Department of Physics
Indian Institute of Technology Guwahati
Ph.D. Thesis

Development of a New Class of Enhanced Kinetic Sampling Methods for Biomolecular Simulations

Susmita Ghosh

Roll No: 136121016

Supervisors: Dr. Swati Bhattarcharya
Prof. Saurabh Basu

January, 2018

Development of a New Class of Enhanced Kinetic Sampling Methods for Biomolecular Simulations

A thesis submitted by

Susmita Ghosh

Roll No: 136121016

to

Indian Institute of Technology Guwahati
in partial fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy in Physics

Department of Physics
Indian Institute of Technology Guwahati
Guwahati-781039, Assam, India

Declaration

I hereby declare that the work in this dissertation entitled "Development of a New Class of Enhanced Kinetic Sampling Methods for Biomolecular Simulations" has been carried out by me under the supervision of Dr. Swati Bhattacharya and Prof. Saurabh Basu in collaboration with others as acknowledged and the contents of this dissertation have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. I am solely responsible for any errors or inaccuracies, albeit inadvertently.

(Susmita Ghosh) January 12, 2018
Department of Physics
Indian Institute of Technology Guwahati
Guwahati - 781039, India

Disclaimer

The bibiliography included in this thesis is, by no means but complete but contains the ones which are consulted throughly by me. I apologize for inadvertently missing out some of the research papers, review articles and other scientific documents pertaining to the focus of this thesis which should also have been cited. For illustration purpose some of figures in this thesis are taken from other sources and properly cited.

Certificate

It is certified that the work contained in the thesis entitled "*Development of a New Class of Enhanced Kinetic Sampling Methods for Biomolecular Simulations*" by Susmita Ghosh (Roll No-136121016), a Ph.D. student of the Department of Physics, Indian Institute of Technology Guwahati is carried out under our supervision and has not been submitted elsewhere for the award of any other degree.

(Dr. Swati Bhattarcharya)
Department of Physics
Indian Institute of Technology, Guwahati
Guwahati - 781039, India

(Prof. Saurabh Basu)
Department of Physics
Indian Institute of Technology, Guwahati
Guwahati - 781039, India

Dedicated to

my loving parents

&

my brother

Acknowledgements

The Ph.D has been a long and strange journey. This thesis would not have been possible without the inspiration, wisdom and assistance provided by some people I met on the way, and it is a pleasure to show my appreciation here.

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Swati Bhattacharya for providing me much needed guidance and assistance throughout my thesis, and helping me get on the road to bio-simulations. I am also deeply indebted to my other supervisor, Prof. Saurabh Basu for his constant support and guidance which was often extended beyond the call of duty.

I would like to acknowledge our collaborators, Prof. Abhijit Chatterjee and Ms. Arti Bhoutekar, Indian Institute of Technology Bombay, India. My thesis work is based on joint work with Prof. Abhijit Chatterjee.

Special mention of appreciation goes to my doctoral committee members. I would very respectfully like to thank for their availability and precious contribution despite their busy schedule.

I've been delighted to have such wonderful professors here in IIT Guwahati. Mr. Basab Bijoy Purkayastha's help regarding computer issues was truly invaluable. The life here would not be so fulfilling without the kind, honest, friendly people around the Department. I would like to thank the entire Physics Department of IIT Guwahati, as every staff member has been always helpful to me. I am grateful to IIT Guwahati, and Government of India, Ministry of Human Resources Development for their financial support.

The PARAM-ISHAN supercomputing facility of IIT Guwahati is acknowledged for providing the valuable computational resources for my research work.

I would like to express my regard to my college teacher Dr. Debopriya Shyam who has been a great teacher with responsibilities and inspiration to me during my college days. Here I would like to also thank two of my science teachers, Sudhanshu

Ghosh and Pranab Chakraborty for inspiring me in physics during my school days.

I am thankful to my friends Priyadarshini, Kajwal, Eshita, Camelia, Bijita, Sangkha, Subhadeep, Noor, Srikrishna, Sanjib and my classmates for always being with me and making my time in IITG so enjoyable.

Last but not least, I want to convey my gratitude to my family for their unflagging support and patience. Their sacrifices and blessing have always encouraged me to take my further steps.

Susmita Ghosh, IIT Guwahati January 12, 2018

Abstract

Gaining a detailed insight into the molecular kinetics of the biomolecular systems is the fundamental step to comprehend the human health. In the present thesis, our focus is on the development of a new class of enhanced kinetic sampling methods for construction of the kinetic network of biomolecular systems using Markov State Model (MSM) approach. Here a new concept of validity time is introduced to address the uncertainty/error in an MSM due to missing states/pathways and a theoretical framework for calculation of validity time is provided to quantify the completeness of an MSM. An efficient and accurate construction of MSM with desired validity time is accomplished by a suite of new algorithmic developments: namely Swarm MD, State-constrained MD (SC-MD) and Programmed state-constrained MD (PSC-MD). The newly developed methods and concepts are used to construct MSM for Single Molecule Force Spectroscopy (SMFS) experiments with the objective of rapidly construction of kinetic network using a stretching force to accelerate rare conformational transitions of biomolecules involving multiple states and kinetic pathways. An idea of master-MSM for constant-probe separation experiments in a Force-Spectroscopy (FS) setup is proposed to predict the connection between topologically different kinetic networks constructed at distinct trap separations. On the basis of the idea of master-MSM, a Time-Dependent MSM (TD-MSM) formalism is developed where the external stretching force to the system is a function of time. The TD-MSM approach enables us to get the new molecular insights into the kinetic, thermodynamics and mechanical properties of the system. Finally, we extend the Master-MSM method at constant-probe separation to constant-force experiment to predict the intrinsic kinetic properties (kinetic rates at zero-force conditions) of slow transitions at lower computational cost by a handful of simulations at various stretching forces.

List of Publications

1. A. Bhoutekar, S. Ghosh, S. Bhattacharya, and A. Chatterjee. A new class of enhanced kinetic sampling methods for building Markov state models. *J. Chem. Phys.*, **147**, 152702 (2017).

2. S. Ghosh, A. Chatterjee, and S. Bhattacharya. Time-Dependent Markov State Models for Single Molecule Force Spectroscopy. *J. Chem. Theory Comput.*, **13**, 957-962 (2017).

3. **Manuscript Submitted:**
   S. Ghosh, A. Chatterjee, and S. Bhattacharya. Accelerated Construction of Kinetic Network Model of Biomolecules Using Steered Molecular Dynamics.

Conference presentation:

1. S. Ghosh and S. Bhattacharya. Dynamics of deca-alanine under extension. 58^{th} XXVII IUPAP Conference on Computational Physics: CCP2015 2-5 December 2015, IIT Guwahati, Assam, India.

Contents

List of Figures

List of Tables

Abbreviations

MSM	Markov State Model
MD	Molecular Dynamics
MC	Monte Carlo
KMC	Kinetic Monte Carlo
SC-MD	State Constrained MD
PSC-MD	Programmed State Constrained MD
SMFS	Single Molecule Force Spectroscopy
FS	Force Spectroscopy
TD-MSM	Time Dependent Markov State Model
BEP	Bell-Evans-Polany
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
PDB	Protein Data Bank
MLE	Maximum Likelihood Estimation
CG	Coarse Grained
TBA	Thrombin Binding Aptamer
NMR	Nuclear Magnetic Resonance
PME	Particle Mesh Ewald
PBE	Periodic Boundary Condition
SMD	Steered Molecular Dynamics
CPU	Central Processing Unit
GPU	Graphical Processing Unit
RMSD	Root Mean Square Deviation
PES	Potential Energy Surface
TST	Transition State Theory
AFM	Atomic Force Microscopy
FE	Free Energy

Chapter 1

Introduction

1.1 General Introduction

Living organisms are made of cells; cell contains many biochemical components such as protein, lipid, nucleic acid and carbohydrate. But living cells are not only the collections of these macromolecules also, they are extraordinarily "self-organized" or ordered and self-replicated. The most obvious thing about living organisms is their astounding diversity. The functions performed by these macromolecules play an essential role in every day life of every living being. The most fascinating questions about living cells are how organisms use energy to do work, communicating with and sensing the environment and how a living system is growing and self-perpetuating. Even today, although scientists have unraveled many mysteries of life, we still do not understand many of the dynamical features of the molecules that are the essence of living systems and play pivotal role in controlling the human health or preventing many human diseases.

To understand the variety and complexity of living world, it is essential to study the various biological phenomena at every energy level, from atoms and molecules to cells, organisms and environment. The amazing pace of advancement in our understanding of biology, chemistry and physics has engendered cumulative specialized fields involving the development of new approaches, both in the laboratory and in computer models and the combination of these different approaches can provide us important insights into the microscopic world of life on the molecular level to unlock the mystery of life. This highly interdisciplinary research area accompanies life-science by accumulating idea, exchanging methods and knowledge from physics, biology, chemistry, computer science and medicine. Among the various interdisciplinary research, Biophysics is the field which integrates physics, biology and even

chemistry by using the physical toolkits, namely, the methods of mathematical analysis and computer modeling to solve the mystery and complexity of life by simple principles of physics and chemistry.

Changes in configurations of proteins and nucleic acids underpin the majority of emergent biological phenomena in daily life. Many of the human diseases are the result of failure of some proteins to remain in their native state, partly due to abnormal protein folding. In many cases, the functioning of a protein involves conformational changes. Elucidating the details of the kinetics of such proteins is essential for deciphering their functions. Examples of such protein systems are kinases activation (enzymes that phosphorylate other protein and are responsible for aberrant cellular signaling in cancer), G-protein coupled receptor signaling (key signaling proteins that sense a wide range of extracellular signals such as hormones, drugs, photons, ions, etc.), Ligand-binding proteins such as Myoglobin and intrinsically disorders proteins such as amyloid-β. Many neurological diseases such as Creutzfeldt-Jakob, Mad Cow and Alzheimer's are also hypothesized to result from protein mis-folding. Therefore, a deeper insight into the kinetics of biomolecules is important to control the human health as well as to uncover more complex questions in molecular biology. The studies on molecular structure and kinetics are mainly focused on the following key questions: the structural definition of conformational states of system, the mechanism of system's conformational change, the structure of transition states, the kinetic pathways between conformational states and the height of barriers connecting these key conformations. Finding answers to these key questions remains a unifying paradigm of biophysical studies.

1.2 Experimental Techniques to Study Biomolecular Systems

Over last decade, the investigation of internal dynamics and function of biomolecules as well as their stability became possible due to advanced and refined experimental methods like Nuclear magnetic resonance (NMR) Spectroscopy[1–3], X-ray crystallography, Electron Microscopy. X-ray crystallography[4] can provide us atomic details of a protein or a nucleic acid along with information of ligands, inhibitors, ions and other molecules that are incorporated into the crystal by examining X-ray diffraction pattern. NMR spectroscopy is unique among the available methods for characterization of the structure of biomolecules as NMR data can be recorded in solution. In

Electron Microscopy, a beam of electrons is used to obtain 3D images directly to determine the structure of large macromolecules. In general, Electron Microscopy combines the information obtained from X-ray crystallography and NMR spectroscopy to sort out the atomic details. Relaxation-dispersion NMR can see transiently populated, partly unfolded state of proteins. Förster resonance energy transfer (FRET) experiments are being used to study protein-protein interactions inside living cell, *in vivo* protein transport or turnover phenomena, the biological structure[4–7] and to seek the information about metabolic or signaling pathway[8]. Single-molecule force spectroscopy (SMFS), like, the Atomic force microscopy (AFM), Optical traps or Magnetic tweezers can detect forces resulting from conformational changes, binding or activation events and hence examines the mechanical properties or force-induced unfolding-folding dynamics of proteins with unprecedented details at the single molecule level[9–15]. The unfolding and refolding of titin and bacteriorhodospin or ligand-dependent fluctuation of calmodulin are well-known examples[16–19]. Surface plasmon resonance (SPR)[20–23] allows the analysis of binding kinetics, affinity and selectivity of proteins, DNA, other small molecules to surface immobilized capture agents by detecting and quantifying interactions between molecules at a surface. However, in many cases biomolecules' structure and dynamics are not amenable for investigation experimentally or expensive to perform using experimental techniques. Moreover, no conventional experiment allows us for access to all timescales of motion with atomic resolution. Molecular Dynamics (MD) simulations can provide a solution to this problem to some extent.

1.3 Molecular Dynamics for Biomolecular Simulations

Molecular dynamics (MD) is a technique for computer simulation of complex systems, modelled at the atomic level. Nowadays, classical Molecular Dynamics (MD) simulation has become an incredibly powerful and frequently used tool in obtaining microscopically resolved information on biomolecules which is not easily accessible by existing experimental tools. MD simulations can be used for inorganic systems as well as for organic or biomolecular systems. MD allows us to study the interaction between molecules on a microscopic level, such as the influence of ions on the stability of a biomolecule and their coordinated behaviour towards a solute. MD simulations are successfully applied to a wide range of problems in different fields including solid state physics, chemistry, material sciences and biology. It involves solving Newton's

equations of motion for a collection of interacting particles. In MD simulation, a potential model is provided by the theorist and then calculations are carried out by the computer, following a strategy based on a physical potential model. Thus computer experiment (MD simulation) serves as a bridge between the experiment and theory. MD simulations provide a picture of microscopic behavior under controlled conditions and statistical mechanics provides the mathematical link between the microscopic behavior and macroscopic properties (thermodynamics). MD was first developed in the late 50s and started its journey with the pioneering applications to the dynamics of liquids by Alder, Wainwright, Rahman in the late 1950s and early 1960s[24–27]. There are two main families of MD methods: classical MD and "quantum" or "first-principles" MD simulations. In classical MD, molecules are treated as classical objects; everything that explicitly involves the laws of quantum physics, e.g., the motion of electrons, is neglected which make it much less computationally expensive than quantum mechanics, whilst still allowing all the atoms of a protein to be simulated. It has been found that neglecting the electrons does not, in general, prevent the method from generating realistic dynamics of molecule. In "first-principle" MD simulations, interatomic potential are calculated from Density Functional Theory (DFT); a different approach proposed by Car and Parinello since the 1980s[28], takes into account the quantum nature of the chemical bond explicitly. Car-Parinello MD (CPMD) method is being used in the case (ex. chemical bonding, the presence of important non-covalent intermediates and tunneling of protons or electrons etc.) where the electronic motions of the system are involved. However, the use of "Quantum" MD is computationally expensive. At present, in this thesis, we consider only classical MD simulations.

1.4 Challenges in MD Simulations

Understanding the kinetics of biomolecules via MD simulations has been an aspiration of computational biologist since the inception of the field. Unfortunately, they are often too short to capture the biologically relevant timescales for the various complex process such as protein-folding, protein-ligand binding, macromolecular aggregation and conformational transitions with sufficient statistical accuracy. The main challenges that must be overcome to achieve efficient simulations to imitate the system correctly are: the sampling problem and analysis of massive amount of MD data. Moreover, for many complex systems the free energy (FE) landscape is rugged and corrugated. In such a case, the FE landscape of the system contains

multiple deep competing free-energy minima separated by large free energy barriers and the system may be trapped within such basins for long-time. The presence of such kinetic traps make the process slow. The understanding of such slow processes is considerably more challenging by using only MD studies.

The major hurdle in MD simulation is the sampling problem since an atomistic millisecond-long MD simulations requires iterating over the calculation cycle 10^{12} times due to the stiffness of the equation of motion (timestep is limited to fs). This expense is compounded by large system sizes ($\sim 10^5$ atoms for explicit solvent simulations) and the necessity to witness many events for statistical confidence. This makes the computational resources required to capture the events of biological significance like the conformational transitions including folding[29,30], complex conformational rearrangements between native protein substates[31,32], and ligand binding process[33] which span a wide range of timescales usually many microseconds, milliseconds or longer via simulation, enormous. An ever increasing amount of computational power[34,35] and advancement in specialized software and hardware have allowed us to reach increasing system sizes and simulation timescales from picoseconds up to a few milliseconds[36]. For example, the present generation of computers takes the amenities of efficient parallelization of MD codes and accelerators (GPU, multi-core CPU) to speed up the process. The most popular simulation codes (AMBER[37], CHARMM[38], GROMACS[39] or NAMD[40]) have long been compatible with the messaging passing interface (MPI). In addition, for truly understanding the dynamics, a single long MD trajectory is inadequate, one needs to generate the sufficient number of trajectories due to the stochastic nature of molecular kinetics. Furthermore, to identify all possible kinetic traps on the folding pathway of slow kinetic processes the length of MD trajectory should be very long. There have been a number of approaches introduced that are aimed to enhance the sampling, which often involve modifying the energy landscape of a system or increasing the temperature to speed up its transitions between different states. Such methods include, for example, Umbrella sampling[41] (US), Transition path sampling[42,43] (TPS), Metadynamics[44] and new variants of Metadynamics[45–48], Adaptive biasing force method[49], Conformational flooding[50], Replica exchange[51], On-The-Fly Kinetic Monte Carlo[52], Self guided MD[53] and many others that were reviewed by Berne and Straub[54]. Yet, for complex biomolecular systems, an automatic and unsupervised realization of such an adaptive sampling procedure is not always trustworthy as they often need the information *a priori* about the system. For example, generalized ensemble methods, such as the replica-exchange method based on the idea of random walk of the system

in temperature space, may lead to the unrealistic dynamics of the complex system, where the entropic barrier dominates over the energetic barrier because entropic barriers will become more inviolable at high temperature. However, an obvious strategy is to perform a series of parallel simulations from several starting conformations which have now been become possible with the availability of the present HPC systems.

Secondly, even if one could run a sufficient number of long simulations, analyzing massive amount of data resulting from multiple MD trajectories and turning them to a meaningful knowledge are the next challenges faced by simulators to gain scientific insight from simulations. One of the most common analysis technique is the projection of free energy landscape onto few order parameters. This method will be valid if the chosen order parameters represent truly reaction coordinates for the process of interest. But for complex system there is always possibility the chosen order parameter may not be correct. Other popular machine learning approaches and quantitative approach like principal component analysis (PCA) are able to identify key conformational states and some features of the dynamics. But these methods fail to be exploited for extracting the kinetic information embedded in the MD data. Markov State Models (MSMs) provide a potential solution to tackle aforesaid problems by using the statistical information encoded in the data. The other available approaches for the kinetic characterization of the system are: Diffusion Map[55], Kinetic distance and Kinetic Map[56], Markov transition Models[57], Time-lagged independent component analysis (TICA)[58,59], and kernel TICA[60]. Noteworthy, in the last ten years, since the inception of MSMs, it has gained considerable attention within the biomolecular simulations community to yield a novel insight into protein realistic kinetics in terms of state-to-state transition, bridging the gap from molecular dynamics on nanoseconds to folding on milliseconds.

1.5 Markov State Model (MSM)

Markov State Model (MSM)[61–65] is, in essence a kinetic network model for representing the knowledge of molecular kinetics. Basically, an MSM depicts the accessible conformational states to the system, the essential (slow) dynamical process and their timescales, the distinct transition pathways between the relevant metastable states, the kinetic rate between the conformational states and occupancy of the system in metastable states. The MSM provides a kinetic map of a biomolecule under study

associated with the informations about what are the likelihood of residing in a state, how fast the system can move from one state to another state and the chance occurrence of the transition between two important intermediate states. Hence MSM facilitates understanding of the essential dynamics or decisions regarding where to run a new simulation to scour the conformational map. Thus two main applications of the MSM are; i) to qualitatively understand the key features of the dynamics and ii) to reconstruct dynamical trajectories of the system. While the former application can be useful for understanding mechanisms or building order parameters/reaction coordinates, the latter application can be useful for computing ensemble-averaged quantities of interest.

1.6 Mathematical Background of the MSM

Markov State Models approximate the dynamics of the system by deriving a discrete-state continuous (or discrete) time master equations where the system passes through a sequence of states $\{x_k \in S\}$ drawn from a model dependent state-space S at transition times $\{t_0 < t_1 < t_2 < t_3 < - - -\}$. Equivalently the sequence of states (x_k) can be viewed as Markov chain associated with an inhomogeneous Poisson process, that is, first-order dynamics (with exponential decay) with rate $Q(t)$ that generates a sequence of waiting times δ_{tk} for the interval between transitions. Then the dynamics of the system under consideration can be modeled by a continuous (or discrete) time master equation, characterized by a matrix of phenomenological rate constants describing the rate (inverse of waiting time, δ_{tk}) of interconversion between states (or transition probability between states), where the local potential energy minima are identified as states and the intrastate transition rates are estimated from transition state theory[66–72]. MSM partitions the configuration space into a number of distinct states, called metastable states, such that the intra-state transitions are fast but the inter-state transitions are slow. Such separation of timescales ensures that the model is Markovian, in that the probability of being in a given state at time $t + \Delta t$ depends only on the state at time t. Here Δt is the lag time or observation interval for which the rate matrix (transition probability matrix) model is constructed and shorter than the timescale of the process of interest with a reasonable number of states. So Δt turns to be an important parameter to determine the quality of an MSM. The key steps while building an MSM at room temperature are: i) identification of states, ii) finding the rate constants for the moves from the states. Once the transition matrix or rate matrix is constructed, many dynamical features and ther-

modynamic information such as stationary probability distribution on microstates, ensemble averages of molecular observables, or the free energy differences between macrostates, relaxation timescales, transition pathways can be extracted by solving master equation.

1.7 Advances in MSM Building Approaches

MSMs have recently become a very successful approach in capturing the long-time scale conformational dynamics of complex system with multitude of metastable states. MSMs have the power to compute many important thermodynamic, kinetic, and mechanistic molecular quantities more directly and unambiguously than with conventional MD analyses. In this paradigm, the dynamics of the system is described by state-to-state transitions by seeding the trajectories from multiple short independent MD runs rather than attempting to generate one realization of an entire process and thus provide better predictions at the microsecond or millisecond scale with nanosecond timescales MD data. There are several recent reviews on discussing the Markov State Model and its application to biological systems[73–76]. The noteworthy example is the fully automated MSM construction by post-analyzing long MD trajectories, has been pioneered by Pandey and co-workers[77]. Adaptive sampling algorithms for MSM construction take this statistical approach a step further[78–81]. In adaptive sampling techniques, one first obtains an initial model of the entire process of interest, then new simulation are performed from states that are least populated. Weber and Pandey have shown that such adaptive sampling results a dramatic reduction in statistical uncertainty in the observable of interest. Once sufficient sampling is obtained, MSM can be used at multiple resolution by varying the degree of coarse-graining[64,82]. Still, all of the MSM building techniques are limited by the state discretization error (systematic error) due to Markov approximation of continuous dynamics on discretized state space and statistical error associated with the uncertainty in rate constant estimation due to limited quantity of data. McGibbon et al.[83] have reported use of Hidden Markov Model (HMM) as an alternative approach of regular MSM which provides a way to discard the Markovian approximation for optimal construction of MSMs.

Already a number of software packages, named by EMMA[63], PyEMMA[84] and MSM-Builder[85] have been developed to facilitate the construction, validation and interpretation of MSMs. In addition, MSMExplorer[86] software is used for visualization

purpose of MSM network. The above two softwares (EMMA and MSM-Builder) follow the same basic operation for MSM construction (data clustering, transition matrix estimation, lumping kinetically similar micro-states to macro-states by PCCA (pairwise criterion comparison approach) method, identification of highest flux pathway by TPT (Transition path theory)). Herein states are obtained by categorizing similar conformations on the basis of some dimension reduction algorithms such as clustering technique, PCA analysis and TICA analysis. A major advantage of clustering is that it is less biased since no reaction coordinate has to be assumed *a priori*. Recently, Pande and his coworkers have suggested that Ward's minimum variance method is the best way for predicting cluster assignments of MD data set among the existing clustering algorithms[87]. The rates of interconversion between states are estimated from simulation trajectories by a statistical approach such as maximum-likelihood estimation (MLE) or Bayesian statistics. In both of the MSM-building softwares, the models are validated by implied timescales (internal timescales of relaxation) and Chapman-Kolmogorov test. The difference is that MSMbuilder is suitable for large number of clusters which has been optimized by rapid RMSD clustering whereas EMMA deals with the smaller MSM and puts focus on the quality of MSM by statistical validation of MSM and reduction of statistical uncertainty of quantities of interest. The availability of these open software tools enable a rapid advancement in this field and these software packages are rapidly evolving with new development. For example, very recently an improved MSM estimator which corrects the error in the transition probability estimation due to short non-equilibrium simulation along with the reduced state decomposition error by exploiting framework of observable operator model, is implemented in PyEMMA[88].

1.8 Applications of the MSM

MSMs have now been successfully used predominantly to study macromolecule's folding mechanism, protein-ligand binding process, peptide dynamics, peptide aggregation, protein conformational changes, self-assembly, intrinsically disordered proteins and connection between simulated and kinetic experimental data such as temperature-jump or fluorescence correlation spectroscopy.[73] One other potential application where MSMs can make a greater impact is look into the complex dynamics of non-equilibrium system where the system is influenced by an external perturbations as non-equilibrium perturbations of protein folding MSMs reveal the dynamically frozen states in their conformational landscape. Very recently, few

approaches have been proposed for constructing MSM of the system under non-equilibrium conditions[89–93]. The notable examples are extraction of kinetic equilibrium properties by building re-weighted MSM from potential biased MD trajectories[93] and building MSM for periodically driven non-equilibrium systems caused by external fields[89]. In this thesis a special attention is paid to such a category of non-equilibrium experiments, which in this case is the single molecule force spectroscopy (SMFS) experiment and is discussed in next paragraph.

1.9 Single Molecule Force Spectroscopy (SMFS) as a Potential Candidate for the Application of MSM

Over the last two decades, SMFS has emerged as a powerful tool by scientific community to probe the energy landscape and interaction of biomolecules. In SMFS a force is applied on a single molecule to induce a structural change and thus enables single protein molecule to be unfolded over a well-defined reaction coordinate which would be otherwise rarely sampled. Force is not only a tool to explore and examine proteins; many biological systems experience force in a physiological environment. Besides, many biological macromolecules take part in various cellular processes ranging from replication, transcription, translation and protein degradation, to cell adhesion and transport by using mechanical force. So it is important to know how the biomolecules respond mechanically to forces exerted on them. To manipulate a single-molecule by force, the most popular experimental techniques[94] are optical tweezers[95], magnetic tweezers[96,97] and the atomic force microscope (AFM)[98–102]. Typically, in these approaches, one end of the molecule is connected to a surface or tip leaving the other end free to interact with either the tip, another molecule or a surface, and the force vs extension of the molecule are measured. This technique is typically used to measure relatively large forces in biological terms (pN-nN) and conformational changes are monitored with sub-nanometer resolution. The study of force-extension curves provides us an exciting opportunity to understand the mechanical and elastic properties of the molecule and hence gained widespread attention to probe a diverse range of biological systems, including proteins, DNA, RNA and their complexes[103–111]. The application of forces in MD simulation is realized by mimicking AFM or optical tweezers experiments. As analogous to mechanical springs, the harmonic potentials are used to manipulate individual atoms, residues or chains. The

combination of MD simulation and experiment enables to identify the key events in the force-induced unfolding of protein and explores the energy landscape of complex systems.

1.10 Motivation and Contribution to the Research

The present thesis is organized into seven chapters. In this chapter we have discussed the theoretical background and present-state of the art of the MSM. In the next chapter, we provide a brief presentation of the computational techniques employed in our studies, namely, molecular dynamics (MD) and kinetic Monte Carlo (KMC) method. The perspective which propels us towards our research goal and the primary contributions to this thesis are discussed briefly in the following .

Despite the substantial advantage of MSM building techniques and their widespread usage, one thing that is often ignored is that MSM is subjected to uncertainty/error due to finiteness in the number of MD trajectories for MSM building; even if they collectively exceed microsecond timescales, there are bound to be rare states and kinetic pathways missing from MD data. In the worst case, missing states can be important to the ensemble-averaged quantities calculated from an "incomplete" MSM. When the missing states and pathways are relevant to the dynamics, they may compromise the MSM accuracy. The lack of a systematic approach to ascertain the completeness of an MSM remains a major bottleneck. So a conceptual framework is required to produce reliable MSM that accounts for missing states/pathways. With this motivation our first study introduces a new concept of validity time of an MSM which quantifies the completeness of an MSM. The validity time of an MSM is the time-scale in which MSM is guaranteed to yield the correct dynamics. Herein, a theoretical framework is proposed to calculate the validity time of an MSM to discard the uncertainty/error within the MSM due to missing kinetic information and identify relevant states/kinetic pathways in the time-scale of interest. The usefulness of the validity time is demonstrated by constructing an MSM of solvated alanine dipeptide. Our first study mainly emphasizes the danger arising from missing relevant states and pathways in an MSM network, and usefulness of validity time of MSMs and other related kinetic network models.

Our second study presents a new class of enhanced kinetic sampling methods for MSM construction along with the concept of validity time. In this work, we

have developed new algorithms, namely, *Swarm* MD, *State-constrained* MD (SC-MD) and *Programmed state-constrained* MD (PSC-MD) method to construct MSMs more efficiently and accurately than ad-hoc seeding of trajectories from different states or ad-hoc pruning of inadequately sampled kinetic pathways. The important contribution in this work is the development of an adaptive method, named by, *Programmed state-constrained* MD (PSC-MD) to accelerate the extension of validity time for directing the simulation in an efficient manner. The developed procedures of constructing MSM of desired validity time are assessed with the help of a prototype network model. Finally, we demonstrate the application of MSMs of a desired validity time to study deca-alanine molecule kept under tension. With the fundamental concept of validity time, we have shown that to predict the dynamics of the system at nanosecond timescale, microsecond long MD data are required and the absence of relevant state/pathways in the MSM can lead to incorrect prediction of the observed kinetic and thermodynamic properties.

The usual analysis of SMFS experiments involves an excessively coarse-grained view. For example, a 1-D reaction coordinate between the "folded" and "unfolded" states invoked in the interpretation of SMFS experiments hides the complexity of a free energy landscape with multiple states and kinetic pathways. Resolving the individual states can avoid complicated force-dependent rate maps arising from incorrectly lumping multiple intermediate states and competing pathways. Unfortunately, the inability to experimentally observe microscopic structural changes in a molecule presents an obstacle towards gaining higher resolution. A potential solution is to fill gaps in our understanding of the experimental force-extension curve with the help of the kinetic models derived bottom-up from molecular simulations. With this context, here our objective is to find out that how the idea of using a stretching force to accelerate rare transitions may be applied to rapidly construct kinetic networks using computer simulations. We have introduced an idea of master-MSM or MSM-0 for constant-probe separation (d) experiments in a Force-Spectroscopy (FS) set up to predict the connection between topologically different kinetic networks constructed at distinct trap separations based on the Bell-Evans-Polanyi (BEP) principle. The method may be adapted to experiments where the probe-separation or the force is varied as a function of time. The time dependence of force is then inherited by the MSM leading to the new concept of time-dependent MSM (TD-MSM). Our proposed TD-MSM model can predict the change in the kinetic network of a single molecule for time-varying anchor separation or pulling force experiment in FS setup. This can open up the possibility of exploring the force induced folding/unfolding dynam-

ics on a complex energy landscape comprising of multiple pathways associated with biomolecules.

Though TD-MSM approach in our previous study enables us to understand the connection between topologically different kinetic networks at various constant probe separations (d), it is not trivial to extract the intrinsic kinetic properties from such experiments. So, we have extended the Master-MSM method to SMFS experiments under constant pulling force (F) for recovering the kinetic rates at zero-force conditions. The similar approach based on BEP principle followed in constant-d ensemble is also adopted in the constant-F ensemble. In our computational study of constant-F experiment we have considered three small systems: peptide, DNA and RNA structures. Here, instead of constant trap separations, the tension is maintained via equal and opposite forces (F) applied to the two ends of the molecule. The programmed state constrained MD (PSC-MD) techniques mentioned previously is used to construct the MSM of desired validity time at various force conditions. Next, we can extract the rates at zero-force conditions by extrapolating the kinetic rate vs force plot based to zero force on the BEP principle and reconstruct the free energy landscape. The reliability of our model is inspected by comparing the intrinsic kinetic rates obtained from the actual simulations at zero-force conditions to that obtained from constant-F experiment.

Finally, I summarize the major conclusions of the work of this thesis and present our future perspectives.

Chapter 2

Overview of Methods

2.1 Molecular Dynamics (MD) Simulations

In general, molecular modeling encompasses several theoretical methods and computational techniques, which are used to model (simulate) the behaviour and features of biomolecules. Among various computational techniques, MD method is the most popular technique applied in the fields of computational physics, chemistry and biology, as well as in bio-informatics and material science to study molecular systems ranging from small chemical compounds to large biological complexes and material assemblies. The basic features of MD simulations is to record the time evolution of the system of N particles interacting via a prescribed potential by integrating Newton's equation of motion,

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = F_i(t) \tag{2.1}$$

where m_i denotes the mass of the i^{th} particle and $\mathbf{F}_i(t)$, the conservative force on i^{th} particle given by the derivative of potential function U is given by,

$$F_i(t) = -\nabla_i \sum_{i=1}^{N} \sum_{j>i}^{N} U(\mathbf{r}_{ij}(t)) \tag{2.2}$$

where $\mathbf{r}_{ij}(t) = |\mathbf{r}_i(t) - \mathbf{r}_j(t)|$.

Molecular dynamics normally employs crystal structure from the Protein Data Bank (PDB) as the starting structure of a multitude of biomolecules and adds velocities and coordinates through a combination of complex algorithms, physical chemistry and physics. The velocity-Verlet algorithm[112] is commonly used in many molec-

ular dynamics software packages to solve N-body problems. In this algorithm, time is discretized into individual "timesteps" and Taylor expansions are used to update the next positions and velocities of each atom at each time t (Frenkel and Smit, 2002[113]):

$$\begin{aligned} \mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\mathbf{F}_i(t)\Delta t^2}{2m_i} \\ \mathbf{v}_i(t+\Delta t) &= \mathbf{v}_i(t) + \left[\frac{\mathbf{F}_i(t+\Delta t) + \mathbf{F}_i(t)}{2m_i}\right]\Delta t \end{aligned} \tag{2.3}$$

where $\mathbf{r}_i(t)$, $\mathbf{v}_i(t)$, and $\mathbf{F}_i(t)$, is the position, velocity and force vectors relevant to atom i at time t respectively, and m_i is the mass. The velocity-Verlet algorithm [Verlet, 1967[112]] is calculated in three consecutive steps; first the new positions are updated, from this the new velocities can be computed, followed by calculation of the new forces. The positions of the particles of the system are evolved in time according to Eq. (2.1). Each timestep provides a snapshot from which the ensemble statistics are averaged and calculated. A simple flow diagram of a standard MD algorithm is shown in Fig. 2.1.

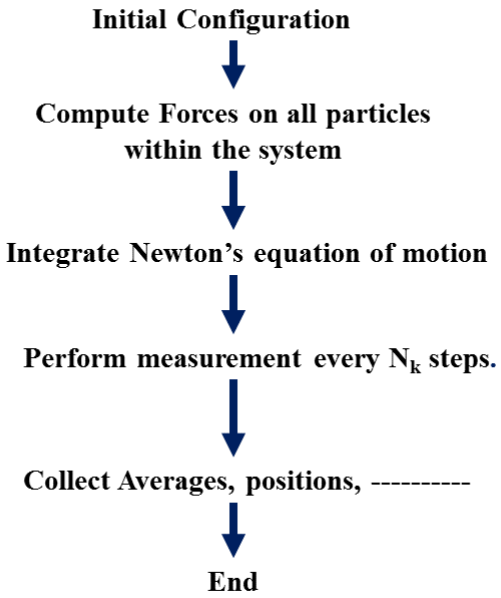

Figure 2.1: A simple diagram of the standard MD algorithm.

We used NAMD[40] molecular dynamics package for all our simulations. This package is a suite of programs that allows users to simulate biomolecule using various MD methods including metadynamics[44], steered molecular dynamics (SMD)[205], Adaptive biasing method (ABF)[49], Umbrella sampling (US)[41] etc.

2.1.1 Connection to Statistical Mechanics

Thermodynamics or equilibrium properties are obtained using statistical mechanics from MD simulation data. The value of an observable Q is calculated as the average $< Q >$ measured over time and space of Q which is produced by different molecular conformations. In order to get the correct value, Q has to be weighted by the probability P of a conformation to occur; integrated over momenta ($\vec{p}$) and positions ($\vec{r}$), namely

$$< Q >= \int \int Q(\vec{p},\vec{r})P(\vec{p},\vec{r})dpdr. \tag{2.4}$$

It is often very difficult to sample the whole conformational space by MD. However, if conformations that are relevant for the average are sampled properly, <Q> can be calculated from this finite set of conformations. The ergodic hypothesis assumes that if the relevant configurations are sampled during a MD simulation that is long enough, then, the phase space average of observables is same as that of time average. Averages are then calculated as

$$< Q >_{MD}= \lim_{\tau \longrightarrow \infty} \frac{1}{\tau} \int_0^\tau Q(t)dt \tag{2.5}$$

where t is the instantaneous time and τ is the total simulation time.

2.1.2 Force Field and the Potential Energy Function

In order to calculate the force felt by each atom, the first step is to compute the potential energy function, U. Although a precise calculation of the potential energy of a N atom system would have to consider the contribution of each individual atom, pair, triplet and so forth, most molecular dynamics programs (including the NAMD package used to perform all simulations reported in this thesis) describe the potential energy using a more simplistic three component picture. In this scheme the potential

has the following basic form:

$$U(\mathbf{r}) = U_{bonded}(\mathbf{r}) + U_{non-bonded}(\mathbf{r}) + U_{special}(\mathbf{r}) \tag{2.6}$$

which is only dependent on the positions of each atom where each atom is represented by a point with position vector $\mathbf{r}$. As we can see the potential energy function of our system consist of three terms. The first term is the bonded term and it considers interaction within a molecule. It comprises of bond stretching, angle bending, torsional-angle rotation, and improper dihedral angle distortion. The latter is responsible for keeping the proper hybridization type for particular groups of atoms. When deviating from an ideal reference value of a length or an angle in these terms, the potential energy rises as,

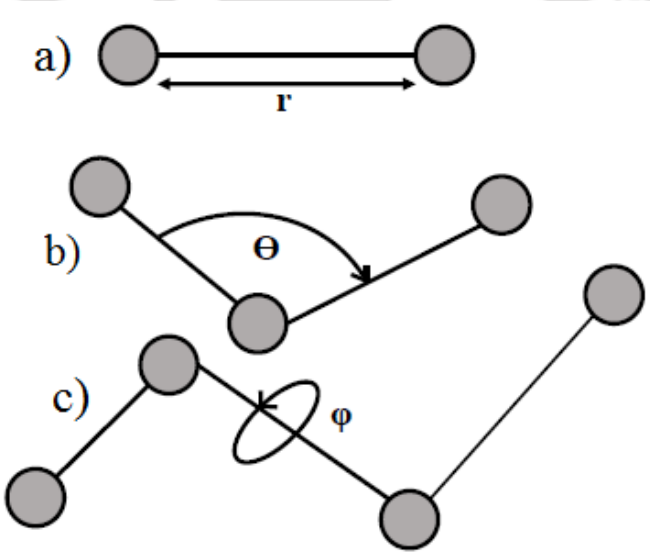

Figure 2.2: Depiction of the coordinates used to describe bonded interactions in the interatomic interaction potential of a MD force field: (a) bond-length (r), (b) angle (θ) and (c) dihedral angle (ϕ) between two atoms connected by three covalent bonds in a molecule.

$$U_{bonded}(\mathbf{r}) = \sum_{bonds} \frac{1}{2}K_b(r - r_0)^2 + \sum_{\substack{bond \\ angles}} \frac{1}{2}K_\theta(\theta - \theta_0)^2 + \sum_{\substack{improper \\ dihedral \\ angles}} \frac{1}{2}K_\xi(\xi - \xi_0)^2 + \sum_{\substack{dihedral \\ angles}} \frac{1}{2}K_\phi[1 + cos(m\phi - \phi_0)]. \tag{2.7}$$

The letter b, θ, ξ, and ϕ represent the bond lengths, bond angles, improper dihedral angles, and torsional dihedral angles. The bond stretching (r), bond bending (θ) and dihedral angle (ϕ) distortion are shown in Fig. 2.2. The variables with a 0 as subscript are the reference or ideal values and they are parameters of the force field. Generally all interactions have a harmonic functional form, except for the torsional dihedral-angle term which has a trigonometric form. The labels K stands for force constants.

Non-bonded energies are calculated using two terms, the first being the Lennard-Jones terms[114] describing the van der Waals interaction[115], and the second one being the Coulomb term[116] dealing with the electrostatic interactions of partial charges of the atoms. The former term describes atomic repulsion due to atom-atom overlap and the attraction due to London dispersion interactions[117]. The non-bonded interaction energy is expressed as,

$$U_{non-bonded}(\mathbf{r}) = \sum_{\substack{atom\\pairs}} \left(\frac{C_{12}}{r^{12}} - \frac{C_6}{r^6}\right) + \sum_{\substack{atom\\pairs}} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1 r}. \tag{2.8}$$

In Eq. (2.8), r is the distance between the two atoms for which the non-bonded energy is calculated, $4\pi\varepsilon_0$ is a constant, and ε_1 denotes the relative dielectric permittivity. C_{12}, C_6, q_i and q_j are the force-field parameters. The constants are taken from standard parameterization schemes such as CHARMM[38] and AMBER[37]. Sometimes potential energy includes the special term ($U_{special}$ in Eq. (2.6)) to ensure proper hydrogen bonding, though this term is superfluous if appropriate van der Walls and Coulombic term are used.

2.1.3 Ensemble Types

An ensemble is a collection of all possible thermodynamic systems, which have diverse microscopic states but the same macroscopic or thermodynamic state. The properties of an ensemble are subjected to specific constraints as listed below:

- **Microcanonical ensemble (NVE):** The thermodynamic state can be described by a fixed number of atoms (N), volume (V), and energy (E).
- **Canonical Ensemble (NVT):** The thermodynamic state is defined with a fixed number of atoms (N), volume (V), and temperature (T).
- **Isobaric-Isothermal Ensemble (NPT)**: This ensemble is characterized by a fixed number of atoms (N), pressure (P), and temperature (T).

- **Grand canonical Ensemble (μVT)**: The thermodynamic state for this ensemble is distinguished by a fixed chemical potential (μ), volume (V), and temperature (T).

2.1.4 Periodic Boundary Conditions for Explicitly Solvated Systems

In order to avoid large time consumption and surface effect reasonably, the simulation must use small sample size. Periodic boundaries conditions (PBC) make this possible for a simulation by using a procedure in which the particles are exposed to forces as if they were in a bulk fluid. As shown in Fig. 2.3, the cubic or non-cubic

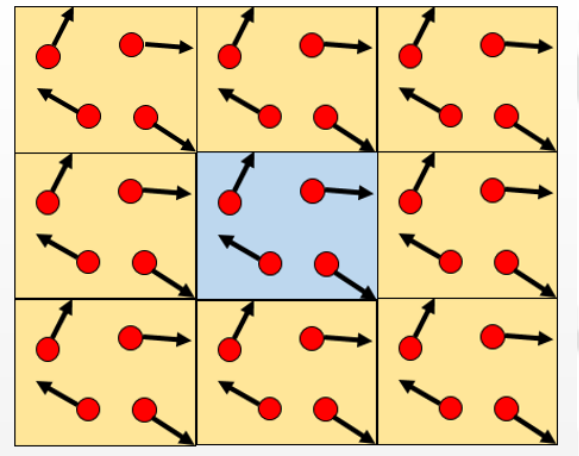

Figure 2.3: Representation of periodic boundary condition (PBC) in two dimension. The central simulation cell is replicated infinitely in each direction.

box is conceptually employed to model the space of an infinite system. During the simulation, the periodic image of each atom moves in exactly the same way each atom moves in the central box, that is, original box. If one atom leaves its box during the simulation, it is substituted with an image particle that comes in from the opposite side. Hence, the topology of the system is correctly represented, while the number of particles inside the center box is conserved. Ensuring that a large enough box is used to minimize interactions between proteins and their images, is generally the most important factor to avoid the "finite" size effect and boundary effect of the simulation box.

2.1.5 Temperature and Pressure Control in Ensembles

In order to control the temperature in NPT/NVT ensembles, several algorithms have been developed, including Berendsen[118], Langevin[119], Andersen[120] and Nosé-Hoover thermostat[121]. We will confine ourselves to a discussion to those implemented in NAMD[40].

Since the velocities of atoms are function of temperature; we can write from the equipartition theorem,

$$\begin{aligned} E =& \frac{3}{2}k_B T = \frac{1}{2}mv^2 \\ T =& \frac{1}{3}\frac{mv^2}{k_B}. \end{aligned} \tag{2.9}$$

Then a common way for controlling temperature is the rescaling of the velocities (up or down) in successive time steps to maintain a constant temperature. Such modification can trigger the system to perform a non-Newtonian behaviour. So the Berendsen thermostat diminishes the effect of velocity scaling by pairing the system to an external heat bath at constant temperature T_{bath} such that the change in temperature is proportional to the difference in temperature between unit cell and heat bath. Then, the rate of change of the temperature can be expressed as

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}(T_{bath} - T(t)) \tag{2.10}$$

where T is the temperature calculated from the simulation, T_{bath} is the temperature of the heat bath and τ_T is the coupling parameter. The Langevin thermostat applies friction and random forces to momenta of particles to overcome the problem as a consequence of "hot solvent and cold solute". In the Langevin thermostat, heat is transferred from heat bath to unit cell via collision between particles of heat bath to those of simulation box instead of a direct heat transfer from heat bath to simulation box. In NAMD, this is implemented via the Langevin equation[122],

$$m_i\frac{d^2\mathbf{r}}{dt} = F_i(\mathbf{r}_i(t)) - \gamma_i\frac{d\mathbf{r}_i(t)}{dt}m_i + R_i \tag{2.11}$$

where a frictional force with coefficient γ_i and a stochastic force, R_i with $<R> = 0$, simulating the thermal noise, are applied to the system. The terms, m_i and F_i represent the mass and force calculated from the potential on the i^{th} particle. The stochastic force $R(t)$ has a Gaussian probability distribution with correlation function

$$< R_i(t)R_j(t') >= 2\gamma k_B T\delta_{ij}\delta(t-t') \tag{2.12}$$

where k_B is the Boltzmann's constant and $\delta(t-t')$ is the Dirac delta function. The choice of γ_i determines whether the frictional or stochastic forces dominate. Clearly, the simulation employing Langevin thermostat for temperature control, is not deterministic due to use of stochastic force. Andersen thermostat assigns the velocity of random particles to new velocity from a Maxwellian distribution. The Nosé-Hoover thermostat applies a thermal reservoir and additional friction expressions to the equations of motion.

The idea of pressure coupling methods (Berendsen barostat[118], Nosé-Hoover barostat[123], Nosé-Hoover Langevin piston method etc.) are very similar to those in temperature coupling thermostats. For controlling the pressure, the dimension of simulation box and atom coordinates are scaled at every time step. The Berendsen barostat employs a pressure bath analogous to the heat bath described above. In fact, the change in instantaneous pressure has a similar form to Eq. (2.10).

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}(P_{bath} - P(t)) \tag{2.13}$$

where $P(t)$ is the instantaneous pressure, P_{bath} the pressure of the bath and τ_P the pressure coupling parameter. The volume of the system is then scaled by a factor μ,

$$\mu = 1 - \kappa\frac{\delta t}{\tau_P}(P - P_{bath}) \tag{2.14}$$

where κ is the isothermal compressibility. The new coordinates, r', are given by

$$\mathbf{r}_i' = \mu^{\frac{1}{3}}\mathbf{r}_i. \tag{2.15}$$

The Langevin piston Nosé-Hoover method is implemented in NAMD by coupling Langevin piston method[124] to Nosé-Hoover method to obtain a constant pressure and constant temperature of the system . This method of pressure coupling is used in MD simulation throughout this thesis. Barostats are key to MD simulations because usually it is not feasible to build simulation system in which the particle density is high enough so that the pressure is close to atmospheric pressure.

2.1.6 Treatment of Short-range and Long-range Interaction

One of the most computationally expensive parts of a molecular dynamics program is the calculation of the non-bonded interaction energies. In a pairwise model, the

computational expenses for this calculation is scaled for a system with N particles as $\mathcal{O}(N^2)$. The Lennard-Jones potential (first term in Eq. (2.8)) is only significant over a very short range and it has very little effect to atoms at distant points. A common way to reduce the computational effort of calculating its effect is to impose a distance cut off beyond which the Lennard-Jones potential is set to zero. But it needs an additional computation of distances between the atoms and comparison with the cut off (introducing $N(N-1)$ calculation). In order to avoid this problem, advantage is taken of the fact that the neighbours of an atom are unlikely to move by large amount of distance over 10-20 timesteps. So, a list of the atoms which fall within the "pairlisidist" (slightly larger than cut off) are generated at the beginning of each cycle and are assumed to contain every atom that passes within the cut off distance during entire cycle, means that distance comparisons need to be calculated much less frequently. Hence, a gain in computational efficiency is achieved.

The truncation of the van der Walls potential energy at cut off distance introduces a discontinuity in potential energy (hence force) at the cut off. In order to prevent problems with energy conservation due to discontinuity in potential energy, most MD codes multiply the real potential by a switching potential factor which goes smoothly to zero at the cut off. This change to the potential is often only introduced a short distance before the cut off.

Unlike the Lennard-Jones contribution, the electrostatic contribution to the potential is significant at long distances as well as short distance owing to slow decay of Coulombic term with distance $\boldsymbol{r}$. So, it requires numerous calculation ($\mathcal{O}(N^2)$) to determine the full-electrostatic potential. In order to diminish the computational workload, Ewald summation (first described in 1921[125]) takes the advantage of infinite effect of periodic boundary condition by splitting the potential into short and long range contributions and has now become a standard tool for macromolecular simulations. In this approach, the long range contribution can be represented as a sum over the Fourier transforms of the potential and the charge density. This sum converges rapidly (as $\mathcal{O}(N^{3/2})$ rather than $\mathcal{O}(N^2)$) and hence can be truncated with little error. However, a significant gain in terms of computational workload is obtained. Further Particle Mesh Method (PME)[126] is implemented to improve the computational performance; scaling as $\mathcal{O}(NlnN)$ and thus permitting the routine calculation of electrostatics without any cut off for periodic system. This method of treatment of electrostatics interaction is employed throughout this thesis.

2.1.7 Constrained Dynamics

The integration timestep used in a simulation is determined by the fastest motions in the system. For a biomolecular system the fastest motions are the vibrations of hydrogen atoms bound to heavy atoms (X-H bond). If one assumes that the X-H bond vibrations do not contribute strongly to the overall dynamics of the system then the lengths of these bonds can be constrained and the integrator allowed to proceed more rapidly. The "SHAKE" algorithm by Ryckaert[127] assumes that the length of the hydrogen bond can be considered constant. In this algorithm, first the unconstrained equation of motion are solved, then the atomic positions are modified. Another analytical variant of SHAKE algorithm, named by "SETTLE", is specially developed for rigid water model to constrain bonds in water molecules[128]. The NAMD code makes use of SETTLE to constrain hydrogen atoms within water molecules and RATTLE[129] for those in all other atoms. This combination allows to extend the timestep to 2 fs, rather than 1 fs as required in unconstrained dynamics.

2.1.8 Single-Molecule Force Probe MD Simulation

Although MD allows the simulations of molecular systems on the nano- to microsecond time scale, many processes in biology require longer time scales. The dynamics can be accelerated by adding biasing potentials or external forces into the calculation. In this thesis, we put special attention on force probe simulations with two pulling protocol, constant velocity and constant force, those are already implemented in NAMD. In force probe simulation, a force is applied to a set of atoms to guide it in a particular direction (or set of directions over time), the technique is known as steered molecular dynamics (SMD).

Constant Velocity

For constant velocity pulling, a harmonic potential V_{spring} is added to the atoms of the pull group,

$$V_{spring} = \frac{k}{2}[\mathbf{r}(t) - \mathbf{r}_{spring}(t)]^2 \tag{2.16}$$

where k is the spring constant, $\mathbf{r}$(t) is the position of the center of mass (COM) of the pulled atoms and $\mathbf{r}_{spring}$ is the position of the spring. The force is given by,

$$F(t) = k[\mathbf{r}(t) - \mathbf{r}_{spring}(t)] \tag{2.17}$$

and is applied by moving the spring position $\mathbf{r}_{spring}(t)$ with constant velocity $\mathbf{v}$ along

a chosen reaction coordinate

$$\mathbf{r}_{spring}(t) = \mathbf{r}_{spring}(t=0) + \mathbf{v}t. \tag{2.18}$$

For weak spring, the loading rate $|\dot{\mathbf{F}}| = \frac{dV_{spring}}{dt} = kv$ is directly proportional to pulling velocity v. The pulling force $\mathbf{F}_{spring}$ on the COM of a group of atoms is redistributed among masses of the atoms weighted to the individual atoms.

