed structural heterogeneity. We found trivial similarities like the early formation of some secondary structure, which we can already expect, since most small proteins retain some secondary structure even in their denatured state [2, 19]. Structural heterogeneity of folding pathways ensures that the protein can switch pathways when conditions in the cell (physiological conditions, presence of chaperones, etc) change. Biological proteins are likely to have been evolutionarily selected for their multiplicity of folding pathways [7]. Although we talk of multiple pathways, this must not be misconstrued to mean that the protein can only fold through fixed "pathways" in conformational space. We merely imply the lack of select routes along which folding

proceeds. Whether or not the notion of a pathway is sensible, is an interesting question in itself.

2. **Local/global structure conflict**: There was competition between local structure formation and global rearrangement in all three villin headpiece trajectories that folded. If all three helices were formed in the beginning, the protein got trapped (for a few hundreds of nanoseconds) in double helix like states. However, such differences are not likely to alter the folding timescales by more than a microsecond or two for small proteins such as the villin headpiece. Since biosynthesis of proteins takes about 1-2 sec in bacteria and longer in Eukaryotes, delays over a few microseconds are irrelevant. In order to study the competition between local and global structure, it is important to study larger proteins ($\sim$100 residues or more). Early consolidation of secondary structures could enhance topological frustrations in large proteins.

3. **Role of hydrophobic collapse**: The intermediate structures in protein folding, resulting from hydrophobic collapse [14] are not only vulnerable to coagulation, but may also be topologically frustrated and may hence delay folding. We demonstrated this for the case of the norleucine mutant, where, in all but one of the trajectories (that reached completion), the protein was trapped in hydrophobically collapsed states. The norleucine residues were more buried in the hydrophobic core compared to their lysine counterparts in the (wild type) villin headpiece. This suggests that forming a tight hydrophobic core may prevent folding. In the case of larger proteins, hydrophobicity plays a crucial role and we discuss this in detail in this chapter [54].

Before I extend our observations to a general theory of protein folding, it is important to review and carefully analyze a prevalent notion (albeit vague) in protein folding today, the Energy Landscape or Folding Funnel theory.

2 The Energy Landscape theory of protein folding

In the early 90s, Leopold, Montal and Onuchic [55] proposed that protein folding is a collective self-organisation process that does not occur by an obligatory series of intermediates or a pathway, but by a multiplicity of routes down a folding funnel [56, 12, 57]. The global energy landscape was viewed as a funnel, which is a progressive organisation of an ensemble of partially folded structures that a protein visits on its way to the native state. The local roughness (or ruggedness) of the funnel reflects the presence of local minima which transiently trap the protein. There is a unique global minimum for the free energy which corresponds to the protein's native state. Appropriate order parameters are necessary to describe such a funnel pathway and the search for such order parameters has received much attention in the recent years.

I will argue below that it is not fruitful to think of folding in terms of funnels, while outlining a possible scenario for the folding.

Folding is usually discussed in terms of the free energy landscape only, but it is more instructive to think of the entropy and the energy terms separately. The denatured state is a higher energy, higher entropy state than the folded state. Therefore, if the energy contour is written in the configuration space, there must be a funnel like structure (phase space funnel). This is a truism. If we wish to reduce the dimension of the space on which the energy contour is drawn, then the landscape slope should not be the energy, but closer to the free energy. That is, if the phase space is coarse-grained, the room in which the trajectories wander must be effectively described by the reduction of the slope by entropy contribution. The free energy difference of unfolded and folded proteins is of the order of a few k_BT, because proteins are soft machines effectively utilizing thermal fluctuations to function. For tuning biological function, it may be necessary for the protein to effectively utilize thermal fluctuations for structural (and hence functional) changes. This "soft machine" picture of proteins tells us that low-lying excited conformations are without large ΔG, so that after

the initial hydrophobic collapse, the ΔG driving force should not be very large. Thus, the funnel slope is not strong and the trajectories are easily hindered by local bumps.