Constant Force

In constant force mode, either a linear potential is applied to tilt the energy landscape in the direction of the reaction coordinates or the force is added directly to the calculated forces for each atom of the pull group. The resulting time-independent tilt of the energy landscape permits a direct insight into the kinetics of the probed system by looking at the unfolding of the system.

2.1.9 Energy Minimization Schemes

The initial conformation of the biomolecules is taken from a published structure for the protein or nucleic acid from which the rest of the system (solvent, ion, lipid) is constructed. The initial geometry of the system does not necessarily correspond to one of the stable/lowest energy conformation. Therefore, it is essential to bring the system in its stable conformation by carrying out an energy minimization before starting the actual simulation. Energy minimization is a numerical procedure of finding the energy minima on the potential energy surface (PES) starting from higher energy state. During an energy minimization, the geometry of the system is changed in a stepwise fashion such that the energy of the system is reduced as shown in Fig. 2.4. After a number of steps, the global or local energy minima on the PES is reached. In general, the energy-minimization is performed by gradient optimization. The most popular energy minimization techniques are: Newton-Raphson Method, Steepest Descent Method and Conjugate Gradient Method.

Newton-Raphson Method:

Newton-Rapshon method is based on the Taylor expansion of the PES at the current geometry. According to this method, the geometry (x) is updated in every step as

$$x_{new} = x_{old} - \frac{E'(x_{old})}{E''x_{old}}. \tag{2.19}$$

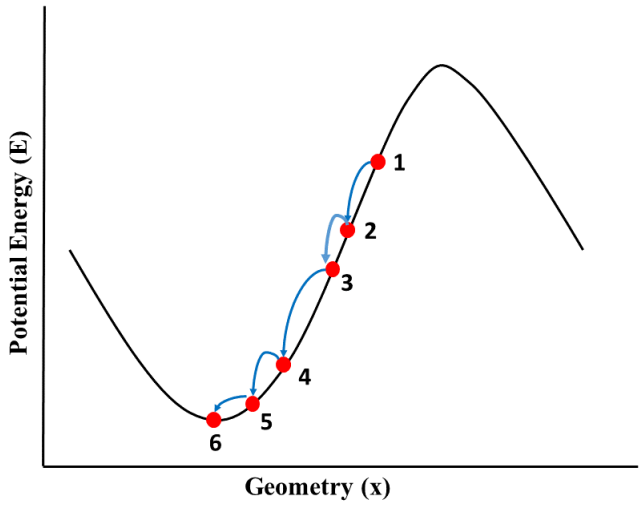

Figure 2.4: Sliding of the molecule towards the energy minima in the process of energy minimization in a stepwise fashion.

The method is computationally expensive as it requires the calculation of first and second derivative of geometry at each point. Yet, usually it takes fewer steps to reach the energy minima.

Steepest Descent Method: The Steepest Descent method does not require the calculation of second derivative, it relies on the approximation that second derivative is constant. Therefore, the equation to update the geometry becomes

$$x_{new} = x_{old} - \gamma E^{'}(x_{old}), \tag{2.20}$$

where γ is a constant. In this method, first the geometry minimization takes place in the direction opposite to the largest (*i.e.* steepest) gradient at the initial point. Once a minimum in first direction is achieved, next minimization is carried out starting from that point and moving in the remaining steepest directions. This process is continued until a minimum has been reached in all directions within a sufficient tolerance. This method is much faster at each step than the Newton-Raphson method but due to the approximation, it needs more steps to find the energy minima. In addition, this method is not very efficient in many dimensions because it easily winds up in a zig-zag pattern which does not move towards the minimum efficiently.

Conjugate Gradient Method: In the Conjugate Gradient method, the first part of the search takes place in the direction of the largest gradient, just as in the

Steepest Descent scheme. However, to avoid some of the oscillating back and forth that often plagues the steepest descent method as it moves toward the minimum, the conjugate gradient method mixes a little of the previous direction in the next search. This allows the method to move rapidly to the minimum as illustrated in Fig. 2.5 for a system with two geometrical coordinates. The equations for the con-

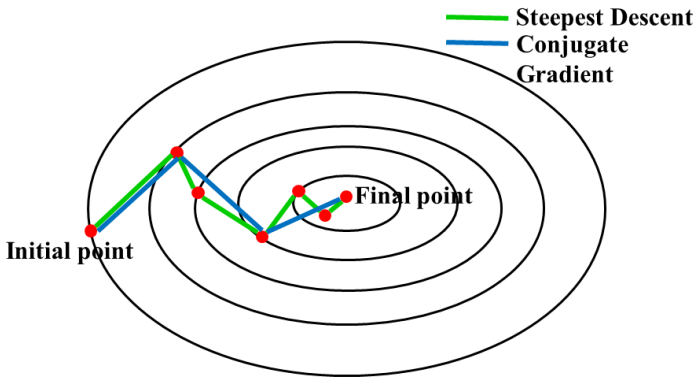

Figure 2.5: The counter plot of a function, with the steps of the steepest descent method in green and of the conjugate gradient method in blue.

jugate gradient method in two or more dimension are more complex than those of the other two methods.

2.1.10 All-atom vs Coarse-Grained Models

The conventional all-atom molecular modeling of biomolecule's structure and dynamics remains difficult for many complex systems mostly due to larger system size and long timescales. Lowering the resolution from all-atom to coarse graining (CG) of biomolecules opens up the possibility of exploring the dynamics of the system at longer timescales. The term "coarse-graining" is used to refer to the simplified model of molecules by grouping several atoms of it into one component, thus reducing the degrees of freedom of system and allowing the simulation to be run for relatively long timescales. The CG model is desirable as it is more manageable computationally in simulations and more amenable to theoretical treatment. In the past few years, CG models have been used successfully for studying the protein folding mechanism and protein-protein interaction, prediction of protein structure, modeling of complex dynamic processes and modeling of membrane protein. The various coarse-graining models and their applications in biomolecular systems are discussed in Ref. 130.

Most popular coarse-grained models are residue-based coarse graining (RBCG) and structure based coarse graining (SBCG). In RBCG model of protein or lipid, several atoms are grouped onto a single "bead". The bead is placed at center of mass of atomic group which defining it and the different bead interact with each other through an effective potential. The SBCG model is used to design the large macromolcule assemblies. In SBCG model, the shapes of the large molecules consisting of thousands of atoms are represented with a small number of beads. Furthermore, the arrangement of the bead is tuned to reproduce the shape of the molecule such as that are available from a X-ray diffraction or NMR data. Then, the interactions between beads are parameterized from all-atom simulations of bead components. The RBSG model has been widely used to lipid-protein system involving large-scale conformational changes while the SBCG model to polymeric systems.

As stated earlier coarse-grained simulations are particularly useful for large systems on long timescales. This kind of simulation gives a semi-quantitative picture of how a large system behaves over time. But in many cases, the important degrees of freedom may be eliminated due to coarse graining, thus may lose important kinetic information which is the main disadvantage of coarse-grained molecular dynamics (MD) simulations.

2.2 Monte Carlo

Monte Carlo (MC) method is a most popular numerical statistical method based upon the repeated sampling of random elements to compute the result. From its definition it is clearly understood that it doesn't give us exact results such as "real" mathemathical analysis might, relies on the game of chance or probability theory. The general approach of MC is to create an experiment with a random element and to perform the experiment repeatedly to estimate the result. Typically, MC methods are used for simulation of complex physical experiment but for also for optimization models in finance. Furthermore, they are suitable for solving analytical problems such as high-dimensional or particular types of different equations with complex boundary condition. In the context of biomolecular simulation, they provide a stochastic approach to explore the molecular level configurational space available to a system at equilibrium. Unlike MD, they are not representative of the true dynamics of the system, but allow the calculation of thermodynamic properties of the system.

The underlying concept is to take a three-dimensional protein structure and use this as the starting point for a random walk in conformational space. At each step along this walk the probability of a given change in conformation is dependent on the change in energy required from the previous state. In order to ensure thermodynamically correct sampling the probability of visiting a particular state $\mathbf{r}$ is proportional to the Boltzmann factor $\rho(\mathbf{r}^N) = e^{-\beta U(\mathbf{r}^N)}/k_B T$. The generation of configurations according to a distribution is called importance sampling. The most commonly used method for achieving the importance sampling is Metropolis Monte Carlo method using Markov Chain process that is discussed below.

2.2.1 Markov Chain

A Markov chain describes a stochastic process in which the state of a system (here, the instantaneous atomic configuration) changes randomly with time and has no memory of previous states. At each step in time, we apply some new randomness to determine the next step in a way that is a function of current time. An important property of Markov processes is that it doesn't matter how the system arrived to its current state, the probability that it will move to any other state at the next time step only depends on where it currently is. In the context of a Metropolis Monte Carlo simulation, the basic procedure to generate a Markov chain are:
(*i*) At some step i, the system has an atomic configuration (state A).
(*ii*) The configuration is randomly perturbed to generate a new atomic configuration (state B). A typical perturbation might be a single-particle displacement. An atom is randomly picked and displaced it by small random amount in x, y and z directions. In general, these perturbations are termed Monte Carlo moves.
(*iii*) The new configuration will be accepted or rejected by choosing an acceptance criterion in such a way that the long-time trajectory correctly generates the configurations according to canonical probability $\rho(\mathbf{r}^N)$ distribution.
(*iv*) The trajectory includes the states of the system after each acceptance or rejection.

In order for this approach to work, we need a way to decide how to accept or reject the proposed configurations in our simulations. The most commonly used method for choosing the next state is the Metropolis algorithm. This algorithm uses random work in the phase space with transition probability to go from state m to state n equal to 1 if the move is downhill in energy ($\Delta U_{nm} < 0$), if the move is uphill in energy ($\Delta U_{nm} > 0$) then the move is accepted with a probability defined by the ratio of probabilities of initial and final states, that is, mathematically can

be written as

$$P(\mathbf{r}'|\mathbf{r}) = min\left(1, e^{-\frac{1}{k_B T}(U(\mathbf{r}')-U(\mathbf{r}))}\right). \tag{2.21}$$

where k_B is the Boltzman constant and T is the temperature.

2.2.2 Detailed Balance Condition

Now imagine the system has evolved for some time randomly according to Markov chain. Suppose $P(m,t)$ is the probability of being in configuration m at time t, $P(n,t)$ the probability of being in configuration n at time t, $w(m \to n, t)$ is the probability of going from state m to state n per unit time (transition probability). Then we have according to Chapman-Kolmogorov (C-K) equation

$$P(m, t+1) = p(m,t) + \sum_n [w(n \to m)P(n,t) - w(m \to n)P(m,t)]. \tag{2.22}$$

In the limit $\Delta t \longrightarrow 0$ (continuous time), the C-K Equation (Eq. (2.22)) can be expressed as

$$\frac{dP(m,t)}{dt} = \sum_n [w(n \to m)P(n,t) - w(m \to n)P(m,t)], \tag{2.23}$$

known as Master equation in continuous time space. At long time limit we want P(m,t) to be p(m). Clearly, a sufficient condition for equilibrium (time independent) known as the so-called detailed balance condition, is given as

$$w(n \to m)P(n,t) = w(m \to n)P(m,t). \tag{2.24}$$

The metropolis acceptance criterion satisfies the detailed balance equation. This can be applied to any probability distribution, but if we choose the Boltzmann distribution we have

$$\frac{w(m \to n)}{w(n \to m)} = \frac{P(n)}{P(m)} = \frac{\frac{1}{Z}\exp(-\frac{U_n}{k_B T})}{\frac{1}{Z}\exp(-\frac{U_m}{k_B T})} = \exp(-\frac{U_{nm}}{k_B T}) \tag{2.25}$$

The partition function is not involved in the above expression, it only involves the known quantity temperature (T). So we can easily calculate the potential energy (ΔU) difference between two conformational states from the above expression.

There is another kind of Monte Carlo algorithm, Kinetic Monte Carlo, (KMC)[52], to address the kinetics of the evolving system, which has been discussed in next

section. In general, the difficulty in efficiently choosing conformational steps for the random walk results in high rejection rates and also slow convergence of $\frac{1}{\sqrt{N}}$ (where N is the number of samples) making Monte Carlo simulations less attractive for biomolecular systems than molecular dynamics.

2.2.3 Kinetic Monte Carlo Method

With MD we can only reproduce the dynamics of the system upto several microseconds. Slow kinetic processes cannot be modeled. Metropolis MC samples configurational space and generates configurations according to the desired statistical-mechanics distribution but can not be used to study the dynamics of system. Kinetic Monte Carlo (KMC) method is an alternative computational technique that can be used to study kinetics of slow processes. In Metropolis MC method, we decide whether to accept a move by considering the potential energy difference between two states whereas in KMC method, we use kinetic rate that depend on the energy barrier between the states. The basic idea behind KMC is to use increment of time which are defined by the transition rates of all processes and formulated such a way that they relate to the microscopic kinetics of the system.

The main trick involved in the KMC simulation is described in the following. As transition between two states for a Markov process depends on only the kinetic rate (or free energy barrier) between them, a stochastic process can be designed to propagate the system correctly from state to state. In that case, the probability of observing a sequence of states and escape times in KMC simulation is same as that obtained from MD data. Hence, the resultant KMC trajectory will be indistinguishable from a long MD trajectory. The steps of KMC algorithm are:
(i) Suppose the system is in initial state i.
(ii) Generate an exponential distribution of time (t_j) for an escape pathway to state j. The actual process takes place along the pathway.
(iii) Select the escape pathway for which t_j is minimum or the escape time of fastest transition.
(iv) Advance the overall clock time by t_j^{min} and discard the other times drawn from exponential distribution.
(iv) Move the system to new state j^{min}.
(v) New simulation begins from new state j^{min} and process repeats.
In this way the KMC algorithm involves cataloguing of all the possible kinetic events, and calculating the escape rates associated with these processes.

Therefore, the couping of KMC method with MD method can be utilized as an accelerated MD method to study the long-timescale dynamics of large biomolecular system.

Chapter 3

Validity Time of a Markov State Model

3.1 Introduction

Markov state models (MSMs)[61–65] and other related kinetic network models are frequently used to study the long-timescale dynamical behavior of biomolecular and material systems. MSMs are detailed kinetic network models wherein the configurational space of a biomolecule under study is partitioned into states. The dynamical evolution of the system is approximated in terms of state-to-state transitions. The number of states can range between tens to thousands depending on the complexity and level of coarse-graining. Each node in the network denotes a metastable state of the system while the connections between the nodes provide rates of interconversion between the states. MSMs have become useful tools for probing the dynamics of nucleic acids and proteins[131–134], for example, folding and unfolding events at long time scales. Though we restrict ourselves to biomolecular systems, MSMs are closely related to kinetic Monte Carlo (KMC) models[135–137] used in the materials and reactions areas for studying catalysis[138], crystal growth[139], material processing[140], and adsorption phenomena[141] to name a few. Both approaches solve a master equation and have benefited from the exchange of ideas between the respective communities. For instance, knowledge of the network structure can be exploited to accelerate the KMC dynamics by eliminating fast degrees of freedom[142–145]. Despite their widespread usage, some aspects of MSM construction are still poorly understood.

A key step in the MSM construction entails determining states and kinetic pathways to be included in the model. The availability of a large number of parallel processors has enabled rapid construction of high fidelity MSMs using brute-force

molecular dynamics (MD) calculations[146–148]. Herein states and kinetic pathways are identified via coarse-graining several independent MD trajectories. The MD trajectories can be seeded from different starting configurations, which allows for better sampling of the configurational space. Other simulation techniques offering resolution greater than the MSMs can also be employed[149–151]. Enhanced thermodynamic-sampling techniques that can sample rare events with large energy barriers can aid in the efficient construction of the model[152–156]. However, often overlooked is an additional challenge associated with building MSMs (and indeed with KMC models as well[157]); namely, a fundamental limitation remains that the entire configurational space cannot be sampled by a finite number of MD trajectories, that is, a MSM is never complete. Even when the MD trajectories used for network-building collectively exceed microsecond time scales, there are bound to be rare states and pathways missing from the MD data. When relevant states and pathways are missing from the constructed MSM, thermodynamic/kinetic quantities being sought can be inaccurate. The main purpose of this chapter to highlight the danger arising from missing relevant states and pathways in a network, develop a strategy to quantify the *completeness* of a kinetic network model, and identify regions of configuration space relevant to the dynamical evolution, which can guide further network construction.

Many network-building procedures entail pruning/lumping of states and kinetic pathways to enforce detailed balance and avoid absorbing states. Although the length of the MD trajectory used to build a network is generally reported, it is not enough to establish the maximum duration for which the dynamics is faithfully predicted by the network model. In the worst case, missing states can be important to the ensemble-averaged quantities calculated from an "incomplete" network model. Given that network models are nowadays generated by seeding the MD calculations starting from different states while using a variety of computational tricks, comparing the dynamics from models for the same system is subject to error/uncertainty resulting from the missing kinetic information. A conceptual framework that accounts for missing states/pathways will bolster endeavors to generate reliable network models.

Estimators for missing rates from a state first developed in Refs. 158 and 159 have been applied to different material systems[157,160–162]. However, more than the missing rates, conceptually, it is the time scale where the missing pathways become relevant to the dynamics that is of interest. The largest time scale where the network model continues to yield the correct dynamics, termed as the validity time for the model, is introduced here. The validity time allows one to systematically compare the behavior

of two models of the same system while accounting for known kinetic rates, topology, and relaxation times of the network, as well as missing pathways and states that have not been included in the model. The main idea is contingent on identifying states that may have a large probability flux into configurational space that is not part of the existing model. One can then compute the validity time where the error in the dynamics is small, that is, the existing network model can be regarded as complete till its validity time and is safe to use. The theoretical underpinning of the validity time is discussed in the present chapter. The usefulness of the validity time is illustrated in Sec. 3.4 by constructing MSM of Alanine dipeptide that has been studied extensively in the past[77,161–163,186–190].

3.2 MSM Methodology: Construction, Validation and Error

The goal of MSM building is to reproduce the long-timescale conformational dynamics of a biomolecular system from the ensemble of relatively short molecular dynamics trajectories by solving Master equation. In MSM paradigm, the stochastic conformational dynamics of a biomolecular system is modeled as discrete-time Markov chain[164] or a continuous-time master equation model with coarse-grained time[165] with the key assumptions that dynamics are Markovian (memoryless), ergodic and reversible with respect to unique stationary distribution. Any model is said to be Markovian if the system reaches local equilibrium within a state before attempting to exit to another state. The construction of an MSM boils down to two key steps: i) discretization of continuous state-space to discrete Markov states. ii) computation of the transition probability matrix, $\boldsymbol{P}$ or rate matrix, $\boldsymbol{T}$ of state-to-state transitions. The transition probability, $P_{ij}(\tau)$ between two state i and j is defined as the conditional probability of transitioning from state i to j in an interval (time resolution) called the lag time, τ of the model whereas the element of rate matrix, k_{ij} is the inverse of time elapsed in state i before escaping to state j. In discrete-time Markov model formalism, the time-evolution of the system is governed by the Chapman-Kolmogorov equation; written as,

$$\boldsymbol{\pi}(t = n\tau) = \boldsymbol{P}(n\tau)\pi(t = 0) = [\boldsymbol{P}(\tau)]^n\boldsymbol{\pi}(t = 0) \tag{3.1}$$

where $\boldsymbol{\pi}$ is the vector of probabilities of occupying any of the Markov states at time t . As the system reaches the global equilibrium, the detailed balance condition will

be satisfied,

$$P_{ji}\pi_{eq,i} = P_{ij}\pi_{eq,j}. \tag{3.2}$$

where $\pi_{eq,i}$ denote the equilibrium probability of state i. Alternatively, the continuous-time master equation is written as

$$\frac{d\pi(t)}{dt} = T\pi(t) \tag{3.3}$$

and the evolution equation has the formal solution of the form,

$$\pi(t) = \exp(-Tt)\pi(0) \tag{3.4}$$

which describes the time evolution of the system among the discrete state in a continuous time. Once the transition probability matrix or rate matrix is constructed, the eigen spectrum gives us the important kinetic and thermodynamic information. Eigenvalues are related to the relaxation timescales of the dynamical processes and the corresponding eigenvectors denotes the associated structural changes. Time dependence of ensemble averaged quantities can be obtained by solving the master equation. A large body of work has focused on the discrete time MSM formalisms where the model is characterized by transition probability matrix.

The specific challenges to build a MSM are: i) what should be the correct state definition to decompose the conformational space in a kinetically meaningful scheme. ii) what should be the state decomposition algorithm to construct a transition matrix or rate matrix in an efficient manner. The state decomposition algorithm should produce a decomposition for which dynamics appear to be Markovian at lag time τ. The common way for partitioning the conformational space into metastable states, followed by most popular MSM building procedures are first by dividing the MD data into an appropriate set of discrete states based on their structural similarity and then lumping the kinetically similar microstates to macrostates. The discrete partitioning of the state space may lead to a deviation of Markov model dynamics from true dynamics, known as discretization or systematic error. Also, due to limited quantity of trajectory data there is always a statistical error in the estimation of each element of transition matrix. An important aspect is to balance the statistical and the systematic error of MSMs; while a fine partitioning will minimize discretization error, the fine partitioning with the limited quantity of trajectory data increases the statistical error. An optimization of both type of errors in MSMs construction may be possible by an approach which is itself adaptive.

Now, the MSM constructed from a given set of MD trajectories depends sensitively on how it was constructed. The lag time, τ is a crucial parameter to access the accuracy of the Markov Model. Therefore, it is necessary step to choose a lag time and validate the model against the simulation data. In conventional MD techniques, the value of τ is picked such that the implied timescale plot is independent on τ. The implied timescales (relaxation timescales of a model) can be computed from the eigenvalues of transition probability matrix (P) as,

$$t_i = -\frac{\tau}{\ln\lambda_i} \tag{3.5}$$

where t_i is the relaxation time, τ is the lag time, and λ_i is i^{th} largest eigenvalue estimated at τ. For a correct choice of the lag time, the implied timescales are plotted as a function of increasing τ. At sufficient large value of τ, the implied timescale will be independent on τ, implying that exponential of transition matrix will be identical to that of transition matrix constructed at longer time interval of observation. The shortest time interval of observation for which the implied timescales start to be constant with τ is chosen as appropriate lag time, that can be correlated as longest internal equilibrium time, τ_{int} of any discrete state and the model should be Markovian for all lag time $\tau >> \tau_{int}$. Once a lag time is selected, the further validation can be done by comparing the dynamics to raw simulation data whether the Chapman-Kolmogorov equality (Eq. (3.1)) holds with statistical uncertainty.

For the rest of thesis we use continuous time Markov Models instead of discrete-time MSM formalisms. Note that, in this chapter we are discussing about the uncertainty due to missing states /pathways which is different from recent considerations of error/uncertainty in MSMs. Also, most MSM-building procedures entail pruning/lumping of states and kinetic pathways to enforce detailed balance and avoid sink states in the MSM to achieve ergodicity. In our approach of MSM construction, the detailed balance condition is not imposed and also, sink states are taken into account. Here we have not carried out the conventional tests such as the implied timescale plots used in discrete time MSM formalisms. Instead, we have checked if the system obeys first order kinetics from the distribution of the escape times from the various states. More details about first order kinetics are discussed in Sec. 4.2.2 of the next chapter.

3.3 Validity Time for a Markov State Model

Consider a MD trajectory of a duration τ_{MD}. Analysis of the trajectory using a combination of distance metrics[166–171] and clustering methods,[64,172] and tests for the Markovian approximation (e.g., implied time scales for discrete-time MSMs[148] and tests for first-order behavior for continuous time MSMs[161]) can yield information about the states of the system and the associated kinetic rates. Although states are randomly visited in the trajectory, the occupation $\pi_S(t)$ for a Markov state S at time t is deterministic and is given by the master equation,

$$\frac{d}{dt}\pi_S(t) = \sum_{S' \neq S} k_{S' \to S}\pi_{S'}(t) - \sum_{S' \neq S} k_{S \to S'}\pi_S(t). \tag{3.6}$$

Here, $k_{S \to S'}$ is the kinetic rate from S to state S', and $k_{S' \to S}\pi_{S'}$ and $k_{S \to S'}\pi_S$ denote the inflow and outflow probability flux for S. Kinetic rates can be obtained using a statistical approach, such as maximum likelihood estimation (MLE)[173]. While Eq. (3.6) forms the basis of a MSM, the MSM constructed using finite MD trajectories is approximate because of various errors[174,175], statistical uncertainty[176,177], as well as missing kinetic information[158,159]. Here, we shall focus on error from the missing kinetic information. Equation (3.6) can be written as the continuous-time MSM,

$$\frac{d\pi}{dt} = T\pi, \tag{3.7}$$

where T is the rate matrix. Equations (3.6) and (3.7) are solved with a specified initial distribution and rate matrix.

The number of states and pathways in the MSM can increase when additional MD data are made available. Pathways with large (small) probability flux are more (less) likely to be selected in the dynamics. Due to a number of factors including randomness inherent in sampling, time scales accessed, the network topology, and the starting conformation, it is possible that certain pathways that have a reasonably large probability flux are still missing in the MD trajectory. As a consequence, states that can be reached only via the missing pathways are also missing in the MSM. It is well known that topologically different MSMs are often generated for the same system when MD calculations are seeded from different starting conformations that are separated by large free energy barriers (even when enhanced conformation sampling techniques are used[152,163,178–182]). The MD time spent in each state is a key parameter that affects the MSM accuracy.

It is convenient to define a time τ_V termed as the MSM validity time such that all kinetic pathways that are likely to be selected within τ_V are present in the existing MSM. Pathways that are less relevant can be missing from the MSM without affecting the accuracy of the kinetic model. MD data pertaining to less relevant pathways do not result in appreciable increase in τ_V. Next, we relate τ_V to the MD time.

3.3.1 Core, Periphery and Missing States

States in Eq. (3.6) can be partitioned into three types: core, periphery, and missing states as shown in Fig. 3.1. A state where the system has resided for a significant time in the MD calculation is termed a core state. The probability flux out of the core states can be estimated from the MD data, that is, they constitute the *source* terms in Eq. (3.6). For the remainder of our discussion, a MSM is the core network model, and we use these terms interchangeably. The MSM is ergodic since

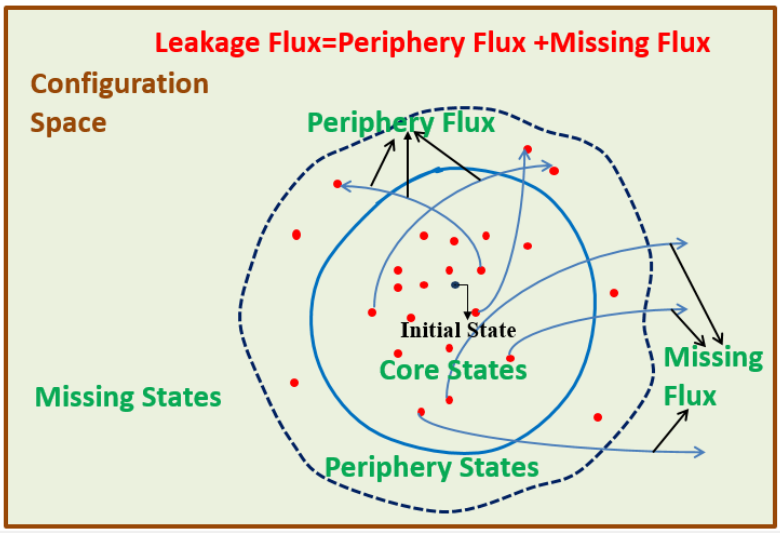

Figure 3.1: Partitioning of whole conformational states into three types: core, periphery, and missing states. The retangular diagram represents the full configuration space. The black dot within inner circle indicates the initial state for the simulation and red dots are the states generated through simulation. Each red dot enclosed by the inner circle represents the core state and periphery states are denoted by red dots within the region between inner and outer circle. The states outside the outer circle are the missing states which are actually not found in the MD simulation. The pathways between the states are shown by arrow lines. The sum of the flux from core states to periphery and missing states constitute the total leakage flux from core region.

it is possible to reach a core state from any part of the network. States that are visited for a short time in the MD calculation preventing estimation of kinetic rates from such states with reasonable confidence are termed as periphery states. Note that the rates from core states to periphery states might be available. The need for periphery states will become clear in Sec. 4.4 of chapter 4. These states are redundant at the MSM validity times. Periphery states correspond to *absorbing* or *sink* states in Eq. (3.6). The dynamics of periphery states can be quite different from the one predicted by the MSM at longer times. We term a network model comprising of core and periphery states as a full network model. The validity time of a MSM can be increased by performing additional MD in the core and periphery states. A periphery state becomes a core state when sufficient time has been spent in the state so that one/more kinetic rates can be estimated. States that have never been visited in the MD trajectory are termed as missing states. Some missing states later become periphery or core states as additional MD is performed. Estimates for the rates from the core states to the missing states are given in Sec. 3.3.2.

Figure 3.2 shows the structure of the rate matrix in Eq. (3.7) constructed with MD where core, periphery, and missing states are considered. Each off-diagonal

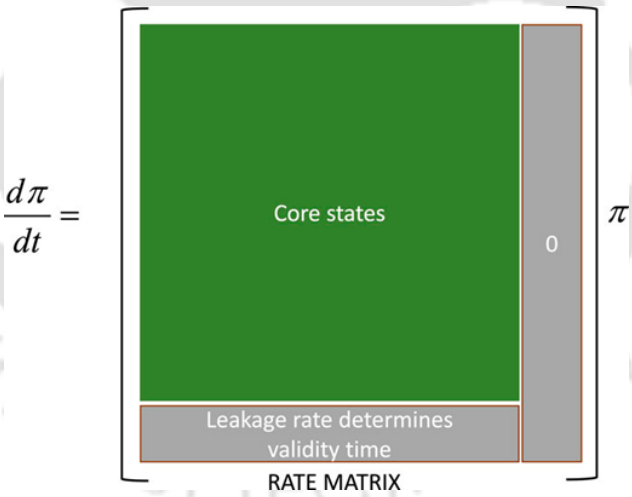

Figure 3.2: Structure of a rate matrix in the master equation [Eq. (3.7)] when independent MD trajectories are used to build a Markov state model (MSM). $\pi(t)$ denotes the occupation column vector.

term in the row (column) j in the matrix denotes the kinetic rates into (from) the state j. Core (periphery and missing) states are placed at the top (bottom) of the occupation vector on the right-hand side of Eq. (3.7). The matrix can be divided into three parts. The top-left (green) corner involves the core states, that is, it forms the rate matrix for the MSM. We lump the periphery and missing states together

as absorbing states. The last column of the matrix contains the rates from the absorbing states. The rates are set to zero. The bottom row of the matrix contains estimates of rates (termed as *leakage rates*) from the core states to the absorbing states. As we shall show next, the validity time of the MSM is determined by the leakage rates, which in turn depends on the time spent in the core states. The validity time for the core network can be made large when all leakage rates are kept small.

3.3.2 Upper Bound for the Missing Rate from a Core State

Consider a collection of pathways from a state S with total rate k. Assuming first order kinetics, the probability of not selecting these paths in time t_S spent in S is exp(-kt_S). The likelihood that the rate equals - lnδ/t_S given that the probability of these paths are not selected in MD is δ. All values of k that result in likelihood greater than δ satisfy $k <$ - lnδ/t_S. Thus, (1 - δ) can be regarded as a confidence associated with the estimate for k. An upper bound for the missing rates for a core state S is given by

$$k_S^{max} = \frac{\ln(1/\delta)}{t_S}. \tag{3.8}$$

Note that the entire MD duration $\tau_{MD} = \sum_S t_S$. Similarly, an upper bound for the missing flux consistent with the MD data is $F_S^{max}(t) = k_S^{max}\pi_S(t)$.

3.3.3 Leakage Flux from Core Network

For a core state S, Eq. (3.6) can be rewritten in terms of the core (C), periphery (P), and missing (M) states as,

$$\frac{d\pi_S}{dt} = \sum_{S' \in C} (k_{S' \to S}\pi_{S'} - k_{S \to S'}\pi_S) - F_S. \tag{3.9}$$

where the leakage flux is,

$$F_S = \sum_{S' \in P \cup M} (k_{S \to S'}\pi_S - k_{S' \to S}\pi_{S'}). \tag{3.10}$$

In the worst-case scenario, using Eq. (3.8) to replace F_S with an upper bound, we define the maximum leakage flux from S as

$$F_S^{leak} = k_S^{leak}\hat{\pi}_S = \hat{\pi}_S \left(\frac{\ln(1/\delta)}{t_S} + \sum_{S' \in P} k_{S \to S'} \right). \tag{3.11}$$

The caret denotes maximum leakage into the missing/periphery states. State occupations are obtained by solving,

$$\frac{d\hat{\pi}_S}{dt} = \sum_{S' \in C} (k_{S' \to S} \hat{\pi}_{S'} - k_{S \to S'} \hat{\pi}_S) - F_S^{leak}. \tag{3.12}$$

The leakage flux can be ignored when $\left| \sum_{S' \in C} (k_{S' \to S} \hat{\pi}_{S'}(t) - k_{S \to S'} \hat{\pi}_S(t)) \right| \gg F_S^{leak}(t)$. Once a state with large leakage is detected, one can analyze the source of leakage. Large leakage due to k_S^{max} implies additional MD in state S would be beneficial. The other possibility is that one or more periphery states have become important to the dynamics. By performing additional MD, a periphery state S' is converted into a core state and its contribution to the leakage flux is eliminated. The flux from the core to periphery states is determined by the network topology. Topologies where the core states are connected to a large number of periphery states generally have a short validity time. The missing flux from state S can be made small by performing MD calculations in S, which extends the time t_S.

The occupation, $\tilde{\pi}_S(t)$ is obtained by solving the MSM,

$$\frac{d\tilde{\pi}_S}{dt} = \sum_{S' \in C} (k_{S' \to S} \tilde{\pi}_{S'} - k_{S \to S'} \tilde{\pi}_S). \tag{3.13}$$

Detailed balance is not assumed, that is, presence of both forward and backward pathways is not required. Dynamics from the core network [Eq. (3.13)] and full network [Eq. (3.12)] models diverge beyond the validity time. Consider a special case where the stationary distribution $\tilde{\pi}_S^{st}$ is attained and leakage in the full network model becomes significant at time scale τ beyond the relaxation timescales. We write

$$\frac{d\hat{\pi}_S}{d\tau} = -k_S^{leak} \hat{\pi}_S; \;\; \hat{\pi}_S(0) = \tilde{\pi}_S^{st}. \tag{3.14}$$

Since the core network dynamics is fast, the ratio of occupations for two core states is fixed, *i.e.*, $\hat{\pi}_S(\tau) = \hat{\pi}_S(0) f(\tau)$ with $f(\tau)$ denoting the probability of residing in the core states at time τ. Equation (3.14) is rewritten as,

$$\frac{df(\tau)}{d\tau} = -k^{leak} f(\tau); \;\; f(0) = 1, \tag{3.15}$$

and

$$k^{leak} = \sum_S k_S^{leak} \tilde{\pi}_S^{st}. \tag{3.16}$$

Here, k^{leak} denotes the total leakage rate for the network. The occupation for a core state S is

$$\hat{\pi}_S(t) = \tilde{\pi}_S^{st} \exp(-k^{leak} t). \tag{3.17}$$

In general, Eqs. (3.12) and (3.13) need to be solved simultaneously to determine the validity time for the core network. For convenience, we can approximate the validity time scale τ_V in terms of the time constant for leakage, namely,

$$\tau_V = \frac{1}{k^{leak}}. \tag{3.18}$$

According to Eq. (3.17), the core state occupations decrease by a factor of exp(-1) at these time scales. The advantage of Eq. (3.18) is that only the core network model needs to be solved to calculate the validity time using Eq. (3.16). The assumption in Eqs. (3.16)–(3.18) that the relaxation time scale within the core network is smaller than the time scale associated with probability leakage is violated when the MD time accumulated in core state(s) is shorter than the relaxation time scale. Note that in this work, a version of Eq. (3.18) using the time-dependent occupations is employed for calculating the validity time.

3.4 Illustration of Usefullness of the Validity Time by Building a Markov State Model of Alanine Dipeptide

A. System Setup

A single molecule of alanine dipeptide (N-acetyl-N-methyl-L-alanylamide) (shown in Fig. (3.3)) was placed with 390 pre-equilibrated TIP3P water molecules in a periodic box of dimension $2.3 \times 2.3 \times 2.3$ nm^3. We employed the CHARMM27 force field[183].

B. Simulation Protocols

Equilibration at constant pressure of 1 atm was performed. Next we performed energy minimization for 1600 steps using conjugate gradient method. MD calculations using Langevin thermostat were performed at 300 K with NAMD[184]. Particles mesh Ewald electrostatics was employed for electrostatics terms. RATTLE[129] and SETTLE[128] algorithms were applied to covalent bond involving hydrogen in water and

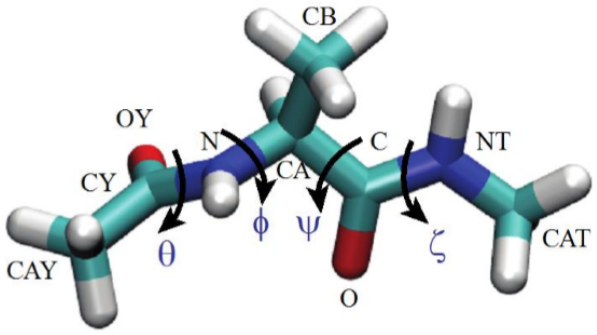

Figure 3.3: Structure of Alanine-dipeptide. θ, ϕ, ψ and ξ are the dihedral angles.

peptide, respectively. A time step of 2 fs was used. Independent MD trajectories were generated using different random seeds. The system configuration was checked for transition after every 100 MD steps by comparing the structure to a database of states. Structures were compared using Kabsch algorithm with a tolerance of 1.5 Å. A match is said to be occurred when non-hydrogen of the alanine dipeptide molecule lie within the tolerance. A short MD calculation for $\tau = 0.8$ ps is additionally performed to avoid counting of the recrossing events as transitions. A MSM of alanine-dipeptide (Fig. 3.4) was constructed from a swarm of independent MD trajectories by using swarm MD calculations (see flow chart in Fig. 4.2 of chapter 4, detailed discussion is provided in the next chapter).

3.4.1 Markov State Model for Alanine Dipeptide

A MSM constructed for solvated alanine dipeptide at 300 K is shown in Fig. 3.4. States and kinetic pathways were found by analyzing $on-the-fly$ several thousand MD trajectories as they were being generated in parallel as discussed in previous section. States 1-5 (Fig. 3.5) form the core network. Detection of states 1-4 is possible within 0.1 μs MD, while detection of state 5 can require longer trajectories. States 1 and 3 closely represent the α_R conformation of alanine dipeptide. States 2 and 4 are associated with β/(PII;ϕ = -85° and ψ = 160°)/($C7_{eq}$;ϕ = -86° and ψ = 79°) structures of the system, while state 5 can be identified as the $C7_{ax}$ (ϕ = 76° and ψ = -62°) conformation (see Ref. 185). Occasionally, the system will visit 11 other states (see Fig. 3.4) only to quickly return to the core states. We consider these 11 states as periphery states. Kinetic rates were estimated using MLE when a pathway is sighted 10 times or more. Standard error in rates shown in Fig. 3.4 is computed using the Bootstrap method.

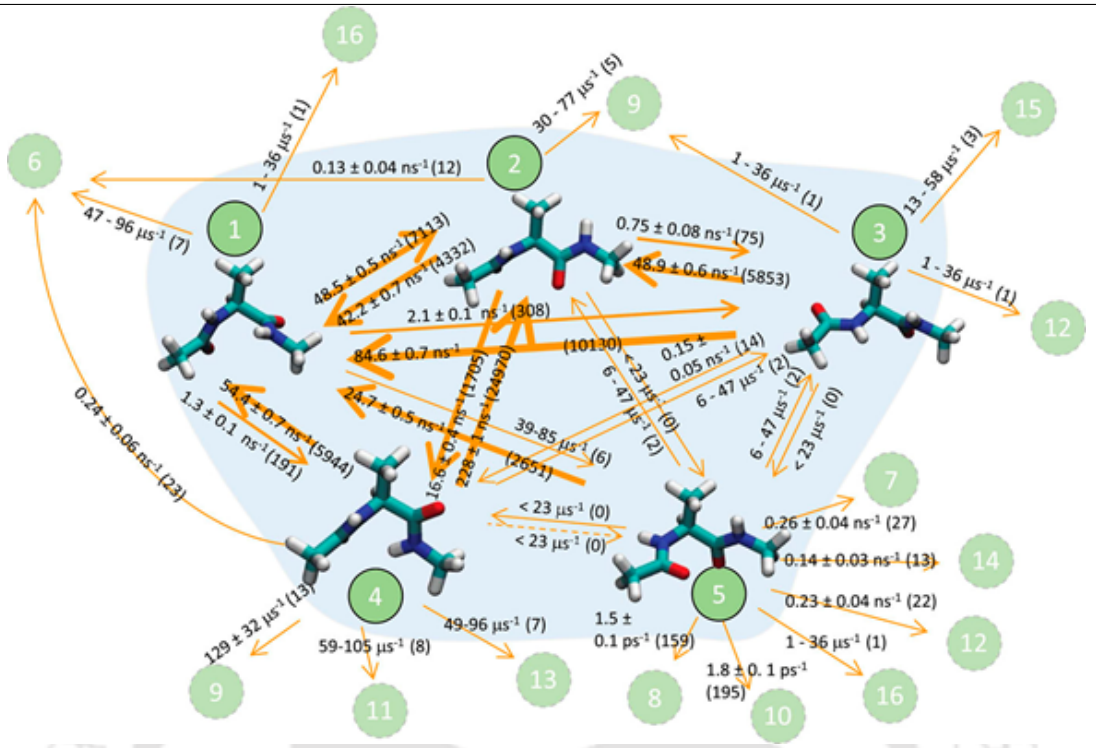

Figure 3.4: Markov state model for solvated alanine dipeptide at 300 K. States are numbered in the order they were discovered with MD. States 1-5 are core states and 6-16 are periphery states. Kinetic rates are shown along with the number of sightings for the event in parentheses.

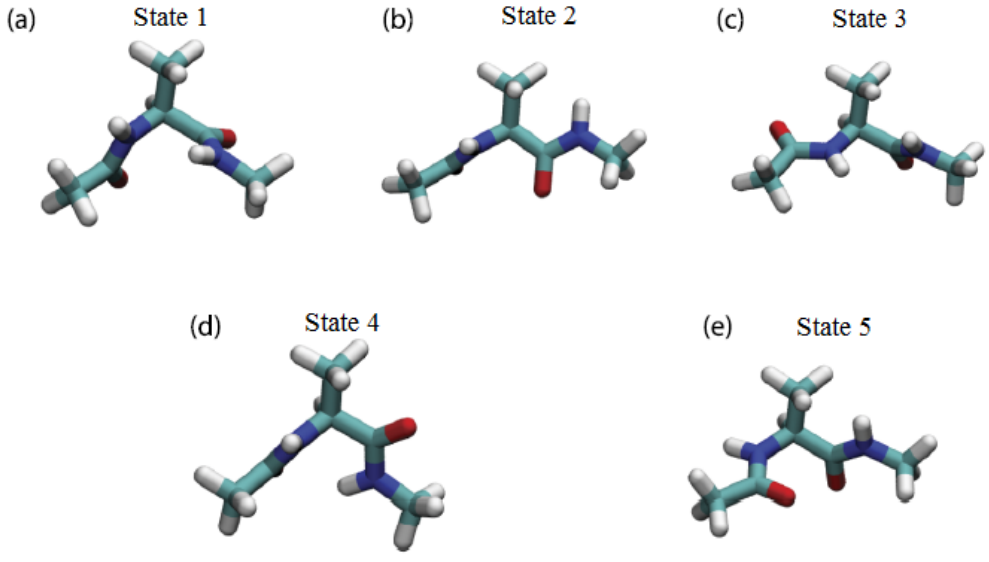

Figure 3.5: Structure of the 5 core states of Alanine-dipeptide.

The free energy map in the in (ϕ, ψ) space obtained using standard MD calculations at 300 K with setup mentioned earlier is shown in Fig. 3.6. The free energy

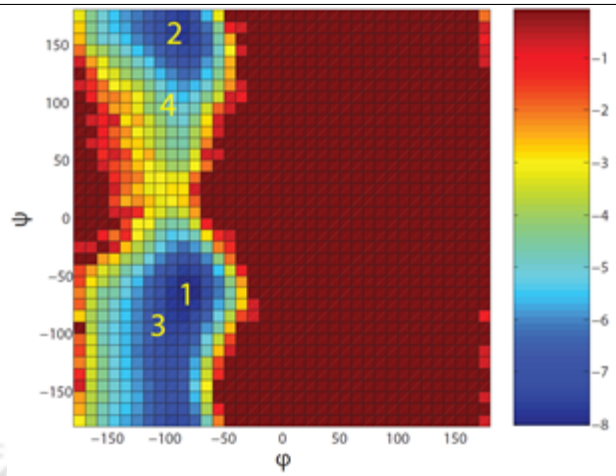

Figure 3.6: Free energy map (in units of $k_B T$) at 300 K from MD simulations. State 5 is not accessible at the timescles accessed.

map is in good agreement with one in the top-left panel of Fig. 3.6 of Ref 186. The free energy of states 2-4 with respect to state 1 obtained by solving the MSM shown in Fig. 3.4 is -0.06, 2.28 and 1.596 kcal/mol, respectively, which is in good agreement with the free energy map shown above as well as free energy difference mentioned in Ref. 187. Transitions between β/PII/$C7_{eq}$ structures are known to be fast (less than ps at 300 K) as reported in Ref. 188, which is found to be true also with our results in Fig. 3.4. The rate for moves from 1 to 2 (alpha-helix to beta strand transitions) is estimated to be 0.045 ps^{-1} (mean escape time of $\sim$ 22 ps) which is comparable to previous estimates (Ref. 189). The activation barriers for the moves 2-5, 5-2, and 5-4 are 6.55, 5.7, and 5 kcal/mol respectively. The corresponding barriers in vacuum[190] are less than 7, 5, and 5 kcal/mol respectively.

Figure 3.7 shows the core state occupation in dashed lines obtained by solving Eq. (3.13). Although the MD trajectory exceeds 0.3 μs, it is conceivable that in the worst-case scenario absorbing-states will be visited via pathways that are missing in the core network model. The leakage flux, calculated based on the rates from core states to periphery states and the missing rate from each core state using $\delta = 0.1$, is included in the full network model of Eq. (3.12). Equations (3.12) and (3.13) agree at short times [Fig. 3.7(a)]; however, they diverge at longer time scales [Fig. 3.7(b)]. In the worst case, all core state occupations will decay exponentially with the same rate consistent with Eq. (3.14).

Suppose the system is trapped/equilibrated in N core states corresponding to deep basins in the energy landscape and periphery states are absent, the MD time

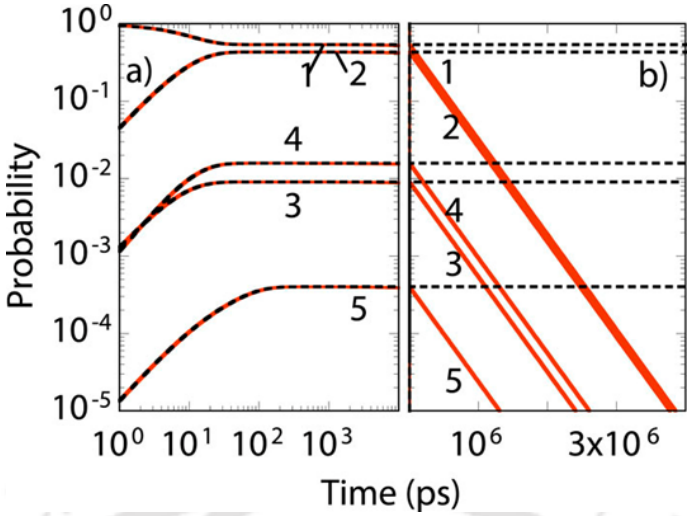

Figure 3.7: Occupation for core states (states 1-5) of Fig. 3.4 found by solving the core network [Eq. (3.13), dashed black line] and full network [Eq. (3.12), red line] models at (a) short (log-log) and (b) long (semilog) time scales. Both models were constructed using MD trajectories. The full network model denotes the worst-case scenario where probability leakage into periphery/missing states occurs because of which the core state occupations decay exponentially at long time scales. The core network model is a compact MSM that does not contain any periphery/missing states.

t_S for a core state S is proportional to $\tilde{\pi}_S^{st}$, that is, $t_S = \tilde{\pi}_S^{st}\tau_{MD}$ where τ_{MD} is the entire MD duration. Equation (3.16) simplifies to

$$\kappa^{leak} = N\frac{\ln(1/\delta)}{\tau_{MD}}. \tag{3.19}$$

Subsequent coarse-graining of states into superstates can result in a more compact MSM, but the validity time is still determined by the total time spent in the superstate.

3.5 Discussion

MSMs have the potential to become even more powerful computational tools in the future for studies of biomolecular, materials, and reacting systems as new advances emerge that enable one to accurately encode the kinetic and thermodynamic information on the multidimensional landscape in terms of state-to-state transitions. Unfortunately, MSMs constructed bottom-up from finite MD trajectories are rarely complete, which has a direct implication on its accuracy. Since the (kinetic) information content in a single trajectory can be different from that of an ensemble at the

same length of time, questions related to validity of the MSM due to missing information arise. We introduce the fundamental concept of validity time of an MSM to quantify its completeness. The concept guarantees that all states and kinetic pathways that are relevant to the dynamics will be present in the MSM provided the time scales accessed by the model are smaller than its validity time. Put differently, this concept helps us understand the time scales where states and pathways missing in the available MD data or the MSM can become relevant to experimentally measurable quantities, and differences between the experimental quantities and MSM predictions might originate from the missing information in the MSM. A general framework that relates the kinetic uncertainty in the model to the validity time, missing states and pthways, network topology, and statistical sampling is provided here. Our methodology is flexible in terms of its ability to handle a wide-range of kinetic rates, number of states, relaxation times in the network, topology of the network, as well as missing pathways and states that have not been found with MD. The fundamental concept of validity time introduced here can be used alongside standard MSM building strategies.

Chapter 4

Methods Developed for Building Markov State Model: Swarm MD, State-Constrained MD and Programmed State-Constrained MD Calculations

4.1 Introduction

MSMs are usually constructed bottom-up using brute-force molecular dynamics (MD) simulations when the model contains a large number of states and pathways that are known *a priori*. However, the resulting network generally encompasses only parts of the configurational space, and regardless of any additional MD performed, several states and pathways will still remain missing. This implies that the duration for which the MSM can faithfully capture the true dynamics, which we term as the validity time for the MSM, is always finite and unfortunately much shorter than the MD time invested to construct the model. In the previous chapter, we have provided a theoretical framework to calculate the validity time. Our focus here lies on a class of methods that build MSM of longer validity time where the starting state is known but the final conformation is unknown. Although various temperature biasing schemes[191–195], applied directly or indirectly to construct MSM, there are certain challenges of existing MSM building approaches, namely, (i) systematically determining the states of the system, (ii) generating the MSM rapidly without the need for extensive MD calculations, that present obstacles while studying more complex

systems. Pandey and his co-workers have pioneered the development of an automated procedure for building of MSMs by using replica exchange MD (REMD) to seed MD trajectories and then post-analyzing long MD trajectories[61,146,147]. These methods[63,172] employ kinetic clustering approaches, which requires the desired number of states as input to generate the MSM. The correct choice of inputs requires some degree of experience. Wales and co-workers have employed energy-minimized configurations of the protein as the Markov state using an implicit solvent force field[196]. In addition, in most approaches conformational sampling of the biomolecular system using standard MD can be inadequate for building an accurate MSM when the system remain trapped in certain states for extended period of time. McCammon and his co-workers have attempted to address this issue by using an accelerated molecular dynamics approach called the hyperdynamics method[197,198]. The hyperdynamics method employs a boost potential that is applied to overcome large energy barrier and escape a potential energy basin. We believe that prior insights required to choose the boost potential for most protein systems are generally lacking.

Here we have developed a different approaches, named by *Swarm* MD calculations to build an MSM on-the-fly while MD calculations are being performed. A *State-constrained* MD (SC-MD) method is developed to search for the kinetic pathways from selective states. An enhanced kinetic-sampling technique called *programmed state-constrained* MD (PSC-MD) calculation is presented that guides selection of states, where additional MD must be performed to extend the validity time.

We apply our PSC-MD method in conjugation with Swarm MD and SC-MD method to construct MSM of desired validity time. First, we describe the rationale for our strategy of MSM construction in Sec. 4.2. We then outline our procedures, namely, Swarm-MD, SC-MD and PSC-MD in Secs. 4.3 and 4.4 respectively. The use of PSC-MD calculation to generate a MSM of desired validity time is illustrated in Sec. 4.5 with the help of a prototype network models that are completely known to us. We demonstrate application of MSMs of desired validity time to study of stretched deca-alanine under tension in Sec. 4.6 employing the methods developed here. Using the concept of validity time, we conclude that a large number of rarely visited states are also relevant to the dynamics.