In the initial phase, secondary structures are formed while hydrophobic cores are made independently. This process eliminates many complicated labyrinth-like impasses equilibrium conformations could wander into. However, topological hindrances set in, especially in large proteins, and make the landscape more like labyrinths than funnels. Note that the "local minima" or metastable states are mostly not energetic, so funnel picture is misleading. In a labyrinth, the slope is only frustrating, just like someone pushing your back in a maze is not helpful at all. Local labyrinths and occasional lowering of energy locally (by formation of secondary structure elements, or local rearrangements) ensures forward motion (i.e., the energy slope is not helpful). Proteins must have been designed so that significant energy decrease does not lead the protein into a topological impasse. Rapid hydrophobic collapse does not lead to energetically metastable states, but topologically trapped states, which lead to deeper labyrinths in the free energy landscape. The main role of chaperones is likely to ease the depth of labyrinth. Then, the only way to do this is, because chaperones do not have any topological insight to solve the labyrinth, to ease the grips of hydrophobic pressure. At the risk of adding more terminology to the field, I call this picture a "punctuated labyrinth" picture as significant moves must be punctuated in the free energy landscape, while the protein is still searching in the labyrinth to minimize topological frustration.

The picture presented above must hold true for all folding scenarios, with or without chaperones. Study of folding of large proteins is not possible or complete without paying attention to the role of chaperones. If we wish to make a serious theory of protein folding, we must understand how protein folding is aided by chaperones.

3 Chaperone mediated protein folding

The folding of most newly synthesized proteins in the cell requires the interaction of a variety of protein cofactors known as molecular chaperones. Molecular chaperones specifically bind to nascent polypeptide chains and partially folded intermediates of proteins, preventing their aggregation and misfolding. Some chaperones only prevent mis-folding, while others also aid directly in folding. There are several families of chaperones; those most involved in protein folding are the 40-kDa heat shock protein (HSP40; DnaJ), 60-kDa heat shock protein (HSP60; GroEL), and 70-kDa heat shock protein (HSP70; DnaK) families.

Members of the HSP60 and HSP70 molecular chaperone families seem to be involved in preventing aggregation of intermediates formed during folding. Current understanding of the role of HSP70 in protein folding suggests that the chaperone sequesters the unfolded or partially folded protein, thereby preventing its aggregation, but does not actively participate in the folding process [4]. Hsp70 (heat shock) proteins can act to protect cells from thermal or oxidative stress. These stresses normally act to damage proteins, causing partial unfolding and possible aggregation. When newly synthesized proteins emerge from the ribosomes, the substrate binding domain of Hsp70 recognizes hydrophobic amino acid residues and binds to them in a reversible way (binding ATP). When ATP is hydrolyzed to ADP the binding pocket of Hsp70 closes, tightly binding the now-trapped peptide chain. By binding tightly to partially-synthesized peptide sequences, Hsp70 prevents them from aggregating. Once the entire protein is synthesized, a nucleotide exchange factor stimulates the release of ADP and binding of fresh ATP, opening the binding pocket.

GroEL/GroES chaperones from the chaperonin family of proteins aid more directly in protein folding. Within the cell, the unfolded protein binds to a hydrophobic patch on the interior rim of the GroEL, which is shaped like a rice cooker with a hydrophobic outside and hydrophillic inside, to form an ATP-GroEL-protein complex. The complex binds with a separate cooker lid, GroES to the open cavity of the chaperonin. This induces the individual

subunits of the chaperonin to rotate such that the hydrophobic substrate binding site is removed from the interior of the cavity, causing the substrate protein to be ejected from the rim into the now largely hydrophilic chamber. The hydrophilic environment of the chamber favors the burying of hydrophobic residues of the substrate, inducing substrate folding. Hydrolysis of ATP and binding of a new substrate protein to the opposite cavity sends an allosteric signal causing GroES and the encapsulated protein to be released into the cytosol. A given protein will undergo multiple rounds of folding, returning each time to its original unfolded state, until the native conformation or an intermediate structure committed to reaching the native state is achieved. GroEL/ES in this manner turn the protein inside out, that is turn their hydrophobic cores outward and this seems to accelerate folding. When GroEL dissociates from the protein, the less hydrophobic patches are released first [54]. It seems that the easing of the hydrophobic core by expansion and confinement inside the GroEL accelerates folding, perhaps not only by avoiding sticky hydrophobic patches coming together, but also by enabling the polar residues to participate in folding [54].

From the above two scenarios, it seems that while hydrophobic forces drive folding, hydrophobic patches can also be problematic for protein folding. While it is known that hey make the protein sticky and lead to aggregation and misfolding, it also seems likely that a very tight hydrophobic core can render the proteins functionally useless and tight cores are undone by chaperones like GroEL in larger proteins. Many chaperones may be working to simply ease the hydrophobic interaction and non specifically unfold the protein, so that it is able to fold correctly.

4 Future of protein folding studies

4.1 Disordered proteins

Firstly, proteins in their functional state, do not need to be at the free energy minimum [5]. Proteins can exist in metastable states that have a lifetime of a few milliseconds, during

which the protein is functionally active. Biology therefore, need not solve the free energy minimization problem [58]. Finding the free energy landscape in detail may hence be irrelevant to a large class of proteins [17]. Many proteins, in fact, exist in an unstructured or amorphous state, i.e., they are able to change their form depending on cellular conditions and functionality required [3]. Intrinsically unstructured proteins are characterized by lack of stable tertiary structure under physiological conditions *in vitro*. They adopt fixed three dimensional structure only after binding to other macromolecules such as other protiens/enzymes. Many unstructured proteins undergo transitions to ordered states upon binding to their targets. The coupled folding and binding may be local, involving only a few interacting residues, or it might involve an entire protein domain. Experimental and computational study of disordered proteins has picked up only in recent times. This is a very promising avenue for future protein folding studies.

4.2 Study of larger proteins and chaperones

Folding of many small proteins are already being widely studied. However, it is likely that most proteins show structural heterogeneity in folding pathways [7] and probing for specific pathways of folding is not a useful approach. Additionally, if the computational study of structural transitions in particular cases is desirable, then all atom MD simulations are the best bet we have. Coarse grained models and Monte Carlo cannot be used to study transitions or kinetics of the folding processes.

To study interesting features like local/global structure conflict and role of chaperones, we must study large proteins, as explained before. Current technology only enables simulations over a few microseconds. Interesting avenues for research would be to examine the role of hydrophobicity in folding. If a hyper-hydrophobic force field is used, it must slow down folding and trap the protein in collapsed states. Comparative studies with Go models may be useful for people wishing to build better force fields in MD. By intelligently dissecting the role of chaperones (e.g., do they reduce hydrophobicity? do they affect secondary

structure?), force fields with biased terms may be used to study chaperone mediated folding computationally. In order to study large proteins through all atom MD simulations, we must attain up to 4 orders of magnitude, faster computing power. Folding larger proteins will also help us validate and better design force fields used in MD. For instance, in a recent long simulation of the WW domain, it was reported that the CHARMM 22 force field was unable to fold the protein [22]. Probably, accurate potentials in MD simulations must be state dependent and include many-body effects, and binary potential decomposition (such as in current MD force fields) is not enough to study proteins over folding timescales.

Appendix A

ICS Survey to find commonalities amongst trajectories

nMDS results on the three trajectories taken together in dihedral angle space, cartesian coordinate space and C_α contact space (described above) do not show significant similarities between the trajectories. We used a stronger method to check for similarities if any. Are there any set of coordinates that vary similarly across all trajectories? To answer this question, we use a method called ICS Survey, developed by Rajaram [50, 49] which we explain below.

The basic idea of this method is summarized schematically in Fig. A.1. Each point in the picture corresponds to a single coordinate and its position is a representation of its movement in the original input space in a trajectory.

Geometrically, if we think of coordinate vectors as points lying in a high dimensional space, closely related coordinates would form a robust constellation that move in a similar way throughout the trajectory. Such a set of coordinates will be called an ICC (Internal Consistency Core).

The method adopted to characterize the data can then be considered to have the following steps:

1. Construct an ICC corresponding to a single trajectory.
2. Create a method to measure the consistency of the value of an arbitrary coordinate with respect to the ICC coordinates.
3. Rank all the coordinates in terms of this consistency score.
4. Use some statistical test to determine how many of the top ranked coordinates should be considered at the desired level of confidence. This extended set of coordinates should