4.2 Rationale Behind our MSM Strategy

The key steps while building an MSM at room temperature are: i) identification of states, ii) finding the rate constant for the moves from the states. The philosophy of our approach, which is general and significantly different from past attempts, is discussed below.

4.2.1 Potential Energy Superbasins and Kinetic Pathways

Solvated proteins can have a large number of degrees of freedom due to the presence of solvent molecules. The potential energy landscape associated with the system is corrugated and contains several potential basins separated by small energy barriers. Although the solvent molecules can be arranged in many possible configurations, the protein molecule prefers a smaller set of configurations due to the presence of strong covalent bonds. Typically, large energy barriers are involved when the protein conformation changes. Based on this premise, we believe that the energy landscape involving solvated protein systems can be represented in terms of potential energy superbasins, where a superbasin is a collection of energy basins that are accessible to each other via fast low barrier moves (e.g., solvent molecule rearrangements), while an escape out of one superbasin to another involves rare event or infrequent event (e.g., a protein conformation change). This concept is illustrated in Fig. 4.1 which shows a schematic of two superbasins in a one-dimensional potential energy surface. It is expected that a superbasin contains a large number of potential energy basins, and that two superbasins can be separated by a collection of higher-energy basins that do not belong to either state.

The presence of large energy barriers between the superbasins A and B implies that the system will remain trapped inside the superbasin A for a long time. Occasionally, the system is able to overcome the energy barrier and ventures into B. An escape from a superbasin can be detected by tracking the protein conformation and ensuring the system remains trapped in the state B long enough for the event to classify as a successful transition. Although multiple pathways may exist between two states, these pathways contain common features such as dihedral angles, end-to-end distance, helix content, native contact etc. that are modified as the system traverses between two states. Assuming that superbasin escape event is a Poisson process, that is, first-order dynamics should be obeyed, a rate constant can be estimated for the lumped pathways comprising of all possible pathways connecting the two states. Once the important states of the system is clear, if required, the mechanism can

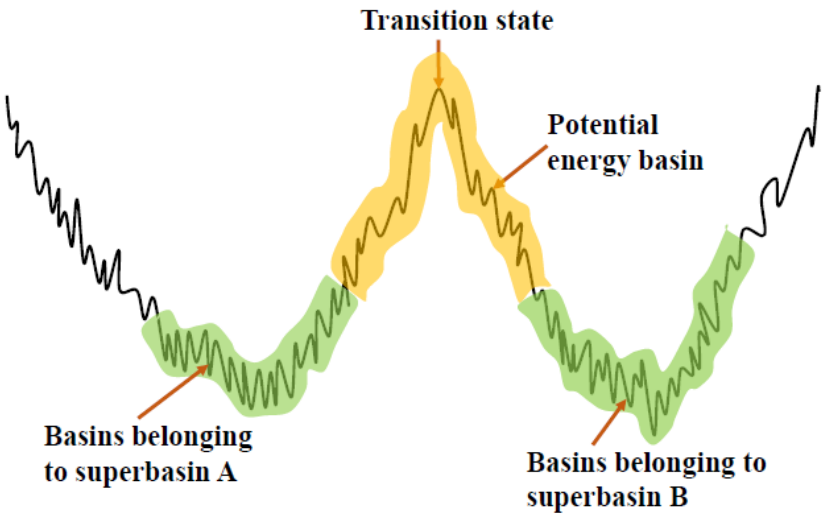

Figure 4.1: Schematic of two superbasins or (coarse) states (outlined in green) in a one-dimensional potential energy surface. The basins in a particular superbasin can be accessed in the dynamics by overcoming small barriers, such as those corresponding to solvent molecule rearrangements. The system can escape from a superbasin to another superbasin by overcoming a large potential energy barrier, such a change in the protein conformation. Two superbasins can be separated by basins (outlined in yellow) that belong to neither superbasin. These outlier basins will depend on the tolerances used in the state detection algorithm.

be investigated separately using other techniques named by *state-constrained* MD calculations which is described in Sec. 4.4.

4.2.2 Estimating Rate Constants

A special property of a discrete state, continuous time Markov process is that the waiting times t for transition from a state S are exponentially distributed. Let us consider a particular type of move α from the state S. According to first order kinetics, the probability $p(t)dt$ associated with the move α in an infinitesimally small time interval $[t, (t + dt)]$ is given by,

$$p(t)dt = \exp(-k_\alpha t)k_\alpha dt, \tag{4.1}$$

where t denotes the time elapsed in the state, and k_α is the rate constant of α at the temperature T. In Eq. (4.1) the probability that the system resides in the state till time t is given by $\exp(-k_\alpha t)$, while the probability for the escape to occur during the time dt is given by $k_\alpha dt$. From Eq. (4.1) the waiting time t is given by the

expression,

$$p(t) = k_\alpha \exp(-k_\alpha t). \tag{4.2}$$

Here $p(t)$ is the probability density for escape (waiting) times of the move α, which can be derived by assuming that the system spends a long time in the state before an escape occurs. The probability of observing at least one escape in the time interval $[0,t]$ is given by the cumulative density function,

$$F(t) = 1 - k_\alpha \exp(-k_\alpha t). \tag{4.3}$$

The average time for escape is $1/k_\alpha$. Since the kinetic pathways from state S are supposed to be independent, escape times or waiting times of kinetic pathways between states is also exponentially distributed, that is, the pathways are first-order processes. With this assumption, we employ the maximum likelihood estimation (MLE)[203] method to estimate the kinetic rate as $k_\alpha = n/t_{MD}$, where n denotes the number of times a move has been observed and t_{MD} is the total MD time elapsed for this move. Although we employ likelihood estimates for the rate constant in our MSMs, the probability density from MLE rate constant should be in good agreement with the observed time distribution[161].

4.3 Swarm MD Claculations

In order to identify the relevant state, we seek a sequence of state-to-state transitions beginning from state S of Fig 4.1 using a swarm of independent MD calculations in parallel. We call this approach the swarm MD method since the trajectories will independently wander on the energy landscape while new states are discovered on-the-fly and the state information is shared between the MD calculations (see flowchart in Fig. 4.2). After a successful transition from a state a thermalization step is performed in the latest state to avoid the recrossing events where the system escapes a state only to quickly return to the same state which are not considered as successful transitions[200–202]. As an escape from a state is detected the system should remain trapped in new state long enough before the next transition to ensure the Markovian behaviour of the process. After thermalization, the time required for the transition is recorded as waiting time or escape time. Each time an escape is observed to a new state of the system that is missing in the state database, the new state is added to the database. The process is continued until a specified number of transitions is detected or the accrued MD time exceeds a predefined limit.

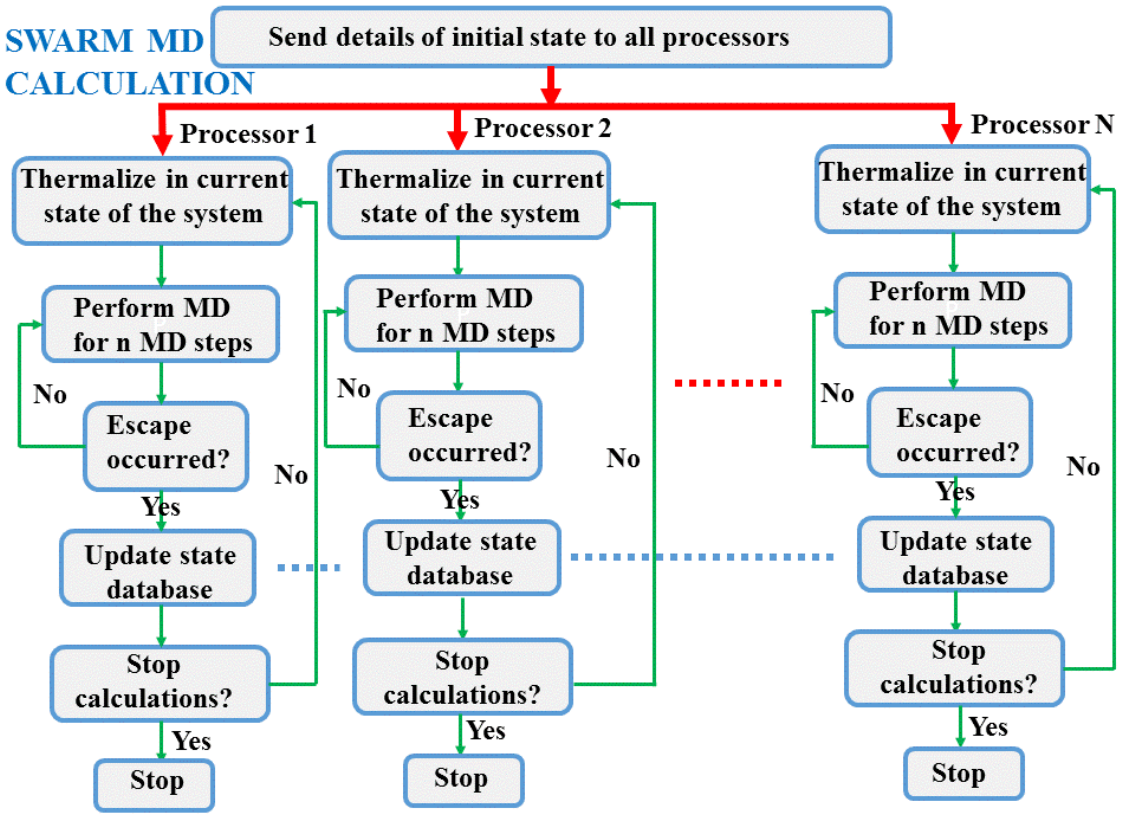

Figure 4.2: Flowchart for swarm MD calculation. A large munber of processors independently perform molecular dynamics (MD) calculations.

Furthermore, the crucial step to this process is that database of states is shared by all processors.

4.4 State Constrained MD and Programmed State Constrained MD

Adaptive sampling methods[177,204] that seek rare-configurations so that new MD trajectories can be seeded from such configurations, are generally used for calculating thermodynamic properties. Relevant kinetic information can be gained by efficiently extending the MSM validity time. If state S is poorly sampled in a dynamical trajectory even though it is relevant, one will require a longer trajectory with the hope that at some point enough transitions from the state will be sampled. Such situations can be tackled using *state-constrained* MD (SC-MD) calculations. In SC-MD calculations, one performs MD in state S while checking for a transition at regular intervals. Once a transition is detected, the MD calculation is stopped, the waiting

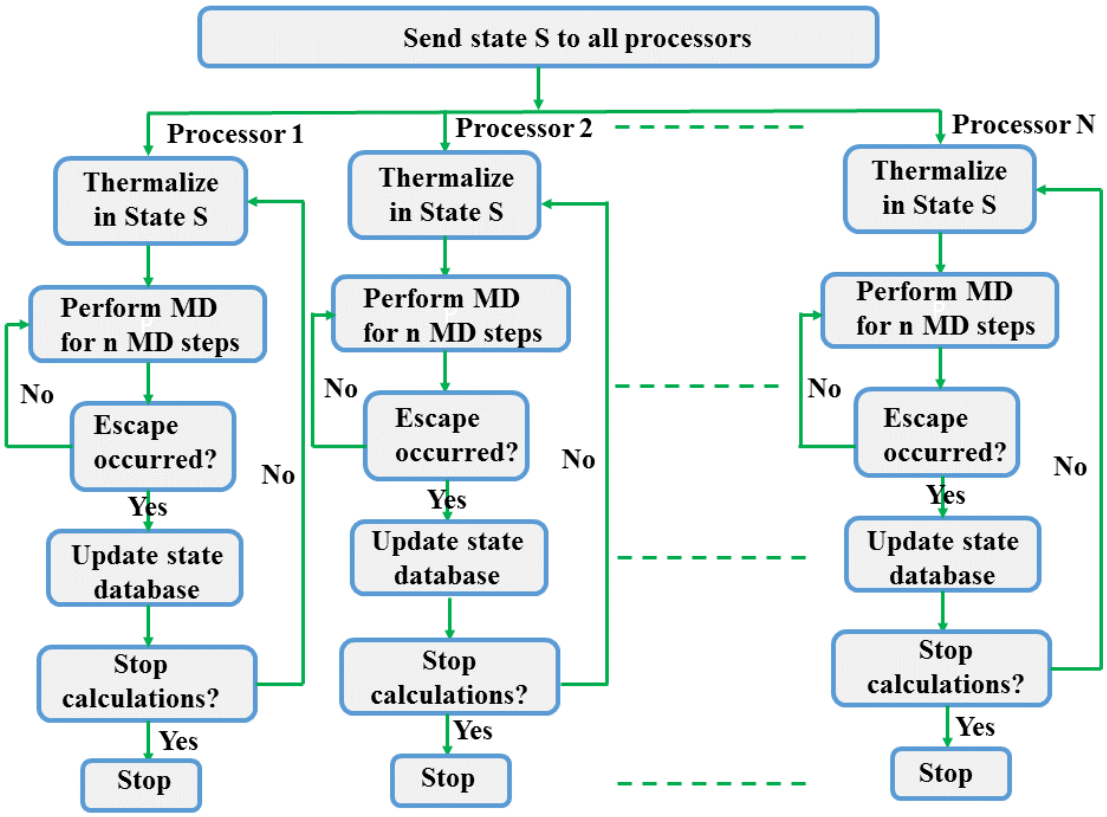

Figure 4.3: Flowchart for a state-constrained calculation. The overall steps are similar to a swarm calculation (see fig. 4.2), except that the system is returned to a chosen state S each time an escape is found.

time and final state are noted. A fresh independent MD calculation is seeded from S such that the system is in thermal equilibrium. The flowchart of SC-MD calculation is shown in Fig. 4.3. More transitions from S are sought. This prevents the system from freely diffusing over the potential energy landscape and confines it to a particular state for the purpose of detecting kinetic pathways from that state, calculating the rates, and lowering leakage flux of state S. Core- and full-network models (Eqs. 3.13 and 3.12) can be constructed efficiently with *programmed* state constrained MD (PSC-MD) by automatically targeting states with the largest leakage flux and performing SC-MD calculations in those states. The flowchart of PSC-MD method is given in Fig. 4.4.

PSC-MD method provides a way to increase the validity time of MSM by lowering the largest contribution to the leakage rate k^{leak} systematically. In this approach, the first step is to calculate the validity time (Eq. 3.18) of an existing MSM constructed from swarm MD calculations starting from initial state S by solving core-(and full-)

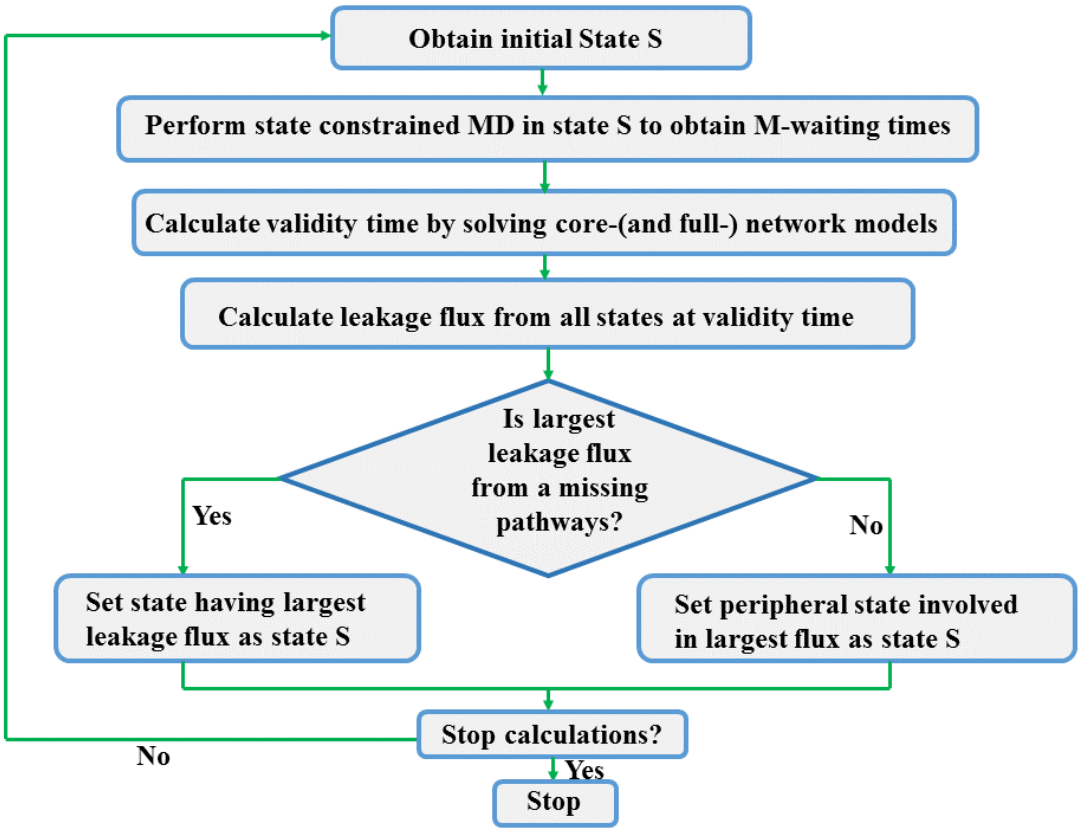

Figure 4.4: Flow chart for programmed state constrained MD (PSC-MD). Note that here MD can be replaced with another dynamical method.

network models (as in Eq. 3.12). Next, the leakage flux (F_S in Eq. 3.11) from all the core states to the missing states and the periphery states at validity time (τ_V) are calculated separately, and the missing pathway or the periphery state associated with the largest leakage flux is identified. If a periphery state S' is found to be accountable for maximum leakage flux from core states then, a state-constrained simulation is performed in periphery state S' until sufficient number of transitions is obtained from state S' and state S' will be considered as a new core state in core-network model. In the other case, if the largest leakage flux is due to a missing pathway from a core state S then, a state constrained is carried out in state S to get desirable number of transitions from state S. Thereafter, next cycle of the process begins with validity time calculation of new core-(and-full) network model. As a consequence, one-by-one the periphery states are included within core-network model. Thus, the validity time is extended systematically by reducing the leakage flux to the periphery/missing states from the core-states.

Let us consider the case where the largest leakage is due to a kinetic pathway from a core state S to a peripheral state S'. The MD time spent in state S' on average in a trajectory of duration τ_{MD}, which is given by,

$$t_{S'} = \pi_{S'}^{avg} \tau_{MD} \tag{4.4}$$

can be made significantly large to ensure that the validity time is not lowered by inclusion of S' as core state. Here, $\pi_{S'}^{avg} = \tau_{MD}^{-1} \int_0^{t_{MD}} \pi_{S'}(t)dt$; the occupancy of the system in state S' on average. Rarely-visited states with small values of $\pi_{S'}$ present a challenge as they require long MD trajectories. Suppose we require that the fastest rate k_f from state S' is selected at least m times in the MD calculation, we conclude that on average $k_f \pi_{S'}^{avg} \tau_{MD} = m$, that is, extending the validity time is limited by the rate k_f which determines the duration of MD as,

$$\tau_{MD} = \frac{m}{k_f \pi_{S'}^{avg}}. \tag{4.5}$$

where m denotes the number of transitions needed for estimating the rate with reasonable statistical accuracy. This issue can be tackled using SC-MD calculations where an MD calculation is started in the state S' and it is returned to the same state once an escape is detected. The duration of MD required is $\tau_{MD} = m/k_f$. The computational speed-up over regular MD given by $1/\pi_{S'}^{avg}$ can be phenomenal in most biomolecular systems due to the typically small values of $\pi_{S'}^{avg}$.

The PSC-MD scheme usually introduces many periphery states in the network model as it is used to build the MSM in patches. Occasionally, states are chosen from one part of the network for SC-MD calculations, and later, another part may be selected based on the calculated leakage flux. When the largest leakage is from core state S to a periphery state S', state-constrained MD is performed in S' until S' becomes a core state. In summary, in PSC-MD method, one identifies configurations where kinetic information appears inadequate as starting states for subsequent MD. While $ad-hoc$ SC-MD similar to such approaches can result in some improvement in computational efficiency, an impressive speedup is expected from PSC-MD where the objective is to determine the core/peripheral state that currently has the largest leakage rate and perform MD in that state for a chosen duration. In this way PSC-MD extends the validity time in an efficient manner. Hence, PSC-MD is more efficient than regular MD.

4.5 Prototype Example

That the validity time is a vital parameter for quantifying the MSM accuracy becomes evident by examining simple networks that are fully known to us from the outset. We consider a landscape containing trapping states, which is representative of deep basins of protein systems.

4.5.1 Network with Trapping States

Energy landscapes of biomolecular systems often contain low lying minima separated by large barriers. The inset of Fig. 4.5 shows a one-dimensional network with 5 states. Initially the system is in state 2. The rate is calculated as $k = 10^5 \exp(-\Delta F/k_B T) \text{ps}^{-1}$, where ΔF is the energy barrier given in the caption of Fig. 4.5, k_B is the Boltzmann constant, and $T = 300$ K is the temperature. Since the barrier for the move from state 2 to 3 is large (0.35 eV), the system remains trapped in states 1-2 at short times. States 4 and 5, which are also kinetic traps, are accessed at longer times.

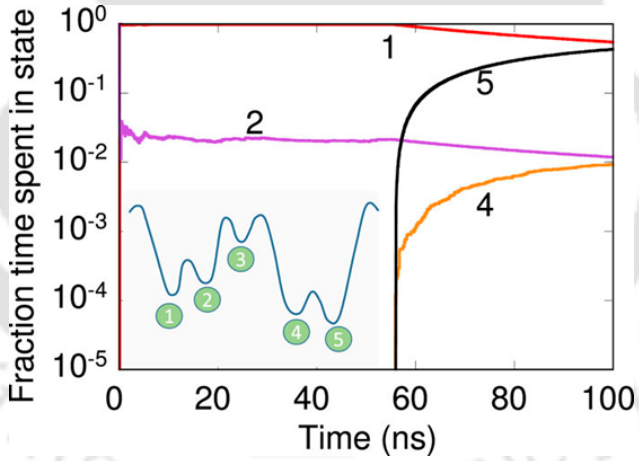

Figure 4.5: Fraction of time spent in states belonging to the 1-D network shown in the inset over the course of the simulation. The x-axis denotes the time elapsed in a dynamical trajectory. Energy barrier for forward moves from left to right are 0.3, 0.35, 0.1, and 0.2 eV. Barriers for backward moves from left to right are 0.2, 0.1, 0.45, and 0.3 eV. The system resides in state 2 at time $t = 0$ ps.

A dynamical trajectory is generated using a kinetic Monte Carlo procedure wherein a move is selected randomly from the current state with a probability proportional to its rate and the time is advanced by $-\ln(\xi)/k_S$. Here, ξ is a uniform random deviate and k_S is the sum of rates from the current state S. An MSM is

constructed with the help of the dynamical trajectory. Kinetic rates are estimated using MLE when a pathway is sighted 10 times or more.

The fraction of time spent in a state in the dynamical trajectory is plotted in Fig. 4.5. After the initial transient, the occupations for states 1-2 have reached a plateau. Based on the short dynamical trajectory, one may correctly conclude that the MSM containing only states 1-2 will suffice at short time scales. The picture changes at the longer time scales. After 50 ns states 4-5 are accessed and the fraction of time spent in states 1-2 decays. State 3 is not shown in Fig. 4.5 for convenience. Similar behavior is expected from other dynamical trajectories started from state 2, although the time required to access states 4-5 will vary. The relevance of state 3 cannot be ignored as it provides access to states 4-5. Thus, the MSM should include states 1-5 at longer time scales. So when should one include states 3-5 to ensure that the MSM remains accurate?

PSC calculations were performed with state 2 as the starting state. The validity time is calculated using $\delta = 0.1$, that is, the confidence associated with the estimate for kinetic rate is 90%. Initially, the MSM contains only two states and the MSM validity is extended by performing state-constrained calculations in states 1-2. The leakage rate from Eq. (3.19) of chapter 3, namely, $2\ln(\frac{1}{\delta}) = \tau_{MD}$, is in close agreement with PSC calculations (Fig. 4.6(a)) where τ_{MD} is the total MD duration. The validity time keeps increasing with the MD time except for an abrupt decrease when the periphery states 3-5 are detected (Fig. 4.6(b)). All states and pathways were available in the MSM at 553 ns. The corresponding validity time was 20.1 ns. Although we are aware of the total number of states and pathways in this toy model, in general for complex systems such as proteins, whether the network model is complete will remain unknown to us. Therefore, one would continue to search for newer states and pathways. At longer times, the leakage rate is given by $\frac{5\ln(1/\delta)}{\tau_{MD}}$ (see Eq. (3.19) of chapter 3).

It can be shown that the fraction of time spent in states 1-2 is identical for the dynamical trajectory of Fig. 4.5 and the state-constrained calculations of Fig. 4.6 at short times. Similar behavior is true for states 4-5 at longer times. Therefore, conclusions from Fig. 4.6(b) can be extended to Fig. 4.5. From Fig. 4.6(b), an MSM generated from a 40 ns long dynamical trajectory is valid only till 8 ns. The MSM of Fig. 4.5 only contains the pathways between states 1-2 at 40 ns, that is, the pathways missing in the MSM are less relevant to the dynamics till approximately

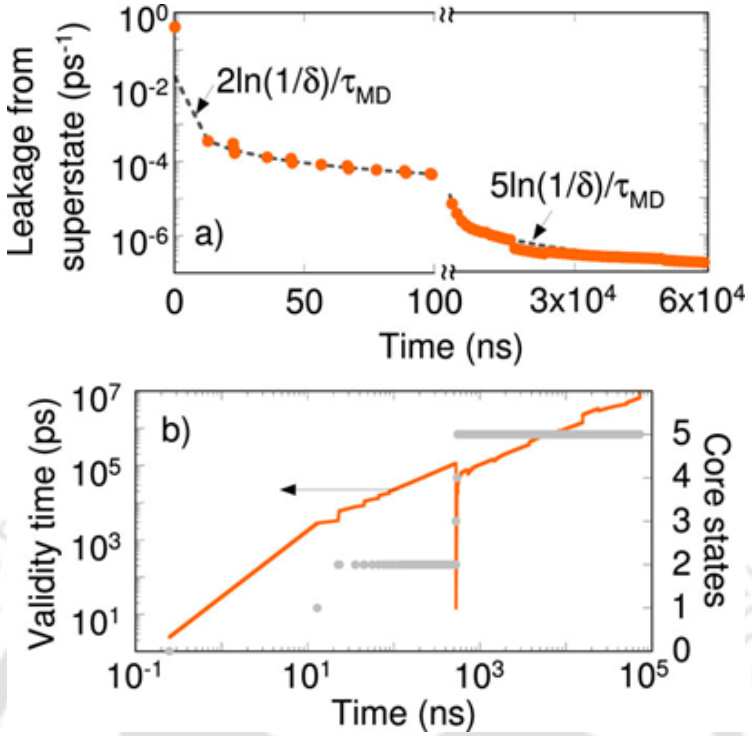

Figure 4.6: Leakage flux calculated using Eq. (3.19) of chapter 3 (dashed line) and state-constrained calculations (filled circles). (b) Validity time calculated for the MSM constructed for the network in Fig. 4.5 using state-constrained calculations (orange line). The number of core states shown in filled grey circles increases as the trajectory grows longer. The x-axis in both panels denotes time elapsed in the dynamical trajectory.

8 ns. This is confirmed in Fig. 4.7, where the time-dependent state occupations are plotted. States 3-5 are visited before 60 ns in Fig. 4.5, which corresponds to a validity time of approximately 10 ns. This implies states 1-5 should be present in a MSM when 10 ns time scales are accessed. Fig. 4.7 shows that roughly in 1 in 10 dynamical trajectories the system will be in state 5 at 10 ns. When 1/10 is set as the tolerance limit, the dynamical behavior can no longer be predicted using a two-state MSM, highlighting the importance of missing pathways.

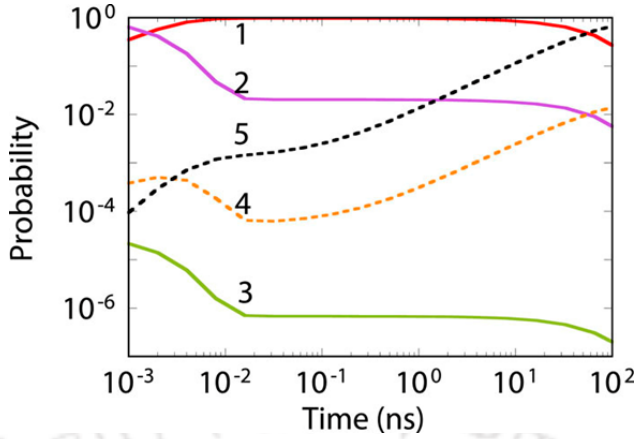

Figure 4.7: State occupation for the network shown in Fig. 3.7 of chapter 3 obtained by solving the MSM with validity time of 10^4 ns. Initial state is 2.

4.6 Markov State Model of Stretched Deca-alanine

Section 3.3 in chapter 3 and Sections 4.4-4.5 lead to the conclusion that the MSM validity time is always shorter than the duration of MD used to construct the MSM. This has serious implications on longtime studies using MSMs. Despite this, crucial insights can be obtained with MSMs. We demonstrate this aspect by building an MSM with a desirable validity time to study the kinetics of a deca-alanine molecule in vacuum under tension. A capped deca-alanine (Ala10) with acetylated N-terminus and amidated C-terminus was selected for the study. The initial configuration was obtained from the 104-atom helical model of Ref. 205. The C_α atoms at the two ends (residues 1 and 10) are tethered to two anchor points by harmonic restraints with a spring constant of 0.86 kcal·mol^{-1}Å^{-2} (Fig. 4.8 inset). The anchor separation d is kept fixed during the construction of a MSM and fluctuations in molecular extension, and forces are measured.

Past studies of deca-alanine demonstrated that the unravelling of the helical structure results in higher free-energy configurations, although the presence of metastable configurations of stretched deca-alanine has not been reported. Hence, specific 3D meta-stable structures encountered in the dynamics are determined as d is varied between 16 and 26 Å. Questions related to appropriateness of employing the same Markov state definitions and pathways across different anchor separations as well as changes in the kinetic rates are examined. Analysis of dominant configurations helps us to probe the importance of multiple pathways for unravelling of the helical structure as the molecule is stretched. Since helix winding/unwinding is a reversible process, helical and stretched configurations coexist, which causes the force expe-

rienced by the AFM tip to fluctuate. Ensemble-averaged forces from MSMs and MD are compared. Through this study, we conclude that forces are inaccurately predicted using MSMs with short validity time, thus highlighting the important role of missing states/pathways in ensemble-averaged quantities.

MSMs are constructed on-the-fly with 90% confidence ($\delta = 0.1$) (sec Sec. 3.3.2 of previous chapter) using thousands of short-MD calculations that are run in parallel as part of the PSC procedure. All MD simulations were performed with NAMD 2.9[184] with the CHARMM36 force fields[206]. Temperature was held at 300K using a Langevin thermostat. Bonds involving hydrogen atoms were constrained to their equilibrium values using RATTLE[129]. An integration time step of 2 fs was used. States were determined by comparing the backbone atoms after aligning the molecule using the Kabsch algorithm. A tolerance of 3 Å was found to be suitable for identifying the states. MD snapshots were collected every 0.2 ps because of rapid interconversion between the states. A transition was said to have successfully occurred when the system continues to reside in the new state for at least 1.2 ps after the transition was detected in the MD trajectory. This prevents recrossing events to be counted as transitions.

Steered MD simulations were performed with the deca-alanine for several nanoseconds to obtain a preliminary collection of unfolded structures. These configurations were provided as inputs to several nanosecond-long regular MD trajectories with chosen anchor separations. A preliminary catalog of states was constructed that could be employed with different anchor separations. States were indexed in the order they are found. State-constrained MD calculations, which are more efficient than regular MD (see Sec. 4.4), were used to confirm Markovian behavior and that the kinetic pathways can be described as a first-order process. Poor MSM validity was achieved when states or MD duration were selected ad-hoc in the state constrained MD calculations. For example, in a preliminary MSM-building attempt we found a 4 μs long MD trajectory resulted in a validity time of 0.064 ns, which is the reason why only PSC-MD calculations are performed.

The full network model consisted of 810 states. Most states are periphery states. The number of core states and their relevance varies with anchor separation. State occupation at $d = 16$ Å obtained by solving an MSM of 65 ns validity is shown in Fig. 4.8. The MSM constructed using a 9 μs trajectory contains 25 core states and 78 pathways. The system was initially present in state 2 shown in Fig. 4.10. Rapid

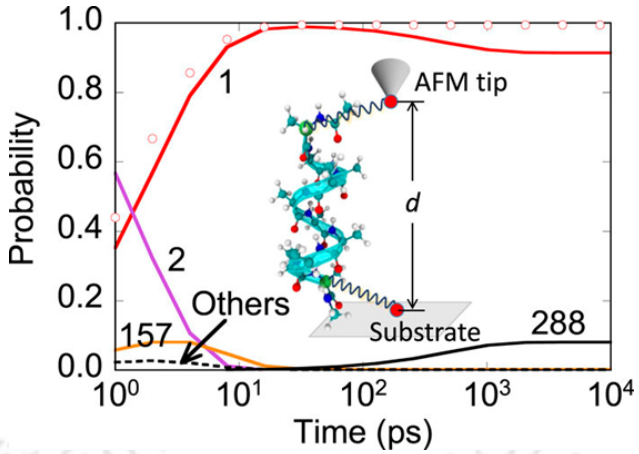

Figure 4.8: Occupation for the top-four states using a MSM with 65 ns validity time when the anchor separation $d = 16$ Å. The force-spectroscopy setup for deca-alanine in vacuum at 300 K is shown in the inset. Harmonic restraint is applied to the light-green colored C_α atoms. Open circles show state 1 occupation from a 5-state MSM with 2 ns validity time (constructed from a 0.57 μs long MD trajectory).

conversion to α-helical configuration (state 1) with a large rate of 0.44 ps^{-1} is observed consistent with previous studies. The average distance between the terminal C_α atoms is 16.4 Å for state 1. Another dominant configuration, namely state 288, is selected beyond 100 ps with nearly 10% probability. State 288 is accessible from state 1 with a small rate of 1.4 ns^{-1}. As a consequence, state 288 is absent in an MSM with a shorter validity time and the time-dependent occupation for state 1 from such an MSM is incorrectly predicted (see open circles in Fig. 4.8). This behavior is analogous to the one observed in Fig. 4.5. State 288 is preferred over state 2 for two reasons. First, it has an end-to-end distance smaller than that of state 2, which is favored at compressive conditions. Second, the average α-helicity for states 1, 2, and 288 are 0.8, 0.2, and 0.74, respectively. Multiple backbone hydrogen bonds impart more stability to state 288 than state 2. It appears that a two-state model (states 1 and 288) might suffice for the calculation of average force at $d = 16$ Å. The AFM set up of deca-alanine and structures of state 1 and state 288 are shown in Fig. 4.9.

The abundance of energetically-favorable stretched-out configurations causes state 288 to lose its relevance at higher anchor separations, but state 1 still continues to be relevant. Figure 4.10 shows the dominant core states for anchor separations 22 and 23 Å. The unravelling of a helical structure to an elongated one proceeds via multiple intermediate states. For instance, one pathway for visiting state 17 from state 1 involves only "local" readjustments. First, a partial opening of lower coils

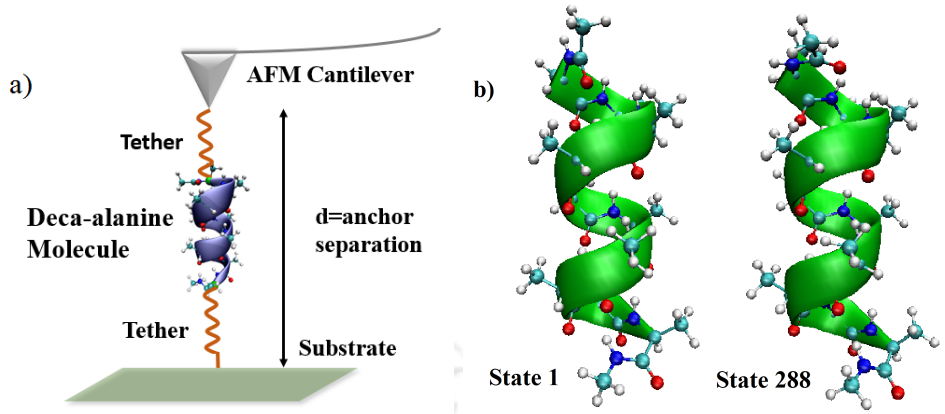

Figure 4.9: (a) Deca-alanine molecule in a AFM set-up where two ends of deca-alanine are connected to anchor points by two harmonic springs of equal spring constant, k_{tether}, at constant anchor separation d. (b) The structure of state 1 and state 288.

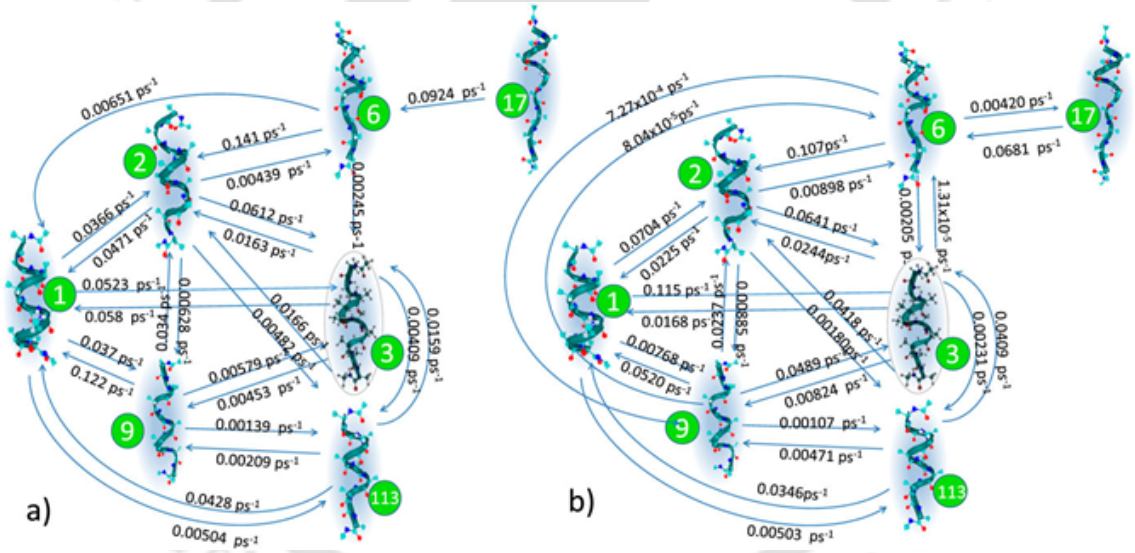

Figure 4.10: Network model obtained when the anchor separation is (a) 22 and (b) 23 Å. Only frequently visited states are shown. State 3 (encircled) is the starting configuration for the subsequent figures.

(C-terminal) is observed (state 1 to 6 via state 2) followed by subsequent stretching of the lower coils (state 6 to 17). Table 4.1 shows the α-helicity and 3_{10} helicity of states along the pathways from states 17→6→2→1 computed from state-constrained MD simulations. State 2 has the largest 3_{10}-helicity, that is, 3_{10}-helical conformations form the intermediates between the folded and the unfolded state. The rates of folding/unfolding depend on the anchor separation as given in Fig. 4.11. Alternatively, fluctuations in the middle residues can frequently cause deformation in the

Table 4.1: Average α-helicity and 3_{10}-helicity of states along a folding pathway.

State	α-Helicity	3_{10}-Helicity
1	0.081	0.08
2	0.201	0.21
6	0.021	0.19
17	0.013	0.15

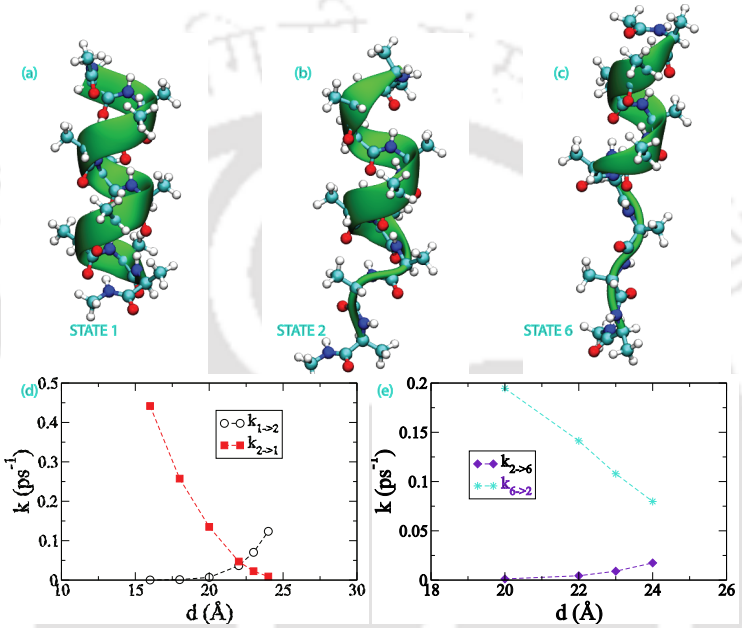

Figure 4.11: States of stretched deca-alanine a) state 1, b) state 2 and c) state 6. Rates of folding and unfolding as a function of the anchor separation along the path d) state 2 to 1 and e) state 6 to 2.

helix (state 1 to 3) that may sometimes lead to an elongated configuration (state 3 to 6). States 2, 3, and 6 have a non-negligible 3_{10}-helicity. State 6, which is an essential intermediate for both pathways, can be reached faster from states 2 and 3 as the anchor separation is increased. The preferred winding/unwinding mechanism proceeds predominantly at the C-terminal, that is, the former pathway.

In the past, the end-to-end distance has been employed as a reaction coordinate for stretched deca-alanine. End-to-end distance distribution for a state tends to be sharply-peaked with a standard deviation of nearly 1-2 Å; however, the distribution is a function of the anchor separation. Large overlap in end-to-end distribution for the core states makes it practically impossible to distinguish states when only end-to-end distances are employed. Inclusion of the 3D structure, which is implicit in

our state description, helps resolve intermediate states and state-specific properties. In particular, we are interested in the stiffness of deca-alanine, which determines the force on the AFM tip. The stiffness, calculated using state-constrained MD calculations (a detailed description of state-specific stiffness calculation is given in Secs. 5.4.1 and 5.4.2 of the next chapter), is found to vary from one state to another depending on the intramolecular interactions. Presence of strong hydrogen bonds in state 1 results in a large spring constant of 39.44 pN/Å. On the other hand, state 3 has a smaller spring constant of 25.38 pN/Å. A natural consequence is that the average force experienced by the AFM tip can be altered by as much as 100 pN in either direction during a state-to-state transition because of the differences in the state-specific spring constants. The average force is given by the sum of force experienced for each state times the state occupations.

MSMs are sensitive to small changes in the anchor separation [see Figs. 4.10(a) and 4.10(b)]. Pathways between states 6 and 9 are absent in the MSM for $d =$ 22 Å, but they are dynamically more relevant at validity times of 8 ns when $d =$ 23 Å. Forward and backward rates are found for many pairs of states. Exceptions in Fig. 4.10(a) include the move from state 1 to 6, although this is not an issue since a stationary solution is obtained without the requirement of detailed balance. The rate constants involving the core states vary over several orders of magnitude between 10^{-1} and 10^{-5} ps^{-1}. Exponential increase/decrease in rates is witnessed between 16 and 26 Å as shown in Figs. 4.11(d) and 4.11(e). While a handful of states can describe the dynamics for small anchor separations, additional states should be included in the MSM when the molecule is stretched extensively. An explosion in the number of core states is witnessed from 36 to 78 states between 22 and 25 Å. Correspondingly, the validity time plummets by almost 10 times from 12 ns at 22 Å separation to 1.5 ns at 25 Å separation for a 0.5 μs long MD trajectory. Note the length of (PSC)-MD trajectory required to reach the nanosecond-long validity time.

Figure 4.12 shows the evolution obtained with different MSMs for anchor separations between 22 and 25 Å. The rise and fall in the relevance of states and kinetic pathways is witnessed. The initial state of the system is state 3. One might expect that the helical structures (state 1) would not be selected at high separations; however, even at 24 Å separation there is a 1% chance of finding state 1. This is attributed to the lower energy of state 1 and the small stiffness of the tethers. In other words, it is possible for the tether to stretch to an extent where state 1 can still be visited in the dynamics at 24 Å separation. State 3 is an important

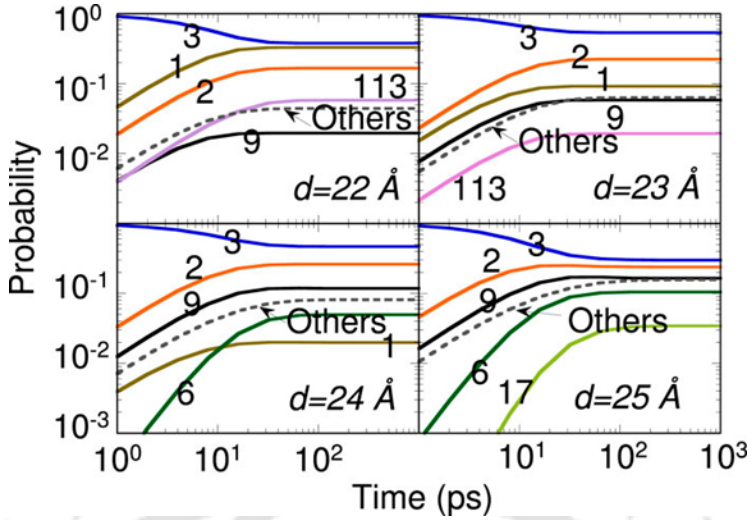

Figure 4.12: Probability evolution for different anchor separations. Numbers denote the state index in Fig. 4.10. The initial state of the system was state 3. The MSM validity time exceeded 1 ns in all cases.

configuration between 22 and 25 Å separation. The maximum state 3 occupation is witnessed around 23 Å. States 9, 6, and 17 compete with each other at large anchor separations. As the number of core states increases, there is an increased chance that the system will visit these states. The occupation in all other core states combined is shown in the dashed lines in Fig. 4.12. The time spent in the top-five states decreases from 96% at 22 Å separation to 30% at 26 Å separation. Based on Fig. 4.12, the stationary solution for deca-alanine is reached at nearly 100 ps, that is, time-dependent forces can be resolved only at sub-100 ps time scales.

The free energy difference for a state A with respect to state B is calculated from the stationary distribution as $\Delta F_{A-B} = -k_B T \ln(p_A/p_B)$, where p denotes the stationary occupation of a state. Since the stationary solution depends on the relevant pathways, absence of one or more pathways can introduce errors in the free energies. Occupations in Fig. 4.12 were used to calculate the free-energy difference between the states. Insights into the separation-dependence of the kinetic rates can be obtained from the Bell-Evans-Polyani principle[207,208]. Consider states 1 and 2. Since the free energy of state 1 increases with reference to state 2 as deca-alanine is stretched, the free barrier for the move from 1 to 2 (2 to 1) decreases (increases), which explains the corresponding shift in the rate constants as shown in Fig. 4.11(d). Similarly, the free energy of state 2 decreases with respect to state 6, causing the rate constant from state 2 to 6 to increase (Fig. 4.11(e)).

Figure 4.13 shows the time-dependent ensemble-averaged force experienced by the AFM tip corresponding to the evolution shown in Fig. 4.12 for different anchor separation. Average forces were computed for the individual core states using

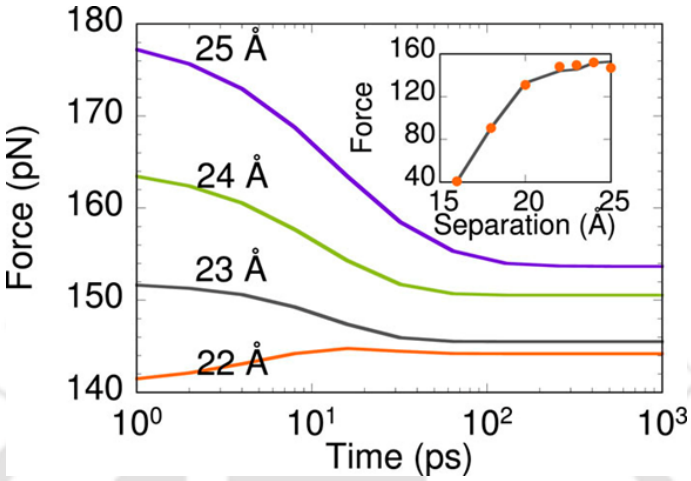

Figure 4.13: Solid lines show the average force acting on the AFM tip when the deca-alanine is stretched (using MSMs with validity time exceeding 4 ns). Behavior for anchor separation 22-25 Å is shown. Steady state forces calculated from the MSM (line) and MD (filled-circles) shown in inset are in good agreement.

state-constrained MD. Since the force experienced with state 3 is smaller than with state 1, the average force increases with time at 22 Å separation as state 1 occupation increases (see Fig. 4.12). Although state 3 plays a minor role in the winding/unwinding of deca-alanine, it is important for calculation of average forces due to its large occupancy. Beyond 23 Å separation, one finds that state 2 is preferred over state 3. The average force decreases with time since the force experienced with state 2 is smaller than with state 3. The average force at steady-state from the MSM and MD are in good agreement (Fig. 4.13 inset) validating our MSM. Such an agreement between MD and the MSM is not witnessed when the validity time is small as many relevant states are missing. The sudden drop in the slope of the force-separation curve is attributed to the smaller stiffness of states encountered at higher separations. As shown in Fig. 4.14, the work done in pulling deca-alanine from a compact to a stretched configuration can be calculated using the MSMs as follows. MSMs constructed for $d = 16$, 18, 20, 22, 23, 24 and 25 Å provide average force and end-to-end distance for the molecule. We assume that molecule is stretched by moving one anchor point infinitely slowly, allowing us to calculate the work done using the MSMs since the molecule reaches steady state for each value of anchor separation d. The calculated work done is in agreement with previous values in the literature[205].

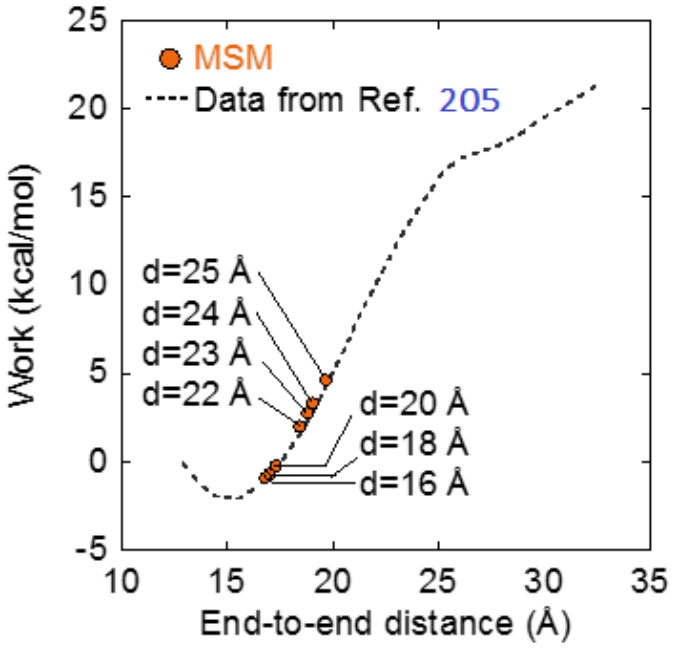

Figure 4.14: Work done while pulling deca-alanine is calculated using MSMs constructed shown in orange filled circles. Good agreement is observed with Ref.[205]. Each symbol is obtained from a separate MSM constructed for the anchor separation d mentioned in the figure.