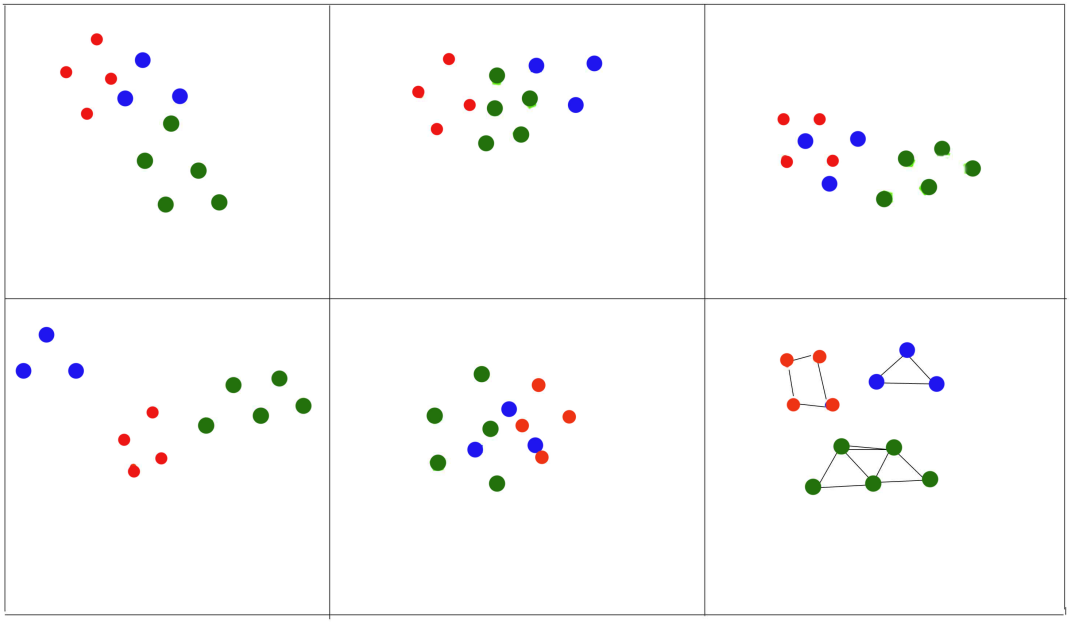

Figure A.1: **ICS Schematic**: Each point in the picture corresponds to a single coordinate, and its position is a representation of its movement in the original input space for one trajectory. The position changes from one rectangle to another (representing its change from one frame of the trajectory to another). We find that some of the coordinates move in a coordinated way throughout the trajectory (geometrically represented). Coordinates that move in a concerted fashion are colored similarly for the purpose of illustration.

now all correspond to the same trajectory as the ICC, and will be called an Internally Consistent Set (ICS). The ICS is the result of the method.

I now describe the implementation of Step 2 above. Suppose we have N_t trajectories and M_c coordinates of interest (cartesian coordinates of C_α atoms, dihedral angles, etc). For a given experiment t and for each coordinate c, construct a coordinate vector, which is its value across all the frames of a trajectory, $\boldsymbol{x_c^t}$. Let $G^t(h1, h2)$ be a correlation between the coordinate vectors for $h1$ and $h2$ for trajectory t. The Pearson correlation coefficient was used for the results in the thesis. Let us assume that by some means (methods to do this will be shown later), we have a set of N coordinates which are believed to constitute an ICC in all the trajectories. Then to find the consistency of a coordinate h, with respect to this ICC, the following steps are followed:

1. Consider the coordinate vector of coordinate h and of the ICC coordinates $(c_1 \dots c_N)$ across all N_t trajectories.

2. For each trajectory, find correlation of h with each ICC coordinate, and construct a vector of length N with these correlations (one such vector will be constructed for each trajectory)

$$(G^t(h, c_1), \cdots, G^t(h, c_N))$$

3. Normalize each vector appropriately (mean zero, unit standard deviation), to allow comparison over trajectories on equal footing

$$\boldsymbol{x_h^t} = \frac{1}{\sigma_h^t}\left(G^t(h, c_1) - m_c^t, \cdots, G^t(h, c_N) - m_c^t\right)$$

4. Compare vectors across trajectories, finding variance of each component across .trajectories

5. Inconsistency score (larger the score, less the consistency) for the coordinate c is given by sum of these variances

$$S_c = \sum_{i=1}^{N} Var_t(x_{h,i}^t),$$

Here, $\boldsymbol{x}_{h,i}^t$ is the i-th component of $\boldsymbol{x}_h^t$ and Var_t denotes the variance over all N_t trajectories. The use of variance ensures that the strength of (anti)correlation per se is not considered, just its consistency.

I now describe how the method ICS Survey finds an ICC across all experiments in an internally consistent way.

1. Select a random set of N coordinates as the ICC candidates.
2. Calculate inconsistency scores S_h for all coordinates(including the ICC coordinates themselves) with respect to the ICC.
3. Rank coordinates according to S_h.
4. N top ranked coordinates are new candidate ICC.
5. If new and old candidate ICCs are not the same, goto step 2.
6. If fixed point ICC has been reached. Use top ranked coordinates based on this fixed point ICC. Number of coordinates that make up an ICC may be decided using nMDS on the coordinate vectors.
7. Repeat from step one with another randomly selected set of coordinates to get sampling of the ICCs supported by the data set.

Thus, essentially the ICS Survey involves starting with a random set of coordinates and using them as an ICC, and then using the top ranked consistent coordinates as the ICC, and repeating this process till a fixed point is reached.To avoid a situation in which the ICC update gets caught in a loop, if the algorithm does not converge within a certain number of steps (chosen to be 500 for our implementation), then that run is terminated.

1 ICS Survey results

We first performed ICS on the dihedral angle space assuming that some of the dihedral angles of residues forming helix 1 and helix 3 fall into the ICC (This was obtained from nMDS analysis on dihedral angle coordinate vectors). We used a small ICC size of 8 (roughly 1/10th of the total number of coordinates in the dihedral angle space) to find the minimum number of coordinates that might be moving in tandem. We then performed ICS Survey comparing this ICC (containing 10 dihedral angles) across all trajectories. After 500 runs,

we still did not reach a fixed point, that is, no consistent ICC emerged across the trajectories. We performed the same test on the C_α contact distances between 7 residues (residues: 5, 7, 8, 15, 23, 27, 30) using an ICC size of 5. We found that no consistent ICC emerged across the trajectories.

References

[1] P. L. Freddolino and K. Schulten. Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophys. J.* , 97:2337–2346, 2009.

[2] T. E. Creighton. *Proteins*. W. H. Freeman and Company, New York, 2nd edition, 1993.

[3] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Bio.* , 6:197–208, 2005.

[4] A. L. Fink. Chaperone-Mediated Protein Folding. *Physiol Rev.* , 79:425–449, 1999.

[5] J. D. Bloom, A. Raval, and C. O. Wilke. Thermodynamics of Neutral Protein Evolution. *Genetics*, 175:255, 2007.

[6] C. Scholz, G. Stoller, T. Zarnt, G. Fischer, and F. X. Schmid. Cooperation of enzymatic and chaperone functions of trigger factor in the catalysis of protein folding. *EMBO J.* , 16:54–58, 1997.

[7] J. Udgaonkar. Multiple routes and structural heterogeneity in protein folding. *Ann. Rev. Biophys.* , 37:489–510, 2008.

[8] M. Ota, M. Ikeguchi, and A. Kidera. Phylogeny of protein-folding trajectories reveals a unique pathway to native structure. *PNAS*, 101:17658–17663, 2004.

[9] C. Levinthal. Are there pathways for protein folding? *J. Chem. Phys.* , 65:44–45, 1969.

[10] P. S. Kim and R. L. Baldwin. Intermediates in the folding reactions of small proteins. *Ann. Rev. Biochem.* , 59:631–660, 1990.

[11] V.I. Abkevich, A.M. Gutin, and Shakhnovich. Specific nucleus as the transition state for protein folding: Evidence from the lattice model. *Biochem.* , 33:10026–10036, 1994.

[12] K. A. Dill and H. S. Chan. From levinthal to pathways to funnels. *Nat. Struc. Biol.* , 4:10–19, 1997.

[13] O. B. Ptitsyn. How molten is the molten globule? *Nat. Stru. Biol.* , 3:488–490, 1996.

[14] V. E. Bychkova and O. B. Ptitsyn. The molten globule state of protein molecules is becoming a rule rather than exception. *Biophys. J.* , 38:58–66, 1993.

[15] M. Karplus and D. L. Weaver. Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Sci.* , 3:650–668, 1994.

[16] A. R. Ferscht. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *PNAS*, 92:10869–10873, 1995.

[17] M. Levitt. Nature of the protein universe. *PNAS*, 106:11079, 2009.

[18] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter. Sub-microsecond protein folding. *J. Mol. Biol.* , 359:546–553, 2006.

[19] J. Kulbeka, J. Hofritcher, and W. A. Eaton. The protein folding speed limit. *Curr. Opin. Struct. Biol.* , 14:76–88, 2004.

[20] P. Maragakis, K. Lindorff-Larson, M. P. Eastwood, R. O. Dror, J. L. Klepeis, I. T. Arkin, M. Jensen, H. Xu, N. Trbovic, R. A. Freisner, A. G. Palmer, and D. E. Shaw. Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *J. Phys. Chem. B.* , 112:6155–6158, 2008.

[21] D. L. Ensign, P. M. Kasson, and V. S. Pande. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.* , 374:806–816, 2007.

[22] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten. Ten-microsecond MD simulation of a fast-folding WW domain. *Biophys. J.* , 94:L75–L77, 2008.

[23] Y. Duan and P. Kollman. Pathways to a protein folding intermediate observed in a 1 microsecond simulation in aqueous solution. *Science*, 282:740–744, 1998.