Weaker electrostatic interactions are expected when deca-alanine is present in water. To mimic the effect of water, we construct MSMs at anchor separations 16 and 24 Å with a dielectric constant of 80. In order to find common features in the dynamical behavior for dielectric constants 1 and 80, the list of 810 states obtained previously was used as the starting known structures for our new calculations. Deca-alanine was initially kept in state 3. The dominant states for anchor separation of 16 Å include states 1, 20, 113, and 288. Figure 4.15(a) shows that the occupation for state 3 decreases continually in time. As in Fig. 4.8, a maximum value of the occupation for state 1 is witnessed. However, the steady state occupation for state 1, namely, 0.401, is much smaller than the one observed in Fig. 4.8. The occupation for other states combined increases to a significant value. We find that 11 states possess steady state occupations greater than 0.01. By considering the full network model, we conclude that the MSM obtained with 10.8 ns long MD calculations has a validity time of nearly 30 ps. At 30 ps, the difference in the state 1 occupations in the MSM and the full network model has exceeded 0.1. More states are observed with the anchor separation of 24 Å. 25 states possessed a steady state occupation in excess of 0.01 with state 3 being the only state common to both Figs. 4.12 and 4.15 with $d = 24$ Å. The total number of core states was found to be 137 using 53 ns

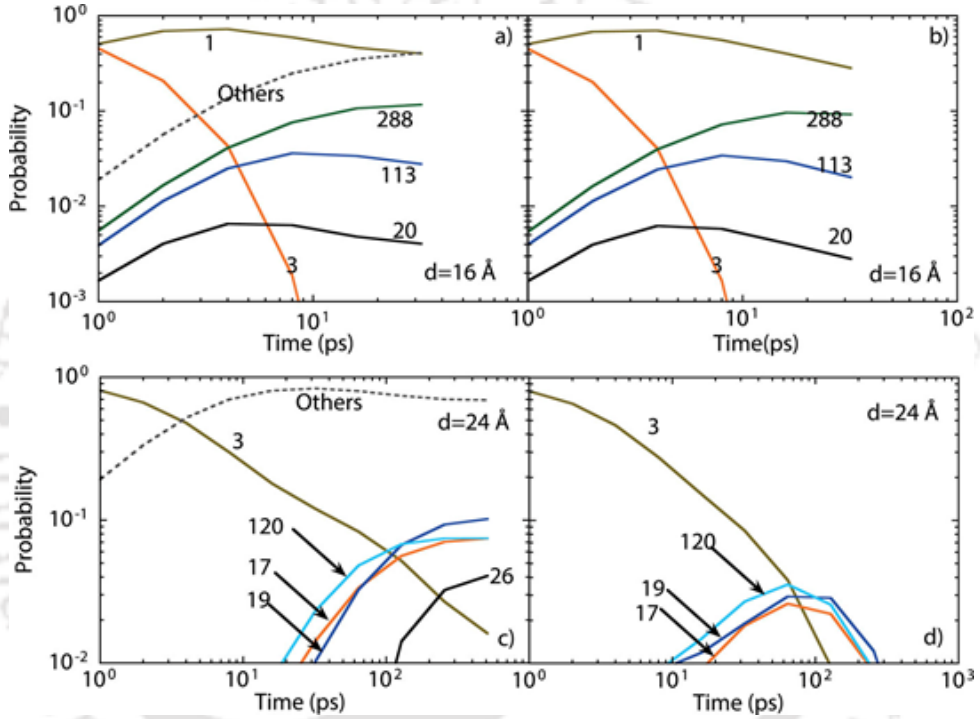

Figure 4.15: Probability evolution for anchor separations 16 Å [panels (a) and (b)] and 24 Å [panels (c) and (d)] using the dielectric constant of 80 to mimic deca-alanine in water. Numbers denote state index. Initial state of the system was state 3. Left panels [(a) and (c)] show results from the MSM while right panels [(b) and (d)] show results obtained with the full network model.

long PSC-MD. Figures 4.15(c) and 4.15(d) show state occupations as a function of time from the MSM and the full network model. Predictions from the two models diverge at nearly 0.1 ns. One can determine whether states and kinetic pathways in the present MSM will continue to remain relevant at longer time scales only by extending the validity time of the MSM.

4.7 Discussion

Unfortunately, the kinetic information embedded in the kinetic network model or Markov State Model (MSM) constructed bottom-up using molecular dynamics (MD) trajectories are rarely complete due to inadequate sampling while studying the long-time dynamics of biomolecular system. The uncertainty due to missing states and rates has a direct bearing on the accuracy of the model. As a consequence, it has generally been difficult to compare kinetic studies of biomolecular systems. Therefore, to ensure the goodness of MSM, an MSM-building procedure is desired which would be longer validity time compared to the time required to build it. So, the development of a suite of new class of methods, namely, Swarm MD, State-constrained MD (SCMD) and Programmed state constrained (PSC-MD) method to construct MSMs more efficiently and accurately than ad-hoc seeding of trajectories from different states or ad-hoc pruning of inadequately sampled kinetic pathways are discussed here. A computational procedure, called the Swarm MD method is used for systematically building an MSM on-the fly by performing the MD simulation and analyzing the MD data simultaneously whereas in the other MSM building procedures, MSMs are constructed generally by post-analyzing the MD data. States-constrained MD calculations are employed for searching the kinetic pathways the quickly from states that are of interest. The advantage of SCMD is that sampling is performed in a confined space which enhances our chances of finding sufficient transitions between two states. Development of PSC-MD calculations in this chapter provides efficient means for extending validity time of an MSM. MSMs with sufficiently large validity time provide following key insights into the stretching of deca-alanine. Unwinding of deca-alanine proceeds mainly via breaking of hydrogen bonds at the C-terminal. Characterization of states might not be possible only using simple reaction coordinates such as end-to-end distance. Each state possesses its own mechanical characteristics, e.g., spring constants, that ultimately determine the force experienced by the AFM tip in the force spectroscopy (FS) setup. To calculate the force, one also needs to estimate the state occupations accurately. It is not straightforward to

guess the relevant states/pathways at different anchor separations. The list of relevant states and kinetic pathways, and associated rates, that is, the network topology, can change dramatically with the anchor separation. The size of the MSM increases for large anchor separations, which has a direct bearing on the amount of MD required to know the state occupations accurately. Our studies demonstrate that the absence of relevant states/pathways in the MSM can lead to incorrect prediction of the kinetic and thermodynamic quantities being sought, which is the main reason why studies employing MSMs and other related kinetic network models should state the validity time of the model.

Chapter 5

Master-MSM for Constant Probe Separation Experiment in Single Molecule Force Spectroscopy Set Up and Time-Dependent Markov State Model (TD-MSM)

5.1 Introduction

A stretching force applied on a biomolecule such as a peptide or a nucleic acid strand can induce structural changes in the molecule as well as radically alter the kinetic depending on how the molecule is stretched[209–211]. Single molecule force spectroscopy (SMFS) setups such as AFMs, optical or magnetic tweezers employed in various modes, constant force, constant extension, force-ramp, force-jump, and more recently, extension clamp experiments[212–217] enable us to study an array of systems under mechanical tension. Recent advances in single-molecule approaches have become powerful tools to provide unprecedented insights into the underlying energy landscape of proteins, probe interactions between molecules, and even explore important structural transitions such as unzipping of DNA. In this chapter, we discuss how a similar idea can be exploited for rapidly constructing a detailed kinetic network model of the biomolecule using computer simulations, that is, kinetic information pertaining to *rare* transitions can be recovered by stretching the biomolecule at conditions where transitions are more frequent.

One potential beneficiary of such an approach would be SMFS experiments themselves. In such force spectroscopy (FS) experiments, typically the force applied to the system is recorded as a function of end-to-end distance, thereby producing a force-extension curve[218–222]. The force extension curve characterizes the molecule's elasticity and may provide valuable new information about the protein kinetics. An applied force can alter the dynamics of the system and transitions are identified from the rips in the force-extension curves. Despite the high precision of force measurements, a quantitative description of the high dimensional dynamics of biomolecule under tension requires utmost care. Moreover, such conventional analysis of SMFS experiments involves an excessively coarse-grained view[223–230]. For instance, a 1-D reaction coordinate between the "folded" and "unfolded" states invoked in the interpretation of SMFS experiments hides the complexity of a rough multi-dimensional free energy landscape with multiple states and kinetic pathways[231]. Resolving the individual states can avoid complicated force-dependent rate maps arising from incorrectly lumping multiple intermediate states and competing pathways. Unfortunately, the inability to experimentally observe microscopic structural changes in a molecule presents an obstacle towards gaining higher resolution. Kinetic models derived bottom-up from molecular simulations may be able to fill gaps in our understanding of the experimental force-extension curve. While this may provide unprecedented insights, it also raises broader questions whether i) detailed microscopic kinetic model can be sufficiently versatile to encompass a wide range of stretching experiments and ii) it is possible to convert between network models applicable for each type of SMFS experiment, for example, constant force or extension, without the need to generate a new model for each new experiment.

Usual analysis of SMFS experiments involves a top-down approach where the goal is to de-convolute the effects of handles, obtain the barrier height along a reaction coordinate between the "folded" and "unfolded" states of the molecule, and probe rare molecular transitions such as the dissociation of a ligand or the unfolding of a domain[223–229]. The classic Bell-Zhurkhov[232] model gives the kinetic rate of rupture of a molecule, $k(F)$, in terms of a force (F) dependent barrier height ($k(F) = k_0 e^{\beta F x^*}$). Here the parameters, x^* is the distance between the free energy minimum and the barrier along the reaction coordinate, k_0 is the intrinsic rate which can be recovered from the experimental data, and $\beta = 1/k_B T$, k_B is the Boltzmann factor and T is the temperature. A modified model by Evans and Ritchie[233] giving the distribution of forces for unfolding a molecule has been widely used to analyze data in force-ramp experiments. However, both models assume that the barrier

does not vary with the applied force. Dudko and coworkers developed a method also based on Kramer's theory that includes a barrier that moves with an applied force to obtain an analytical expression for the force distribution for certain landscape profiles[224], namely, the cusp or linear cubic single-well free energy surfaces as function of the pulling coordinate. More recently Zhang and Dudko[227] extended the formalism to treat systems with multiple barriers subjected to time dependent force to recover the force dependent rate-maps from the experimental data. The methods described above are applicable when all the intermediate states lie on a single pathway. However, when multiple states and competing pathways are involved which may have been triggered differently due to the stretching force, these will be discussed with a simple example in this chapter and there exists no general formalism to describe the landscape.

Complementing the top-down approaches *in silico* studies have provided valuable insights into the mechanics of molecules under tension[229,234,235]. Wales *et al.*[236] applied geometry optimization techniques to study the energy landscape of two proteins (L and G) as a function of a static pulling force using coarse-grained models. Methods such as steered molecular dynamics simulations[205] can provide a window into the dynamics of a molecule under stretching in atomistic detail, however, finding the relevant events in folding/unfolding pathways and interpreting the calculated energy barriers encountered by the molecule remains an important challenge. Markov state models (MSMs) parametrized by molecular dynamics (MD) simulations have been used widely to study folding/unfolding/dynamics of peptides in the absence of external pulling forces[62]. Our objective is to develop a method to predict the change in the network model at a given stretching condition (probe separation or static force) based on underlying connections between the thermodynamics and kinetics of the system. Analogous to kinetic theories in electrochemistry (the Butler Volmer model[237,238]), we find that the forward and backward rates between two states depends on an intrinsic rate that quantifies the kinetic facility available for the pathway and a quantity that we term the thermodynamic disposition for the move. Upon stretching, as in a force ramp experiment, the barrier for a particular kinetic pathway can change by an amount that can be understood in terms of a simple theory if the molecule has sufficient time to visit multiple states while the average force/extension varies slowly. A time-dependent MSM based on this formalism is the outcome of this study. Actually, the connection between MSMs at different stretching conditions via the Bell-Evans-Polanyi (BEP)[207,208] principle is the cornerstone of this study. To get a detailed understanding of how the dynamics

is influenced by a pulling force (static or dynamic), one does not require separate MD calculations at various stretching conditions rather the time-dependent MSM adapted to recover the dynamics from different FS protocols. Here we propose a theory that can be used to analyse and predict changes in the kinetic network of a single molecule under a pulling force as in a SMFS setup and most significantly, provide a microscopic picture of the dynamics of the system. One straightforward implication is that parts of the kinetic network and the free energy landscapes that are challenging to sample in the absence of an applied force can then be recovered.

The assumptions made above are tested by comparing force-extension curves of deca-alanine in vacuum under a SMFS position clamp setup from the theory to the ones obtained with all-atom MD using the same force-field. The deca-alanine molecule, a classic example of an alpha-helix with minimal interference from side-chain interactions is chosen as a platform to test the framework. Markov state model(s) describing the kinetics of deca-alanine stretched to different anchor separations are first constructed bottom-up using molecular dynamics calculations. These measurements yield the elastic, thermodynamic and kinetic parameters of the time-dependent MSM. Also, we have attempted to predict the dynamics in the Force-ensemble experiment without actual simulations at constant force conditions to show that it is possible to convert between network models applicable for each type of SMFS experiment without the need to generate a new model for each new experiment.

5.2 Theoretical Basis

5.2.1 Markov State Models of Force Sepctroscopy Setup

Consider a molecule tethered at both ends by harmonic springs to two fixed anchor points mimicking a dual optical trap setup (see Fig. 5.1). The spring constant of the device molecule interaction used in this study is 0.86 kcal·mol^{-1}Å^{-2} which is significantly higher than typical experimental values. In this analysis, we assume a position-clamp setup in which the instantaneous force varies with the anchor separation as in the absence of a feedback loop. The anchor separation denoted by d, is the control parameter. We assume that the molecule can exist in stable conformations corresponding to basins in the free energy landscape such that transitions between states are infrequent, that is, the Markovian assumption is valid, a Markov State

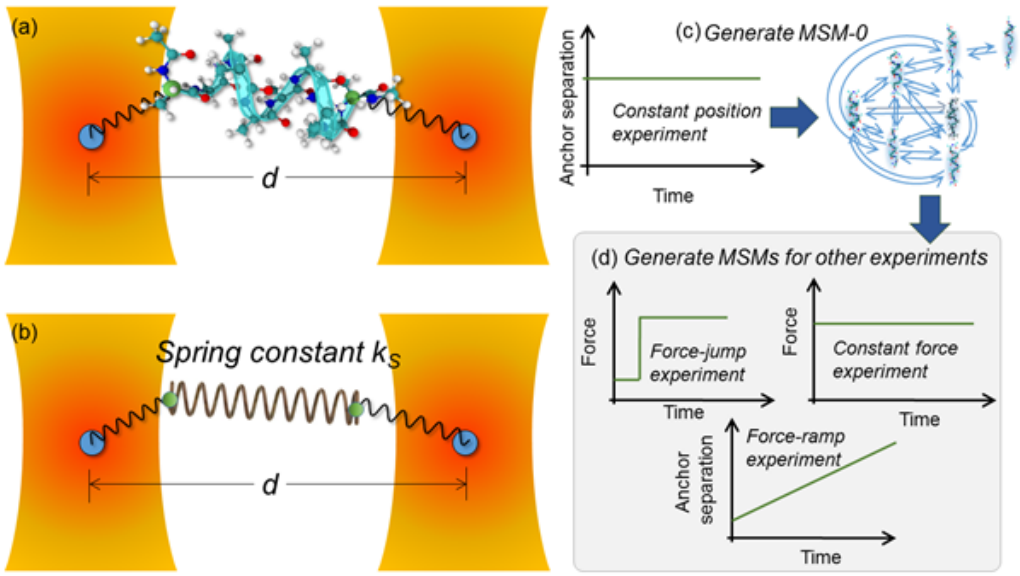

Figure 5.1: (a) Illustration of an alanine decapeptide in a dual optical trap setup. The oligopeptide is connected by harmonic springs to two static anchor points at the two ends. In the simulation setup, the C_α atoms of the first and last residues are harmonically restrained to two fixed points to mimic the setup. The elastic rod model corresponding to the peptide conformation in (a) is depicted in (b). An elastic cylinder of the same length (distance between the first and last C_α atoms) as the peptide is suspended by springs between two fixed points. (c) MSM-0 constructed for constant position experiment can be applied to other stretching experiments like (d) constant force, force-ramp and force-jump experiment.

Model (MSM) given by

$$d\pi(d)/dt = T(d)\pi(d) \tag{5.1}$$

offers a means to approximate the dynamics in terms of state-to-state transitions where $\pi(d)$ is the occupation probability vector for the states and $T(d)$ is the rate matrix. We assume that the same state definitions can be applied consistently across a range of stretching conditions. At equilibrium at a given anchor separation, d, the occupation $\pi_S(d)$ of state S relative to that of a reference state R given by $\pi_R(d)$ at separation d is,

$$\pi_S(d)/\pi_R(d) = \exp(-\beta \Delta A_{S,R}(d)) \tag{5.2}$$

where $\Delta A_{S,R}(d) = A_S(d) - A_R(d)$ is the free energy difference between the two states at the anchor separation d, and $A_{S/R}(d)$ is the free energy of state S/R at d.

Molecular dynamics simulations at specified values of the anchor separation can be used to obtain a catalogue of states along with the rate matrix and equilibrium occupancies. The occupancies and kinetic rates depend on separation d. Thus a

kinetic theory is required that can provide this information if we want to model a pulling experiment where d is a parameter that varies with time as it is not feasible to recalibrate the MSM at every possible value of the anchor separation. We therefore, seek a model that can be used to predict the changes in the kinetic network over a range of anchor separations using a small number MSMs at different spacings.

5.2.2 Bell-Evans-Polanyi Principle

In physical chemistry, the Bell-Evans-Polanyi (BEP)[207,208] principle states that the difference in activation energy between two reactions of the same family is proportional to the difference of their enthalpy of reaction. The relationship can be expressed as,

$$E_a = A + \alpha \Delta H \tag{5.3}$$

where E_a is the activation energy of a reference reaction of the same class, ΔH is the enthalpy of reaction and α characterizes the position of the transition state along reaction coordinate. The BEP model is a linear energy relationship that serves as an efficient way to calculate activation energy of many reactions within a distinct family. The activation energy may be used to characterize the kinetic rate parameter of a given reaction through application of the Arrhenius equation. According to this, a rate constant k is the product of a pre-exponential factor A and an exponential term, written as,

$$k = Ae^{-E_a/RT} \tag{5.4}$$

where R is the gas constant, E_a is the activation energy and T is the temperature.

The BEP model assumes that the pre-exponential factor of the Arrhenius equation and the position of the transition state along the reaction coordinate are the same for all reactions belonging to a particular reaction family. Actually, we test the validity of the BEP principle in the context of MD simulations of peptide under tension.

5.2.3 Estimating the Kineic Rate of Conformational Transition from BEP Principle

We now seek a relation between the kinetic rates and the anchor separations along the lines of the BEP principle[207,208]. Let k_f and k_b denote the forward and reverse rates for transition between states S and R. Using transition state theory (see

Appendix A.1), the forward and reverse rates are given by

$$k_f(d) = \nu_f \exp(-\beta \Delta A_f(d))$$
$$k_b(d) = \nu_b \exp(-\beta \Delta A_b(d)) \tag{5.5}$$

respectively, $\Delta A_f(d)$ and $\Delta A_b(d)$ are the forward and reverse free energy barriers for the transition from state S to R, that is, the difference between the free energy at the saddle point and the basins of S and R (see Fig. 5.2(a)). The pre-factors

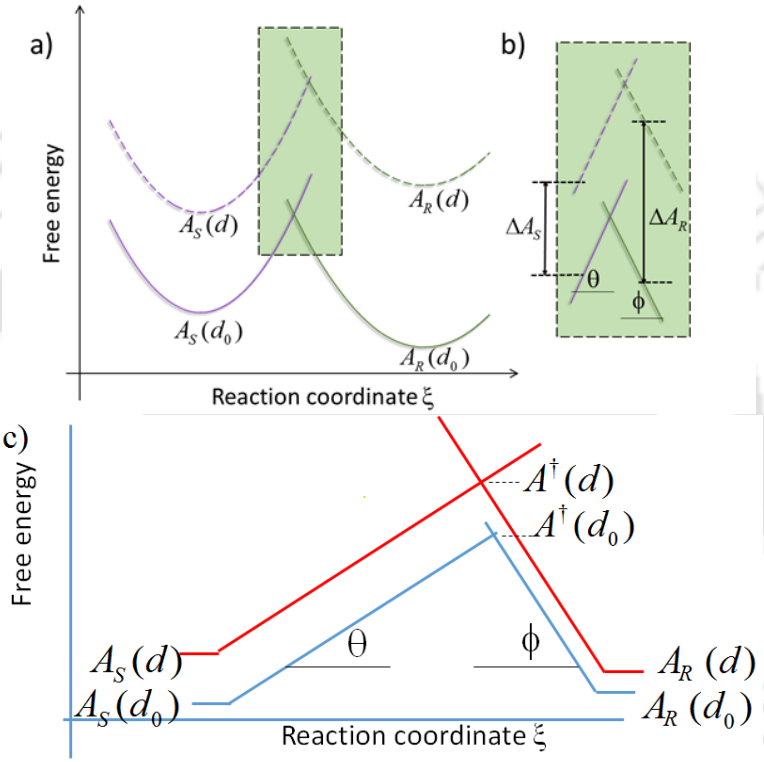

Figure 5.2: (a) Free energy profile against the reaction coordinate ξ for the transition from state S to R at anchor separation d_0 and d. (b) The shift in the location of the saddle point for the transitions is obtained assuming the free energy to be a linear function of ξ in each basin. (c) The amplified picture of the energy profile along reaction coordinate ξ near the saddle point.

ν_f and ν_b are assumed to be constant, independent of d and the temperature T. We suppose that the states S and R lie on a one-dimensional reaction coordinate, ξ which, unlike previous theories does not necessarily represent the pulling direction in general. Figure 5.2(b) depicts the variation of the free energy as a function of ξ in the vicinity of the saddle point. We assume the free energy to have a linear dependence on the coordinate in the vicinity of the saddle point, while $\tan\theta$ and $\tan\phi$ correspond to the effective gradient of the free energy along ξ for the states S

and R respectively. In Fig. 5.2(c) $A^{\dagger}(d)$ is the saddle point free energy and $A_{S/R}(d)$ is the free energy of the basins S/R at trap separation d respectively. From Fig. 5.2 we see that $\Delta A_f(d) = A^{\dagger}(d) - A_S(d)$ and $\Delta A_b(d) = A^{\dagger}(d) - A_R(d)$. Now, at equilibrium, the ratio of the forward to reverse kinetic rates is given by,

$$\frac{k_f(d)}{k_b(d)} = \frac{\nu_f \exp(-\beta \Delta A_f(d))}{\nu_b \exp(-\beta \Delta A_b(d))} = \frac{\exp\big(-\beta(A^{\dagger}(d) - A_S(d))\big)}{\exp\big(-\beta(A^{\dagger}(d) - A_R(d))\big)} = \exp\big(-\beta(A_R(d) - A_S(d))\big). \tag{5.6}$$

To get the thermodynamic consistency in Eq. (5.6), we infer that prefactors for the forward and backward reaction should be equal, that is, $\nu_f = \nu_b$. Defining $\Delta A_{SR}(d) = A_R(d) - A_S(d)$, this equation can be rewritten as,

$$\frac{k_f(d)}{k_b(d)} = \exp(-\beta \Delta A_{SR}(d)) \tag{5.7}$$

where $\Delta A_{SR}(d)$ denotes the change in free energy for a transition from state S to R at anchor separation d.

Now, let us evaluate the change in free energy barrier with the change in anchor separation d. We treat the MSM at a particular anchor spacing, d_0 (= 16 Å in our example) as the reference MSM. The value of the free energy for the state S along ξ is approximated as,

$$A_S(\xi; d_0) = \tan(\theta)\xi + C_S \tag{5.8}$$

for reference MSM. The effect of increasing the anchor separation to d is to lift the free energy profile vertically by an amount ΔA_S leading to the approximation,

$$\begin{aligned} A_S(\xi; d) &= \tan(\theta)\xi + C_S + (A_S(d) - A_S(d_0)) \\ &= \tan(\theta)\xi + C_S + \Delta A_S(d) \end{aligned} \tag{5.9}$$

at anchor separation d. Similarly for the state R, the free energy as a function of ξ is given by,

$$\begin{aligned} A_R(\xi; d) &= -\tan(\phi)\xi + C_R + (A_R(d) - A_R(d_0)) \\ &= -\tan(\phi)\xi + C_R + \Delta A_R(d). \end{aligned} \tag{5.10}$$

Now, let us focus on the saddle point. The free energies of the state S at the location of the saddle point (denoted as $\xi^{\dagger}$) with anchor separation d_0 and d are given by,

$$A_S^{\dagger}(\xi^{\dagger}(d_0)) = A_S(d_0) + \Delta A_f(d_0) \quad = (\tan\theta)\xi^{\dagger}(d_0) + C_S \tag{5.11}$$

and

$$A_S^{\dagger}(\xi^{\dagger}(d)) = A_S(d) + \Delta A_f(d) \quad = (\tan\theta)\xi^{\dagger}(d) + C_S + (A_S(d) - A_S(d_0)) \tag{5.12}$$

respectively. Similarly, for anchor spacing d and d_0, the free energies for the energy landscape R at the saddle point $\xi^{\dagger}$ are,

$$A_R^{\dagger}(\xi^{\dagger}(d_0)) = A_R(d_0) + \Delta A_b(d_0) \quad = (-\tan\phi)\xi^{\dagger}(d_0) + C_R \tag{5.13}$$

and

$$A_R^{\dagger}(\xi^{\dagger}(d)) = A_R(d) + \Delta A_b(d) \quad = (-\tan\phi)\xi^{\dagger}(d) + C_R + (A_R(d) - A_R(d_0)). \tag{5.14}$$

From Eq. (5.11)-Eq. (5.14) we have the following four terms

$$A_S(d_0) + \Delta A_f(d) = (\tan(\theta)\xi^{\dagger}(d) + C_S \tag{5.15a}$$
$$A_S(d_0) + \Delta A_f(d_0) = (\tan\theta)\xi^{\dagger}(d_0) + C_S \tag{5.15b}$$
$$A_R(d_0) + \Delta A_b(d) = (-\tan\phi)\xi^{\dagger}(d) + C_R \tag{5.15c}$$
$$A_R(d_0) + \Delta A_b(d_0) = (-\tan\phi)\xi^{\dagger}(d_0) + C_R. \tag{5.15d}$$

The terms in Eq. (5.15) are used to obtain the shift in the saddle point location due to varying anchor separation, $\xi^{\dagger}(d) - \xi^{\dagger}(d_0)$ as,

$$\begin{aligned} \Delta A_f(d) &= \Delta A_f(d_0) + (\tan\theta)(\xi^{\dagger}(d) - \xi^{\dagger}(d_0)) \\ \Delta A_b(d) &= \Delta A_b(d_0) + (-\tan\phi)(\xi^{\dagger}(d) - \xi^{\dagger}(d_0)). \end{aligned} \tag{5.16}$$

By rearranging the terms in Eq. (5.16), finally, we have

$$\frac{\Delta A_f(d) - \Delta A_f(d_0)}{\tan\theta} = \frac{\Delta A_b(d) - \Delta A_b(d_0)}{-\tan\phi}. \tag{5.17}$$

Then by using Eq. (5.17) we can derive the barrier expression as,

$$\begin{aligned} &\Delta A_f(d) - \Delta A_b(d) = [A^{\dagger}(d) - A_S(d)] - [A^{\dagger}(d) - A_R(d)] = A_R(d) - A_S(d) = \Delta A_{SR}(d) \\ &\Rightarrow \Delta A_b(d) = \Delta A_f(d) - \Delta A_{SR}(d) \\ &\Rightarrow \Delta A_f(d) = \Delta A_b(d) + \Delta A_{SR}(d). \end{aligned} \tag{5.18}$$

Finally,

$$\begin{aligned}
&\frac{\Delta A_f(d) - \Delta A_f(d_0)}{(\tan\theta)} = \frac{\Delta A_b(d) - \Delta A_b(d_0)}{(-\tan\phi)} = \frac{\Delta A_f(d) - \Delta A_{SR}(d) - \Delta A_f(d_0) + \Delta A_{SR}(d_0)}{(-\tan\phi)} \\
&\Rightarrow \frac{\Delta A_f(d) - \Delta A_f(d_0)}{(\tan\theta)} = \frac{\Delta A_f(d) - \Delta A_f(d_0)}{(-\tan\phi)} - \frac{\Delta A_{SR}(d) - \Delta A_{SR}(d_0)}{(-\tan\phi)} \\
&\Rightarrow \left\{\Delta A_f(d) - \Delta A_f(d_0)\right\}\left[\frac{1}{\tan\theta}\right] = \frac{\Delta A_f(d) - \Delta A_f(d_0)}{(-\tan\phi)} + \frac{\chi(d)}{\tan\phi} \\
&\Rightarrow \left\{\Delta A_f(d) - \Delta A_f(d_0)\right\}\left[1 + \frac{\tan\theta}{\tan\phi}\right] = \chi(d)\frac{\tan\theta}{\tan\phi} \\
&\Delta A_f(d) = \Delta A_f(d_0) + \frac{\tan\theta}{\tan\theta + \tan\phi}\chi(d) \\
&\Delta A_f(d) = \Delta A_f(d_0) + \alpha_f\chi(d).
\end{aligned} \tag{5.19}$$

Similarly for backward pathways,

$$\begin{aligned}
&\frac{\Delta A_f(d) - \Delta A_f(d_0)}{(\tan\theta)} = \frac{\Delta A_b(d) + \Delta A_{SR}(d) - \Delta A_b(d_0) - \Delta A_{SR}(d_0)}{(\tan\theta)} = \frac{\Delta A_b(d) - \Delta A_b(d_0)}{(-\tan\phi)} \\
&\Rightarrow \frac{\Delta A_b(d) - \Delta A_b(d_0)}{\tan\theta} + \frac{\Delta A_{SR}(d) - \Delta A_{SR}(d_0)}{\tan\theta} = \frac{\Delta A_b(d) - \Delta A_b(d_0)}{-\tan\phi} \\
&\Rightarrow \frac{\Delta A_b(d) - \Delta A_b(d_0)}{\tan\theta} + \frac{\Delta A_b(d) - \Delta A_b(d_0)}{\tan\phi} = -\frac{\Delta A_{SR}(d) - \Delta A_{SR}(d_0)}{\tan\theta} \\
&\Rightarrow \left\{\Delta A_b(d) - \Delta A_b(d_0)\right\}\left[1 + \frac{\tan\theta}{\tan\phi}\right] = -\chi(d) \\
&\Rightarrow \Delta A_b(d) - \Delta A_b(d_0) = -\chi(d)\left[\frac{\tan\phi}{\tan\theta + \tan\phi}\right] = -\chi(d)\left[1 - \frac{\tan\theta}{\tan\theta + \tan\phi}\right] = -\chi(d)(1 - \alpha_f) \\
&\Rightarrow \Delta A_b(d) = \Delta A_b(d_0) - (1 - \alpha_f)\chi(d) \\
&\Rightarrow \Delta A_b(d) = \Delta A_b(d_0) - \alpha_b\chi(d).
\end{aligned} \tag{5.20}$$

In compact notation we write,

$$\begin{aligned}
\Delta A_f(d) &= \Delta A_f(d_0) + \alpha_f\chi(d) \\
\Delta A_b(d) &= \Delta A_b(d_0) - \alpha_b\chi(d)
\end{aligned} \tag{5.21}$$

where

$$\chi(d) = \Delta A_{SR}(d) - \Delta A_{SR}(d_0). \tag{5.22}$$

The term $\chi(d)$ which we call the mechanical disposition for the transition, gives the difference between the free energy change that occurs for the transition between the states S and R as the anchor separation is varied. The free energy barrier increases

when $\chi > 0$. The χ term measures the thermodynamic preference for the states upon application of a force relative to the rate at separation d_0. Other two terms in Eq. (5.21), $\alpha_f = \frac{\tan\theta}{\tan\theta+\tan\phi}$ and $\alpha_b = \frac{\tan\phi}{\tan\theta+\tan\phi}$ depends on the geometry of the barrier in the vicinity of the saddle point.

Now, from Eqs. (5.5) and (5.21) we can relate the kinetic rate at separation at d to the rate at reference separation d_0 as,

$$\begin{aligned} & k_f(d) = \nu_f \exp(-\beta \Delta A_f(d)) = \nu_f \exp(-\beta\{\Delta A_f(d_0) + \alpha_f \chi(d)\}) \\ & \Rightarrow k_f(d) = \nu_f \exp(-\beta\{\Delta A_f(d_0)\}) \exp(-\beta \alpha_f \chi(d)) \\ & \Rightarrow k_f(d) = k_f(d_0) \exp\big(-\beta \alpha_f \chi(d)\big). \end{aligned} \tag{5.23}$$

$$\begin{aligned} & k_b(d) = \nu_b \exp(-\beta \Delta A_b(d)) = \nu_b \exp(-\beta\{\Delta A_b(d_0) - (1-\alpha_f)\chi(d)\}) \\ & \Rightarrow k_b(d) = \nu_b \exp(\beta\{\Delta A_b(d_0)\}) \exp(\beta(1-\alpha_f)\chi(d)) \\ & \Rightarrow k_b(d) = k_b(d_0) \exp\big(\beta(1-\alpha_f)\chi(d)\big) \\ & \Rightarrow k_b(d) = k_b(d_0) \exp\big(\beta \alpha_b \chi(d)\big). \end{aligned} \tag{5.24}$$

Finally, the forward kinetic rate of the transitions between the two states at separation, d related to that at the reference separation d_0 is given as

$$k_f(d) = k_f(d_0) \exp(-\beta \alpha_f \chi(d)) \tag{5.25}$$

and the backward kinetic rate is given as,

$$k_b(d) = k_b(d_0) \exp(\beta \alpha_b \chi(d)). \tag{5.26}$$

For sake of brevity, the pair of states has not been explicitly mentioned in the notation for k_f, k_b, ν_f, ν_b, $\Delta A_f(d)$, $\Delta A_b(d)$, α_f, α_b and $\chi(d)$. When the trap separation d is a function of time, the mechanical disposition and the rate matrix become time-dependent. The kinetic rate parameters $k_f(d_0)$, $k_b(d_0)$ and the symmetry parameter α are estimated from MSMs constructed at constant d, as described next.

In practice, to get the kinetic rates at any specified value of the anchor separation, one needs the factor α and the mechanical disposition (χ). Here it is convenient for us to consider the forward and backward pathways separately instead of reporting α_f and α_b for a pair of states to check the validity of the proposed kinetic model. In order to accurately predict the kinetic rates within a range of anchor separation, we need to generate a few MSMs from which the co-factor α_f may be calculated using

the equation,

$$\ln k_f(d) = \ln k_f(d_0) + \frac{\alpha_f}{k_B T} \left(\Delta A_{SR}(d) - \Delta A_{SR}(d_0)\right) \tag{5.27}$$

from the plot of the kinetic rates as a function of the free energy differences $\left[\chi(d) = \Delta A_{SR}(d) - \Delta A_{SR}(d_0)\right]$ between two selected states. Here $\chi(d)$ at d can be obtained using the quadratic dependence of free energy on d as discussed in the next section.

Choosing a reference value of anchor separation, d_0, a detailed MSM is constructed that contains relevant states alongside the kinetic parameters of Eq. 5.25. We term this *master*-MSM as MSM(d_0) or simply MSM-0. Note that MSM-0 contains a list of states and kinetic pathways that would be relevant for certain/entire range of extensions to be sampled with the model, the associated kinetic parameters include $k_f(d_0)$ and α_f, and the thermodynamic parameter (χ). MSM-0 is a precursor for Time-Dependent MSM (TD-MSM) at a variety of stretching conditions.

5.2.4 Model for Calculating Free-energy Dependence on Anchor Separation

A kinetic model should be consistent with thermodynamics. In a constant-d experiment, the probability of residing in a state S, $\pi_S(d)$, is given by

$$\pi_S(d) = \frac{\int_{r \in S} \exp(-\beta E) dr \Big|_d}{Z(d)} \tag{5.28}$$

where $Z(d)$ is the partition function at trap separation d involving the integral over entire phase space that are accessible at d and E is the energy term. Hence we can express the free energy (A_S) of the states (S) as a function of the probe separation (d) as,

$$A_S(d) = -k_B T \ln \pi_S(d) - k_B T \ln Z(d). \tag{5.29}$$

At equilibrium, the detailed balance requires,

$$k_f(d) \pi_S^{eq}(d) = k_b(d) \pi_R^{eq}(d). \tag{5.30}$$

Combining Eqs. (5.7) and (5.30) we have, $\pi_R(d)/\pi_S(d) = \exp(-\beta \Delta A_{SR}(d))$ which is also consistent with Eq. (5.29). Thus, the free energy difference ($A_{SR} = A_R - A_S$)

can be calculated on the basis of the equilibrium state occupations from the MSMs at the different values of d as,

$$\Delta A_{SR}(d) = -k_B T \ln(\pi_R^{eq}(d)/\pi_S^{eq}(d)). \tag{5.31}$$

Hence the free energy difference (note that, $\Delta A_{S,R} = A_S - A_R$) between two states S and R is expected to vary with a change of the anchor separation from d to d_0 as,

$$A_{S,R}(d) - A_{S,R}(d_0) = -k_B T \left(\ln \frac{\pi_S(d)}{\pi_R(d)} - \ln \frac{\pi_S(d_0)}{\pi_R(d_0)} \right). \tag{5.32}$$

When the molecule possesses a large stiffness, in the simplest scenario, it is reasonable to assume that the energetic contributions to the free energy will dominate over entropy. The main contribution to the free energy change arises from the potential energy associated with the states as the molecule is pulled. In the simplest case, we assume that as in the case of energy, the free energy of state S has a quadratic dependence on the anchor separation. Also, since $\Delta A_{S,R} = \Delta A_{S,P} - \Delta A_{R,P}$ suggests that we may write an expression for $A_S(d)$, the free energy associated with a state at an anchor separation d as,

$$A_S(d) = c_0^S + c_1^S d + c_2^S d^2 - k_B T \ln Z(d). \tag{5.33}$$

More generally the quadratic dependence of free-energy on d can be written as a function of time,

$$A_S(d) = c_0^S(d,t) + c_1^S(d,t) d + c_2^S(d,t) d^2 - k_B T \ln Z(d,t) \tag{5.34}$$

Here the coefficients c_0^S, c_1^S, c_2^S may be obtained from fitting -$k_B T \ln \pi_S(d)$ *vs* d curve obtained from the MSMs as shown in Fig. 5.8 of Sec. 5.4.3.

5.3 Methods

5.3.1 System Setup

The model of a capped deca-alanine with acetylated N-terminus and amidated C-terminus was generated using the 104-atom helical model of Ref. 205 as the initial configuration. The C_α atoms at the two ends (residues 1 and 10) are tethered to two anchor points by harmonic restraints with a spring constant of 0.86 kcal/mol/Å^2 (Fig. 5.1). The anchor separation d is kept fixed during the construction of a MSM

and fluctuations in molecular extension and forces are measured. The MSM was generated using the adaptive algorithm, programmed state constrained MD (PSC-MD) as described in Sec. 4.4 of previous chapter.

5.3.2 Simulation Protocols

All MD simulations were performed with NAMD 2.9[184] with the CHARMM36[206] force fields. Temperature was held at 300 K using a Langevin thermostat. Bonds involving hydrogen atoms were constrained to their equilibrium values using RATTLE[129]. An integration time step of 2 fs was used. Steered MD simulations were performed with the deca-alanine (without the tethers) for several nano-seconds to obtain a preliminary collection of unfolded structures. These configurations were provided as inputs to the Swarm MD procedure. In addition SMD simulations with various pulling speeds (1, 0.1 and 0.01 Å/ns) were performed with selected initial states to compare with the TD-MSM results.

5.3.3 MSM Construction Protocols Using Swarm MD and Programmed State Constrained (PSC) MD

MSMs (Fig. 5.3) are constructed on-the-fly with 90% confidence using thousands of short-MD calculations that are run in parallel as part of the PSC-MD method. PSC-MD is used to initiate trajectories from rare states until the transitions from the state have been mapped adequately, ensuring that the MSMs are complete with 90% confidence (or, $\delta = 0.1$) (see Sec. 3.3.2 of Chapter 3). The independent trajectories share and contribute to a common catalogue of states but are otherwise not coupled. States are determined by comparing the backbone atoms after aligning the molecule using Kabsch algorithm. A tolerance of 3 Å was found to be suitable for identifying the states. Standard practices, such as assessing the quality of the Markov approximation[161] were employed (discussed in Sec. 4.2 of previous chapter) besides verifying the presence of a sharp single-peaked distribution for force and end-to-end distance for each Markov state. MD snapshots were collected every 0.2 ps because of rapid interconversion between the states. A transition was said to have successfully occurred when the system continues to reside in the new state for at least 1.2 ps after the transition was detected in the MD trajectory. This prevents recrossing events to be counted as transitions. Kinetic rates were determined using a maximum likelihood analysis. PSC-MD procedure was used to build and progres-

sively refine the MSMs at various anchor separation $d = 16, 18, 20, 22, 23, 24, 26$ Å by increasing the validity time associated with it as explained in previous chapter.

5.4 Results and Discussions

Master-MSM (MSM-0) - Markov State Model at Reference Anchor Spacing d_0

Regular MSMs for the deca-alanine system constructed using MD calculations for a range of anchor separations from 16 to 26 Å demonstrate a large topological variation with d. As expected, the dominant state(s) were found to change across the range of anchor separations studied. These topological differences arise due to missing states and pathways as they become relevant to the dynamical evolution at the new stretching condition. Although the total number of states discovered through the self-learning algorithm was found to be large (810), the actual number of states relevant at any stretching condition was quite small. In particular, we found that ten states (numbered 1, 2, 3, 4, 5, 6, 9, 17, 113 and 288) that dominate the dynamics of the system over the entire range of anchor separations studied. However, if the folding/unfolding process is under investigation, low occupancy states numbered 6 and 17 are found to play an important role. At small extensions (d<20 Å), the compact states 1 and 288 (see Fig. 4.9(b) in Chapter 4) have high occupancies while at higher extensions, the stretched configurations shown in Fig. 5.3 dominate. Figure 5.3 shows a detailed MSM constructed that contains 10 states, choosing $d_0 =$ 16 Å, alongside the kinetic parameters of Eq. (5.25). We term this master-MSM as MSM(d_0) or simply MSM-0. This master-MSM forms the basis for generating the TD-MSM.

The Composition of the MSM is Found to Vary Significantly with the Anchor Separation

At an anchor separation of 16 Å, only two states accounted for 95% of the total occupancy, whereas at 24 Å anchor separation more states were required to capture the same occupation probability. A total of 10 states were required to construct the TD-MSM in the range 16 Å to 26 Å anchor separations, while maintaining a 95% threshold of occupancy at any anchor separation. The network comprising of all the ten relevant states is displayed in Fig. 5.3 with representative images for all the states.

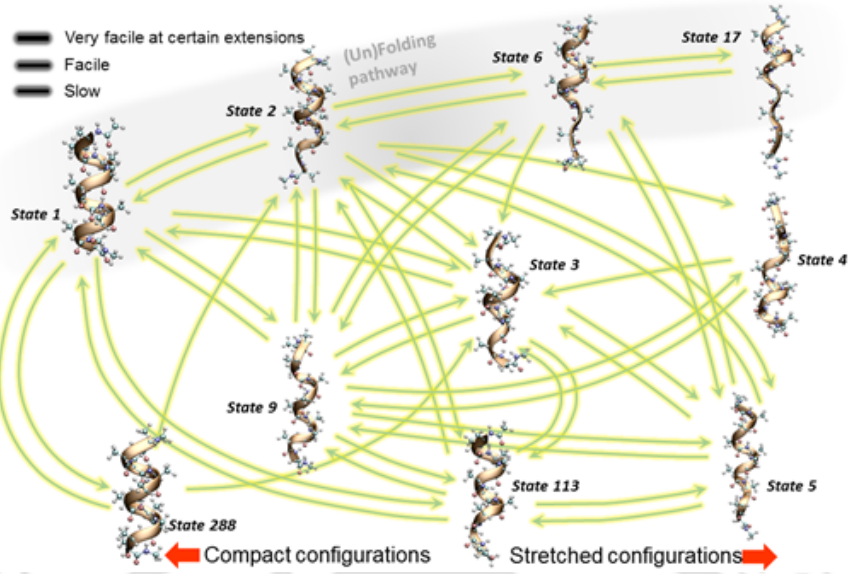

Figure 5.3: The Markov state model (MSM-0) for the deca-alanine system (N- terminus at the top) at the anchor spacing of 16 Å. The arrows represent the pathways observed in the MSMs constructed at anchor spacings between 16 to 26 Å.

Figure 5.4 shows the variation of the probability of occupation of the four dominant states (1, 2, 3 and 9) with increasing validity time of MSM for different anchor spacing. The probabilities are saturated/stabilized between 10 and 100 ps validity time. At anchor separation 22 Å, states 2 and 3 show significant occupation probability (comparable to state 1). At higher anchor spacing state 1 is almost absent. State 3 is the dominant configuration for spacing between 22 and 25 Å. Indeed, states 2, 3 and 9 may be considered to be intermediate hubs in any unfolding pathway.

An explosion in the number of core states is witnessed from 36 to 78 states between 22 to 25 Å. Most of the dominant states are those that had very low probabilities of occupation when the anchor spacing was small. Many of these structures were found to deviate significantly from the α-helical structure of the peptide. With increasing anchor spacing, the net number of hydrogen bonds was found to decrease, with the hydrogen bonds at the C-terminal end being disrupted first, an exception being state 4 in which the N-terminal unravels. However, a part of the decrease in the number of hydrogen bonds from the α-helical structure (state 1) was offset by the formation of new hydrogen bonds, commensurate with the 3_{10} helix structure. Correspondingly, the validity time plummets by almost 10 times from 12 ns at 22 Å separation to 1.5 ns at 25 Å separation for 0.5 μs long MD trajectory. The

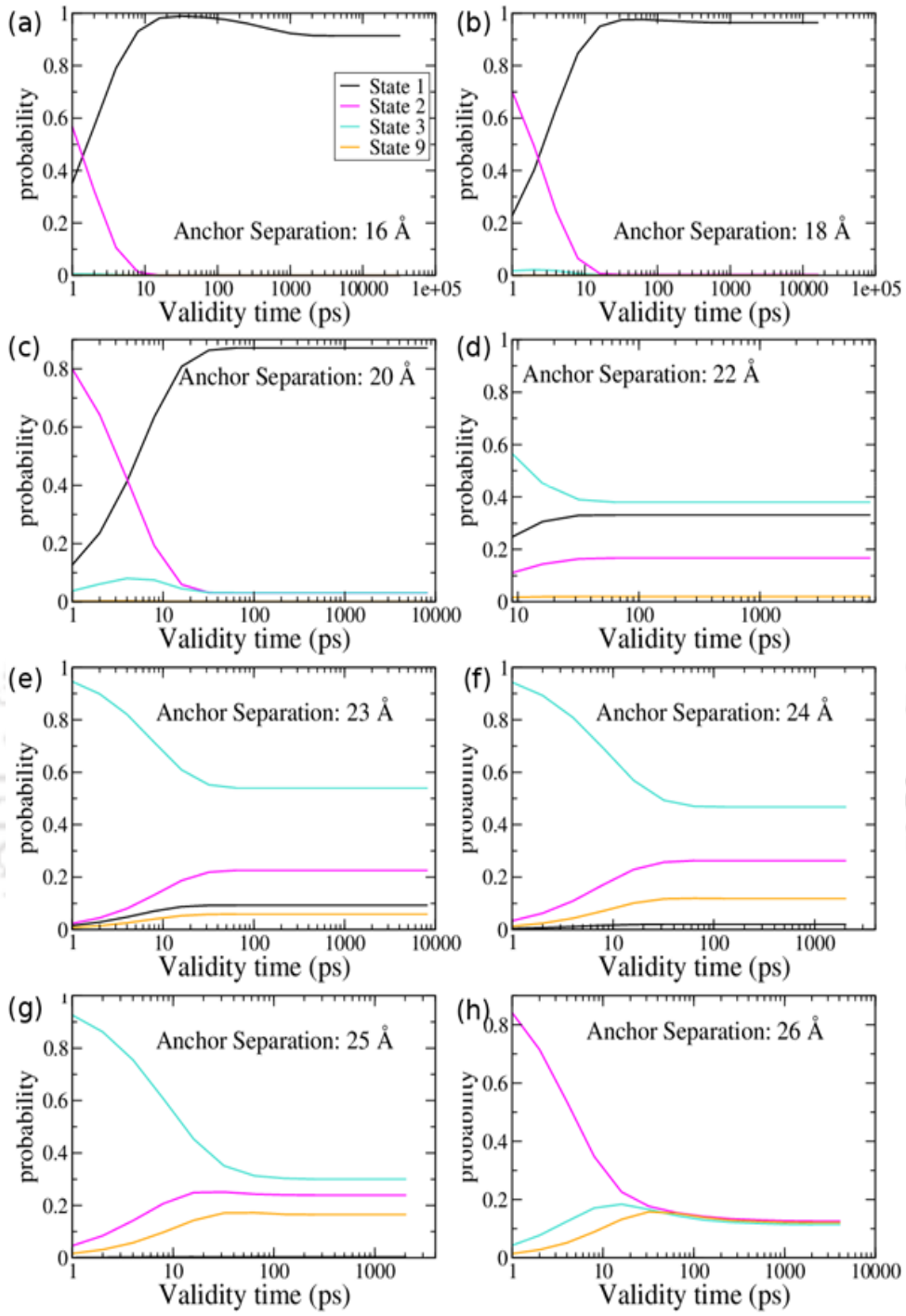

Figure 5.4: Variation of the occupation probability of four dominant states with increasing validity time of the MSM for a range of anchor separations (between 16 and 26 Å).

rate constants involving the core states vary over several orders magnitude between 10^{-1}-10^{-5} ps^{-1}.

Even with the simple example chosen in this study, we find that the (coarse-grained) two-state model usually invoked in experimental analysis of folding/unfolding transitions may not effectively capture the dynamics.

5.4.1 Force Model

We assume that the deca-alanine molecule behaves as a harmonic spring with a spring constant associated with a state that depends intrinsically on the chemical interactions, namely the structure and interactions. However, in several previous studies of biomolecular systems, the native or extended state of the system have been approximated as harmonic oscillators[239–241]. Now, we require a force-separation relation for each state to compute the average forces acting on the molecule. Let us consider the molecule in state S to be modelled as an elastic rod with spring constant, k_S and an equilibrium length, l_S^{eq} in the absence of pulling forces, as shown in Fig. 5.1(b). The molecule extends due to the applied force according to the spring model while it resides in a particular state. SC-MD calculations can provide the average end-to-end distance versus d and the average force $F_S(d)$ experienced in state S. Overall, the stiffness of deca-alanine was found to lie in the range 20-80 pN/Å across all the states visited in our calculations. The potential energy ($E_S(d)$) of the molecule-device system comprises of the sum of the molecular potential energy in a state S and the spring energy of the tethers. The former is given by $U_S(d) = U_0 + (1/2)k_S x_S(d)^2$ where d is the extension with respect to the equilibrium length, l_S^{eq}, in the absence of applied force and the latter is given by, $T_S(d)$ represented as harmonic springs with a spring constant of k_{tether} at the two ends of the molecule. The force may be evaluated using the average elongation of the molecule in state S along the direction of stretching, $l_S(d)$, as a function of the anchor spacing d, in the expression $F_S(d) = k_S(l_S(d) - l_S^{eq})$. We assume that all the terms, the force, the fluctuating elongation of the molecule as well as the equilibrium length, have been projected in the direction of stretching (along z direction in our case). The difficulty in using the expression above arises because both the force and the elongation are fluctuating quantities and are not controlled directly in the setup described. An alternate view of the setup is to consider the entire molecule-tether system as an effective spring. The net force vanishes when the anchor spacing $d = l_S^{eq}$. Hence, any deviation of the anchor separation d from the equilibrium length of the molecule

will result in a force having a form,

$$F = k_S^{eff}(d - l_S^{eq}) \tag{5.35}$$

where k_S^{eff} represents the effective stiffness of the molecule-tether system while the molecule resides in state S (not simply spring constant of the molecule).

However, we must exercise caution in applying the model of Eq. (5.35) since the actual setup as shown in Fig. 5.1(b) is clearly not 1-dimensional and hence an equivalent spring constant such as $1/k_S^{eff}=1/k_S+2/k_{tether}$ can not be applied here as it is relevant for 1-models in the absence of thermal fluctuations. Also temperature-

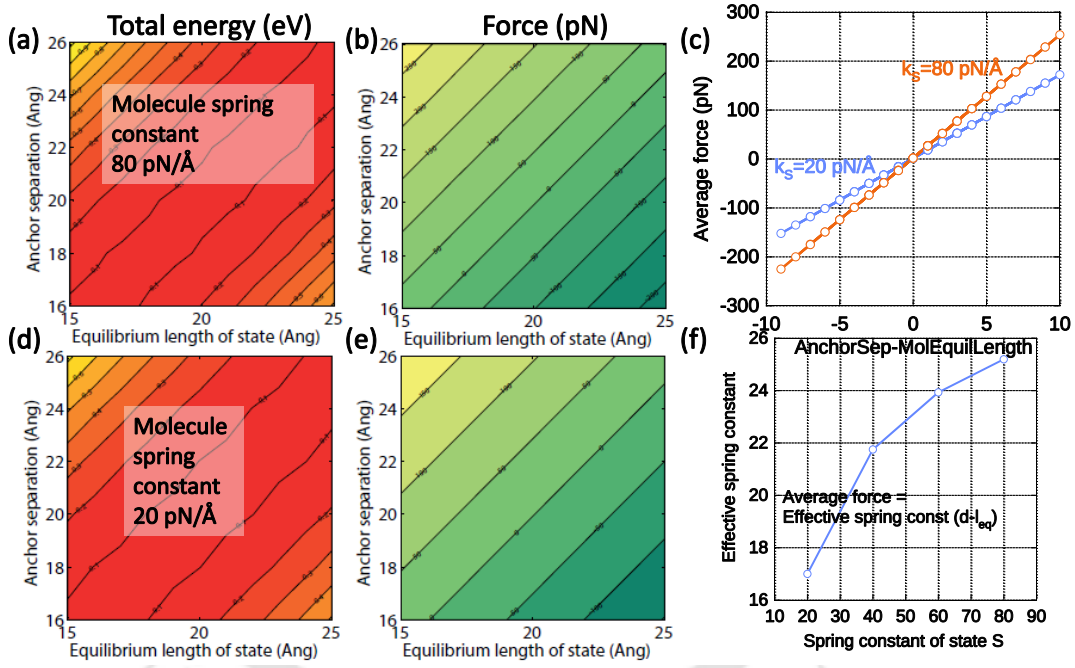

Figure 5.5: Contour plot of total energy (panels a,d)/force (panels b,e) as a function of the equilibrium length of the state (abscissa) l_S^{eq} and the anchor separation distance (y-axis) d for two pre-specified molecular spring constants, 80 pN/Å and 20 pN/Å respectively. (e) Average force plotted as a function of $d - l_S^{eq}$ for the two given molecular state spring constants k_S. The effective stiffness of the system (comprising of three springs in series: the two tethers and the molecule) is obtained from the slope of the force vs. $d - l_S^{eq}$ curve. (f) The plot of the effective spring constant vs. spring constant of the molecular state S. We obtain a relation between the effective spring constant (k_S^{eff}) and molecular spring constant (k_S).

related fluctuations in the end-to-end distance and 3-D orientation both determine $F_S(d)$. Furthermore, the relation between the effective spring constant of the system, k_S^{eff}, and the state spring constant k_S, is not clear. In the absence of closed integral

forms for the average force and energy experienced by the molecule, we resort to Metropolis Monte Carlo calculations for the force and energy in terms of three parameters, k_S, l_{eq}^S, and d describing the system in Fig. 5.1(b) where the three springs are connected in series. The stiffness of the springs of the tethers is kept fixed in the Metropolis calculations at 0.86 kcal/mol/Å^2. Figure 5.5(a-d) shows the force and energy response surface as a function of the equilibrium length of the state l_S^{eq} and the anchor separation d for two specified molecular spring constants k_S (20 and 80 pN/Å). The Metropolis calculations shows that the average force varies linearly with the difference between the anchor separation and the state equilibrium length (Fig. 5.5(c)), thus validating the one-dimensional model proposed in Eq. (5.35). In addition the energy (E) of the system was found to vary as $E \propto (d - l_S^{eq})^2$.

5.4.2 Elastic Properties of the States

Force *vs.* anchor spacings curves (Fig. 5.6) obtained from SC-MD simulations were used to obtain the effective spring constants, k_S^{eff} (from the slopes) and equilibrium lengths l_S^{eq} (from the intercepts) associated with the relevant states (Fig. 5.3) as discussed above. The correlation coefficients of the linear fits confirm that the effective spring model is a good approximation for the system for each of the states considered. As shown in Fig. 5.6, the compact configurations such in state 1 and 288 are associated with a higher spring constant than more facile structures. State 4, which shows unraveling at the N terminal also leads to a higher effective stiffness. Interestingly, the extended structure represented by state 17 also has a high effective stiffness which may be attributed to the tadpole like structure. The part of the molecule that is unwound from the helix is almost fully stretched.

Since we know the relationship between effective spring constant (k_S^{eff}) and state spring constant (k_S) from Fig. 5.5(f) we can now obtain the state spring constant (k_S) for the state S. Determining the state spring constant from the effective spring constant becomes an inverse problem. The effective spring constants, state spring constants and the equilibrium lengths of the molecule for some of the significant states are provided in Table 5.1. The state-specific elastic parameters display an interesting behaviour. The stiffness of deca-alanine was found to lie in the range 20-50 pN/Å across the 10 states in Fig. 5.1. Large α-helicity in state 1 results in $k_S = 40$ pN/Å. State 3, which has a 3_{10}-helicity content of 0.21 has a smaller spring constant of 25 pN/Å. On the other hand, the unfolded state 17 has $k_S = 48$ pN/Å as the dihedral angles are stretched to achieve such a configuration. The state equi-

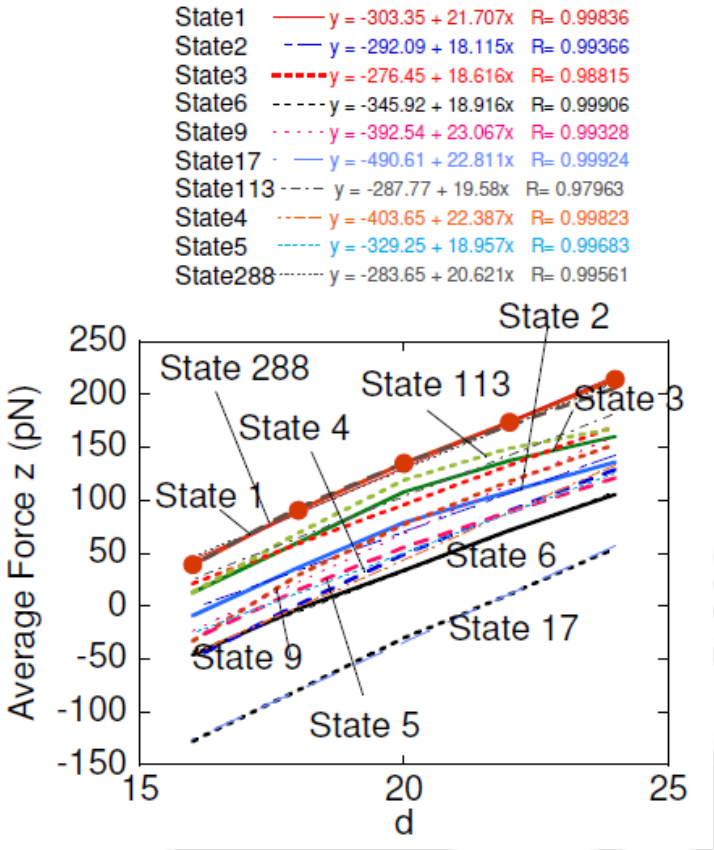

Figure 5.6: Average force obtained from state-constrained molecular dynamics calculation plotted at constant values of d. Using $f_z = k_S^{eff}(d - l_S^{eq})$ to describe the relation between the force and anchor separation d, we obtain the values of k_S^{eff} and l_S^{eq} for state S from the slope and intercept of the best linear fit respectively.

Table 5.1: State-specific spring constants and the equilibrium lengths.

State No.	k_S^{eff}(pN/Å)	k_S(pN/Å)	l_S^{eq}(Å)
1	21.707	39.44168	13.97475
2	18.115	23.74025	16.12421
3	18.616	25.37929	14.85013
6	18.916	26.37949	18.28716
9	23.067	50.7662	17.01738
17	22.811	48.27667	21.50761
113	19.58	28.73651	14.69714
4	22.387	44.31409	18.03055
5	18.957	26.32888	17.36825
288	20.621	32.97698	13.75539

librium length and the effective stiffness are used in the following discussion on free energy dependence on the anchor separation d, that is, Energy $\propto (d - l_S^{eq})^2$. We also note that at high extension, the effective harmonic spring model is not adequate and anharmonic correction terms are required to calculate the energy.

Force-Anchor Separation Curve from Direct MD is in Excellent Agreement with that from MSM

The force *vs.* anchor separation plot in Fig. 5.7 compares the variation of the force

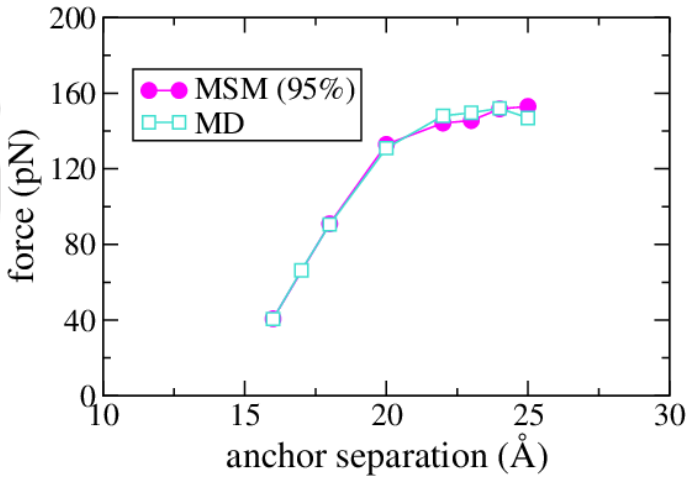

Figure 5.7: Plot of the force vs. the anchor separation from (i) plain MD (sky colour) and (ii) MSM (margenta colour) constructed with dominant states accounting for 95 % of total occupation probability at a given stretching condition.

with anchor separation from two sources: MD simulations with fixed anchor separations and from the Markov state model constructed with states accounting for 95% of the total occupation probability of the system. The stiffness and equilibrium lengths in the Table 5.1 are used in the expression (5.35) in the Force model to calculate the force for each state, which is then weighted by the appropriate probabilities from the MSM to obtain the total force at a give anchor separation. The plot shows that the MSM prediction is in excellent agreement with direct MD simulations with harmonic restraints to keep fix the anchor separation at specified d (deviations lie within 3%).

5.4.3 Thermodynamic Properties of the States

The free energy, calculated as the negative of the logarithm of the relative probabilities of the top ten dominant states ($-k_B T \ln(\pi_S(d))$, from the MSMs at various anchor separations between 16 to 28 Å are shown in Fig. 5.8. In each case, a

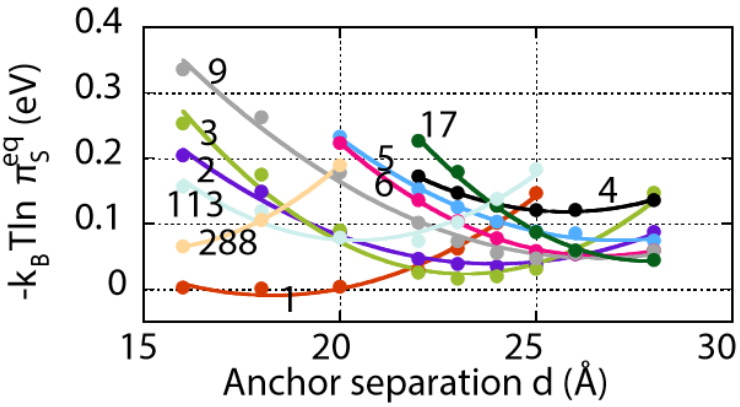

Figure 5.8: The free energy as a function of the anchor separation for the various states. The circles represent the values obtained from the MD simulations while the lines represent quadratic fits of the data.

quadratic fit was performed to obtain the coefficients of the quadratic Eq. (5.34). Excellent quadratic polynomial fit for the free energy of state S as a function of d is obtained in Fig. 5.8, which in turn yields the mechanical disposition factor $\chi(d)$ from Eq. (5.22) for any trap separation, d, within the range considered.

5.4.4 The kinetic Rates of Relevant Moves from the MSM

The variation of the kinetic rates with the anchor separation, as obtained from the MSMs constructed at the selected anchor spacings are shown in Fig. 5.9. Coupled with the coefficients of Eq. (5.34) obtained from Fig. 5.8, the plots are used to calibrate the model in the Eq. (5.25), thus enabling a direct calculation of the kinetic rates as a function of the anchor separation. We note that in several cases, the kinetic rates varied over several orders of magnitude with changing anchor separation. In most cases, the fits as per Eqs. (5.34) and (5.25) are excellent thus validating the kinetic model proposed. The move from state 3 to 2 (Fig. 5.9(c)) is an outlier, in which the fit is poor. In the absence of an alternative theory to predict the kinetic rates, it may be useful to apply a linear fit for such cases where the data does not obey the model.

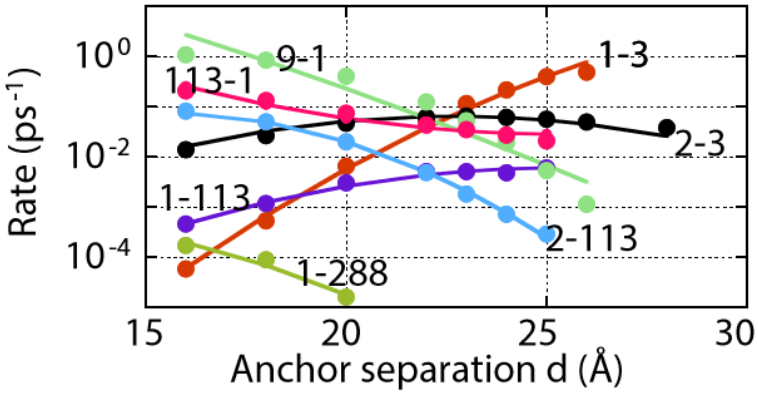

Figure 5.9: The kinetic rates for the relevant moves plotted against the anchor separation d. The circles represent the rates obtained from the MD simulations using MLE. The lines represent the fits according to Eq. (5.25) which incorporates the free energy fits as per Eq. (5.34).

A list of the values of kinetic rate calculated by using Eq. (5.27) for the relevant pathways is shown in Table 5.2. Instead of reporting α_f and α_b for a pair of states, we report the $k_f(d_0)$ and α_f values for the forward and backward pathway separately.

Table 5.2: Kinetic parameters for pathways.

	Initial state	Final state	$k_f(d_0)(\mathrm{ps}^{-1})$	α_f
1	1	2	0.00034839	-0.54709
2	1	3	6.1015e-005	-0.59274
3	1	9	6.8156e-006	-0.49303
4	1	113	0.00046631	-0.53674
5	1	288	0.00019553	-0.48526
6	2	1	0.82791	-0.45990
7	2	3	0.016057	-0.40825
8	2	4	2.313e-008	-1.17633
9	2	5	0.00025552	-0.53338
10	2	6	1.7435e-005	-0.73261
11	2	9	0.0013838	-0.38311
12	2	113	0.073205	-0.80283
13	3	1	3.1789	-0.48172
14	3	2	0.030502	-0.08193
15	3	5	0.00015933	-0.66092
16	3	9	0.0038348	-0.46981

17	3	113	0.037626	-0.38923
18	4	3	0.11062	-0.34478
19	4	9	0.030867	0.03810
20	5	2	1.9037	-0.35666
21	5	3	0.13986	-0.10560
22	5	6	0.043155	0.11139
23	5	9	0.000755	1.38952
24	5	113	0.49061	-0.35890
25	6	2	3.123	-0.42062
26	6	3	0.018211	-0.42137
27	6	5	0.0077344	0.15267
28	6	9	0.00015996	0.35658
29	6	17	6.5916e-006	-0.69357
30	9	1	2.7062	-0.34448
31	9	2	0.10142	-0.32410
32	9	3	0.066837	-0.25349
33	9	4	0.035147	1.85045
34	9	5	1.5273e-007	-3.93712
35	9	6	5.8563e-006	-0.91877
36	17	6	7.0329	-0.49613
37	113	1	0.25134	-0.45849
38	113	2	0.0009167	-0.71395
39	113	3	6.6838e-005	-0.77097
40	113	5	1.4377e-006	-0.61275
41	113	9	5.621e-005	-0.48036
42	288	1	0.0012743	-0.35428
43	288	2	0.00012598	-0.39950
44	288	3	2.0084e-005	-0.42861

Note that the sum of the geometrical terms for the forward (α_f) and reverse (α_b) for a pathway between a pair of state is ~ 1 which also verifies our proposed kinetic model.

5.5 Time-Dependent Markov State Model (TD-MSM)

Our interest lies in the application of MSMs for time-dependent anchor separations d (or forces). Although we have discussed all the basic concepts underpinning the time-dependent Markov state model in the theory sections, one critical aspect that has to be considered are the relevant timescales in the system. The fastest times that are relevant are the vibrational timescales of the atomic motions. Next are the internal relaxation timescales of the states which depends on the definitions of the states, that is, the level of coarse-graining involved in the system and also on the external forces applied. For the moment, we consider static pulling conditions (that is, fixed anchor separations). The MSM will be valid only at timescales larger than the internal relaxation timescales of the states, a necessary condition to satisfy Markovian property. Finally, the longest relevant timescales correspond to the equilibration times for the full network models, that is, the times at which thermodynamic equilibration is achieved over the full network. The time-dependent MSM that we propose correspond to the case where the anchor separation d varies slowly allowing the system to remain Markovian at all times, that is, the timescales associated with pulling are larger than the internal relaxation timescales involved. So in our time-dependent MSM (TD-MSM) formalism, we assume that vibrational timescales, $\tau_{vib} \ll$ relaxation timescales within a Markov state, $\tau_S \ll$ network relaxation timescales and timescales with pulling. In this work we ignore both the short timescales at which the system is non-Markovian as well as the very long timescales where there are deeper questions regarding the propagation of uncertainty and the validity-time of an MSM. The derivation presented in the previous section for constant anchor position mode can be extended to time-dependent extension of the molecule at small stretching rates. The key step involved includes specifying the range of stretching conditions where the MSMs are valid, determining the list of relevant states/pathways over the entire range, and allowing rate matrix (T) to be a function of time. We term the resulting MSM,

$$\frac{d}{dt} = T(t)\pi(t) \tag{5.36}$$

as a TD-MSM. The transition rates and the rate matrix need to be evaluated from MSM-0 for the prevailing value of the extension in a force ramp experiment. Using Eq. (5.27), to obtain the kinetic rates for the relevant moves for anchor separations at which MSMs are not available, the evolution of the state probabilities with

time for stretching (constant rate pulling experiments) are obtained by solving Eq. (5.36) alongside an equation specifying the rate of change of anchor separation using a Runge-Kutta 4-5 ODE solver with an initial condition.

5.5.1 Results of Pulling Expriment

The corresponding plots of force vs. time generated from the TD-MSM are shown in Fig. 5.10(c-d) alongside similar plots generated from direct pulling MD simulations

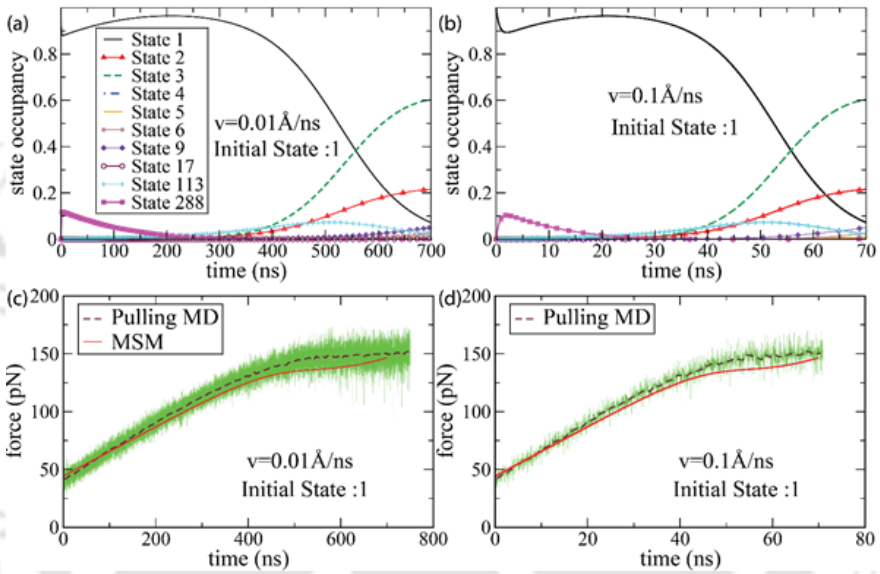

Figure 5.10: (a-b) shows the evolution of probabilities of the 10 major states over 700 (70) ns at a pulling speed of 0.01(0.1) Å/ns, with the system initially in state 1 (null occupancies of all other states) starting with an anchor separation of 16 Å. The corresponding force is plotted as a function of time in panels (c) and (d) from the MSM and the direct SMD simulations. The force calculated from the MSM by employing Eq. (5.35) with the effective spring constants listed in Table 5.1. The force for each state is weighted by the probabilities from panels (a) and (b) to obtain the total effective force on the system.

in which the N-terminal C_α was pulled using a moving constraint (with stiffness 0.86 kcal/mol/Å^2) at a velocity of 0.1 and 0.01 Å/ns. Five independent SMD trajectories were generated starting from the same initial state and initial position of dummy atom while the C-terminal C_α atom was constrained by a harmonic tether. We find that the TD-MSM generated results are in excellent agreement with the results of direct SMD calculations. In both cases, the system was initially in state 1. As the system is pulled the occupancy of state 1 undergoes a slight dip, while the system transits between states 1 and 288. But a higher extensions, the occupancy of state

288 diminishes to zero even though state 1 registers a high probability even at intermediate stretching (with $d \sim 20$Å). Eventually, the compact structures are replaced by the stretched out configurations such as states 2, 3 and 9.

5.6 Work Done Calculation in Pulling Experiment

It is possible to recover the free-energy (FE) difference of system from non-equilibrium work measurement in a thermodynamics transformation by applying Jarzynski relationship[242]. According to second law of thermodynamics, the average work done ($\overline{W}$) on the system can not be smaller than the free energy difference (ΔF) between the initial and final state in a thermodynamics transformation,

$$\overline{W} \geq \Delta F \tag{5.37}$$

where $\overline{W} - \Delta F$ is the dissipated work associated with the increase in entropy in a irreversible process. However, Jarzynski discovered a relationship between the free energy difference between two states and the irreversible work done along an ensemble of trajectories connecting the two states which holds the equality as,

$$e^{-\beta \Delta F} =< e^{-\beta W} > \tag{5.38a}$$

or equivalently,

$$\Delta F = -\beta^{-1} \overline{\ln \exp(-\beta W)} \tag{5.38b}$$

regardless of the speed of the process. In the steered MD (SMD)[205] simulation, one end of the molecule is fixed and the other end of the molecule is pulled by a dummy atom with constant velocity v and spring constant k. A guiding potential $\phi(\mathbf{r};\lambda) = (k/2)[\xi(\mathbf{r}) - \lambda]^2$ is added to control the end-to-end distance $\xi(\mathbf{r})$ where λ is a control parameter which distinguish the different states of system as a function of time. The work done on the system is calculated as,

$$W_{0 \longrightarrow t} = -kv \int_0^t dt' [\xi(r_{t'}) - \lambda_0 - vt'] \tag{5.39}$$

where the control parameter is fixed to λ_0 for starting configuration of the SMD simulation. The free energy difference ΔF between two configurations A and B of a classical parameter-dependent system characterized by two different values of

the control parameter $\lambda(\lambda_A, \lambda_B)$ can be expressed in terms of canonical partition function (Z) as,

$$\Delta F = -k_B T \ln \frac{Z_{\lambda_B}}{Z_{\lambda_A}}. \tag{5.40}$$

Thus the free energy difference $(\Delta F = -\beta^{-1}\ln(Z(d_{S'})/Z(d_S))$ between two states correspond two different values of anchor separations, namely, d_S and $d_{S'}$, can be obtained from non-equilibrium measurement of work performed on the system to switch from one ensemble to another by using Jarzynski equality.

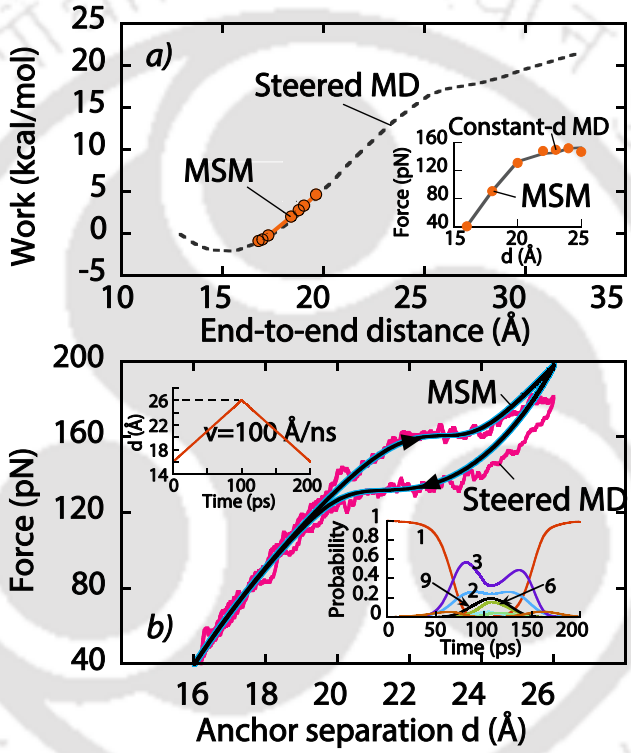

Figure 5.11: (a) Work done calculated at separations d = 16-26 Å with pulling velocity 0.1 Å/ns. Inset shows force at constant-d values. (b) Force versus anchor separation for a cyclic-pulling experiment (d-vs-t in top-left inset). Bottom-right inset shows state occupations as a function of time. States 17, 4, and 5 are not shown here because of their small probabilities. States 1, 2, 3, 6, and 9 are depicted by red, blue, purple, green and black lines, respectively.

The work done (shown in Fig. 5.11(a)) in pulling the deca-alanine from a compact to a stretched configuration is calculated using MSM-0 assuming that the molecule is stretched infinitely slowly. With the state-specific forces and end-to-end distances known for the selected values of d as well as the corresponding equilibrium state occupations from MSM-0, we obtain the average force and end-to-end distance for

the molecule (see inset Fig. 5.11(a)) for average force versus d. The work done was calculated using numerical integration from the starting from end-to-end distance $l(d_0) = 16$ Å, that is, $W(d) = \int_{l(d_0)}^{l(d)} F(d)dl$. Dashed line in Fig. 5.11(a) shows the work done calculated using SMD calculations with a pulling velocity 0.1 Å/ns such that the process was reversible (as in Park *et al.*[205]). In this case, the work done in stretching the system between any two reference separations is equal to the free-energy difference between the initial and final states. Thus, an attractive feature of the kinetic model generated from MSM-0 is that it yields the thermodynamic behavior.

5.6.1 The Behaviour of the System in a Cyclic-Pulling Experiment at Different Pulling Rates

We attempted to extend the MSM-0 to other stretching experiments. Let us consider cyclic pulling experiments where the anchor separation is increased at constant rate v (Å/ns) from a minimum to a maximum separation, and subsequently the separation is decreased with a rate -v (Fig. 5.11(b) top-left inset). At slow pulling rates less than 1 Å/ns the F-d plot in stretching and compression cycle overlap. The scenario changes when the molecule is pulled at timescales shorter than the MSM relaxation timescales leading to the hysteresis shown in Fig. 5.11(b). The origin of the hysteresis in the F-d plot in Fig. 5.11(b) lies in the difference in the relaxation rates from compressed to extended configurations and rates for the reverse direction, as evident from the state occupations in Fig. 5.11(b) inset. These results are in good agreement with the average force calculated using steered MD trajectories in Fig. 5.11(a) inset.

5.7 Connection between the kinetics in the Force and Constant Trap-Separation Ensembles

Let us now consider constant-force ensemble experiments. Since the force F simultaneously depends on trap separation (d), the end-to-end distance and the orientation, the value of d fluctuates about a certain value in a particular state. A sudden jump in d can be experienced in a MD trajectory when the molecule visits a new state. States in MSM-0 can be directly employed with constant-F ensemble as long as states missing in MSM-0 remain inconsequential to the dynamics. This aspect can be tested by straightforward MD calculations that ascertain the dominance of

MSM-0 states. Obtaining a rate constant in terms of F is more challenging. Here our attempt is to get the rate constants of pathways in constant-F ensemble experiments.

5.7.1 Theory

A constant force (F) is maintained on deca-alanine. As a result, the value of d fluctuates about a certain value in a particular state. The fluctuations in are expected because F depends not only on d and the end-to-end distance but also the orientation, that is, there are several degrees of freedom. This behavior is similar to what happens with constant d experiments where the end-to-end distance, orientation and force can fluctuate. In a constant-F ensemble, the probability of residing in a state S is given by,

$$\pi_S(F) = \frac{\int\limits_{r\in S} \exp(-\beta E)dr\Big|_F}{Z(F)}. \tag{5.41}$$

Z is the function involving the integral over entire phase space. We can define the free energy (A) of the state as,

$$\pi_S(F) = \frac{\exp(-\beta A_S(F))}{Z(F)}. \tag{5.42}$$

such that the partition function is written in terms of the discretized state space as,

$$Z(F) = \sum_{S'} \exp(-\beta A_{S'}). \tag{5.43}$$

An expression for the free energy as a function of F is desired. Since the average force and d are related to each other through the equation,

$$F = k_S^{eff}(d - l_S^{eq}) \tag{5.44}$$

which is already stated in the Eq. (5.35), let us assume that the configurational space sampled in a constant F experiment in state S closely matches the landscape sampled in the constant d experiment where the anchor separation is $d^S(F)$, that is, the d value that will result in average force F with state S. Thus Eq. (5.41) becomes,

$$\pi_S(F) = \frac{\int\limits_{r\in S} \exp(-\beta E)dr\Big|_{d^S(F)}}{Z(F)}. \tag{5.45}$$

The numerator in Eq. (5.45) appears in the probability of occupying state S for constant d experiment, that is,

$$\pi_S(d^S(F)) = \frac{\int\limits_{r\in S} \exp(-\beta E)dr\Big|_{d^S(F)}}{Z(d^S(F))}. \tag{5.46}$$

For brevity, we drop F dependence from $d^S(F)$. This expression can be written in terms of the free energy (A) as a function of d, that is,

$$\pi_S(d^S) = \frac{\exp(-\beta A_S(d^S))}{\sum\limits_{S'} \exp(-\beta A_{S'}(d^S))}. \tag{5.47}$$

where A has been calculated from the MSMs constructed for constant d and a numerical fit is given by,

$$A_S(d) = -k_BT\ln\pi_S(d) - k_BT\ln Z(d) = c_0^S + c_1^S d + c_2^S d^2 - k_BT\ln Z(d). \tag{5.48}$$

Combining Eqs. (5.45) and (5.47) we write

$$\pi_S(F) = \frac{\pi_S(d^S)}{Z(F)} Z(d^S). \tag{5.49}$$

The probability $\pi_S(d^S)$ is already known to us from expression (5.48) and written as,

$$-k_BT\ln\pi_S(d^S) = c_0^S + c_1^S(d^S) + c_2^S(d^S)^2. \tag{5.50}$$

So we rewrite Eq. (5.49) as,

$$\pi_S(F) = \frac{\exp(-\beta(c_0^S + c_1^S(d^S) + c_2^S(d^S)^2)}{Z(F)} Z(d^S) \tag{5.51}$$

Also noting that $d^S{=}(k_S^{eff})^{-1}F + l_S^{eq}$ we can write

$$A_S(d^S) = p_0^S + p_1^S F + p_2^S F^2 - k_BT\ln Z(d). \tag{5.52}$$

Where,

$$\begin{aligned} p_0^S &= c_0^S + c_1^S l_S^{eq} + c_2^S (l_S^{eq})^2 \\ p_1^S &= c_1^S (k_S^{eff})^{-1} + 2c_2^S (k_S^{eff})^{-1} l_S^{eq} \\ p_2^S &= c_2^S (k_S^{eff})^{-2} \end{aligned} \tag{5.53}$$

Thus finally Eq. (5.49) becomes

$$\pi_S(F) = \frac{\exp(-\beta(p_0^S + p_1^S F + p_2^S F^2))}{Z(F)} Z(d^S). \tag{5.54}$$

Finally, we write,

$$\frac{\pi_S(F)}{\pi_R(F)} = \frac{\exp(-\beta A_S(F))}{\exp(-\beta A_R(F))}. \tag{5.55}$$

or,

$$\begin{aligned} A_S(F) - A_R(F) &= -k_B T \ln \left(\frac{\pi_S(F)}{\pi_R(F)} \right) \\ &= -k_B T \ln \left[\frac{\exp(-\beta(p_0^S + p_1^S F + p_2^S F^2))}{\exp(-\beta(p_0^R + p_1^R F + p_2^R F^2))} \frac{Z(d^S)}{Z(d^R)} \right] \end{aligned} \tag{5.56}$$

We can now write,

$$\begin{aligned} A_S(F) - A_R(F) &= A_S(d^S) - A_R(d^R) \\ &= -\beta^{-1} \ln(\pi_S^{eq}(d^S)/\pi_R^{eq}(d^R)) - \beta^{-1} \ln(Z(d^S)/Z(d^R)) \end{aligned} \tag{5.57}$$

This allows us to build the free energy of a state as a function of F as,

$$A_S(F) = -k_B T \ln \pi_S(d^S) - k_B T \ln \frac{Z(d^S)}{Z(d_0)} \tag{5.58}$$

We write Eq. (5.58) as,

$$A_S(F) = -k_B T \ln \pi_S(d^S) + W(d^S) \tag{5.59}$$

where $W(d^S)$ is the work done to bring the anchor separation to d^S from d_0. Finally, recognizing that the kinetic model measures the change in the rates in terms of the relative free energies of the states with respect to a reference, we write,

$$k_f(F) = k_f(d_0) \exp(-\beta \alpha_f \eta(F)); \tag{5.60}$$

the equivalent version of Eq. (5.25) in the constant-F ensemble. Here $\eta(F)$ is the mechanical disposition at constant force F. The rate is $k_f(d_0)$ when anchor separation is d_0. The mechanical disposition ($\eta(F)$) at constant force F is given by,

$$\begin{aligned} \eta(F) &= \Delta A_{SR}(F) - \Delta A_{SR}(d_0) \\ &= (A_R(F) - A_S(F)) - (A_R(d_0) - A_S(d_0)). \end{aligned} \tag{5.61}$$

Since,

$$A_R(d_0) - A_S(d_0) = (c_0^R - c_0^S) + (c_1^R - c_1^S)d_0 + (c_2^R - c_2^S)d_0^2, \tag{5.62}$$

we finally write,

$$\eta = -k_BT\ln\left[\frac{\pi_R(d^R)}{\pi_S(d^S)}\right] - k_BT\ln\left[\frac{Z(d^R)}{Z(d^S)}\right] - \left((c_0^R - c_0^S) + (c_1^R - c_1^S)d_0 + (c_2^R - c_2^S)d_0^2\right). \tag{5.63}$$

Or since total free energy of system in a state S at trap separation d_0 is $A(d_0) = -k_BT\ln\pi_S(d_0) - k_BT\ln Z(d_0)$, from Eq. 5.59 we can write the mechanical disposition factor as,

$$\eta = -k_BT\ln\left[\frac{\pi_R(d^R)/\pi_R(d_0)}{\pi_S(d^S)/\pi_S(d_0)}\right] + (W(d^R) - W(d^S)). \tag{5.64}$$

5.7.2 Result

The Eq. (5.60) is the equivalent version of Eq. (5.25) in the constant-F ensemble. Equations (5.22, 5.60, 5.36, 5.59) can be solved to obtain the occupation as a function

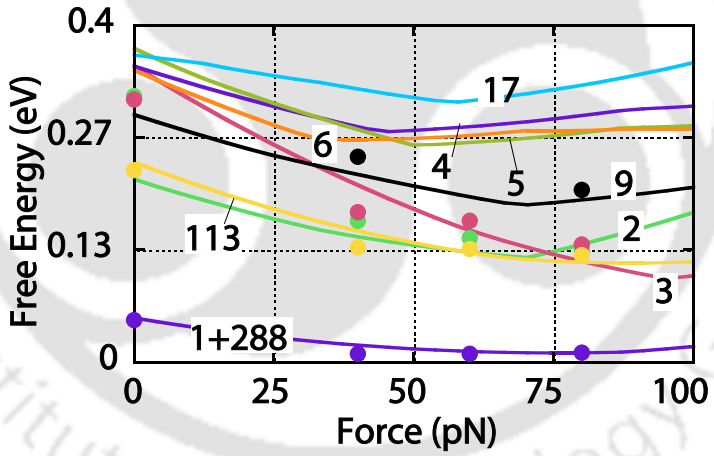

Figure 5.12: Equilibrium occupations at constant-force values between 0 and 100 pN are predicted using constant-d occupations (Eq. (5.59)). Symbols are MD values at constant-F. The combined state (1 + 288) is denoted by purple while states 2, 3, 113, 9, and 17 are denoted by green, red, yellow, black, and blue symbols/lines, respectively.

of time. Fig. 5.12 shows the free energy of the 10 states calculated using Eq. (5.59) for constant force values between 0 and 100 pN. State 1 and 288 dominate throughout this range. States 2, 3 and 113 are also important. Relative free energies for selected states calculated directly from constant-force MD calculations are in good agreement

with the MSM (Fig. 5.12). States not included in MSM-0 become relevant beyond $F = 100$ pN. Figure 5.13(a) shows a comparison between the relaxation modes from MSMs predicted using constant-d ensemble and the MSMs constructed using MD calculations at constant force. Eigenvalues/eigenvectors of the rate matrix are

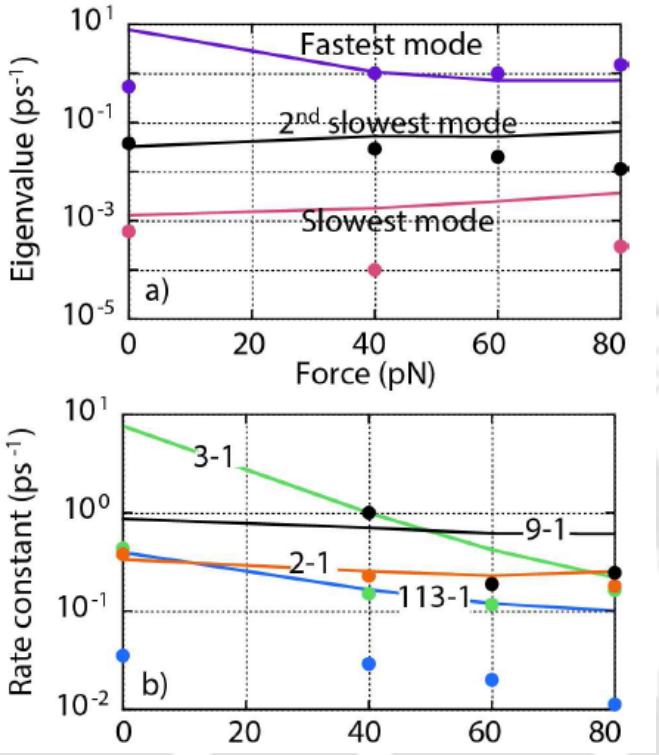

Figure 5.13: Eigenvalues (absolute values) for the fastest, second slowest, and slowest relaxation modes for the MSM predicted from constant d ensemble (lines) and MSMs constructed using MD at different constant force values (symbols). (b) Selected kinetic rate at constant force values between 0 and 80 pN predicted from constant-d ensemble (lines). Symbols denote the corresponding values from direct MD simulations at constant force.

calculated. One eigenvalue is zero since the sum of probabilities is constant while the remaining eigenvalues are negative and indicate the timescales required to reach equilibrium. Good agreement is observed with the fastest, second slowest and slowest eigenvalues/eigenvectors from both MSMs. The kinetic rates obtained from direct MD at 0-80 pN are compared with the predictions from constant-d ensemble in Fig. 5.13(b). Good agreement is observed for kinetic pathways 2-1 and 9-1. Overall the rates are within one order of magnitude within the MD rates.

5.8 Discussion

We have shown that the time-dependent MSMs constructed using MD can be a useful tool for obtaining molecular-scale insights into the dynamics of a molecule being stretched. An obvious difficulty in applying MSMs in studies of pulling experiments is that a single MSM constructed at a particular trap separation or constant force will not suffice to describe the kinetic and thermodynamic behaviour over a range of separations. We demonstrate that the essence of the dynamics is captured by the construction of a master-MSM (MSM-0) comprising of a list of states that are relevant over the range of trap separations of interest alongside a rate parametrization that relates the changes in the kinetic network to thermodynamic changes (Eq. (5.25)). The utility of the MSM-0 formalism lies in the fact that one can, in principle, extrapolate the network to low or even zero force conditions (as shown in Fig. 5.13), to predict the kinetics and thermodynamics of the molecule in its original, unbiased energy landscape. Stretching experiments that control separation or force by keeping it constant or changing it as a function of time can directly employ MSM-0 without generating new MSMs from scratch using MD. We have demonstrated the TD-MSM method for cyclic pulling experiments that reveal hysteresis at high pulling speeds. Our attempt to predict changes in the kinetic network with stretching have revealed that the rates depend on both initial and final states via the free energy differences, that is, the kinetic model also obeys thermodynamics. We have also provided a theoretical basis for relating force spectroscopy experiments in different ensembles (e.g., constant force or trap separation) which may provide a deeper understanding of single molecule experiments. Note that instead of comparing the ensemble averaged forces, one can generate stochastic realizations of state-to-state transitions from the time-dependent MSM using kinetic Monte Carlo[52]. In such a case, the average force experienced in the current state is recorded as the force $F(t)$.

Finally, in this chapter, we have dealt exclusively with constant trap separations and variations thereof. The connection with constant force ensemble was addressed indirectly. In the next chapter, we will directly apply the MSM-0 formalism to constant force experiments.

Chapter 6

Extension of Master-MSM to Constant Force Experiment in Force-Spectroscopy Setup

6.1 Introduction

Decoding the dynamics of biomolecules remains one of the grand challenges in science. Typically, the kinetics of folding/unfolding is described as a diffusive search for the global minimum in a corrugated free energy landscape. However, even small molecules may sample an extraordinarily large number of conformations on an inherently multidimensional landscape owing to the large number of degrees of freedom, which is why the problem continues to be so daunting even in the age of supercomputers. In recent years, single molecule experiments have been used to probe the energy landscape of biomolecules with notable success. In particular, the 1D landscape profiles have been investigated experimentally using single molecule force spectroscopy techniques such as AFMs and optical tweezers. In the previous chapter we described how the idea of using a stretching force to accelerate rare transitions may be applied to rapidly construct kinetic networks using computer simulations. The approach may be considered as a kind of accelerated MD method that allows the rapid sampling of the landscape. The primary challenge addressed in the study was to find a connection between topologically distinct Markov State Models constructed at different probe separations based on the thermodynamics of the system. The theoretical basis of the study was the Bell-Evans-Polyani principle (BEP)[207,208] that provided the connection between the kinetics and underlying thermodynamics, enabling the construction of the MSM-0 (or master-MSM) comprising of a list of

states that are relevant over the range of trap separations of interest alongside a rate parametrization relating the changes in the kinetic network to thermodynamic changes. Furthermore, we showed that if the anchor separation was varied as a function of time, the model inherits the time dependence, resulting in a time-dependent network that can be predicted based on a handful of simulations at a few probe separations. While the method was found to be effective in the case of the d-ensemble, that is, with the trap separation held constant, the question arises if a similar approach can be adopted in the force-ensemble, that is, with a constant force applied at both ends. *A priori* it is not obvious that such an underlying relation between the network models at various forces may exist. However, if such a relation is found to hold, it would not just provide a symmetric counter-part to the Master-MSM formalism at constant trap separations, but provide an opportunity to directly measure the rates at zero-force conditions, that is, reconstruct the original free-energy landscape. Here, we extend the Master-MSM method to SMFS experiments under constant pulling force to recover the kinetic rates at zero-force conditions.

6.2 Theoretical Basis

Here, we extend the Master-MSM method to SMFS experiments under constant pulling force or force-ramp conditions. In experimental setups, the biomolecule is attached via a soft polymer linker to the cantilever or a large bead. The total extension varies stochastically while a constant force can be maintained using a computer controlled feedback loop. Variations of the method includes probing the response of a molecule under a steadily increasing force (force-ramp technique), or suddenly changing the force-level (force-jump technique)[223–230]. Let us consider a time homogeneous Markov process for a system with n-states where π represents the probability of residence of the system in each of its n states. The time evolution of such a system is given by the continuous time master equation:

$$\frac{\mathrm{d}\pi(t)}{\mathrm{d}t} = T(F)\pi(t). \tag{6.1}$$

approximates the dynamics in terms of state-to-state transitions while providing control over the model resolution. Here $T(F)$ is the rate matrix at applied constant force F. Suppose $\pi_S(F)$ is the equilibrium residence probability associated with state S at a constant stretching force F and we assume that the same state definitions can be applied consistently across a range of stretching forces (F). Free energy

differences between any pair of states can be estimated from the equilibrium state occupations from the MSM at force F, that is, $\Delta A_{SS'}(F) = A_{S'}(F) - A_S(F) = \beta^{-1}\ln(\pi_S^{eq}(F)/\pi_{S'}^{eq}(F))$ where $\beta = \frac{1}{k_B T}$; k_B is the Boltzmann constant and T is the temperature.

6.2.1 Calculate Kinetic Rate of Transition at Various Constant Force

Consider a kinetic pathway $S \rightleftharpoons S'$ connecting the states S and S' (as shown in Fig. 6.1). Applying the transition state theory, the kinetic rates for the forward and reverse pathways are given by $k_f(F) = \nu_f \exp^{-\beta \Delta A_f(F)}$ and $k_b(F) = \nu_b \exp^{-\beta \Delta A_b(F)}$, respectively. The kinetic rates are dependent on the applied force F. The pre-exponential factors ν_f and ν_b are assumed to be constant, and $\Delta A_f(F)$ and $\Delta A_b(F)$ are the forward and reverse energy barriers, which also depend on the applied force.

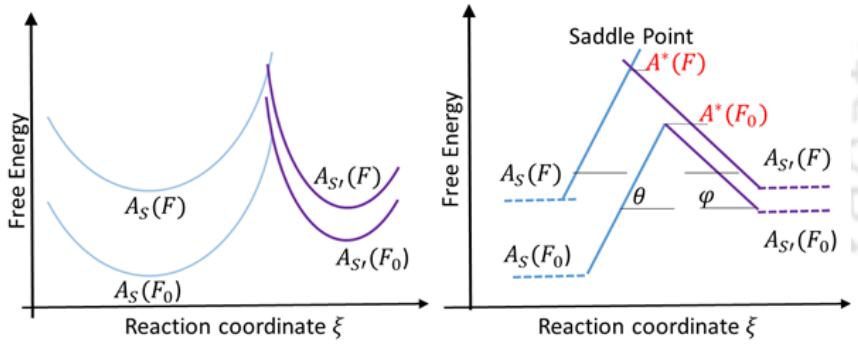

Figure 6.1: Schematic diagram showing how the energy landscape around the basins S and S' are altered when the force is changed from F_0 to F.

As in our previous study, following the Bell-Evans-Polyani principle, we assume that in the vicinity of each basin, the free energy profile along the reaction coordinate (ξ) is linear as shown in Fig. 6.1. Suppose the free energy in the S basin at applied force F_0 may be given by the expression $A_S(\xi, F_0) = (\tan\theta)\xi + C_S$ where C_S is a constant. At a different force, F, the expression is changed to $A_S(\xi, F) = (\tan\theta)\xi + C_S + (A_S(F) - A_S(F_0))$, that is, the profile is vertically shifted by a constant independent of ξ. A similar expression is obtained for the free energy profile in the S' basin is given by $A_{S'}(\xi, F_0) = -(\tan\phi)\xi + C_{S'}$ and $A_{S'}(\xi, F) = -(\tan\phi)\xi + C_{S'} + (A_{S'}(F) - A_{S'}(F_0))$ at F_0 and F respectively. Now consider the

shift in location of the saddle point along the reaction coordinate ξ as the force is changed from F_0 to F. At the reference force F_0, the free energy at the saddle point (ξ^*) as measured in the S basin, is given by,

$$A_S^*(\xi^*(F_0)) = A_S(F_0) + \Delta A_f(F_0) = (\tan\theta)\xi^*(F_0) + C_S. \tag{6.2}$$

where $\Delta A_f(F_0)$ is the forward barrier at F_0. As the force is changed to F, the saddle shifts to $A_S^*(\xi^*(F)) = A_S(F) + \Delta A_f(F) = (\tan\theta)\xi^*(F) + C_S + (A_S(F) - A_S(F_0))$. Hence,

$$A_S(F_0) + \Delta A_f(F) = (\tan\theta)\xi^*(F) + C_S. \tag{6.3}$$

Similarly, when calculated from the S' basin, the saddle point free energy may be estimated as,

$$A_{S'}^*(\xi^*(F_0)) = A_{S'}(F_0) + \Delta A_b(F_0) = -(\tan\phi)\xi^*(F_0) + C_{S'} \tag{6.4}$$

at the reference F_0. At a force F, this is shifted to $A_{S'}^*(\xi^*(F)) = A_{S'}(F) + \Delta A_b(F) = -(\tan\phi)\xi^*(F) + C_{S'} + (A_{S'}(F) - A_{S'}(F_0))$. Hence,

$$A_{S'}(F_0) + \Delta A_b(F) = -(\tan\phi)\xi^*(F) + C_{S'} \tag{6.5}$$

Eqs. (6.2) and (6.3) give the shift in the forward barrier due to the change in the applied force,

$$\Delta A_f(F) = \Delta A_f(F_0) + (\tan\theta)(\xi^*(F) - \xi^*(F_0)). \tag{6.6}$$

Similarly, the change in the barrier for the reverse transition may be expressed as,

$$\Delta A_b(F) = \Delta A_b(F_0) - (\tan\phi)(\xi^*(F) - \xi^*(F_0)) \tag{6.7}$$

Finally, we have,

$$\frac{\Delta A_f(F) - \Delta A_f(F_0)}{\tan\theta} = \frac{\Delta A_b(F) - \Delta A_b(F_0)}{-\tan\phi} \tag{6.8}$$

Now, the difference between the forward and the reverse barriers at a tension, F, is equal to the free energy difference between the two states. $\Delta A_f(F) - \Delta A_b(F) = \big(A^*(F) - A_S(F)\big) - \big(A^*(F) - A_{S'}(F)\big) = A_{S'}(F) - A_S(F) = \Delta A_{SS'}(F)$. Here, the free energy change along the forward pathway is denoted by $\Delta A_{SS'}(F) = A_{S'}(F) - A_S(F)$

where $A_S(F)$ denotes free energy of state S. Hence, Eq. (6.8) can be cast as,

$$\frac{\Delta A_f(F) - \Delta A_f(F_0)}{\tan\theta} = \frac{\left(\Delta A_f(F) - \Delta A_{SS'}(F)\right) - \left(\Delta A_f(F_0) - \Delta A_{SS'}(F_0)\right)}{-\tan\phi} \tag{6.9}$$

or

$$\left(\Delta A_f(F) - \Delta A_f(F_0)\right)\left(\frac{1}{\tan\theta} + \frac{1}{\tan\phi}\right) = \frac{\left(\Delta A_{SS'}(F) - \Delta A_{SS'}(F_0)\right)}{\tan\phi} \tag{6.10}$$

Finally, defining $\eta(F) = \Delta A_{SS'}(F) - \Delta A_{SS'}(F_0)$, the free energy barrier ΔA_f at applied forces F and F_0 are related through the equation,

$$\Delta A_f(F) = \Delta A_f(F_0) + \alpha_f \eta(F) \tag{6.11}$$

where $\eta(F)$ is termed as the mechanical disposition at force F, measures the thermodynamic preference for states S and S' upon application of a force relative to a reference condition. $\alpha_f = \tan\theta/(\tan\theta + \tan\phi)$ is a symmetry parameter related to the slopes of the free energy profiles in the two basins. Each pair of states that are kinetically connected is characterized by its own mechanical disposition as well as the symmetry parameter. The free energy difference increases from that at the reference force F_0 to F when $\eta(F) > 0$. The kinetic rate, $k_f(F)$ is related to the rate at stretching force F_0 as,

$$k_f(F) = k_f(F_0)\exp(-\beta\alpha_f\eta(F)). \tag{6.12}$$

Similarly, the backward rate $k_b(F) = k_b(F_0)\exp(\beta(1-\alpha_f)\eta(F))$. Please make a note that η corresponds to χ term used in the previous chapter. For sake of brevity, the pair of states has not been explicitly mentioned in the notation for k_f, k_b, ν_f, ν_b, $\Delta A_f(d)$, $\Delta A_b(d)$, α_f, α_b and $\eta(F)$. Generally, α_f and α_b are constants for a given pair of states. Eq. (6.12) gives a prescription for calculating the kinetic rate for a given transition at a force, in the absence of data at that force, provided that the other parameters are known. An interesting revelation of this analysis is the relation between the kinetics and thermodynamics of the system. The transition rates between two states depend on the free energy difference between them. When the force F is a function of time, such as in a force ramp experiment, the time-dependence is inherited by the mechanical disposition and rate matrix.

We next illustrate how the kinetic rate parameter $k_f(F_0)$, the symmetry param-

eter α and $\eta(F)$ are estimated from MSMs constructed at constant F. The basic building blocks for the model are: a) definitions of Markov states, b) a linear relation between the logarithm of a transition rate and the free-energy difference between the end-states of the respective transition, c) a model for the dependence of the free energy of each Markov state on the applied tension. In the absence of a reliable model predicting the relation between free energy and applied force, we perform a linear or quadratic fit the data from the calculation at various forces. Next, the kinetic parameters $k_f(F_0)$ and α_f are estimated by fitting Eq. (6.12) to rates in the MSMs at different values of F. The kinetic rates at the reference tension, $k_f(F_0)$, are directly obtained from the MSM constructed at the force F_0. Then the corresponding kinetic rate at zero force condition can be predicted by extrapolating the Eq. (6.12) to $F = 0$ pN.