[24] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* , 26:1781–1802, 2005.

[25] A. D. MacKerell, Jr, M. Feig, and C. L. Brooks III. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* , 25:1400–1415, 2004.

[26] H. C. Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Chem. Phys.* , 52:24–34, 1983.

[27] S. Miyamoto and P. A. Kollman. Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. *PSFG*, 16:226–245, 1993.

[28] M. E. Karpen, D. J. Tobias, and C. L. Brooks. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochem.* , 32:412–420, 1993.

[29] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich. Understanding ensemble protein folding at atomic detail. *PNAS*, 103:17747–17752, 2006.

[30] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.* , 38:236–240, 1999.

[31] A.K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys*, 31(3), 1999.

[32] D. J. Watts and S. H. Strogatz. Small world. *Nature*, 393:440–442, 1998.

[33] G. Punj. and D. W. Stewart. Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*, pages 134–148, 1983.

[34] H. J. Zeng, Q. C. He, Z. Chen, E. Y. Ma, and J. Ma. Learning to cluster web search results. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217. ACM New York, NY, USA, 2004.

[35] G. B. Coleman and H. C. Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979.

[36] G. Jayachandran, V. Vishal, and V. S. Pande. Using massively parallel simulation and markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J. Chem. Phys.* , 124(16):164902–164914, 2006.

[37] D. van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* , pages 1701–1718, 2005.

[38] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus. One-Dimensional Barrier-Preserving Free-Energy Projections of a beta-sheet Miniprotein: New Insights into the Folding Process. *J. Phys. Chem. B.* , 112:8701–8714, 2008.

[39] J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21:3201–3212, 2005.

[40] O. F. Lange, N. . A. Lakomek, C. FarŔs, G. F. SchrŽder, K. F. A. Walter, S. Becker, J. Meile, H. Grubm§ller, C. Griesinger, and B. L. de Groot. Recognition dynamics up to microseconds revealed from an RDC-derived Ubiquitin ensemble in solution. *Science*, 320:1471–1475, 2008.

[41] LW. Yang, E. Eyal, I. Bahar, and A. Kitao. Principal component analysis of native ensembles of biomolecular structures PCA_NEST : insights into functional dynamics. *Bioinformatics*, 25:606–614, 2009.

[42] SchÃűlkopf, Bernhard, Smola, Alexander, and KR. MÃijller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[43] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten. Principal Component Analysis and long time protein dynamics. *J. Phys. Chem.* , 100:2567–2572, 1996.

[44] R. N. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function I. *Psychometrika*, 27:125–139, 1962.

[45] R. N. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function II. *Psychometrika*, 27:219–246, 1962.

[46] J. B. Kruskal. Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika*, 29:1–29, 1964.

[47] J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, 1964.

[48] Y. H. Taguchi and Y. Oono. Relational patterns of gene expression via non-metric multidimensional scaling analysis. *Bioinformatics*, 21:730–740, 2005.

[49] S. Rajaram. *Phenomenological approaches to the analysis of high-throughput biological experiments*. PhD thesis, University of Illinois, Dept. of Phys. , Urbana-Champaign, IL, 2009.

[50] S. Rajaram. A novel meta-analysis method exploiting consistency of high-throughput experiments. *Bioinformatics*, 25:636–642, 2009.

[51] Y. H. Taguchi and Y. Oono. Nonmetric multidimensional scaling as a data-mining tool: new algorithm and new targets. *Geometrical Structures of Phase Space, Multidimensional Chaos, Special Volume of Adv. Chem. Phys.* , 130:315–351, 2004.

[52] P.E. Green, F. J. Carmone Jr. , and S.M. Smith. *Multidimensional Scaling: Concepts and Applications*. Allyn and Bacon, Boston, MA, 1970.

[53] S. Rajaram and Y. Oono. R implementation of nmds algorithm., 2009. Submitted.

[54] F. U. Hartl and M. H-Hartl. Converging concepts of protein folding *in vitro* and *in vivo*. *Nat. Struc. Biol.* , 16:574–581, 2009.

[55] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *PNAS*, 89:8721–8725, 1992.

[56] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Ann. Rev. Phys. Chem.* , 48:545–600, 1997.

[57] J. N. Onuchic and P. E. Wolynes. Theory of protein folding. *Curr. Opin. Struct. Biol.* , 14:70–75, 2004.

[58] Y. Oono. Integrative Natural History. Lecture Notes, 2009.