6.3 Deca-alanine as Test Model for Master-MSM Method in Constant-Force Experiments

6.3.1 System Setup

In our computational study, we consider a deca-alanine molecule in vacuum. The model of a capped deca-alanine with acetylated N-terminus and amidated C-terminus was generated using the 104-atom helical model of Ref. 205 as the initial configuration. Equal and opposite forces (F) applied to C_α atoms at the two ends as shown in Fig. 6.2. MSMs of the deca-alanine system were constructed at F = 30, 40, 50, 60, 70, 80 and 90 pN.

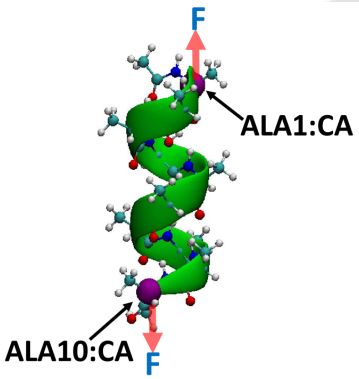

Figure 6.2: The deca-alanine molecule in constant-Force experiment set up. Equal and opposite forces (F) are applied on CA atoms at N-terminal and C-terminal ends of deca-alanine.

6.3.2 Simulation Protocols

All MD simulations were performed with NAMD 2.91[184] and the CHARMM36[206] force-field parameters. Temperature was held at 300 K using a Langevin thermostat. Bonds involving hydrogen atoms were constrained to their equilibrium values using RATTLE[129]. An integration time step of 2 fs was used.

6.3.3 MSM Construction Protocols using Programmed State Constrained (PSC) MD

The programmed state constrained MD (PSC-MD) technique (see Sec. 4.4 of chapter 4) is used to identify the states and pathways on-the-fly. The PSC-MD method entails launching independent MD trajectories and tracking the conformational changes of the evolving systems. Each time a transition to a different state is detected, the transition time and the final state are recorded and the system is returned to the initial state, that is, a new trajectory is initiated with the original configuration to sample additional transitions. Hence, instead of allowing the system to freely evolve dynamically, it is confined to specific states to record sufficient transitions. The PSC-MD methods enables one to systematically increase the "validity time" of an MSM as described in Sec. 4.4 of chapter 4. A catalogue of states of deca-alanine is generated by comparing the backbone atoms after aligning the molecule using the Kabsch algorithm using a tolerance of 3 Å. MD snapshots were collected every 0.2 ps due to the rapid interconversion between states. A transition is said to have occurred when the system remains in the new state for at least 1.2 ps after the transition is detected in the MD trajectory. Kinetic rates for the detected pathways are calculated using a maximum likelihood estimation (MLE) analysis. Only those transitions that were sampled at least ten times were included in the construction of the MSMs.

6.3.4 Results and Discussions

We have constructed MSMs of the deca-alanine system at various forces between 30 and 90 pN. In each case, the accrued MD time was approximately 2.1 μs while the corresponding validity time was between 8-16 ns. Although, about 1300 states were detected, only about 6 states (shown in Fig. 6.3) were found to have a combined occupancy 0.99 in all out calculations. The MD time, validity time, number of relevant states with 99% occupancy at the various values of force ranges from 30 pN

Table 6.1: Validity time and number of relevant states of MSM of deca-alanine at various forces ranging from 30 pN to 90 pN.

Force (pN)	Accrued MD (μs)	Validity Time (ns)	States Required for 99% Occupancy
30	2.047	16.384	1 (0.8966), 288 (0.097)
40	2.347	8.192	1 (0.925), 288 (0.0669)
50	2.336	16.384	1 (0.9503), 288 (0.04)
60	2.324	16.384	1(0.957), 288(0.028), 2(0.004), 113(0.0038)
70	1.894	8.192	1(0.951),288(0.027), 2(0.0063), 113(0.0049)
80	1.863	8.192	1 (0.955), 288(0.0115), 2 (0.0109),113(0.0072), 3(0.006)
90	2.33	8.192	1(0.939), 2(0.0194), 3(0.0125), 113 (0.0102, 288(0.0052), 20(0.0036)

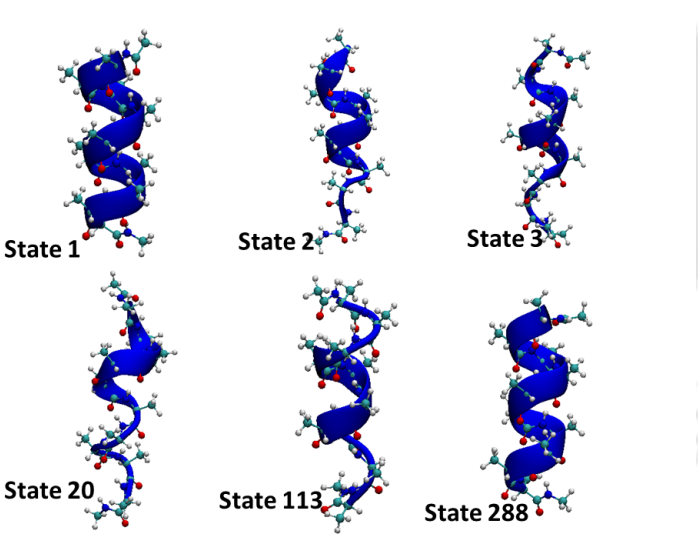

Figure 6.3: The structures of the top six relevant states of deca-alanine in the constant-F ensembles ranging between 30 pN and 90 pN.

to 90 pN are listed above in table 6.1.

6.3.5 Thermodynamic Properties of States

Figure 6.4(a) shows the variation of the free energy, defined as $A_S(F) = -k_B T \ln \pi_S^{eq}(F)$ with the applied force F. Here, $\pi_S^{eq}(F)$ is the probability of residence in a given state at equilibrium at a constant stretching force F. The lines denote quadratic fits to

the data. One of the key ingredients of the methodology is the variation of the free energy associated with relevant states with applied force.

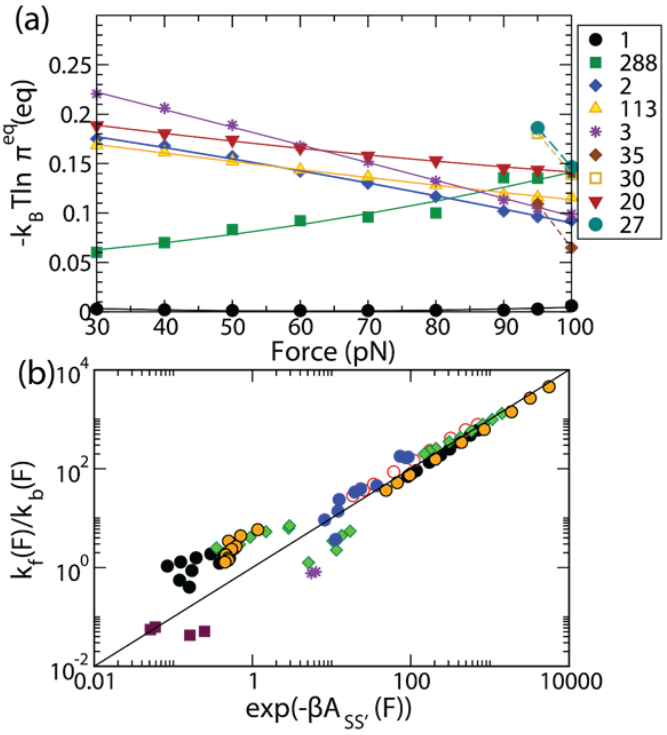

Figure 6.4: (a) Plot of the free energy vs. the applied force for the various states. Symbols represent the value computed from PSC-MD calculations. The lines represent a quadratic fit applied to the data in each case. (b) Parity plot used to verify detailed balance for the kinetic pathways detected in the calculations. Each point represents a kinetic pathway where the ordinate gives the ratio of the forward to reverse rates while the abscissa represents $\exp(-\beta A_{SS'}(F))$. Here $A_{SS'}(F)$ represents the free energy difference between the two states S and S' when a stretching force, F, is applied and β is the reciprocal of k_BT where k_B is the Boltzmann constant and T is the temperature.

Thermodynamic consistency requires $\frac{k_f(F)}{k_b(F)} = \exp(-\beta \Delta A_{SS'}(F))$. Since detailed balance is not inbuilt in the MSM construction method, it is verified in the parity plot in Fig. 6.4(b) which compares the ratio of forward to backward kinetic rates, $\frac{k_f(F)}{k_b(F)}$, for the transitions detected in the simulations to the corresponding Boltzmann factor $\exp(-\beta \Delta A_{SS'}(F))$. Overall, the detected transitions obey detailed balance.

6.3.6 Extracting Rates at Zero-Force

Table 6.2 provides values for $k_f(F_0)$ and α_f for pairs of states.

Table 6.2: The kinetic rate parameter for transitions.

Initial state	Final state	α_f	Rate $k_f(F_0)(\mathrm{ps}^{-1})$
1	288	0.456596	1.23E-04
1	2	0.764375	9.52E-04
1	113	0.61237	9.23E-04
1	3	0.741206	5.94E-04
1	20	0.842141	6.13E-04
288	1	0.389748	2.41E-03
288	2	0.43566	8.61E-04
288	113	0.362568	5.29E-04
288	3	0.507158	3.55E-04
288	20	0.459932	4.79E-04
2	1	0.298804	1.66E-01
2	288	0	
2	113	1.24497	2.88E-02
2	3	0.515991	3.76E-02
2	20	1.10849	1.42E-02
113	1	0.342567	2.866E-01
113	288	0	
113	2	1.10053	2.42E-03
113	3	0.418856	2.60E-03
113	20	0	
3	1	0.257173	5.04E-01
3	288	0	
3	2	-0.02487	2.33E-02
3	113	0.277862	1.63E-02
3	20	0	
20	1	0.392022	3.59E-01
20	288	0	
20	2	0.89773	5.00E-03
20	113	0.209582	2.72E-02
20	3	0.567472	1.99E-02

Choosing $F_0 = 60$ pN, a detailed MSM is constructed that contains six states, alongside the kinetic parameters of Eq. (6.12). We term this master-MSM as MSM(F_0) or simply MSM-0. Note that MSM-0 contains a list of states and kinetic pathways that would be relevant for certain/entire range of forces to be sampled (30-80 pN), the associated kinetic parameters include $k_f(F_0)$ and α_f, and the thermodynamic parameter $\eta(F)$. MSM-0 is a precursor for TD-MSMs at a variety of stretching conditions.

The kinetic rate parametrization was used to calculate the corresponding rates at force-free conditions. The symbols in Fig. 6.5 indicate the kinetic rates directly calculated from the MSMs at 30, 40, 50, 60, 80, 90 and 100 pN, while the broken lines indicate the fit in according to Eq. (6.12). The predicted rates at 0 pN, ob-

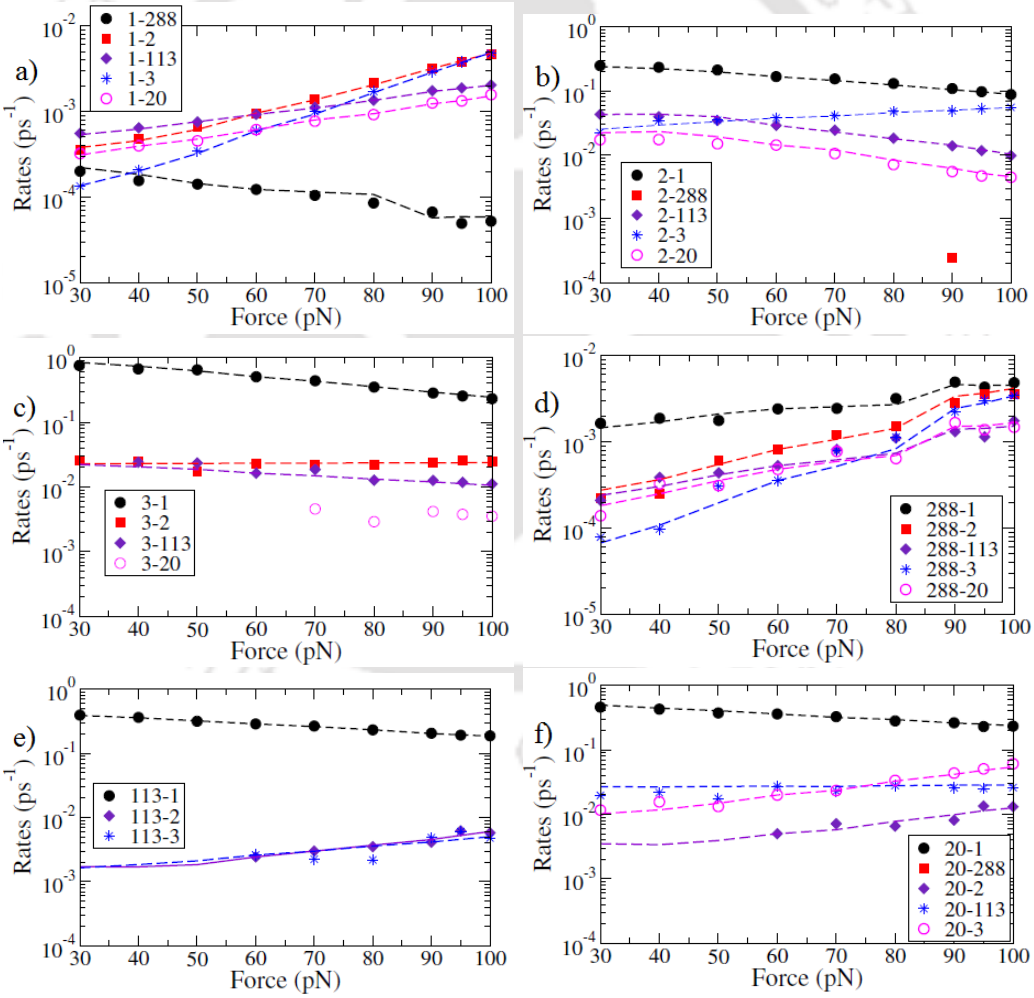

Figure 6.5: Kinetic rates of relevant pathways obtained from the MSMs at various forces. The symbols represents the kinetic rates obtained from the PSC-MD calculations. The lines refer to Eq. (6.12).

tained from the intercepts on the y-axis, are compared to the actual rates obtained from direct simulations in the duration of 0 pN runs. A good agreement between the predicted and directly calculated rates are observed. The estimated rates for slow transitions, such as that from state 1 to 3 (Fig. 6.5(a)), which are of the order 10^{-5} ps^{-1}, were in good agreement with the predicted rates.

For slow transitions, with rates $\sim 10^{-5}$ ps^{-1} or lower, an accurate determination of kinetic rates would require long trajectories ($\sim$ several microseconds or longer). However, the method outlined here shows how an estimate for such rates may be obtained at lower computational cost by a handful of simulations at various stretching forces. Hence the method described here may be classified as an enhanced kinetic sampling technique.

6.3.7 Comparison of the Kinetic Rates Predicted from the force (F) and trap separation (d) Ensembles for the Deca-alanine Model

In our previous study (Sec. 5.7 of chapter 5) of the kinetics in the d-ensemble (that is, constant trap separations), we had attempted to predict the dynamics in the F-ensemble without actual simulations at constant force conditions. The free energy associated with a state, S, in the F-ensemble was given by,

$$A_S(F) = -\beta^{-1}\ln\left(\pi_S^{eq}(d^S)\right) - \beta^{-1}\ln(Z(d^S/Z(d_0))) \tag{6.13}$$

(as mentioned in Eq. (5.58) of chapter 5) where $Z(d^S)$ is the partition function in the constant-d ensemble with trap separation d^S. Crucial to this formulation is a mechanical model for the molecule that can relate the trap separation to the force. Based on the force-extension relations obtained from the constant-d MD calculations, a harmonic spring model characterized by distinct equilibrium lengths and effective spring constants (encompassing the spring constant of the tethers with that of the molecule) for each state, was deemed suitable. We further assumed that the energy landscape sampled by the molecule when it is in state S at constant F is the same as the one at the corresponding constant separation $d^S = l_S^{eq} + F/k_S^{eff}$. Hence, in Eq. (6.13), d^S is the trap separation corresponding to an applied force, F for the system in state S. Finally, we obtain a relation for the kinetic rate at force F based on parametrization in the constant-d ensemble, $k_f(d_0)\exp(-\beta\alpha_f\zeta(F))$; the equivalent version of Eq. (6.12). Here $\zeta(F) = \Delta A_{S,S'}(d) - \Delta A_{S,S'}(d_0)$ is the mechanical disposition estimated for constant-d conditions.

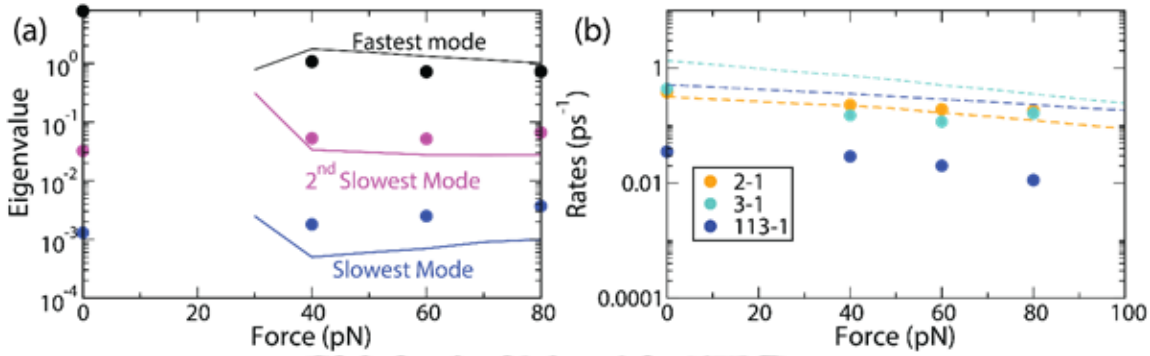

Figure 6.6: (a) Eigenvalues (absolute values) for the fastest, second slowest and slowest relaxation modes are shown for the MSM predicted from constant-F ensemble (lines) and MSMs constructed using constant trap-separations (d) and extended to constant-F using Eq. (6.13) (symbols). (b) Selected kinetic rates at constant force values between 0-80 pN predicted from constant-F ensemble (lines). Symbols denote the corresponding values from the constant force predictions using constant-d ensemble MSMs.

Here, we compare the predictions from constant–F ensemble MSMs parametrized by constant-d data with the results directly obtained from the MSM formalism in the constant-F ensemble. Figure 6.6(a) shows a comparison between the relaxation modes in the two setups. The eigenvalues/eigenvectors of the rate matrix are calculated from the MSMs from the two formalisms. Since the sum of the probabilities is constant, one eigenvalue is zero and the remaining eigenvalues are negative indicating the timescales required to reach equilibrium. Good agreement is observed with the fastest, 2^{nd} slowest and slowest eigenvalues/eigenvectors from both MSMs. The kinetic rates obtained from constant-F MSMs at 0-80 pN are compared with the predictions from constant-d ensemble in Fig. 6.6(b). Good agreement is observed for kinetic pathway 2-1. Overall the rates predicted from the d-ensemble parametrization are within one order of magnitude within the rates directly predicted in the F-ensemble.

6.4 Network Model of TBA under Tension in Explicit Solvent: Prediction of Kinetic Rates at Force-free Conditions

As our second example, we consider a Thrombin Binding Aptamer (TBA) model. Thrombin binding aptamer (TBA) is one of the most studied synthetic oligonu-

cleotides, a single stranded DNA with the Guanine rich sequence 5′-GGTTGGTGTG-GTTGG-3′[244], can fold into G-quadruplex (G4) structure that is an anti-parallel orientation with a chair-like conformation[245] (as shown in Fig. 6.7). An aptamer is a DNA/RNA oligonucleotide, exhibits a specific binding activity towards a protein target[244,246]. It modulates the action of target biomolecule and therefore, serves as a drug candidate in many diseases[247]. TBA binds to thrombin and has interesting anticoagulant properties[244,248] against thrombin. The anti-thrombin activity of TBA has spurred investigations into the structural and kinetic properties of TBA. TBA also serves as a testbed for exploring DNA-protein interactions. Furthermore, the G4 structures are commonly found in the human telomere region and oncogene promoter region and the formation of such structure at the telomere region by any means may be able to treat cancer by blocking the action of telomerase[249–252]. Therefore, a deeper insight into the molecular kinetics of TBA system carries implication not only for cardiovascular therapy, but also can predict the mechanism of G4 structure formation in Guanine-rich nucleic acids.

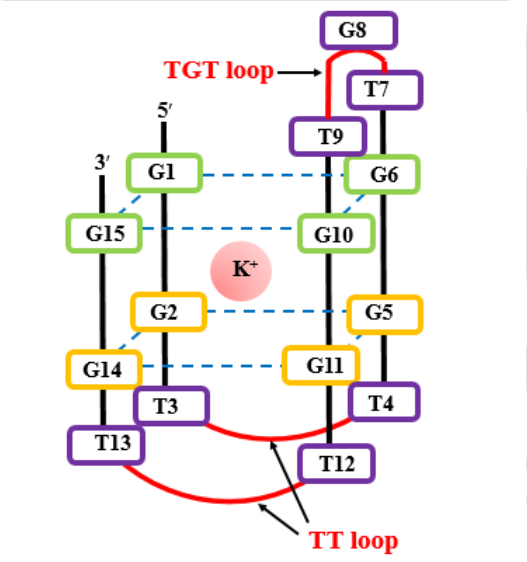

Figure 6.7: Schematic representation of 15-mer TBA and K^+ complex.

The G-quadruplex structure are formed by the self-assembly of guanine-rich oligonucleotides. The underlying repeating motif of G-quaruplexes is the G-tetrad (also termed as G-quartet). Each G-tetrad has four guanines arranged in square planar manner, stabilized by Hoogsteen-like hydrogen bond in which each guanine base both accepts and donates two hydrogen bonds[253]. The G-quadruplex is further stabilized through a cation (typically Na^+ and K^+) binding in its central binding

site located between the two quartets which reduces the repulsion between the aromatic oxygen atoms (O6) in neighbouring guanines[254,255]. The G4 structure of TBA consists of two planar Guanine-quartets connected by two intervening TT and one TGT loop[256–259]. The two lateral TT loops of TBA bind with exosite-I of Thrombin whereas the TGT loop has been associated with exosite-II[260]. The small size of the 15-mer TBA molecule makes it a good target for investigations into folding/unfolding dynamics of G-quadruplexes.

Several MD studies have reported investigations of the fundamental structural conformations and dynamics of TBA in presence of cations (K^+, Na^+, Sr^+ etc.)[258,261–264]. For example, Kim et al.[261] carried out all atom replica exchange molecular dynamics (REMD) simulation to generate the the folding mode of TBA and free energy map[261]. Extremely long MD simulations of the order of a few microseconds concluded that the TGT loop of TBA stabilizes the entire structure of the complex in presence of the K^+ ion and the TT loops are directly involved in the binding with the thrombin[258]. The study on TBA by NMR experiment in Ref. 265 revealed that the unfolding of TBA molecule takes place by uncoupling of the base pairs-G1-G15, G2-G14 and G5-G11 based on hydrogen exchange rate, then following by opening of TGT loop and finally opening through TT loop. However, the folding of G-quadruplex structure is far slower and more complex than fast-folding protein process as reported by experiments on different quadruplex structure[266,267] including TBA[265,268] and MD simulation studies. The presence of several kinetic traps on the folding pathway of TBA presents obstacles towards the construction of detailed kinetic network model of TBA. Recently, Zeng et al. employed MD and MSM method to understand the unfolding mechanism of TBA[269]. In this chapter, we construct Master-MSM in a constant-force setup as previously described for the deca-alanine molecule. The goal of the study is to verify if the method outlined may be applied to more complex solvated systems to efficiently elucidate long-timescale conformational changes and kinetic rates.

6.4.1 System Setup

As a starting point, we took a NMR structure of 15-mer TBA from Protein Data Bank (PDB) code: 148D[270]. The system was solvated in a ($58 \times 58 \times 68$ Å^3) water box with 6919 water molecules and neutralized with 14 K^+ atoms. Then additional 19 K^+ and 19 Cl^- atoms were placed to keep 150 mM KCl concentration of the system. Consequently, the total number of atoms in solvated TBA system were 21141. Since no metal ion are not included in the native PDB file we placed a K^+

ion manually at the center of G-quadruplex core. We performed steered MD (SMD) simulations by applying an equal and opposite constant force F to atom C5′ atom at 5′ end and C3′ atom at 3′ end of the TBA molecule as shown in Fig. 6.8.

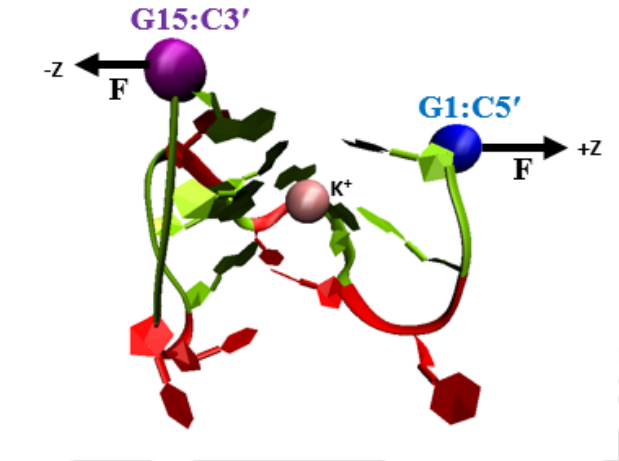

Figure 6.8: The folded G-quadruplex structure of a TBA molecule. Guanine and Thymine residues are represented by yellow-green and red colors respectively. The constant force F is applied on the C5′ atom (blue circle) of residue 1 and C3′ atom (purple circle) of residue 15 in opposite direction. A potassium ion (K^+) denoted by pink circle is kept at the centre.

6.4.2 Simulation Protocols

All the MD simulation were performed using NAMD version 2.11[184] with the force field CHARMM36[206]. Initially the system was relaxed using the conjugated gradient energy minimization algorithm by 10000 steps and equilibrated for 500 ps first in NPT ensemble. We kept the temperature at 310 K by coupling the system to a Langevin heat bath. The atomic configurations were saved every 4 ps for further analysis. The time step of integration was set to 2 fs and the non-bonded cut-off distance was set to 10 Å to treat the long-range interaction by applying Particle Mesh Ewald (PME) method. After NPT simulation a 5 ns long NVT simulations were performed to select initial conformations (that were slightly different from native structure) for parallel long SMD simulations with constant force.

6.5 MSM Construction Protocols using Parallel Long Constant-Force SMD Trajectories

For construction of MSM with constant-F ensembles, we carried out parallel MD simulations at different force F = 0, 10, 20, 30, 40 pN from different starting conformations. The number of trajectories and total time length of MD simulations generated at various constant force conditions are listed in Table 6.3. The nine

Table 6.3: List of SMD simulations at various forces.

Constant Force (pN)	Number of MD trajectories	Length of each MD trajectories (ns)	Total length Total MD duration (μs)
0	98	20 (91 trajectories) 300 (9 trajectories)	3.9
10	96	20	1.9
20	96	20	1.9
30	96	20	1.9
40	96	20	1.9

MD trajectories at force free conditions were continued upto 300 ns due to the slow dynamics observed. Note that, unlike the previous example, the network models of the TBA molecule were constructed via post-analysis of MD trajectories instead of on-the-fly.

6.5.1 Construction of MSM-0 for the TBA under Tension

Due to the slow dynamics observed in our preliminary analysis, a coarse-grained network model was generated using a one-dimensional reaction coordinate, the distance between C5′ and C3′ atoms of the first and last residues respectively, termed the extension length. The configurational states were identified by binning the extension length of the molecule in equispaced windows of width 6 Å. A total of 13 states were identified with extension range given in Table 6.4. Representative configurations for the states, named in order of discovery, are provided in Fig. 6.9.

Table 6.4: List of the states of the TBA molecule and the range of extension lengths for each state. The range includes the lower bound but not the upper bound in each case. The states are labeled in order of detection in the MSM construction method.

State	Extension Length Range (Å)	State	Extension Length Range (Å)
1	12-18	8	48-54
2	18-24	9	54-60
3	6-12	10	60-66
4	24-30	11	66-72
5	30-36	12	72-78
6	36-42	13	>78
7	42-48		

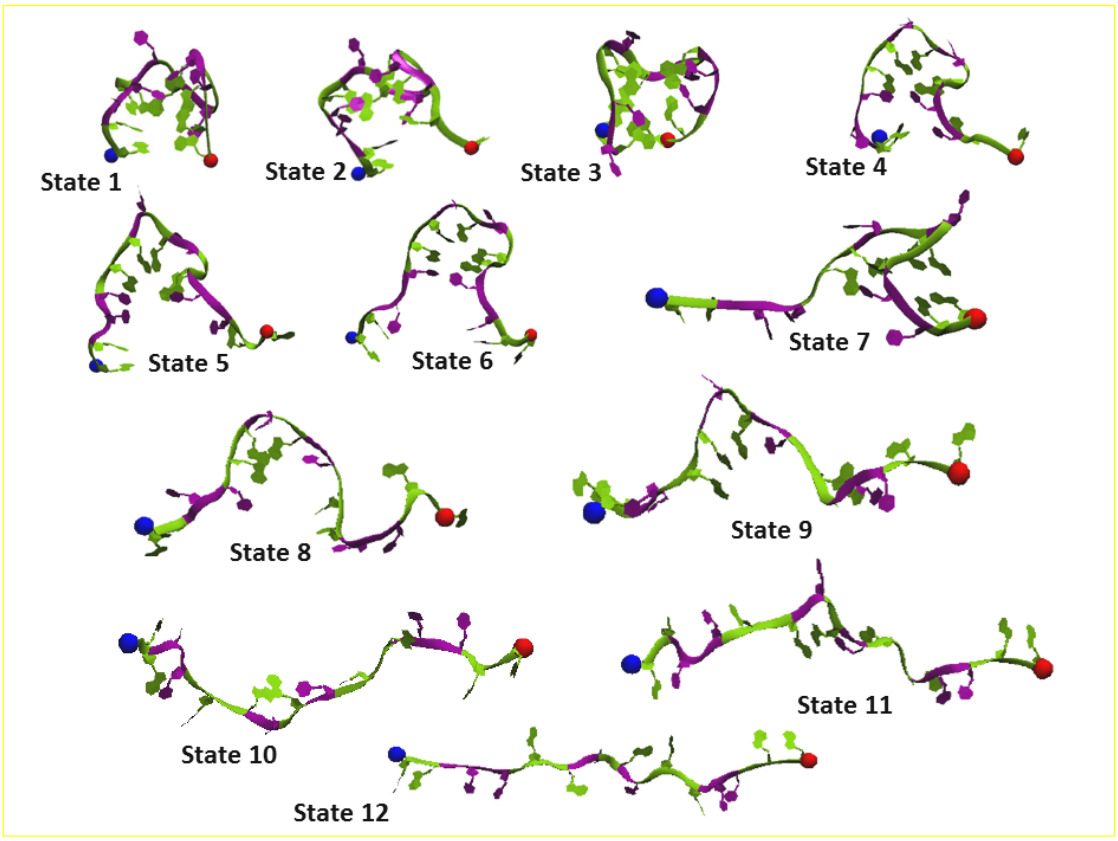

Figure 6.9: Typical structures of the intermediate states on the unfolding pathway. The residues Guanine (G) and Thymine (T) are represented by yellow-green and purple colors respectively. The blue and red circles at the two ends of the molecules indicate C5′ atom at 5′ end and C3′ atom at 3′ end respectively at which opposite forces are applied.

Figure 6.10(a) shows the variation of the free energy, defined as $A_S(F) = -k_B T \ln \pi_S^{eq}(F)$ associated with state S at an applied force F. Here, $\pi_S^{eq}(F)$ is the probability of residence in a given state at equilibrium at a constant stretching force F, k_B is the Boltzmann constant and T is the temperature. The lines denote quadratic fits to the

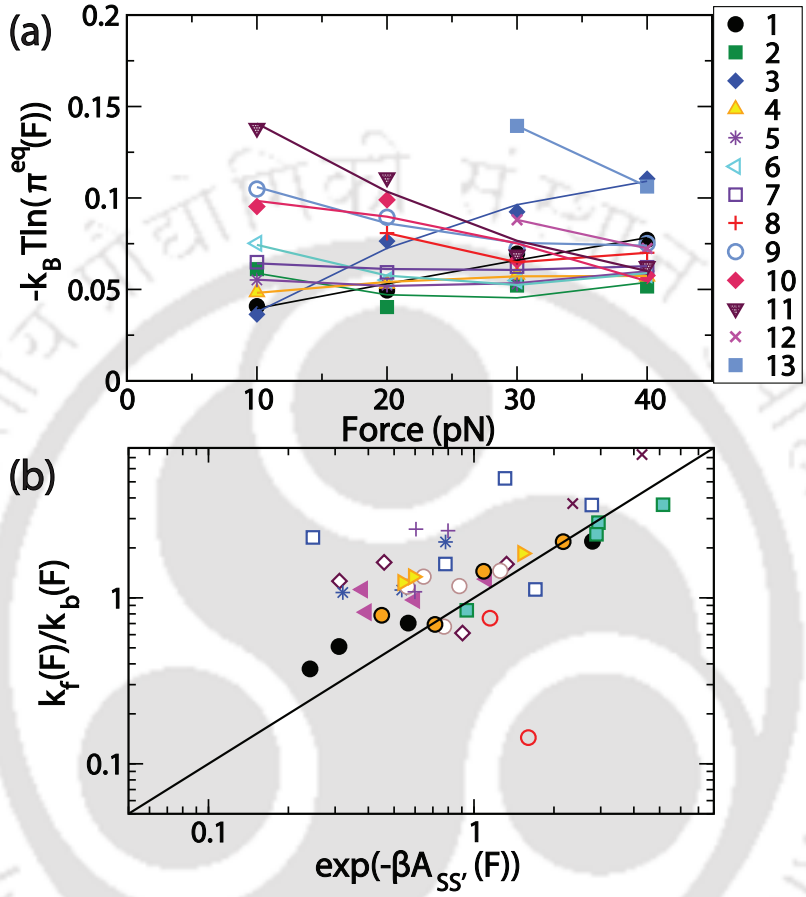

Figure 6.10: (a) Plot of the free energy vs. the applied force for the various states. Symbols represent the value computed from PSC-MD calculations. The lines represent a quadratic fit applied to the data in each case except for state 13. (b) Parity plot used to verify detailed balance for the kinetic pathways detected in the calculations. Each point represents a kinetic pathway where the ordinate gives the ratio of the forward to reverse rates while the abscissa represents $\exp(-\beta A_{SS'}(F))$. Here $A_{SS'}(F)$ represents the free energy difference between the two states S and S' when a stretching force, F, is applied and β is the reciprocal of $k_B T$ where k_B is the Boltzmann constant and T is the temperature.

data for all cases except state 13 (state with extension length > 78 Å), where a linear fit was performed since the state was detected at only two of the forces. States 1, 3 and 4, seen to have compact structures, were found to have diminishing free energies at lower forces, that is, increased occurrence unlike the remaining states. Kinetic rates of transition between states were calculated using MLE. Only those transitions

that were detected at least ten times were considered for generating the MSM. The parity plot in Fig. 6.10(b) comparing the ratio of the forward to backward rates of transition between pairs of states to the corresponding ratio of residence probabilities of the states indicates that detailed balance is satisfied. At each of the four forces, 10, 20, 30 and 40 pN, a 1-d network model was generated using the detected transitions. 20 pN was selected as the reference force. The network model constructed at 20 pN is shown in Fig. 6.11. As in the case of the deca-alanine

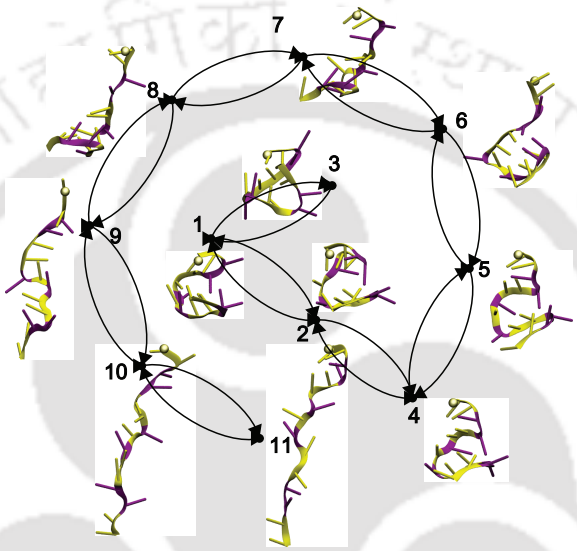

Figure 6.11: Markov state model generated at 20 pN.

molecule in Sec. 6.3, the kinetic parameters $k_f(F_0)$ and α_f are estimated by fitting Eq. (6.12) to rates in the MSMs at different values of F in Fig. 6.12. Figure 6.12 shows kinetic rates of relevant pathways obtained from the MSMs at various forces. The kinetic rates at the reference tension, 20 pN, $k_f(F_0)$, are directly obtained from the MSM constructed at the force F_0. Table 6.5 provides values for $k_f(F_0)$ and α_f for pairs of states. The catalogue of states and the kinetic parameters of Eq. (6.12) comprise the MSM-0 of the TBA system.

Table 6.5: The parameters for the kinetic rates of transition for the TBA molecule.

Initial State	Final State	Rate $k_f(F_0)$ (ps^{-1})	$\alpha_f = \frac{\tan\theta}{(\tan\theta+\tan\phi)}$
1	2	1.29E-03	0.459847
1	3	3.78E-04	0.251809
2	1	7.32E-04	0.072973

2	4	9.12E-04	0.082613
3	1	1.11E-03	0.114201
4	2	8.26E-04	0.038677
4	5	1.39E-03	0.170062
5	4	8.34E-04	0.463098
5	6	1.42E-03	0.009504
6	5	1.11E-03	-0.11232
6	7	1.54E-03	0.062182
7	6	1.36E-03	-0.00709
7	8	8.78E-04	-0.06927
8	7	1.34E-03	-0.02777
8	9	1.15E-03	0.043072
9	8	9.20E-04	0.00912
9	10	1.53E-03	0.04283
10	9	1.66E-03	0.103562
10	11	1.52E-03	-0.00936
11	10	1.99E-03	-0.05168

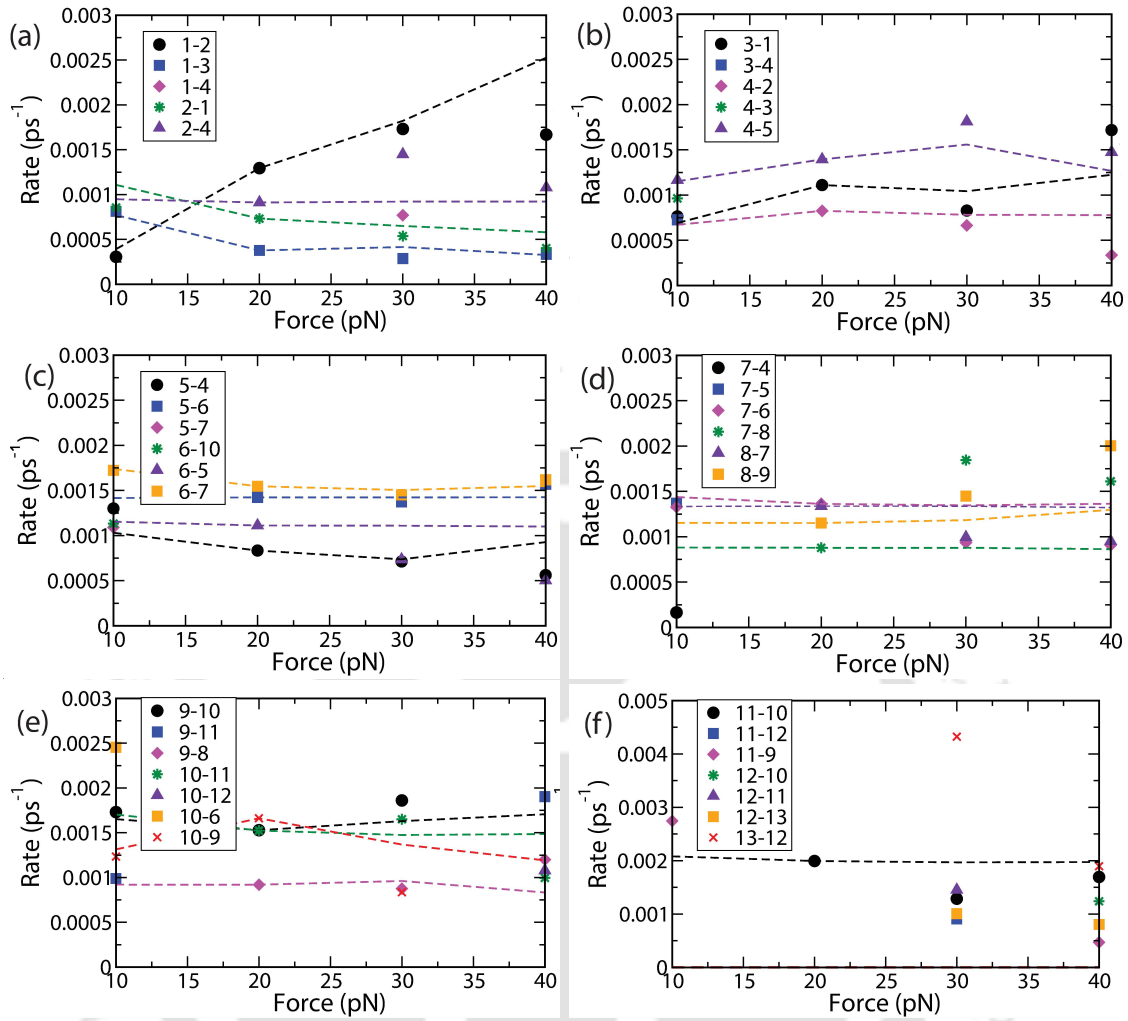

Figure 6.12: (a)-(f) Kinetic rates of relevant pathways obtained from the MSMs at various forces. The symbols represent the kinetic rates obtained from the analysis of MD trajectories. The lines refer to Eq. (6.12). In panel (f) connecting lines (from Eq. (6.12)) are not available for several pathways which were observed only at high forces (>20 pN).

The MSM-0 formalism was then used to predict the relevant kinetic rates at zero-force conditions. The predicted rates were then compared to the rates obtained from direct simulations at force-free conditions in Fig. 6.13. Although, the rates for every transition detected in the force range 10-40 pN, could be extrapolated via Eq. (6.12) to zero-force conditions, the actual simulations yielded only a handful of rates, between states 1-2 and 5-6. The free energy differences at zero-force conditions required for the calculation of the rates were obtained from the extrapolation of the quadratic fits to the free energy plot in Fig. 6.10(a). Although several other states and associated transitions were detected in the zero-force MD calculations, the rates

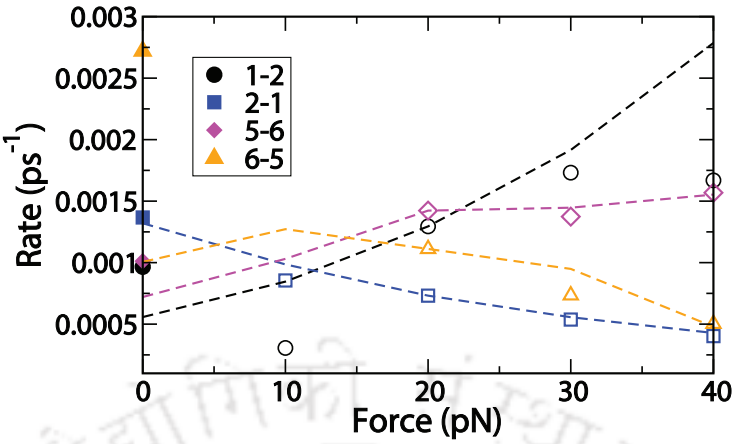

Figure 6.13: Kinetic rates of relevant pathways at zero-force. Solid symbols represent the rates obtained directly from zero-force MD simulations. Empty symbols of the same type indicate the rates at the forces 10, 20, 30 and 40 pN obtained from SMD calculations. The dashed lines of the same color indicate the rates predicted by Eq. (6.12). The predicted rates at zero-force conditions are indicated by the y-intercepts of the dashed lines.

are not presented here due to insufficient sampling. The estimated rates (dashed lines in Fig. 6.13) were found to be in reasonably good agreement with the observed rates (solid symbols in Fig. 6.13).

6.6 Molecular Simulations of RNA hairpin under tension: Prediction of Kinetic Rates at a Specified Force

As our third example, we consider a RNA hairpin model. Ribonucleic acid (RNA) has been found to be associated with a variety of cellular processes[271] including regulation of gene expression and protein synthesis[272,273]. To execute their function RNA molecules need to be folded into specific three dimensional structure. There are two aspects of RNA folding. The first one is the prediction of 3-D structure from the RNA sequences[274,275]. The second aspect concerns the kinetic mechanism by which the formation of assembly of 3D functionally competent structure takes place starting from unfolded conformations. At a first glance, it would seem that RNA folding mechanism should be simple at least in comparison to the better investigated protein folding problem. But the rugged nature of energy landscape arising due to polyelectrolyte character of the phosphate backbone, ion-RNA interaction *in vivo*, base stacking and base pair formation by Watson-Crick hydrogen bond-

ing[276]; contribute to the complexity of RNA folding. Furthermore, their propensity to be misfolded due to the formation of various stable base pairs and base stacks is a major folding problem encountered by RNA molecules. The RNA folding mechanism has been studied extensively *in vitro* and several ribozyme folding paradigms have been revealed[277–290]. Despite the significant progress in understanding RNA folding, many aspects of RNA folding remain unsolved, for example, the identification of intermediate states and unfolding pathways of small RNA molecules, RNA hairpins or larger RNA structures is not easy as they require rigorous analysis to study the slow dynamical process of RNA. A few single-molecule force-spectroscopy measurements[291–295] and pulling simulations[296–299] were carried out on RNA hairpin to probe folding energy landscape of RNA molecule but they provide a little information about kinetic properties. We apply the method described in the preceding sections in order to illuminate possible intermediate states and kinetic pathways using MSM-0 constructed with a handful of SMD simulations of a target RNA hairpin molecule at various constant forces.

6.6.1 System Setup

Initial coordinates for the RNA hairpin structure with sequence UCUUCGGG (Fig. 6.14) were taken by extracting eight residue from the X-ray structure 1C0O[300] which has a sequence-GGGUCUUCGGGUCC. We deleted the first 3 and last 3 nucleotides

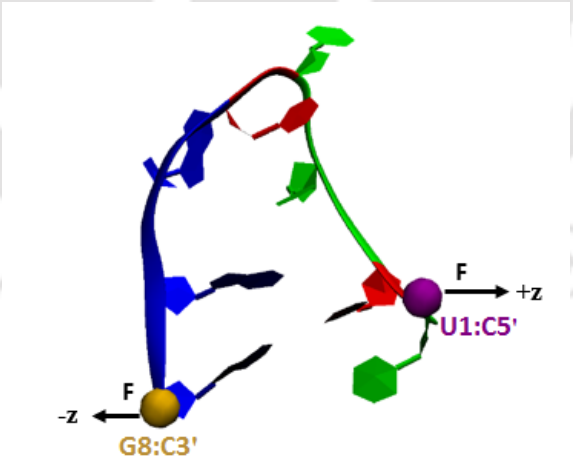

Figure 6.14: Hairpin structure of RNA molecules with sequence UCUUCGGG. The color code for the residue URA, CYT and GUA are green, red and blue respectively.

of the initial RNA hairpin model with a 4-bp helix and retained the 2-bp at the middle. The RNA system was solvated in a periodic water box of 1622 TIP3P water

molecules with a box size $30 \times 32 \times 56$ Å^3, and additional 5 sodium (Na^+) ions were added to neutralize the system. Then, 5 Na^+ and 5 Cl^- ions were placed to ionize the system with a salt concentration of 150 mM. The total number of atoms in the solvated RNA hairpin model was 5088. As a part of SMD simulation with constant force, equal and opposite forces (F) were applied at the C5′ atom of the first residue and C3′ atom of the last residue as shown in Fig. 6.14.

6.6.2 Simulation Protocols

To perform all the simulations, we used NAMD 2.11[183] and CHAMM36[206] force field parameters. The system temperature was kept at 310 K by coupling the system to a Langevin heat bath. The integration time step was set to 2 fs. The PME method was used to calculate the long range electrostatics with a charge grid spacing of ∼1 Å and a cutoff of 10 Å was applied to the Lennard-Jones and direct space electrostatic interactions. The solvated system was first minimized (by 10000 conjugate gradient cycles) at 310 K. Then, 1 ns NPT simulation at 310 K with 1 atm constant pressure was carried out to equilibrate the density of the solvated system. This was followed by 5 ns long NVT simulations under equivalent conditions to collect the starting configuration of PSC calculations (see Sec. 4.4 of chapter 4).

6.6.3 Construction of Kinetic Networks for the RNA Hairpin

The PSC algorithm was applied to generate the MSMs with specified validity times with constant forces varying between 90 and 140 pN. The network models were generated using a one-dimensional reaction coordinate, the distance between C5′ and C3′ atoms of the first and last residues respectively, termed the extension length. In this case, a bin width of 5 Å was selected for classifying the configurational states. A total of 7 states were identified with extension range given in Table 6.6. Representative configurations for the states, named in order of discovery, are provided in Fig. 6.15.

Table 6.6: List of the states of the RNA hairpin and the range of extension lengths for each state. The range includes the lower bound but not the upper bound in each case. The states are labeled in order of detection in the MSM construction method.

State	Extension Length Range (Å)
1	15-20
2	20-25
3	25-30
4	30-35
5	35-40
6	40-45
7	>45

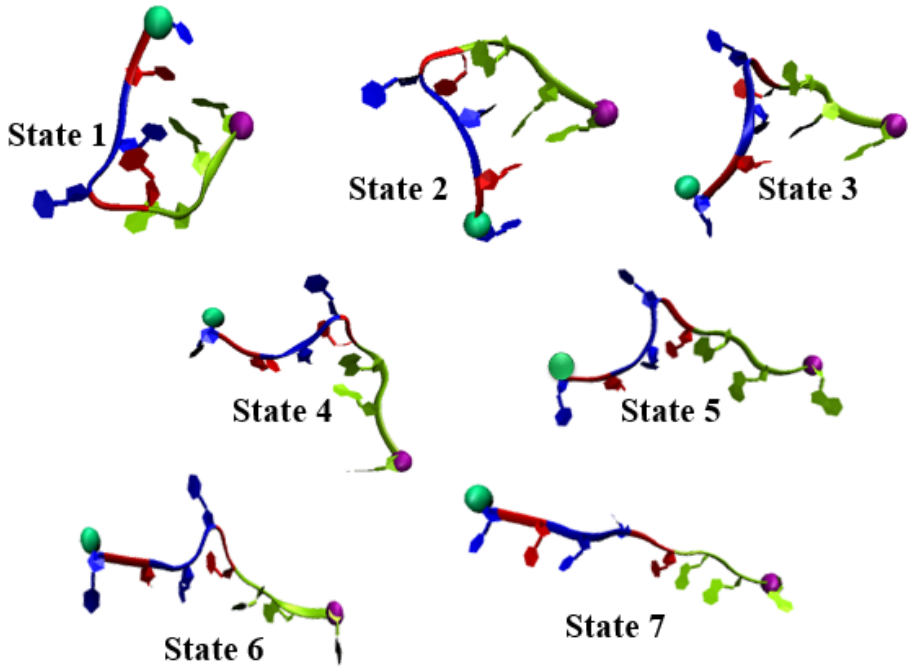

Figure 6.15: Typical structures of seven states of RNA hairpin. The color code for the residue GUA (G), URA (U) and CYT (C) are blue, yellow-green and red respectively. The green and purple circles indicate the C5′ atom of residue 1 and C3′ atom of residue 8.

The parity plot in Fig. 6.16(a) comparing the ratio of the forward and backward rates of transition between pairs of states to the corresponding ratio of residence probabilities of the state indicate detailed balance to be satisfied. Figure 6.16(b) shows the variation of the free energy, defined as $A_S(F) = -k_B T \ln \pi_S^{eq}(F)$ associ-

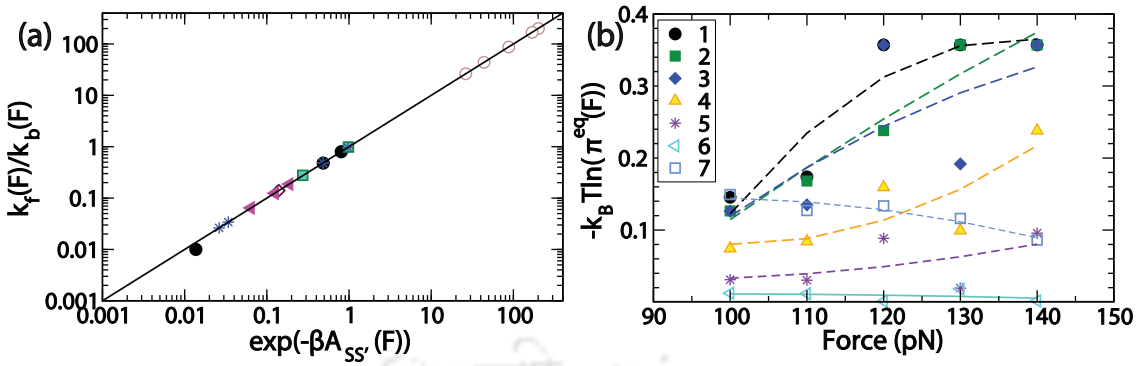

Figure 6.16: (a) Parity plot used to verify detailed balance for the kinetic pathways detected in the calculations. Each point represents a kinetic pathway where the ordinate gives the ratio of the forward and backward rates while the abscissa represents $\exp(-\beta A_{ss'}(F))$. (b) Plot of the free energy vs. the applied force for the various states. Symbols represent the value computed from PSC-MD calculations. The lines represent a quadratic fit applied to the data in each case.

ated with state S at an applied force F. As in the previous examples, $\pi_S^{eq}(F)$ is the probability of residence in a given state at equilibrium at a constant stretching force F, k_B is the Boltzmann constant and T is the temperature. The lines denote quadratic fits to the data. States 1, 2, 3 and 4, seen to have compact structures, were found to have diminishing free energies at lower forces, that is, increased occurrence unlike the remaining states. Kinetic rates of transition between states were calculated using MLE. As in the previous examples, only those transitions that were detected at least ten times were considered for generating the MSMs and are presented in the discussion.

At each of the six forces, 90, 100, 110, 120, 130 and 140 pN, an MSM was generated using the detected transitions. 110 pN was selected as the reference force. MSM-0 was constructed using the catalogue of states and rate parameters obtained from the MSMs constructed at the forces between 100-140 pN. Note that 90 pN MSM was not used to calculate the rate parameters. The network model constructed at 110 pN is shown in Fig. 6.17(a). As in the case of the deca-alanine molecule in Section 6.3, the kinetic parameters $k_f(F_0)$ and α_f are estimated by fitting Eq. (6.12) to rates in the MSMs at different forces (between 100 and 140 pN) in Fig. 6.17(b-c). Table 6.7 provides values for $k_f(F_0)$ and α_f for pairs of states.

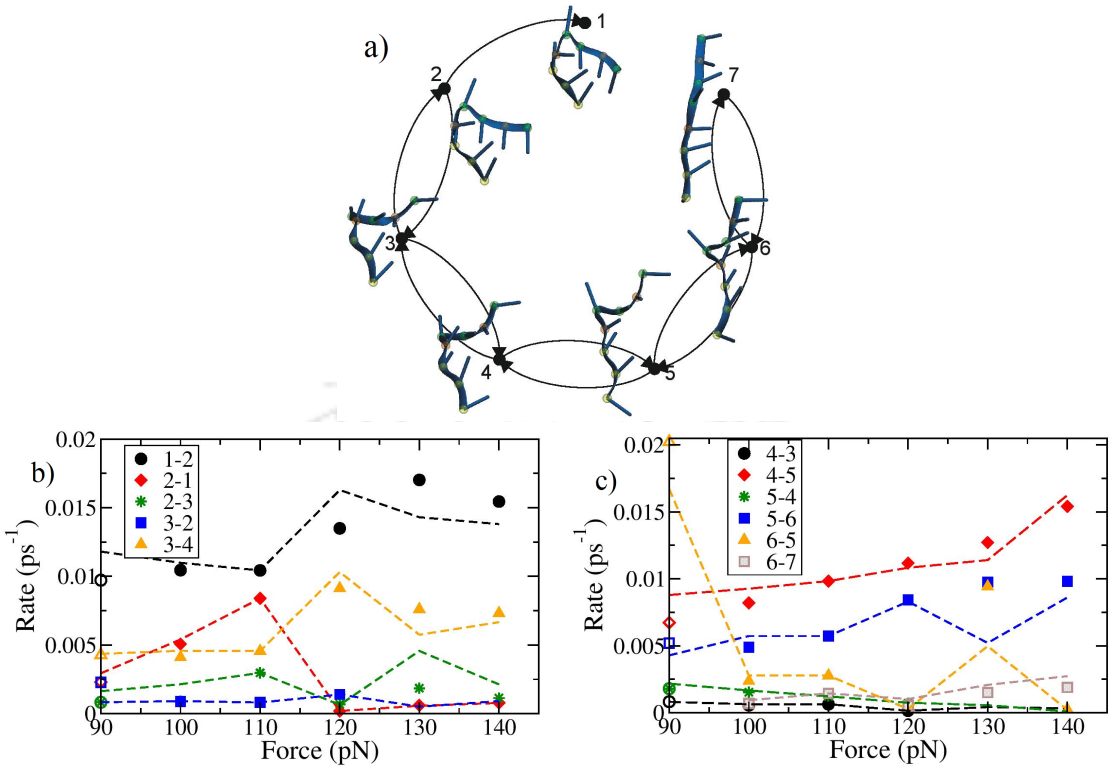

Figure 6.17: (a) Markov state model generated at 110 pN. (b)-(c) Kinetic rates of relevant pathways obtained from the MSMs at various forces. The symbols represent the kinetic rates obtained from the analysis of MD trajectories. The lines refer to Eq. (6.12) generated with the MSMs at 100, 110,120, 130 and 140 pN. The rates at 90 pN, obtained by extrapolating via Eq. (6.12) (as represented by the dashed lines) are compared to the directly computed kinetic rates at 90 pN PSC calculations (empty symbols).

Table 6.7: Kinetic parameters for the RNA molecule.

Initial State	Final State	Rate $k_f(F_0)$ (ps^{-1})	$\alpha_f = \frac{\tan\theta}{(\tan\theta+\tan\phi)}$
1	2	1.04E-02	0.101699
2	1	8.39E-03	0.862183
2	3	2.97E-03	0.260046
3	2	8.13E-04	0.091414
3	4	4.56E-03	0.14365
4	3	6.40E-04	0.251428
4	5	9.83E-03	0.146387
5	4	1.22E-03	0.753261

5	6	5.73E-03	0.139405
6	5	2.78E-03	0.861036
6	7	1.45E-03	0.528748
7	6	1.27E-01	0.474352

The master-MSM or MSM-0 thus constructed for the RNA hairpin system was used to predict the network model at 90 pN through Eq. (6.12) and the quadratic fit to the free energies of the relevant states shown in Fig. 6.16(b). The predicted rates, indicated by the dashed lines in Fig. 6.17(b) and (c), are in good agreement with the rates computed directly from PSC-MD calculations at 90 pN (empty symbols in Fig. 6.17(b)-(c)).

6.7 Discussion

Using an extended approach of the Master-MSM method for constant trap separation experiments in FS set up, here we construct a master-MSM for constant-force experiments to predict the intrinsic kinetic properties of the system from constant-F ensemble MD data. The underpinning concept behind our methodology is the relation between the kinetic rates at different force conditions based on the BEP principle and extrapolating the kinetic rate *vs* force plot to zero-force. For this purpose, first we construct a few kinetic network models at various constant forces within a range to capture the fully extended configuration of the system. Next, a master-MSM (MSM-0) is constructed at reference force (F_0) along with a parametrization of kinetic rates which is the basis of our methodology. The methodology presented here is demonstrated by taking three examples of deca-alanine, TBA and RNA-hairpin systems. The thermodynamic consistency of the system is ensured by checking detailed balance condition for each relevant kinetic pathway. The reliability of our model is verified by a comparison of the relaxation timescales and kinetic rates obtained from constant-F and constant-d ensemble data of deca-alanine. The method presented here can be considered as an enhanced kinetic method, can help us to predict the kinetic rates of transition between relevant states at zero force. Thus, the present approach may provide a overall insight into the unfolding mechanism of kinetically slow systems and unfolding free energy profile of the system under consideration at force-free conditions with less computational effort by taking the advantage of the fact that application of force induce the transitions to occur more quickly.

Chapter 7

Conclusions

In this thesis, we have developed a new class of enhanced kinetic sampling methods for biomolecular simulations. Our focus is on kinetics unlike most other methods are centered on thermodynamic aspects. Thermodynamic studies of systems mainly involve the description of free energy landscapes and equilibrium properties of the system, but it does not help us to get kinetic properties such as transition rates and barrier heights. In this thesis, data for the conformational dynamics of a biomolecular system governed by the classical equation of motion was generated using molecular dynamics simulations. The main focus of the research is on the efficient construction of kinetic network from MD trajectories that is human comprehensible and provides a deeper view of molecular kinetics distilled from a massive amount of data. Markov State Model (MSM) is such an example of molecular kinetic network. So, the first part of our thesis places a special attention to the construction of MSM much accurately and efficiently. To ensure the accuracy of MSM, a theoretical framework of calculation of validity time (τ_V) of an MSM has been proposed. For efficient generation of MSM and extension of validity time, a self-learning algorithm has been developed. The second part of our thesis focuses on the construction of MSM for Single-molecule Force Spectroscopy (SMFS) taking the advantage of the fact that application of force can induce conformational transitions that are otherwise rarely sampled. However, the application of force on the system perturbs the dynamics of the system and the extraction of equilibrium kinetic properties from perturbed dynamics is not straight-forward. Moreover, probing kinetics of biomolecular systems under tension is complicated due to the presence of multiple intermediate state connected by multiple kinetic pathways. In order to address these challenges, we have introduced the idea of master-MSM for finding a connection between topologically distinct MSMs (at different stretching conditions) based on Bell-Evans-Polanyi principle. Then a Time-Dependent Markov State Model (TD-MSM) has been pro-

posed by using the concept of master-MSM to predict the non-equilibrium dynamics of the system in SMFS. There are various modes of SMFS: force-ramp, force-jump, constant-force etc. We have attempted to show that our model is sufficiently versatile to encompass various kind of SMFS experiments. We have employed the concept of master-MSM directly to constant-force experiments to recover the intrinsic kinetic properties of the system.

The primary challenges for building an MSM are choosing a good state definition for generation of conformational states and computation of kinetic rates of transition between conformational states. We used different methods in the various systems those we have probed. In case of the deca-alanine molecule, the states were generated by comparing the backbone atoms after aligning the molecule using the Kabsch algorithm where two consecutive conformations in MD frames were categorized as different states if their structural similarity differ by a tolerance value. For determination of correct tolerance value, we carried out a preliminary analysis with multiple swarm MD calculations using various tolerances. An examination of the state-wise histograms for various parameters, such as distances between specified atoms, helped us to choose the tolerance. In contrast, for TBA and RNA-hairpin systems, we adopted a coarse-grained approach where states were distinguished by taking into account only one reaction coordinate. In this approach, the distance between the $C5'$ and $C3'$ atoms of the first and last residues was defined as extension length (ξ) and splitted into several windows up to maximum value of ξ to cover a full extension range of the system. The interval per window was kept 6/5 Å for TBA/RNA system. Then we extracted the representative conformation for each window of extension. In future, we can take advantage of various clustering techniques for better decomposing of the conformational space. In this thesis, we used a Maximum Likelihood Estimation (MLE) method to calculate the kinetic rate. For the MLE method to be valid, the observed kinetic rates of each kinetic pathway are required to obey exponential distribution according to first order kinetics. Therefore, first order kinetics was verified for several pathways.

Next, we addressed the uncertainty in MSM due to missing kinetic information by introducing the concept of validity time. Within the timescale of validity time the dynamics of the system can be predicted faithfully with a chosen accuracy. Beyond the validity time, the MSM-predicted kinetics may diverge from the true kinetics of the system due to the occurrence of rare transitions that were missing in the MSM. A theoretical framework was provided for the calculation of validity time and using

an MSM of solvated Alanine dipeptide we concluded that it is important to state the validity time of an existing MSM to describe the dynamics accurately.

With the object of efficiently increasing the validity time of an MSM, we developed swarm MD and state-constrained MD (SC-MD) methods for an accurate and efficient construction of MSM. An adaptive method, named by programmed state constrained MD (PSC-MD) method was developed to accelerate the extension of the validity time of MSM. The new methods and concepts were demonstrated by a prototype example and deca-alanine under tension as test model. The study on deca-alanine reveals that the missing states may be relevant to the dynamics at longer timescales.

Pursuing an idea first developed in SMFS experiments, we then developed an enhanced kinetic sampling strategy that is based on applying forces on molecules to enhance kinetic rates. We developed the idea of a master-MSM or MSM-0 for constant-probe separation (d) experiments to predict the connection between topologically different kinetic networks constructed at distinct trap separations. In addition, the availability of state-specific force models enable calculation of force-extension behaviour in a variety of ensembles. Changes in the network topology upon stretching is related through a thermodynamic quantity termed as mechanical disposition (χ). On the basis of the idea of constructing master-MSM at a reference extension, a time-dependent MSM (TD-MSM) was proposed to study the dynamics of time-dependent pulling force experiments where the time dependence is inherited in the model. Here the exploitation of the concept of SMFS in construction of MSM can help us to get a detailed insight into kinetic, thermodynamics and mechanical properties of the system.

Although the constant-d (trap separation) ensemble allows one to understand the time evolution of non-equilibrium system, it is not trivial to predict the intrinsic kinetic rates. Hence, we extended the technique to the constant force ensemble with the aim of recovering the zero-force network. We demonstrated how the method can provide meaningful information about long timescale dynamics without actually performing simulation at force-free conditions. Our approach was demonstrated by previously well-studied system-deca alanine and applied to TBA and RNA-hairpin systems.

At this point of our methodology for construction MSM in SMFS experiments

needs further investigation and improvement. We have started with a simpler example of deca-alanine system which has only one secondary component (helix structure). Our approximation about the linearity of potential energy surface (PES) near saddle point and quadratic approximation of free-energy difference on anchor separation in TD-MSM approach were found to be valid for deca-alanine system. Real protein systems have both helical and beta strand components and exhibit tertiary structure. Hence, the feasibility of using these methods on more complicated peptides need to be tested. Furthermore, we can improve our methodology by taking parabolic form of PES near saddle point instead of linear form. In our study of elastic properties we have considered the deca-alanine molecule as harmonic spring associated with a spring constant and equilibrium length for each state. It is not obvious whether such models would hold for more complicated structures. Systems of increasing complexity may require new models to explain their elastic behaviour. In bigger systems, kinetic trapping remains a major challenge. Therefore, for the TBA and the RNA system, we opted to verify our ideas using a coarse-grained 1-dimensional reaction coordinate. However, this entails loss of valuable kinetic information. The complexity of the dynamics on a multi-dimensional landscape may be uncovered by constructing MSM using multiple degrees of freedom. Finally, we believe that the method can be extended to investigate the various biomolecular process including protein/nucleic acid folding, protein-ligand binding, peptide aggregation, dynamics of intrinsic disordered proteins etc.

Appendix A

Basic Theory and Analysis Techniques

A.1 Theory

A.1.1 Rate Constant and Transition State Theory

The rate constant normally depends on the absolute temperature, and the functional form of this relationship was first proposed by Arrhenius in 1889 to be:

$$k = \bar{A}\exp[-E_a/RT] \tag{A.1}$$

where the activation energy, E_a and the pre-exponential or frequency factor, $\bar{A}$, both do not depend on the absolute temperature. The Arrhenius form of the reaction rate constant is an empirical relationship. However, transition-state theory provides a justification for the previously used Arrhenius formulation, as is discussed below.

The discussion on reaction rate involves two important theories: (1) collision theory and (2) transition-state theory. The collision theory tells us how frequently the molecules involved actually have to bump into each other, and allows the calculation of the efficiency of a reaction, that provides the maximum possible rate. Unfortunately, collision theory tells us nothing about steric factors or the activation energy, and these parameters must be determined experimentally. To calculate the rate constants, we need to consider the chemical properties of the reactants and the activated complex. This is done with either a statistical mechanical approach or chemical thermodynamic approach, named by "Transition State Theory" (TST), developed by Henry Eyring in 1935. The TST is also known as activated complex

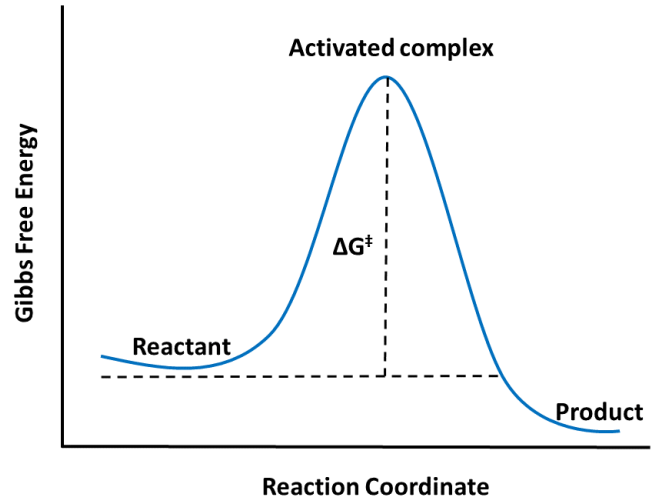

Figure A.1: Illustration of the Gibb's free energy of activation.

theory.

In a chemical reaction, all the structures in the vicinity of this transition state may be considered as the "activated complex", which is very reactive. Although the terms "transition state" and "activated" complex are often used synonymously, the transition state does not have a chemically significant life time. In a chemical reaction, a motion along the "forward" direction will lead to the products. The fundamental assumption of activated complex theory is that transition state is in equilibrium with the reactant and product. According to Eyring equation the transition state rate constant is:

$$\begin{aligned} k &= (k_B T/h) e^{-\Delta G^{\ddagger}/RT} \\ &= \frac{k_B T}{h} e^{-\frac{\Delta H^{\ddagger}}{RT}} e^{\frac{\Delta S^{\ddagger}}{R}} \end{aligned} \tag{A.2}$$

where $\Delta G^{\ddagger}$ ($=\Delta H^{\ddagger}$-T$\Delta S^{\ddagger}$) is the Gibb's free energy of activation, $\Delta H^{\ddagger}$ is the enthalpy of activation and $\Delta S^{\ddagger}$ is the entropy of activation. Here, k_B is the Boltzmann constant, h is the Planck constant, T is the thermodynamic temperature and R is the universal Gas constant. As shown in the Fig. A.1 above, free energy of activation represents the difference in energy between the reactant state and the activated complex (transition state), or,

$$\Delta G^{\ddagger} = G(\text{activated complex}) - G(\text{reactants}). \tag{A.3}$$

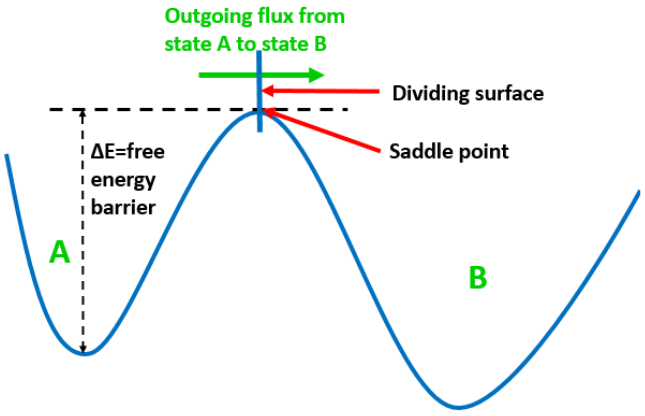

Figure A.2: Illustration of the transition state theory rate constant.

Note that this activation enthalpy quantity, $\Delta H^{\ddagger}$, is analogous to the activation energy quantity, E_a, when comparing the Arrhenius equation (Eq. (A.1)) with the Eyring equation (Eq. (A.2))

The TST formalism for a chemical reaction that proceeds along the reaction coordinate over the transition state has been also extended to study the time evolution of a system undergoing conformational transitions. In the context of conformational transition, the TST formalism approximates the classical rate constant for escape from state A to some adjacent state B as equilibrium outgoing flux through the dividing surface between A and B (Fig. A.2). The dividing surface is a 3N-1 dimensional hyperplane, a Transition state (TS), that represents a bottle neck for going from an initial to a final state. If there are no correlated dynamical events, the TST rate is the exact rate constant otherwise TST can be corrected for recrossing effects to give the exact rate. The TST rate from a state A ($k_{A\rightarrow}^{TST}$) through a dividing surface at $x_1 = q$ is simply proportional to the Boltzmann probability of being at the dividing surface relative to the probability of being anywhere in state A. Mathematically,

$$\kappa_{A\rightarrow}^{TST} = <\mid \frac{dx_1}{dt} \mid \delta(x_1 - q) >_A . \tag{A.4}$$

Here the angular bracket indicates the ratio of Boltzmann-weighted integrals over 6N-dimensional phase space (configuration space $\mathbf{r}$ and momentum space $\mathbf{p}$). Dirac delta function picks out the probability of the system being at the dividing surface, relative to everywhere it can be in state A. For simplicity here, the dividing surface is at $x_1 = q$, involving only the reaction coordinate x_1 ($x_1 \in \mathbf{r}$).

A.1.2 Harmonic Transition State Theory

The harmonic approximation of TST is often used to calculate TST rate constants. Harmonic TST (HTST) is often referred as Vineyard theory[301]. In HTST, we require that transition pathway is characterized by a saddle point on the potential energy surface. One assumes that the potential energy near the basin minimum is well described (out to displacements sampled thermally) with a second-order energy expansion, that is, the vibrational modes are harmonic and that same is true for the modes perpendicular to the reaction coordinate at the saddle point. The dividing surface is taken to be the saddle plane (the hyperplane perpendicular to the reaction coordinate at the saddle point), and evaluation of the average in Eq. (A.4) for a system with N moving atoms gives the simple form

$$\kappa^{HTST} = \frac{\prod_{i}^{3N} \nu_i^{min}}{\prod_{i}^{3N-1} \nu_i^{sad}} exp(-\Delta E/k_B T). \tag{A.5}$$

Here ΔE is the static barrier height (energy difference between the saddle point and the minimum) and k_B is the Boltzmann constant. In the pre-exponential factor, $\{\nu^{min}\}$ are the 3N normal mode frequencies at the minimum and $\{\nu^{sad}\}$ are the 3N-1 non-imaginary normal mode frequencies at the saddle. The computation of k^{HTST} thus requires information only about the minimum and the saddle point for a given pathway. The analytic integration over the whole phase space thus leaves a very simple Arrhenius temperature dependence. Although the exponent depends only on the static barrier height, there is no assumption that the trajectory passes exactly through the saddle point. To the extent that there are no recrossing and the modes are truly harmonic, this is an exact expression for the rate. We have employed this HTST expression of kinetic rate in our TDMSM approach.

A.2 Analysis Techniques

Root Mean Square Deviation (RMSD)

The RMSD distinguishes the extent that a particular particle translates from a defined reference point in the simulation system. The RMSD is calculated according

to,

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} \left(r(t) - r(t_0)\right)^2}{N}} = \sqrt{\langle \Delta r^2 \rangle}$$

where $r(t_0)$ is the reference position, r(t) is the location of a particle in timestep t, and N is the total number of atoms in the molecule. Before calculating the RMSD value for a single snapshot, each snapshot is rotated and translated to superpose itself on the reference coordinates by Kabsch algorithm, in order to filter out large-scale motions that are not indicative of conformational changes.

A.2.1 Kabsch Algorithm

Kabsch algorithm is a method used to calculate the optimal rotation matrix which minimizes the RMSD between two paired set of points. This algorithm works in three step: a translation, the computation of a covariance matrix, and the computation of the optimal rotation matrix.

Translation: Both sets of coordinates must be translated first, so that their centroid coincide with the origin of the coordinate system. This is done by subtracting the point of origin from the coordinates of the respective centroid. The centroids are just the average point and can be calculated as foloows:

$$T = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$centroid_T = \frac{1}{N} \sum_{i=1}^{N} T^i$$

where T^i are points in dataset P.

Computation of the covariance matrix: The second step consist of calculating a cross-covariance matrix between two set of paired points, P and Q. In matrix notation,

$$A = P^T Q,$$

or, using summation,

$$A_{ij} = \sum_{k=1}^{N} P_{ki} Q_{kj}$$

Computation of the optimal rotation matrix:
First, calculate the Singular value decomposition (SVD) of the covariance matrix A as

$$A = V\Sigma W^T$$

where V is orthogonal, Σ is diagonal, and W is orthogonal matrix. If A is a square matrix then V, Σ and W are the same size as well.

Next, decide whether we need to correct our rotation matrix to ensure a right-handed coordinate system. It is possible by computing d as

$$d = \text{sign}(\det(WV^T))$$

Finally the optimal rotation matrix is calculated as:

$$U = W \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix} V^T.$$

We can express total squared distance (E) between two set of conformations with n points vectors as a function of the optimal rotation matrix, U:

$$E = \sum_{i=1}^{n} ||p_i - q_i U||^2.$$

The RMSD is then

$$RMSD = \sqrt{E}/N.$$

A.2.2 H-bond

Hydrogen bonding is a special type of dipole-dipole attraction between molecules, not a covalent bond to a hydrogen atom. H-bond interaction occurs when a hydrogen atom bounded to a strongly electronegative (ex. N, O, F etc.) atom exists in the vicinity of another electronegative atom with a lone pair of electrons. This bonds are generally stronger than ordinary dipole-dipole and dispersion forces, but weaker than true covalent and ionic bonds. Formation of H-bonds between donors and acceptors

is defined by two parameters: the distance between the donor (D) and acceptor (A), and 2) the angle defined by D-H-A. In our measurements, the distance cutoff was set at 3.4 Å and the maximum angle was set to 30°.

A.2.3 Helicity

The calculation of α-helix, 3_{10} and π-helical contents of a protein are done by VMD package.

A.2.4 Runge-kutta-Fehlberg Method (RKF45)

Runge-Kutta is a numerical solver providing an efficient explicit method solve Ordinary Differential Equation (ODE) initial value problems. We have used the standard Runge Kutta 4-5 method to compute the approximate solution of Eq. (5.36). It has a procedure to determine if the proper step size h is being used. At each step, two different approximations (step size h and $h/2$) for the solution are made and compared. If the two answers are in close agreement, the approximation is accepted. If the two answers agree to more significant digits than required, the step size is increased.

Each step requires the use of the following six values:

$$\begin{aligned}
k_1 &= hf(x_k, y_k),\\
k_2 &= hf\left(x_k + \frac{1}{4}h, y_k + \frac{1}{4}k_1\right),\\
k_3 &= hf\left(x_k + \frac{3}{8}h, y_k + \frac{3}{32}k_1 + \frac{9}{32}k_2\right),\\
k_4 &= hf\left(x_k + \frac{12}{13}h, y_k + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right),\\
k_5 &= hf\left(x_k + h, y_k + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right),\\
k_6 &= hf\left(x_k + \frac{1}{2}h, y_k - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104} - \frac{11}{40}k_5\right).
\end{aligned} \tag{A.6}$$

Then an approximation to the solution of the ODE is made using a Runge-Kutta method of order 4:

$$y_{k+1} = y_k + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4101}k_4 - \frac{1}{5}k_5. \tag{A.7}$$

where the four function values k_1, k_2, k_3, k_4 and k_5 are used. Notice that k_2 is not used in formula (Eq. (A.7)). A better value for the solution is determined using a

Runge-Kutta method of order 5:

$$z_{k+1} = y_k + \frac{16}{135}k_1 + \frac{6656}{12,825}k_3 + \frac{28,561}{56,430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6. \tag{A.8}$$

The optimal step size sh can be determined by multiplying the scalar s times the current step size h. The scalar s is

$$s = \left(\frac{\text{tol } h}{2|z_{k+1} - y_{k+1}|}\right)^{1/4} \approx 0.84\left(\frac{\text{tol } h}{|z_{k+1} - y_{k+1}|}\right)^{1/4}. \tag{A.9}$$

where *tol* is the specified error control tolerance.

A.2.5 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) method is the procedure of seeking the value of one or more parameters for a given statistics which maximizes the known likelihood distribution. The likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model.

Let us consider a random sample X_1, X_2, X_3,.....,X_n from an unknown population whose assumed probability depends on some unknown parameter θ. The goal of data analysis in MLE method is to find a point estimator $(u(X_1, X_2, X_3, ..., X_n)$ such that $u(x_1, x_2, ---x_n)$ is a 'good' point estimate where $x_1, x_2, ----, x_n$ are the observed values of the random sample. Suppose the probability density function of each X_i is $f(x_i; \theta)$ we can define the likelihood function $(Lik(\theta))$ as

$$Lik(\theta) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot f(x_3; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{A.10}$$

The maximum likelihood estimate $(\hat{\theta})$ for the parameter θ is that value of θ that maximizes $Lik(\theta)$. For computational convenience, the MLE estimate is obtained by maximizing the log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \log(f(x_i|\theta)) \tag{A.11}$$

For example, a random sample $X = (X_1, ..., X_n)$ is assumed to follow Poisson distribution:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{A.12}$$

with parameter λ, then our goal will be a good estimate of λ using the data $x_1, x_2, ..., x_n$ that we obtained from our specific random sample. Now the log likelihood will be:

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n}(x_i \log\lambda - \lambda - \log x_i!) \\ &= \log\lambda \sum_{i=1} n x_i - n\lambda - \sum_{i=1}^{n} \log x_i! \end{aligned} \tag{A.13}$$

We need to find the maximum by finding the derivative:

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \tag{A.14}$$

which implies that the estimate should be

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \tag{A.15}$$

A.3 Softwares and Programming Languages Used

- **VMD**: Package for visualizing, animating and analyzing the biomolecular system.
- **NAMD**: MD simulation package.
- **Xmgrace**: 2-d plotting program.
- **Matlab:** Software for programming language and plotting data.
- **Gephi:** Software for visualization and exploration of networks.
- **Programming Lanuguage:** Fortran, tcl and SH.

Appendix B

Biomolecules' Structure

Most of the molecules found in the cell are able to intricate itself into folded, well-defined three-dimensional structure to perform its function correctly and efficiently. The main classes of biomolecules are lipids, carbohydrates, protein and nucleic acid. Each type of macromolecules possesses distinct chemical properties that suit it for the functions which it serves in the cell. In a range of standard molecular biology textbooks[302,304] detailed overviews of biomolecule's structure and function has been obtained. In this appendix we will give the brief description of the structure and function of protein and nucleic acid.

B.1 Protein

Protein are the most versatile biological macromolecules present in the cell of all living organisms and responsible for most of the complex functions that make life possible. Much of the contexture of human body are made from poly-proteins; muscle, hair, cartilage, ligaments, nails, feathers - these are all mainly protein materials (keratine, actin and myosin, etc.). It perform an astonishing variety of function spanning every level of cellular processes. The catalysis of chemical reactions, metabolic regulation, selective transport of small molecules, intra and inter cellular messaging, the maintenance of cell shape as well as DNA replication, repair and translation, hormones such as insulin, antibodies, defense and many more are the examples of protein functions. Despite their phenotypic variation all proteins share a common underlying composition; all are polymer chain formed from the 20 amino acid monomers. An enormous diversity of sizes and structure arises in nature from the possible different combination of amino acids. Protein function is almost dependent on protein structure. So a good understanding of the nature of protein structure and the conformational dynamics of the specific protein being formulated are essential

for pharmaceutical target based drug design.

B.1.1 Amino Acid

Amino acids plays a central role as building block of protein containing a central carbon atom (α-carbon, CA or C), an amine group (-NH_2), a carboxylic acid group (-COOH), a hydrogen atom and a distinctive side-chain group (-R). The typical chemical structure of an amino acid is shown in Fig. B.1. The amino group attached to the central carbon atom immediately adjacent to the carboxylic acid group. Those are the most common alpha-amino acid where side-chain constituent R-group connected to α-carbon. There are 20 different types of amino acid and protein consist of a unique combination of amino acids drawn from this 20-members library. The reason amino acids can be linked together because of strong interaction of the acidic -NH2 group at the left side of one to the acidic group -COOH group at the right side of another and in this way they are stick together to form a peptide bond. The chain is said to run from its amino (or N) terminus to its carboxyl (or C) terminus as shown in Fig. B.2(a).

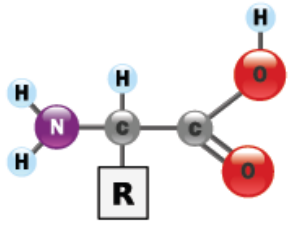

Figure B.1: The structure of an amino acid. (Taken from WikiDoc[305])

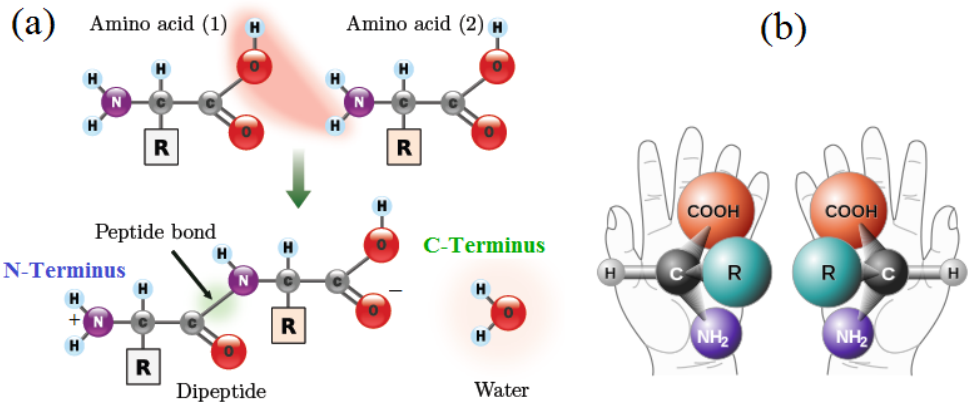

Figure B.2: (a) The condensation of two amino acids to form a peptide bond. (b) The mirror image of asymmetric isomer of Amino-acids. The L-form is shown on the left an the D-form on the right. ((a) Taken from WikiDoc[305] and (b) taken from Wikipedia[306])

All alpha-amino acids except glycine have achiral center at their α-carbons. Glycine has two hydrogens on its α-carbon, and therefore it is achiral. Due to its size and achiral property glycine is a very flexible amino acid and tends to break helical and other secondary structure of proteins. Besides glycine, all proteogenic amino acids can exist in either of two optical isomers, called L (levorotaty) or D (dextro-rotary) amino acids, which are mirror images of each other (Fig. B.2(b)). Biological systems have evolved to use the L-form almost exclusively. D-amino acids are very rare and found in some proteins produced by enzyme posttranslational modifications.

The side-chain denotes the physical and chemical properties of protein. According to charge and polarity of side-chain the structures shown in Fig. B.3 can be

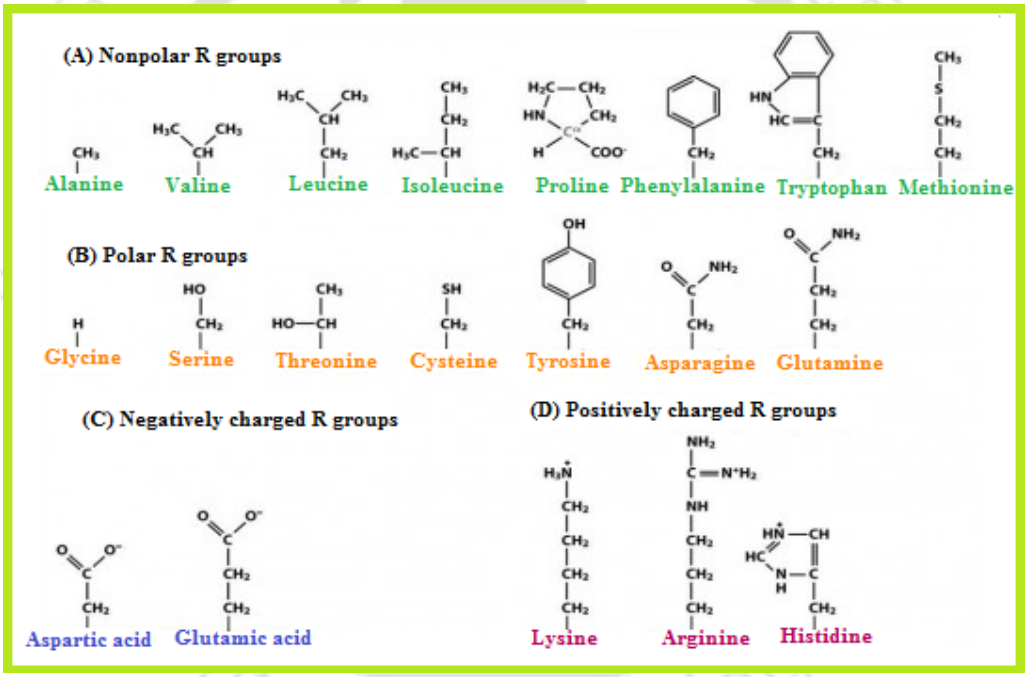

Figure B.3: Classification of 20 amino acids. (Taken from Biology Exams 4 U[307])

grouped into the following categories: (1) Nonpolar or hydrophobic (low propensity to be in contact with water) amino acids as Alanine, Valine, Leucine, Isoleucine, Proline, Phenyalanine and Glycine. (2) neutral (uncharged) but polar amino acids as Serine, Threonine, Glutamine, Asparagine, Tyrosine, Tryptophan and Methionine. The polar amino acid usually participate in hydrogen bonds as proton donor and acceptor. (3) Charged amino acid (energetically favorable contact with water) include acidic amino acids (which have a net negative charge at pH 7.0), and basic amino acids (which have a net positive charge at neutral pH). The table B.1 shows

the abbreviation and properties of amino acids. The side chains in charged amino acid often make salt bridge (a non-covalent interaction among ionized sites). The acidic amino acid are Aspartic acid and glutamic acids and Lysine, arginine and histidine are the example of acidic amino acid. Cysteine and methionine are the Sulphur containing amino acids. Tyrosine, Tryptophan, phynyalanine contains an aromatic group in their side chain. Most of the naturally occurring amino acids are

Table B.1: Proteinogenic amino acids, with corresponding one-letter symbols, the three-letter symbols and the properties of the side-chains.

Amino Acid	3-Letter	1-Letter	Mass (Da)	polarity	Charge (pH 7)
Alanine	Ala	A	89.98	nonpolar	neutral
Arginine	Arg	R	174.20	polar	positive
Asparagine	Asn	N	132.12	polar	neutral
Aspartic acid	Asp	D	133.10	polar	negative
Cysteine	Cys	C	121.15	non-polar	neutral
Glutamic acid	Glu	E	147.13	polar	negative
Glycine	Gly	G	75.07	nonpolar	neutral
Glutamine	Gln	Q	146.15	polar	neutral
Histidine	His	H	155.16	polar	positive
Isoleucine	Ile	I	131.17	nonpolar	neutral
Leucine	Leu	L	131.17	nonpolar	neutral
Lysine	Lys	K	146.19	polar	positive
Methionine	Met	M	149.21	nonpolar	neutral
Phenylalanine	Phe	F	165.19	nonpolar	neutral
Proline	Pro	P	115.13	nonpolar	neutral
Serine	Ser	S	105.09	polar	neutral
Threonine	Thr	T	119.12	polar	neutral
Tryptophan	Trp	W	204.23	nonpolar	neutral
Tyrosine	Tyr	Y	181.19	polar	neutral
Valine	Val	V	117.15	nonpolar	neutral

indispensable. Human body can produce only 11 of total 20 proteinogenic amino acids, are named as non-essential amino acid. These are the Tryptophan, Histidine, Arginine, Leucine, Isoleucine, Lysine, Valine, methionine, phenylalanine, Threonine. For the remaining nine external supplementation is needed on the basis of food, are called essential amino acids.

There are four identified level of protein structure: primary, secondary, tertiary and quaternary (shown in Fig. B.5). The level of complexity increases from primary to quaternary structure. Proteins are synthesized as a primary structure, and secondary, tertiary, and quaternary structures are arising from interactions between progressively more distant amino acids in the primary structure as a result of protein folding.

B.1.2 Peptide Chains and Primary Structure

The simplest level of protein structure, primary structure is the unique sequence of amino acids in each polymer chain which make protein. The sequence is provided by a series of steps called transcription (the use of a DNA strand to make a complimentary messenger RNA strand - mRNA) and translation (the mRNA sequence is used as a template to guide the synthesis of the polypeptide). In translation process the mRNA carries information and is translated in 3-letter sequences called codons (1 codon = 1 amino acid) with the help of ribosomes (rRNA) and tRNA molecules. RNA codons are read by our biological machinery and turned into polypetide through translation. Often, post-translational modifications, such as glycosylation or phosphorylation, occur which are necessary for the biological function of the protein. The code for translation is degenerate and is shown in Fig. B.4. A

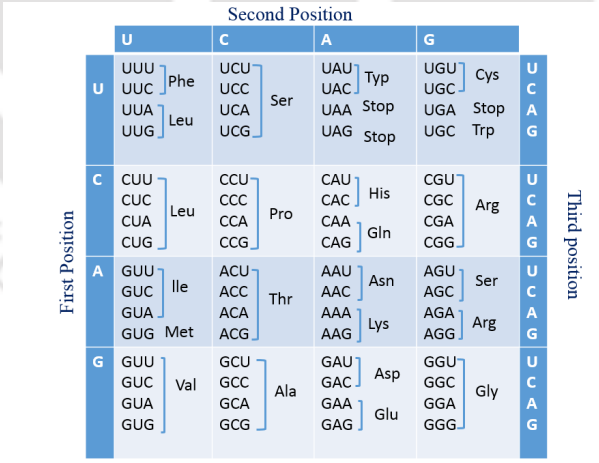

Figure B.4: The RNA code that specifies which amino acids to be included in a protein four bases (U, A, C and G) to make up one codon during a protein synthesis.

change in the gene's DNA sequence may lead to a change in the amino acid sequence of the protein. Even changing just one amino acid in a protein's sequence can effect the protein's overall structure and function. The primary structure is usually shown

using abbreviations for the amino acid residues. Using three letter abbreviations, a bit of a protein chain might be represented as- **NH_3^+-LYS-Ala-His-Gly-Lys-Lys-Val-Leu-Gly-Ala-COO^-.**

B.1.3 Secondary Structure

Secondary structure is the local spatial arrangement of a polypeptide's backbone atom without consideration of the conformation of its side chain. Protein's secondary structure includes the regular repetitive pattern arises from interactions between near-by amino acid (within about of 10 units of each other) as the polypeptide starts to fold into its functional 3-D form. The common structural elements in secondary structure are α-helix, β-sheets and turn which are discussed in below.

Helices:

The most common motif in the helical secondary structure of proteins, the α-helix (Fig. B.5(b)). The α-helix is right-handed, has 3.6 residues per turn formed by hydrogen bonding of C=O group of the n^{th} residues to the amino group of the $(n+4)^{th}$ and a pitch of 5.4 Å. So the terminal NH and CO groups of peptide chain are involved in hydrogen bond formation and the side chains are not involved in the H bonds that maintain the α-helix structure. As a result, the ends of α helices are polar and consequently they are most frequently found on the surface of proteins. The α helices of proteins have an average length of $\sim$12 residues, which corresponds to over three helical turns, and a length of $\sim$18 Å and consecutive residues make an angle of 100^0 around the helical axis.

Apart from α-helix there are two other helix structures: 3-turn helix (3_{10} helix) and 5-turn helix (π helix). In 3_{10} helix, carbonyl group of residue n interacts with the nitrogen of the amide group in residue n+3 (H bond pattern between n and n+3). There are 3 residues per turn, consecutive residues make an angle of 120^0 around the helical axis, a helical rise per residue of 2 Å, and a helical pitch of 5.8-6 Å. Hydrogen bonds within a π-helix display a repeating pattern in which the backbone C=O of residue n hydrogen bonds to the backbone NH of residue n+5. The 3_{10} helix is more tightly coiled and π helix is more loosely coiled than α-helix. Both of them are rarely found within stable configuration of protein because they are not energetically favorable due to their structural variation from general α-helix.

β-Sheets and Turns:

Unlike the α-helix, which is formed of one continuous region, β sheets form a series

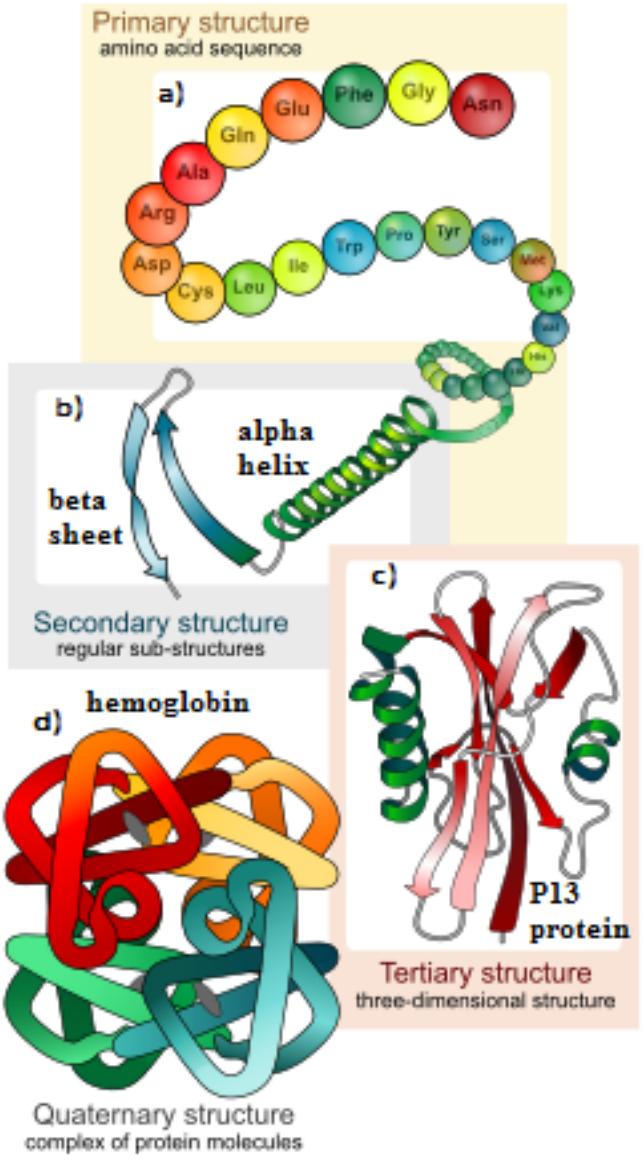

Figure B.5: Levels of protein organization from primary to quaternary structure. in cartoon representation of Tertiary structure, α and β subunits are shown in green and red respectively. (Taken from Wikipedia[308])

of adjacent strands separated by turn regions. These strands are usually around five residues long and can either run parallel or anti-parallel to one another (Fig. B.5(b)). The residues in the strands adopts an extended conformation which allows the adjacent CO and NH groups to hydrogen bond. The β-turns are usually defined by four amino acids turning back on themselves.

Random Coil:

That which cannot be classified as one of the standard three (helix, β-sheet and β-turn) classes is usually grouped into a category called "random coil". In a random coil, the only relationship between amino acids is that they are linked to the adjacent amino acid through the peptide bond.

B.1.4 Tertiary Structure

Tertiary structure (Fig. B.5(c)) refers to three-dimensional structure of a single protein molecule, when α-helices and β-sheets are folded into a compact globule. These secondary structure elements are joined together by regions called loops. Loop regions rarely contain hydrogen bonds between residues but often hydrogen bond with surrounding water molecules. The lack of internal bonding results in these regions being much less well ordered than the structural elements and consequently they exhibit greater flexibility.

B.1.5 Quaternary Structure

The quaternary structure (Fig. B.5(d)) is a larger assembly of several polypeptide chains or protein molecules, stabilized by the same non-covalent interactions and disulfide bonds like the tertiary structure. Although, many proteins do not have the quaternary structure and function as monomers.

The binding of either another protein or a small molecule in a location other than any active site can alter either the tertiary or quaternary structure of a protein (or complex of proteins). These changes underly the phenomenon of allosteric regulation in which they act to either increase or decrease activity. Small molecule binding to regions other than the active sites of a complex can also impact upon protein function.

B.1.6 Protein Folding

The process by which the protein arrives at its final native conformation is known as folding. The process of folding occurs on a timescale ranging from microseconds to milliseconds. The folding in this state is driven by the hydrophobic interactions, formation of salt bridges, tight packing of side chains and disulfide bonding.

B.1.7 Torsion Angles between Peptide Groups and the Ramachandran Diagram:

The conformation of the backbone of peptide chain can be described by the torsion angles (also called dihedral angles or Ramachandran angle). These two angles describe the rotation around the C_α-N bond (ϕ) and the C_α-C bond (ψ) (Fig. B.6(a)). These angles, ϕ and ψ, are both defined as 180^0 when the polypeptide chain is in its fully extended conformation and increase clockwise when viewed from C_α atom[309].

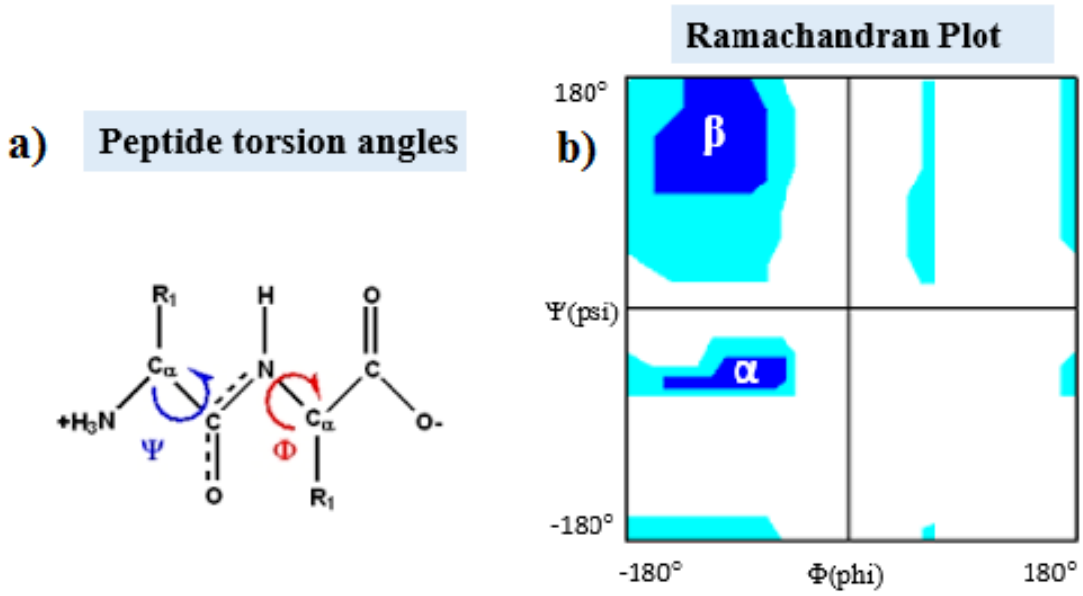

Figure B.6: (a) The ϕ, ψ dihedral angles in a single amino acid. (b) A Ramachandran plots the observed ϕ and ψ angles on the x and y axes respectively. (Taken from UCSF Computer Graphics Lab[310])

Due to the steric constrained imposed upon conformational freedom and therefore the torsional angles of a polypeptide backbone by the attached sidechains, the accessible ϕ/ψ conformational space are limited. Sterically forbidden conformations, such as the one shown in Fig. B.6(b) have the dihedral angles values that would bring atoms closer than the corresponding van der Walls distance. This type of plot known as a Ramachandran plot which is named after its inventor, G. N. Ramachandran. Most areas of Ramachandran plot represent forbidden conformations of a polypeptide chain.

B.2 Nucleic Acid

The nucleic acids are the molecular repositories for genetic information and are referred to as the "molecules of heredity". The structure of every protein, and ultimately of every cell constituent, is a product of information programmed into the nucleotide sequence of a cell's nucleic acids. The two main classes of nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA is carrier of genetic information and RNA is the genetic material of certain viruses, but it is also found in all living cells, where it plays an important role in certain processes such as the making of proteins. In fact, the central dogma of modern biology is

$$\textbf{DNA} \longrightarrow \textbf{mRNA} \longrightarrow \textbf{Protein.}$$

Information coded in DNA directs the synthesis of different RNA molecules. RNA molecules fall into several different categories:
Ribosomal RNA (rRNA): It is required for building Ribosomes, which are structures necessary for protein synthesis.
Transfer RNA (tRNA): It serves to transfer individual amino acid molecules from the general cytoplasm to their appropriate location in a growing polypeptide during protein synthesis.
Messenger RNA (mRNA): It carries the specific instructions for building a specific protein.
DNA directs protein synthesis through a multi-step process. First, DNA is copied to mRNA through the process of transcription. Then translation produces a polypeptide with an amino-acid sequence that is completely specified by the sequence of nucleotides in the RNA. A simple code, the same for all living things on this planet, governs the synthesis of protein from mRNA instructions.

Nucleic Acids are linear polymer, composed of four different types of nucleotides. It is in the sequence of the nucleotides in the polymers where the genetic information is located. Nucleotides are made up of three structural subunits: Nitrogenous bases, Pentose Sugar and phosphate. Figure B.7 shows the schematic diagram of nucleic acids and the structure of nucleotide. The base covalently conjugated to the first carbon of a pentose sugar (ribose or deoxyribose) but without the phosphate group, called Nucleoside. The chemical linkage between monomer units in nucleic acids is a phosphodiester bond.

1. **Pentose Sugar:** A pentose sugar is a monosaccharide with five carbon atoms, D-ribose in RNA and D-2-deoxyribose in DNA. A perusal of the structure of the two

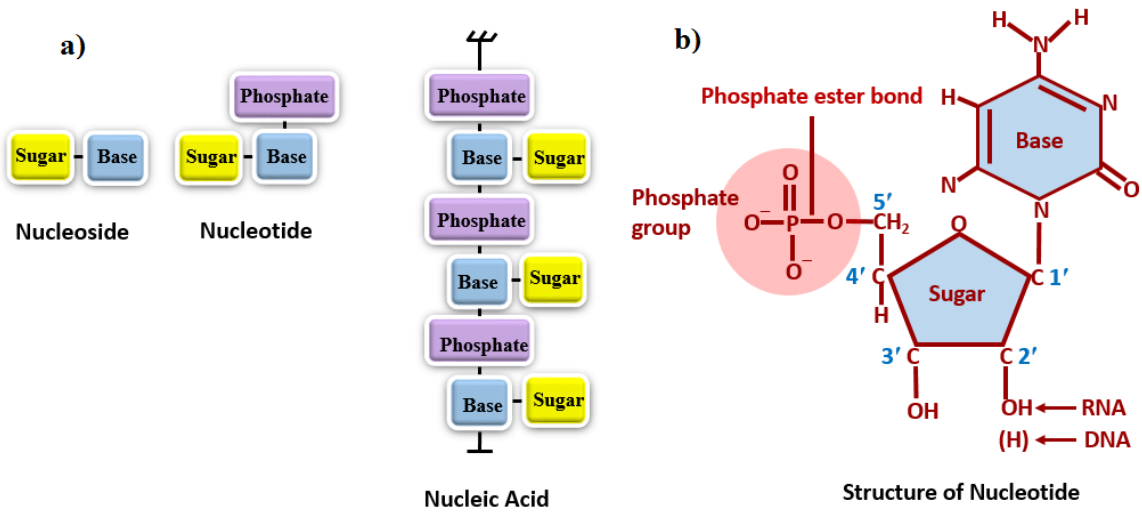

Figure B.7: (a) Schematic diagram of Nucleosides, Nucleotides, and Nucleic acids. (b) Structure of Nucleotide.

types of sugars reveals that D-ribose is the parent sugar while D-2-deoxyribose is a derivative in which OH group on C2′ has been replaced by an H atom. [Note that in chemical nomenclature of Fig. B.7(b), the carbon atoms of sugars are designated by primed numbers, that is, C-1′, C-2′, C-3′ etc., while the various atoms in the bases lack the prime (′) sign and are designated by the cardinal numbers, that is, 1, 2, 3 etc.] An important property of the pentoses is their capacity to form esters with phosphoric acid. In this reaction the OH groups of the pentose, especially those at C3′ and C5′, are involved forming a 3′, 5′- phosphodiester bond between adjacent pentose residues. This bond, in fact, is an integral part of the structure of nucleic acids.

2. **Nitrogeneous base:** The bases are the Heterocyclic-"molecule with at least one ring containing an atom other than carbon", either derivatives of pyrimidine or purine. The purines are adenine (A) and guanine (G), and the pyrimidines are cytosine (C) and thymine (T). A fifth pyrimidine base, called uracil (U), usually takes the place of thymine in RNA and differs from thymine by lacking a methyl (-CH_3) group on its ring. The schematic structures of Nucleotide bases are shown in Fig. B.8.

3. **Phosphate.**: The Phosphate group (-PO_4) is the derivative of Phosphoric acid (H_3PO_4) and the repeating part of the nucleic acid backbone. A phosphate group is attached to the sugar molecule in place of -OH group on the 5′ carbon.

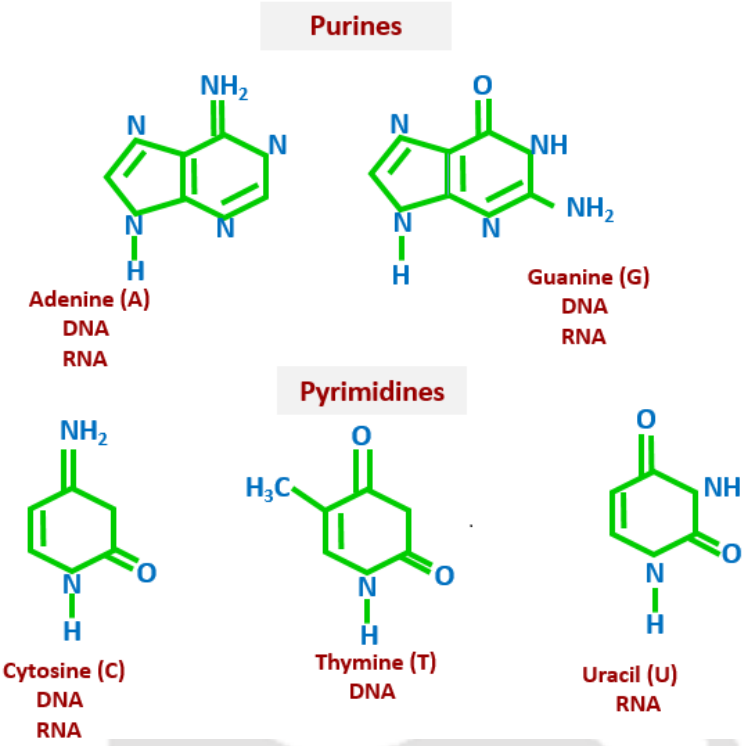

Figure B.8: Schematic structures of Nucleotide base.

B.2.1 DNA and RNA Structure

A molecule of DNA is formed by million of nucleotides joined together by a chain. In fact, DNA is typically consist of double strand of nucleotides while RNA is usually a single-strand helix consisting of shorter chains of nucleotides. In DNA structure, the sugar-phosphate chain are outside and the bases lie on the inside, where they are linked to complementary bases on the other strand through hydrogen bonds (shown in Fig. B.9(a)), known as Watson-Crick base pairing. ADE (A) always pairs with THY (T) through two hydrogen bonds, and GUA (G) always pairs with CYT (C) through three hydrogen bonds. The spans of A:T and G:C hydrogen-bonded pairs are nearly identical, allowing them to bridge the sugar-phosphate chains uniformly. Unlike DNA, RNA can fold upon itself with the folds stabilized by short areas of complementary base pairing within the molecule, forming a three-dimensional structure (shown in Fig. B.9(b)). In addition, DNA can fold into variety of structures which include G-quadruplex (G4 structures) DNA, three-stranded triplex DNA etc.

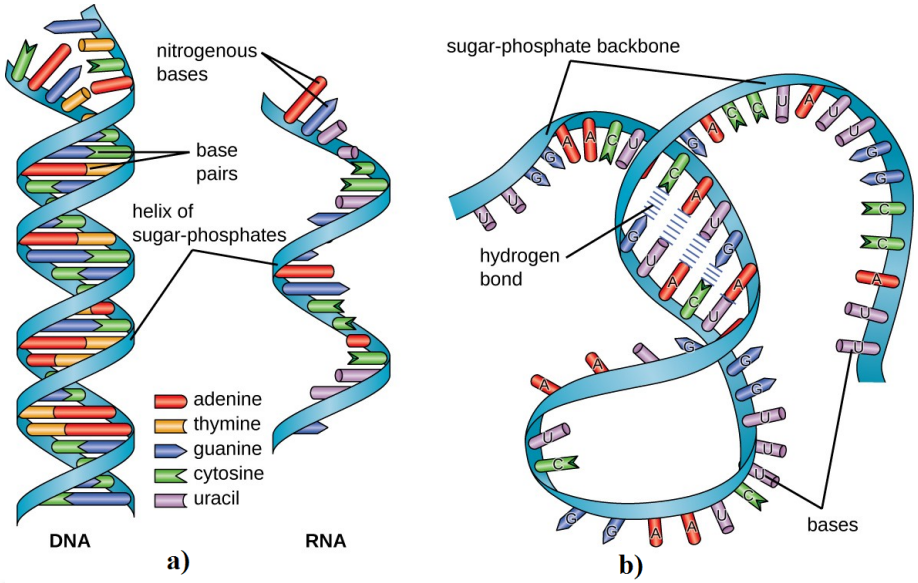

Figure B.9: (a) Structure of double-stranded DNA and single-stranded RNA. (b) Illustration of single-stranded RNA folding by hydrogen boding between complementary bases. (Taken from Lumen Microbiology[311])

B.2.2 Watson-Crick and Hoogsteen Base Pairing

In all canonical duplex structures including B-DNA, A-DNA, and Z-DNA, the so-called Watson-Crick base pairs are found. Figure B.10 shows the Watson-Crick base-pairing of guanine with cytosine and adenine with thymine. In A-T base pair three two hydrogen bonds are found between A-N6 and T-O4, and A-N1 and T-N3 positions, and in a G-C base pair three hydrogen bonds are found between G-O6 and C-N4, G-N1 and C-N3, and G-N2 and C-O2 positions. In a Hoogsteen base pair, the nucleotide base (A) is "flipped" in comparison to the Watson-Crick pairing as shown in Fig B.11. In Hoogsteen pairing the N7 of adenine is bonded to the N3 of thymine. Also, there are a difference in geometry between the two types of base pairing. In a normal Watson-Crick base pair, the two C1 atoms are equidistant at about 10.5 Å, whereas in Hoogsteen pair the distance is 8.65 Å. Majority of nucleotide bases in DNA are linked together with Watson-Crick pairing.

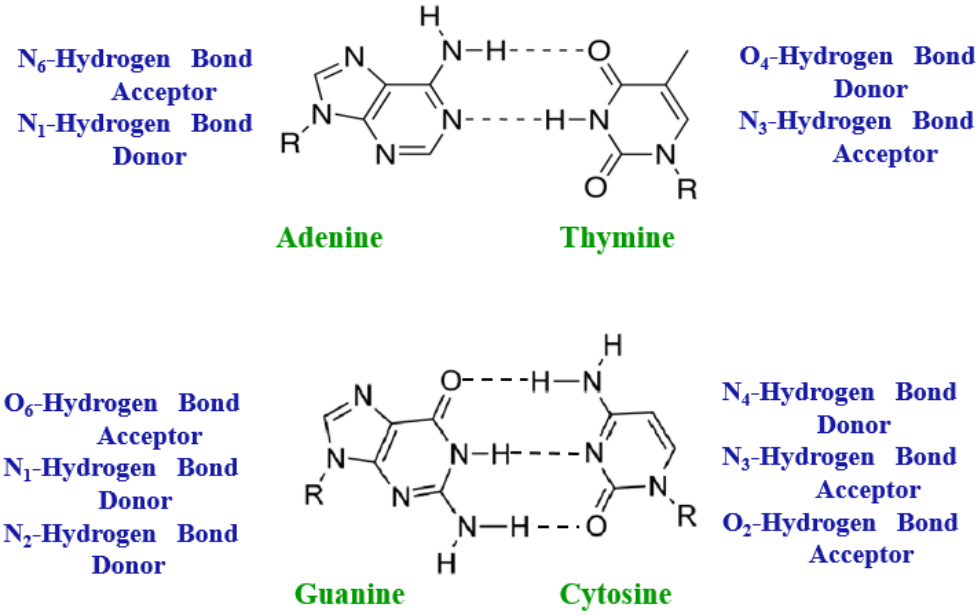

Figure B.10: Schematic illustration of Watson-Crick base pairing. Hydrogen bonds are shown as dashed lines. (Taken from atdbio[312])

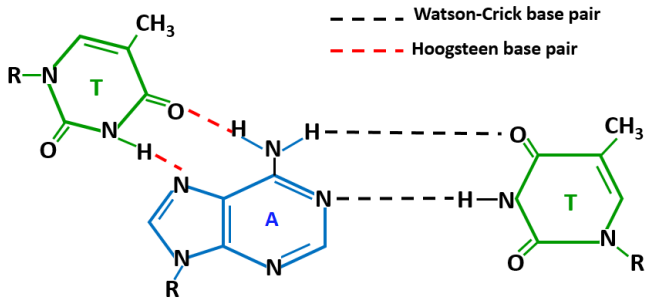

Figure B.11: Schematic illustration of Hoogsteen base pairing in comparison to Watson-Crick base pairing. Hydrogen bonds are shown as dashed lines.

Bibliography

[1] A. Mittermaier and L. E. Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, **312 (5771)**, 224-228 (2006).

[2] R. Brüschweiler. New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins. *Curr. Opin. Struct. Biol.*, **13 (2)**, 175-183 (2003).

[3] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, **438 (7064)**, 117-121 (2005).

[4] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, **450**, 964-972 (2007).

[5] K. D. G. Pfleger and K. A. Eidne. Illuminating insights into protein-protein interactions using bioluminescence resonance energy transfer (BRET). *Nat. Methods*, **3 (3)**, 165-174 (2006).

[6] B. N. Giepmans, S. R. Adams, M. H. Ellisman, and R. Y. Tsien. The fluorescent toolbox for assessing protein location and function. *Science*, **312 (5771)**, 217-224 (2006).

[7] B. Schuler and W. A. Eaton. Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.*, **18 (1)**, 16-26 (2008).

[8] Q. Ni. and J. Zhang. Dynamics visualization of cellular signaling. *Adv. Biochem. Eng. Biotechnol.*, **119**, 79-97 (2010).

[9] A. Borgia, P. M. Williams, and J. Clarke. Single-molecule studies of protein folding. *Annu. Rev. Biochem.*, **77**, 101-125 (2008).

[10] D. J. Brockwell. Force denaturation of proteins - an unfolding story. *Curr. Nanosci.*, **3 (1)**, 3-15 (2007).

[11] Y. Cao and H. Li. Engineered elastomeric proteins with dual elasticity can be controlled by a molecular regulator. *Nat. Nanotechnol.*, **3**, 512-516 (2008).

[12] G. Žoldák and M. Rief. Force as a single molecule probe of multidimensional protein energy landscapes. *Curr. Opin. Struct. Biol.*, **23 (1)**, 48-57 (2013).

[13] E. M. Puchner and H. E. Gaub. Force and function: probing proteins with AFM-based force spectroscopy. *Curr. Opin. Struct. Biol.*, **19 (5)**, 605-614 (2009).

[14] R. B. Best, D. J. Brockwell, J. L. Toca-Herrera, A. W. Blake, D. A. Smith, S. E. Radford, and J. Clarke. Force mode atomic force microscopy as a tool for protein folding studies. *Anal. Chim. Acta.*, **479 (1)**, 87-105 (2003).

[15] T. E. Fisher, A. F. Oberhauser, M. Carrion-Vazquez, P. E. Marszalek, and J. M. Fernandez. The study of protein mechanics with the atomic force microscope. *Trends Biochem. Sci.*, **24 (10)**, 379-384 (1999).

[16] F. Oesterhelt, D. Oesterhelt, M. Pfeiffer, A. Engel, H. E. Gaub, and D. J. Müller. Unfolding pathways of individual bacteriorhodopsins. *Science*, **288 (5463)**, 143-146 (2000).

[17] M. Kessler, K. E. Gottschalk, H. Janovjak, D. J. Mueller, and H. E. Gaub. Bacteriorhodopsin folds into the membrane against an external force. *J. Mol. Biol.*, **357 (2)**, 644-654 (2006).

[18] M. S. Kellermayer, S. B. Smith, H. L. Granzier, and C. Bustamante. Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science*, **277 (5329)**, 1117 (1997).

[19] J. P. Junker, F. Ziegler, and M. Rief. Ligand-dependent equilibrium fluctuations of single calmodulin molecules. *Science*, **323 (5914)**, 633-637 (2009).

[20] A. Otto. Excitation of nonradiative surface plasma waves in silver by the method of frustrated total reflection. *Z. phys. A Hadrons Nucl.*, **216 (4)**, 398-410 (1968).

[21] V. Owen. Real-time optical immunosensors - a commercial reality. *Biosens. Bioelectron.*, **12 (1)**, 1-2 (1997).

[22] L, Novotny and B. Hecht. Principles of Nano-optics, 2nd edn. *Contemp. Phys.*, **54 (2)**, 123-124 (2013).

[23] J. Homola. Present and future of surface plasmon resonance biosensors. *Anal. Bioanal. Chem.*, **377 (3)**, 528-539 (2003).

[24] B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. I. General Method. *J. Chem. Phys.*, **31 (2)**, 459-466 (1959).

[25] A. Rahman and F. H. Stillinger. Molecular Dynamics study of liquid water. *J. Chem. Phys.*, **55 (7)**, 3336-3359 (1971).

[26] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, **267 (5612)**, 585-590 (1977).

[27] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions:dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme. *J. Mol. Biol.* **103 (2)**, 227-249 (1976).

[28] R. Car and M. Parrrinello. Unified approach for molecular dynamics and Density-Functional theory. *Phys. Rev. Lett.*, **55 (22)**, 2471-2474 (1985).

[29] M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. M. Noel, M. Gruebele, and J. W. Kelly. Structure-function-folding relationship in a WW domain. *Proc. Natl. Acad. Sci. U. S. A.*, **103 (28)**, 10648-10653 (2006).

[30] D. Wales. Energy landscapes: applications to clusters, biomolecules and glasses (Cambridge Molecular Science). *Cambridge: Cambridge University Press*, doi:10.1017/CBO9780511721724 (2004).

[31] D. Shukla, Y. Meng, B. Roux, and V. S. Pande. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat. Commun.*, **5**, 3397 (2014).

[32] K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman, V. S. Pande. Cloud-Based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.*, **6**, 15-21 (2014).

[33] A. Ostermann, R. Waschipky, F. G. Parak, and G. U. Nienhaus. Ligand binding and conformational motions in myoglobin. *Nature*, **404 (6774)**, 205-208 (2000).

[34] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, **38 (8)**, 114-117 (1965).

[35] G. E. Moore. Lithography and the future of moore's law. *Proc. SPIE 2438, Advances in Resist Technology and Processing XII*, **2437**, 1-8 (1995).

[36] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande. Molecular simulation of ab initio protein folding for a millisecond folder ntl9(1-39). *J. Am. Chem. Soc.*, **132 (5)**, 1526-1528 (2010).

[37] D. A. Case, T. A. Darden, T. E. Cheatham I, et al. AMBER 12. *San Francisco, CA: University of California*, (2012).

[38] B. R. Brooks , C. L. Brooks 3rd, A. D. Mackerell Jr, et al. CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30 (10)**, 1545-1614 (2009).

[39] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4 (3)**, 435-447 (2008).

[40] M. T. Nelson, W. Humphrey, A. Gursoy, et al. NAMD: a parallel, object oriented molecular dynamics program. *Int. J. Supercomput. Appl. High Perform Comput.*, **10 (4)**, 251-268 (1996).

[41] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, **23**, 187-199 (1977).

[42] C. Dellago, P. Bolhuis, and P. L. Geissler. Transition path sampling. *Adv. Chem. Phys.*, **123 (1)** (2002).

[43] P. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, **53**, 291-318 (2002).

[44] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.*, **99 (20)**, 12562-12566 (2002).

[45] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello. Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.*, **128 (41)**, 13435-13541 (2006).

[46] G. A. Tribello, J. Cuny, H. Eshet, and M. Parrinello. Exploring the free energy surfaces of clusters using reconnaissance metadynamics. *J. Chem. Phys.*, **135 (11)**, 114109 (2011).

[47] J. F. Dama, M. Parrinello, and G. A. Voth. Well-Tempered Metadynamics converges asymptotically. *Phys. Rev. Lett.*, **112 (24)**, 240602 (2014).

[48] S. Awasthi and N. N. Nair. Exploring high dimensional free energy landscapes: temperature accelerated sliced sampling. *J. Chem. Phys.*, **46 (9)**, 094108 (2017).

[49] E. Darve, D. Rodriguez-Gómez, and A. Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.*, **128 (14)**, 144120 (2008)

[50] H. Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E*, **52 (3)**, 2893-2906 (1995).

[51] Y. Sugita Y and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, **314 (1-2)**, 141-151 (1999).

[52] A. F. Voter. Introduction to the Kinetic Monte Carlo Method, in. Radiation Effects in Solids, edited by K. E. Sickafus and E. A. Kotomin. *Springer, NATO Publishing Unit, Dordrecht, The Netherlands*, 1-23 (2007).

[53] X. Wu and S. Wang. Enhancing Systematic motion in molecular dynamics simulations. *J. Chem. Phys.*, **110**, 9401-9410 (1999).

[54] B. J. Berne and J. E. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struc. Biol.*, **7 (2)**, 181-189 (1997).

[55] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, **134 (12)**, 124116 (2011).

[56] F. Noé and C. Clementi. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.*, **11 (10)**, 5002-5011 (2015).

[57] H. Wu and F. Noé. Gaussian markov transition models of molecular kinetics. *J. Chem. Phys.*, **142**, 084104 (2015).

[58] G. Pérez-Hernández, F. Paul, T. Giorgino, G. de Fabritiis, and Frank Noé. Identification of slow molecular order parameters for markov model construction. *J. Chem. Phys.*, **139 (1)**, 015102 (2013).

[59] C. R. Schwantes and V. S. Pande. Improvements in Markov state model construction reveal many non-native interactions in the folding of ntl9. *J. Chem. Theory Comput.*, **9 (4)**, 2000-2009 (2013).

[60] C. R. Schwantes and V. S. Pande. Modeling molecular kinetics with tica and the kernel trick. *J. Chem. Theory Comput.*, **11 (2)**, 600-608 (2015).

[61] C. R. Schwantes, R. T. McGibbon, and V. S. Pandey. Perspective: Markov models for long-timescale biomolecular dynamics. *J. Chem. Phys.*, **141 (9)**, 090901 (2014).

[62] V. S. Pande, K. Beauchamp, and G. R. Bowman. Everything you wanted to know about Markov state models but were afraid to ask. *Methods*, **52 (1)**, 99-105 (2010).

[63] M. Senne, B. Trendelkamp-Schroer, A. S. J. S. Mey, C. Schütte, and F. Noé. EMMA: a software package for Markov model building and analysis. *J. Chem. Theory Comput.*, **8 (7)**, 2223-2238 (2012).

[64] N.-V. Buchete and G. Hummer. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B*, **112 (19)**, 6057-6069 (2008).

[65] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden. Markov state models based on milestoning. *J. Chem. Phys.*, **134 (20)**, 204105 (2011).

[66] H. Eyring. The activated complex in chemical reactions. *J. Chem. Phys.*, **3 (2)**, 107 (1935).

[67] K. J. Laidler and M. C. Klng. The development of Transition-State theory. *J. Phys. Chem.*, **87 (15)**, 2657-2664 (1983).

[68] K. D. Ball and R. S. Berry. Realistic master equation modeling of relaxation on complete potential energy surfaces: kinetic results. *J. Chem. Phys.*, **109 (19)** 8557-8572 (1998).

[69] Y. Levy, J. Jortner, and O. M. Becker. Dynamics of hierarchical folding on energy landscapes of hexapeptides. *J. Chem. Phys.*, **115 (22)**, 10533-10547 (2001).

[70] P. N. Mortenson and D. J. Wales. Energy landscapes, global optimization and dynamics of the polyalanine Ac(ala)$_8$NHMe. *J. Chem. Phys.*, **114 (14)**, 6443-6454 (2001).

[71] P. N. Mortenson, D. A. Evans, and D. J. Wales. Energy landscapes of model polyalanines. *J. Chem. Phys.*, **117 (3)**, 1363-1376 (2002).

[72] D. A. Evans and D. J. Wales. Folding of the GB1 hairpin peptide from discrete path sampling. *J. Chem. Phys.* **121 (2)**, 1080-1090 (2004).

[73] J. D. Chodera and F. Noé. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, **25** 135-144 (2014).

[74] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.*, **23 (1)**, 58-65 (2013).

[75] J.-H. Prinz, B. Keller, and F. Noé. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.*, **13 (38)**, 16912-16927 (2011).

[76] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.*, **131 (12)**, 124101 (2009).

[77] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, **126 (15)** 155101 (2007).

[78] X. Huang, G. R. Bowman, S. Bacallado, and V. S. Pande. Rapid equilibrium sampling initiated from nonequilibrium data. *Proc. Natl. Acad. Sci. U. S. A.*, **106 (47)**, 19765-19769 (2009).

[79] G. R. Bowman, D. L. Ensign, and V. S. Pande. Enhanced modeling via network theory: adaptive sampling of Markov state models. *J. Chem. Theory Comput.*, **6 (3)**, 787-794 (2010).

[80] N. S. Hinrichs and V. S. Pande. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J. Chem. Phys.*, **126 (24)**, 244101 (2007).

[81] S. Roblitz. Statistical error estimation and grid-free hierarchical refinement in conformation dynamics. (thesis, Freie Universitat Berlin) (2008).

[82] X. Huang, Y. Yao, J. Sun, et al. Constructing multi-resolution Markov state models (MSMs) to elucidate RNA hairpin folding mechanisms. *Pac. Symp. Biocomput.*, **15**, 228-239 (2010).

[83] R. T. McGibbon, B. Ramsundar, M. M. Sultan, G. Kiss, and V. S. Pande. Understanding protein dynamics with L1-regularized reversible hidden Markov models. *Proceedings of the 31st ICML'14, PMLR*, **32 (2)**, 1197-1205 (2014).

[84] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé. PyEMMA 2: a software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.*, **11 (11)**, 5525-5542 (2015).

[85] K. A. Beauchamp, G. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande. MSMBuilder2: modeling Conformational dynamics at the picosecond to millisecond scale. *J. Chem. Theory Comput.*, **7 (10)**, 3412-3419.10.1021/ct200463m (2011).

[86] B. Cronkite-Ratcliff and V. S. Pande. MSMExplorer: visualizing Markov state models for biomolecule folding simulations. *Bioinformatics.* **29 (7)**, 950-952 (2013) [PubMed: 23365411].

[87] B. E. Husic and V. S. Pande. Ward clustering improves cross-validated Markov state model of protein folding. *J. Chem. Theory. Comput.*, **13 (3)**, 963-967 (2017).

[88] F. Nüske, H. Wu, J.-H. Prinz, C. Wehmeyer, C. Clementi, and F. Noé. Markov state models from short non-equilibrium simulations - analysis and correction of estimation bias. *J. Chem. Phys.*, **146 (9)**, 094104 (2017).

[89] H. Wang and C. Schütte. Building Markov state models for periodically driven non-equilibrium systems. *J. Chem. Theory Comput.*, **11 (4)**, 1819-1831 (2015).

[90] P. Koltai, G. Ciccotti, and C. Schütte. On metastability and Markov state models for non-stationary molecular dynamics. *J. Chem. Phys.*, **145 (17)**, 174103 (2016).

[91] F. Pellegrini, F. P. Landes, A. Laio, S. Prestipino, and E. Tosatti. Markov state modeling of sliding friction. *Phys. Rev. E.*, **94 (5)**, 053001 (2016).

[92] F. Knoch and T. Speck. Nonequilibrium Markov state modeling of the globule-stretch transition. *Phys. Rev. E.*, **95**, 012503 (2017).

[93] J. K. Weber and V. S. Pandey. Potential-based dynamical reweighting for Markov states models of protein dynamics. *J. Chem. Theory Comput.*, **11 (6)**, 2412-2420 (2015).

[94] K. C. Neuman and A. Nagy. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods*, **5**, 491-505 (2008).

[95] A. Ashkin, J. M. Dziedzic, and T. Yamane. Optical trapping and manipulation of single cells using infrared laser beams. *Nature*, **330**, 769-771 (1987).

[96] S. B. Smith, L. Finzi, and C. Bustamante. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science*, **258 (5085)**, 1122-1126 (1992).

[97] L. Chen, A. Offenhäusser, and H.-J. Krause. Magnetic tweezers with high permeability electromagnets for fast actuation of magnetic beads. *Rev. Sci. Instrum.*, **86 (4)**, 044701 (2015).

[98] K. Mitsui, M. Hara and A. Ikai. Mechanical unfolding of a(2)-macroglobulin molecules with atomic force microscope. *FEBS Lett.*, **385 (1-2)**, 29-33 (1996).

[99] J. Zlatanova, S. M. Lindsay, and S. H. Leuba. Single molecule force spectroscopy in biology using the atomic force microscope. *Prog. Biophys. Mol. Biol.*, **74 (1-2)**, 37-61 (2000).

[100] H. Clausen-Schaumann, M. Seitz, R. Krautbauer, and H. E. Gaub. Force spectroscopy with single bio-molecules. *Curr. Opin. Cell Biol.*, **4 (5)**, 524-530 (2000).

[101] T. Hoffmann and L. Dougan. Single molecule force spectroscopy using polyproteins. *Chem. Soc. Rev.*, **41 (14)**, 4781-4796 (2012).

[102] V. Barsegov, D. K. Klimov, and D. Thirumalai. Mapping the energy landscape of biomolecules using single molecule force correlation spectroscopy: theory and applications. *Biophys. J.*, **90 (11)**, 3827-3841 (2006).

[103] M. Wolny, M. Batchelor, P. J. Knight, E. Paci, L. Dougan, and M. Peckham. Stable single α-helices are constant force springs in proteins. *J. Biol. Chem.*, **289 (40)**, 27825-27835 (2014).

[104] Y. Chen, S. E. Radford, and D. J. Brockwell. Force-induced remodelling of proteins and their complexes. *Curr. Opin. Struct. Biol.*, **30**, 89-99 (2015).

[105] P. Zheng and H. Li. Direct measurements of the mechanical stability of zinc-thiolate bonds in rubredoxin by single-molecule atomic force microscopy. *Biophys. J.*, **101 (6)**, 1467-1473 (2011).

[106] P. E. Marszalek and Y. F. Dufrêne. Stretching single polysaccharides and proteins using atomic force microscopy. *Chem. Soc. Rev.*, **41 (9)**, 3523-3534 (2012).

[107] T. Bornschlögl and M. Rief. Single-molecule dynamics of mechanical coiled-coil unzipping. *Langmuir*, **24 (4)**, 1338-42 (2008).

[108] Z. N. Scholl, Q. Li, and P. E. Marszalek. Single molecule mechanical manipulation for studying biological properties of proteins, DNA, and sugars. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.*, **6 (3)**, 211-229 (2014).

[109] F. Berkemeier, M. Bertz, S. Xiao, N. Pinotsis, M. Wilmanns, F. Gräter and M. Rief. Fast-folding alpha-helices as reversible strain absorbers in the muscle protein myomesin. *Proc. Natl. Acad. Sci. U. S. A.*, **108 (34)**, 14139-14144 (2011).

[110] A. E. M. Beedle, A. Lezamiz, G. Stirnemann, and S. Garcia-Manyes. The mechanochemistry of copper reports on the directionality of unfolding in model cupredoxin proteins. *Nat. Commun.*, **6**, 7894 (2015).

[111] J. Perales-Calvo, A. Lezamiz, and S. Garcia-Manyes. The mechanochemistry of a structural zinc finger. *J. Phys. Chem. Lett.*, **6 (17)**, 3335-40 (2015).

[112] L. Verlet. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, **159 (1)**, 98-103 (1967).

[113] D. Frenkel and B. Smit. Understanding molecular simulation:from algorithms to applications. *Academic Press, San Diego, CA.*, (1996).

[114] J. E. Jones and D. Sc. On the determination of molecular fields.-II. from the equation of state of a gas. *Proc. R. Soc. Lond. A.*, **106 (738)**, 463-477 (1924).

[115] J. D. van der Waals. The thermodynamic theory of capillarity flow under the hypothesis of a continuous variation in density. *J. Stat. Phys.*, **20 (2)**, 200-244 (1979).

[116] C. A. Coulomb. Collection de mémoires relatifs á la physique. *Gauthier-Villars*, **A**, 569-638 (1884).

[117] F. London. Über einige Eigenschaften und Anwendungen der Molekularkräfte. *Z. Phys. Chem.*, **B11**, 222-251 (1930).

[118] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81 (8)**, 3684-3690 (1984).

[119] G. S. Grest and K. Kremer. Molecular-dynamics simulation for polemers in the presence of a heat bath. *Phys. Rev. A*, **33 (5)**, 3628-3631 (1986).

[120] H.C. Andersen. Molecular dynamics at constant pressure and/or temperature. *J. Chem. Phys.*, **72 (4)**, 2384-2393 (1980).

[121] S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, **81 (1)**, 511-519 (1984).

[122] S. A. Adelman. Generalized Langevin theory for many-body problems in chemical dynamics: general formulation and the equivalent harmonic chain representation. *J. Chem. Phys.*, **71 (11)**, 4471-4486 (1979).

[123] S. Nosé and M.L. Klein. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.*, **50 (5)**, 1055-1076 (1983). 50, 1055 (1983)

[124] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks. Constant pressure molecular dynamics simulation: the Langevin piston method. *J. Chem. Phys.*, **103 (11)**, 4613-4621 (1995).

[125] P. P. Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.*, **64 (3)**, 253-287 (1921).

[126] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An $N - \log(N)$ method for Ewald sum in large systems. *J. Chem. Phys.*, **98 (12)**, 10089-10092 (1993).

[127] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23 (3)**, 327-341 (1977).

[128] S. Miyamoto and P. A. Kollman. Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, **13 (8)**, 952-962 (1992).

[129] H. C. Andersen. Rattle: a "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.*, **52 (1)**, 24-34 (1983).

[130] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski. Coarse-Grained protein models and their applications. *Chem. Rev.*, **116 (14)**, 7898-7936 (2016).

[131] S. Gnanakaran, H. Nymeyer, and J. Portman. Peptide folding simulations. *Curr. Opin. Struct. Biol.* **13 (2)**, 168-174 (2003).

[132] G. G. Dodson, D. P. Lane, and C. S. Verma. Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO Rep.* **9 (2)**, 144-150 (2008).

[133] K. Klenin, B. Strodel, D. J. Wales, and W. Wenzel. Modelling proteins: conformational sampling and reconstruction of folding kinetics. *Biochim. Biophys. Acta* **1814 (8)**, 977-1000 (2011).

[134] M. T. Woodside and S. M. Block. Reconstructing folding energy landscapes by single-molecule force spectroscopy. *Annu. Rev. Biophys.* **43**, 19-39 (2014).

[135] A. Chatterjee and D. G. Vlachos. An overview of spatial microscopic and accelerated kinetic Monte Carlo methods. *J. Comput.-Aided Mater. Des.* **14 (2)**, 253-308 (2007).

[136] B. P. Uberuaga and A. F. Voter. Accelerated molecular dynamics methods. *in Radiation Effects in Solids, edited by K. E. Sickafus, E. A. Kotomin, and B. P. Uberuaga, Springer, NATO Publishing Unit*, (2006).

[137] D. J. Wales. Energy landscapes: calculating pathways and rates. *Int. Rev. Phys. Chem.*, **25 (1-2)**, 237-282 (2006).

[138] R. M. Ziff, E. Gulari, and Y. Barshad. Kinetic phase transitions in an irreversible surface-reaction model. *Phys. Rev. Lett.*, **56 (24)**, 2553 (1986).

[139] G. H. Gilmer, H. C. Huang, T. D. de la Rubia, J. Dalla Torre, and F. Baumann. Lattice Monte Carlo models of thin film deposition. *Thin Solid Films*, **365 (2)**, 189-200 (2000).

[140] P. Haldar and A. Chatterjee. Connectivity-list based characterization of 3D nanoporous structures formed via selective dissolution. *Acta Mater.*, **127**, 379-388 (2017).

[141] A. Chatterjee, M. A. Katsoulakis, and D. G. Vlachos. Spatially adaptive grand canonical ensemble Monte Carlo simulations. *Phys. Rev. E*, **71 (2)**, 026702 (2005).

[142] A. Chatterjee and A. F. Voter. Accurate acceleration of kinetic Monte Carlo simulations through the modification of rate constants. *J. Chem. Phys.*, **132 (19)**, 194101 (2010).

[143] A. Chatterjee and D. G. Vlachos. Multiscale spatial Monte Carlo simulations: Multigriding, computational singular perturbation, and hierarchical stochastic closures. *J. Chem. Phys.*, **124 (6)**, 64110 (2006).

[144] E. Weinan, B. Engquist, and Z. Y. Huang. Heterogeneous multiscale method: a general methodology for multiscale modeling. *Phys. Rev. B*, **67 (9)**, 092101 (2003).

[145] M. A. Snyder, A. Chatterjee, and D. G. Vlachos. Net-event kinetic Monte Carlo for overcoming stiffness in spatially homogeneous and distributed systems. *Comput. Chem. Eng.*, **29 (4)**, 701-712 (2004).

[146] G. R. Bowman, X. Huang, and V. S. Pande. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods*, **49 (2)**, 197-201 (2009).

[147] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, **126 (15)**, 155101 (2007).

[148] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.*, **134 (17)**, 174105 (2011).

[149] L. Xu and G. Henkelman. Adaptive kinetic Monte Carlo for first-principles accelerated dynamics. *J. Chem. Phys.*, **129 (11)**, 114104, (2008).

[150] D. Konwar, V. J. Bhute, and A. Chatterjee. An off-lattice, self-learning kinetic Monte Carlo method using local environments. *J. Chem. Phys.*, **135 (17)**, 174103 (2011).

[151] A. F. Voter. A method for accelerating the molecular dynamics simulation of infrequent events. *J. Chem. Phys.*, **106 (11)**, 4665-4677 (1997).

[152] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.*, **99 (20)**, 12562-12566 (2002).

[153] B. Ensing, M. De Vivo, Z. Liu, P. Moore, and M. L. Klein. Metadynamics as a Tool for exploring free energy landscapes of chemical reactions. *Acc. Chem. Res.*, **39 (2)**, 73-81 (2006).

[154] A. F. Voter, F. Montalenti, and T. C. Germann. Extending the time scale in atomistic simulation of materials. *Annu. Rev. Mater. Res.*, **32**,, 321-346 (2002).

[155] S. Divi and A. Chatterjee. Accelerating rare events while overcoming the low-barrier problem using a temperature program. *J. Chem. Phys.*, **140 (18)**, 184115 (2014).

[156] V. Imandi and A. Chatterjee. Estimating Arrhenius parameters using temperature programmed molecular dynamics. *J. Chem. Phys.*, **145 (3)**, 034104 (2016).

[157] P. Haldar and A. Chatterjee. Seeking kinetic pathways relevant to the structural evolution of metal nanoparticles. *Modell. Simul. Mater. Sci. Eng.*, **23 (2)**, 025002 (2015).

[158] V. J. Bhute and A. Chatterjee. Building a kinetic Monte Carlo model with a chosen accuracy. *J. Chem. Phys.*, **138 (24)**, 24411 (2013).

[159] V. J. Bhute and A. Chatterjee. Accuracy of a Markov state model generated by searching for basin escape pathways. *J. Chem. Phys.*, **138 (8)**, 084103 (2013).

[160] S. T. Chill and G. Henkelman. Molecular dynamics saddle search adaptive kinetic Monte Carlo. *J. Chem. Phys.*, **140 (21)**, 214110 (2014).

[161] A. Chatterjee and S. Bhattacharya. Uncertainty in a Markov state model with missing states and rates: Application to a room temperature kinetic model obtained using high temperature molecular dynamics. *J. Chem. Phys.*, **143 (11)**, 114109 (2015).

[162] A. Chatterjee and S. Bhattacharya. Probing the energy landscape of alanine dipeptide and decalanine using temperature as a tunable parameter in molecular dynamics. *J. Phys.: Conf. Ser.*, **759 (1)**, 012024 (2016).

[163] C. A. F. de Oliveira, D. Hamelberg, and J. A. McCammon. Estimating kinetic rates from accelerated molecular dynamics simulations: Alanine dipeptide in explicit solvent as a case study. *J. Chem. Phys.*, **127 (17)**, 175105 (2007).

[164] N. Singhal, C. D. Snow, and V. S. Pandey. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan beta hairpin. *J. Chem. Phys.*, **121 (1)**, 415-425 (2004).

[165] S. Sriraman, I. G. Kevrekidis, and G. Hummer. Coarse master equation from bayesian analysis of replica molecular dynamics simulations. *J. Phys. Chem. B*, **109 (14)**, 6479-6484 (2005).

[166] Y. Mu, P. H. Nguyen, and G. Stock. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins*, **58 (1)**, 45-52 (2005).

[167] E. H. Kellogg, O. F. Lange, and D. Baker. Evaluation and optimization of discrete state models of protein folding. *J. Phys. Chem. B*, **116 (37)**, 11405-11413 (2012).

[168] T. Zhou and A. Caflisch. Distribution of reciprocal of interatomic distances: A fast structural metric. *J. Chem. Theory Comput.*, **8 (8)**, 2930-2937, (2012).

[169] R. T. McGibbon and V. S. Pande. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. *J. Chem. Theory Comput.*, **9 (7)**, 2900-2906 (2013).

[170] G. Pérez-Hernández, F. Paul, T. Giorgino, G. de Fabritiis, and F. Noé. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.*, **139 (1)**, 015102 (2013).

[171] C. R. Schwantes and V. S. Pande. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.*, **9 (4)**, 2000-2009 (2013).

[172] K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande. MSMBuilder2: modeling conformational dynamics on the picosecond to millisecond scale. *J. Chem. Theory Comput.*, **7 (10)**, 3412-3419 (2011).

[173] D. L. Ensign and V. S. Pande. Bayesian single-exponential kinetics in single-molecule experiments and simulations. *J. Phys. Chem. B*, **113 (36)**, 12410-12423 (2009).

[174] M. Sarich, F. Noé, and C. Schütte. On the approximation quality of Markov state models. *Multiscale Model. Simul.*, **8 (4)**, 1154-1177 (2010).

[175] N. Djurdjevac, M. Sarich, and C. Schütte. Estimating the eigenvalue error of Markov state models. *Multiscale Model. Simul.*, **10 (1)**, 61-81 (2012).

[176] P. Metzner, F. Noé, and C. Schätte. Estimating the sampling error: distribution of transition matrices and functions of transition matrices for given trajectory data. *Phys. Rev. E*, **80 (2)**, 021106 (2009).

[177] G. R. Bowman. Improved coarse-graining of Markov state models via explicit consideration of statistical uncertainty. *J. Chem. Phys.*, **137 (13)**, 134111 (2012).

[178] T. Zhou and A. Caflisch. Free energy guided sampling. *J. Chem. Theory Comput.*, **8 (6)**, 2134-2140 (2012).

[179] A. K. Faradjian and R. Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, **120 (23)**, 10880 (2004).

[180] L. Maragliano and E. Vanden-Eijnden. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.*, **426 (1-3)**, 168-175, (2006).

[181] A. Laio and F. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.*, **71 (12)**, 126601 (2008).

[182] A. C. Pan, D. Sezer, and B. Roux. Finding transition pathways Using the string method with swarms of trajectories. *J. Phys. Chem. B*, **112 (11)**, 3432-3440 (2008).

[183] A. D. Mackerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, et al. All-Atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102 (18)**, 3586–3616 (1998).

[184] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26 (16)**, 1781-1802 (2005).

[185] R. Vargas, J. Garza, B. P. Hay, and D. A. Dixon. Conformational study of the Alanine dipeptide at the MP2 and DFT levels. *J. Phys. Chem. A*, **106 (13), 3213-3218** (2002).

[186] Paul E. Smith. The Alanine dipeptide free energy surface in solution. *J. Chem. Phys.*, **111 (12)**, 5568-5579 (1999).

[187] D. J. Tobias and C. L. Brooks III. Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: a comparison of theoretical results. *J. Phys. Chem.*, **96 (9)**,3864–3870 (1992).

[188] B. Strodel and D. J. Wales. Implicit solvent models and the energy landscape for aggregation of the Amyloidogenic KFFE peptide. *J. Chem. Theory Comput.*, **4 (4)**, 657–672 (2008).

[189] D. S. Chekmarev, T. Ishida, and R. M. Levy. Long-time conformational transitions of alanine dipeptide in aqueous solution: continuous and discrete-state kinetic models. *J. Phys. Chem. B*, **108 (50)**, 19487–19495 (2004).

[190] W. Ren, E. Vanden-Eijnden, P. Maragakis, and Weinan E. Transition pathways in complex systems: application of the finite-temperature string method to the alanine dipeptide. *J. Chem. Phys.*, **123 (13)**, 134109 (2005).

[191] D. Van Der Spoel and M. M. Seibert. Protein folding kinetics and thermodynamics from atomistic simulations. *Phys. Rev. Lett.*, **96 (23)**, 238102 (2006).

[192] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 6801 (2005).

[193] N. V. Buchete and G. Hummer. Peptide folding kinetics from replica exchange molecular dynamics. *Phys. Rev. E*, **77 (3)**, 030902(R) (2008).

[194] S. Muff and A. Caflisch. ETNA: equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations. *J. Phys. Chem. B*, **113 (10)**, 3218-3226 (2009).

[195] D. D. Sancho and R. B. Best. What Is the Time Scale for α-Helix Nucleation? *J. Am. Chem. Soc.*, **113 (17)**, 6809-6816 (2011).

[196] B. Srodel and D. J. Wales. Free energy surfaces from an extended harmonic superposition approach and kinetics for alanine dipeptide. *Chem. Phys. Lett.*, **466 (4-6)**, 105-115 (2008).

[197] D. Hamelberg, C.A. de Oliveria, and J.A. McCammon. Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J. Chem. Phys.*, **127 (15)**, 155102 (2007).

[198] Y. Wang, C. B. Harrison, K. Schulten, and J. A. McCammon. Implementation of accelerated molecular dynamics in NAMD. *Comput. Sci. Discov.*, **4 (1)**, 015002 (2011).

[199] W. Zheng, M. Andrec, E. Gallicchio, and R.M. Levy. Simulating replica exchange simulations to protein folding with a kinetic network model. *Proc. Natl. Acad. Sci. U. S. A.*, **104 (39)**, 15340-15345 (2007).

[200] J. C. Keck. Variational theory of reaction aates. *Adv. Chem. Phys.*, **13**, 85 (1967).

[201] J. C. Keck. Statistical investigation of dissociation cross-sections for diatoms. *Discuss. Faraday Soc.* **33**, 173-182 (1962).

[202] A. F. Voter and J. D. Doll. Dynamical corrections to transition state theory for multistate systems: surface self-diffusion in the rare event regime. *J. Chem. Phys.*, **82 (1)**, 80-92 (1985).

[203] D. E. Ensign and V. S. Pande. Bayesian single-exponential kinetics in single-molecule experiments and simulations. *J. Phys. Chem. B*, **113 (36)**, 12410-12423 (2009).

[204] X. Huang, G. R. Bowman, S. Bacallado, and V. S. Pande. Rapid equilibrium sampling initiated from nonequilibrium data. *Proc. Natl. Acad. Sci. U. S. A.*, **106 (47)**, 19765-19769 (2009). ,

[205] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J. Chem. Phys.*, **119 (6)**, 3559-3566 (2003).

[206] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. Mackerell. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.*, **8 (9)**, 3257-3273 (2012).

[207] M. G. Evans and M. Polanyi. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans. Faraday Soc.*, **31**, 875-894 (1935).

[208] R. A. Marcus. Theoretical relations among rate constants, barriers, and Broensted slopes of chemical reactions. *J. Phys. Chem.*, **72 (3)**, 891-899 (1968).

[209] M. T. Woodside and S. M. Block. Reconstructing folding energy landscapes by single-molecule force spectroscopy. *Annu. Rev. Biophys.* **43**, 19-39 (2014).

[210] K. C. Neuman and A. Nagy. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods*. **5 (6)**, 491-505 (2008).

[211] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, **276 (5315)**, 1109-1112, (1997).

[212] Y. Suzuki and O. K. Dudko. Single molecules in an extension clamp: extracting rates and activation barriers. *Phys. Rev. Lett.*, **110 (15)**, 158105 (2013).

[213] X. Hu and H. Li. Force spectroscopy studies on protein-ligand interactions: a single protein mechanics perspective. *FEBS Lett.*, **588 (19)**, 3613-3620 (2014).

[214] T. Hoffmann and L. Dougan. Single molecule force spectroscopy using polyproteins. *Chem. Soc. Rev.*, **41 (14)**, 4781-4796 (2012).

[215] W. J. Greenleaf, M. T. Woodside, and S. M. Block. High-resolution, single-molecule measurements of biomolecular motion. *Annu. Rev. Biophys. Biomol. Struct.*, **36**, 171-190 (2007).

[216] I. De Vlaminck and C. Dekker. Recent advances in magnetic tweezers. *Annu. Rev. Biophys.*, **2012**, **41**, 453-472 (2012).

[217] C. Bustamante, Z. Bryant, and S. B. Smith. Ten years of tension: single-molecule DNA mechanics. *Nature*, **421**, 423-427 (2003).

[218] A. N. Gupta, A. Vincent, K. Neupane, H. Yu, F. Wang, and M. T. Woodside. Experimental validation of free-energy-landscape reconstruction from non-equilibrium single- molecule force spectroscopy measurements. *Nat. Phys.*, **7**, 631-634 (2011).

[219] C. Lv, X. Gao, W. Li, B. Xue, M. Qin, L. D. Burtnick, H. Zhou, Y. Cao, R. C. Robinson, and W. Wang. Force-enhanced binding of calcium ions by gelsolin. *Nat. Commun.*, **5**, 1-9 (2014).

[220] J. C. M. Gebhardt, T. Bornschlögl, and M. Rief. Full distance-resolved folding energy landscape of one single protein molecule. *Proc. Natl. Acad. Sci. U. S. A.*, **107 (5)**, 2013-2018, (2010).

[221] H. Yu, X. Liu, K. Neupane, A. N. Gupta, A. M. Brigley, A. Solanki, I. Sosova, and M. T. Woodside. Direct observation of multiple misfolding pathways in a single prion protein molecule. *Proc. Natl. Acad. Sci. U. S. A.*, **109 (14)**, 5283-5288, (2012).

[222] K. Neupane, H. Yu, D. A. N. Foster, F. Wang, and M. T. Woodside. Single-molecule force spectroscopy of the add adenine riboswitch relates folding to regulatory mechanism. *Nucleic Acids Res.*, **39 (17)**, 7677-7687 (2011).

[223] G. Hummer and A. Szabo. Kinetics from nonequilibrium single-molecule pulling experiments. *Biophys. J.*, **85 (1)**, 5-15 (2003).

[224] O. K. Dudko, G. Hummer, and A. Szabo. Intrinsic rates and activation free energies from single-molecule pulling experiments. *Phys. Rev. Lett.*, **96 (10-17)**, 108101 (2006).

[225] O. K. Dudko. Decoding the mechanical fingerprints of biomolecules. *Q. Rev. Biophys.*, **49**, 1-14 (2015).

[226] O. K. Dudko, G. Hummer, and A. Szabo. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proc. Natl. Acad. Sci. U. S. A.*, **105 (4)**, 15755-15760 (2008).

[227] Y. J. Zhang and O. K. Dudko. A transformation for the mechanical fingerprints of complex biomolecular interactions. *Proc. Natl. Acad. Sci. U. S. A.*, **110 (41)**, 16432-16437 (2013).

[228] G. Hummer and A. Szabo. Free energy profiles from single-molecule pulling experiments. *Proc. Natl. Acad. Sci. U. S. A.*, **107 (50)**, 21441-21446 (2010).

[229] M. Hinczewski, J. C. M. Gebhardt, M. Rief, and D. Thirumalai. From mechanical folding trajectories to intrinsic energy landscapes of biopolymers. *Proc. Natl. Acad. Sci. U. S. A.*, **110 (12)**, 4500-4505 (2013).

[230] D. J. Wales and T. Head-Gordon. Evolution of the potential energy landscape with static pulling force for two model proteins. *J. Phys. Chem. B*, **116 (29)**, 8394-8411 (2012).

[231] O. K. Dudko, T. G. W. Graham, and R. B. Best. Locating the barrier for folding of single molecules under an external force. *Phys. Rev. Lett.* **107 (20-11)**, 208301 (2011).

[232] G. I. Bell. Models for the specific adhesion of cells to cells. *Science*, **200 (4342)**, 618-627 (1978).

[233] E. Evans and K. Ritchie. Dynamic strength of molecular adhesion bonds. *Biophys. J.*, **72 (4)**, 1541-1555 (1997).

[234] D. D. Minh and J. A. McCammon. Springs and speeds in free energy reconstruction from irreversible single-molecule pulling experiments. *J. Phys. Chem. B*, **112 (19)**, 5892-5897 (2008).

[235] S. Marsili and P. Procacci. Free energy reconstruction in bidirectional force spectroscopy experiments: the effect of the device stiffness. *J. Phys. Chem. B.*, **114 (17)**, 2509–2516 (2010).

[236] D. J. Wales and T. Head-Gordon. Evolution of the potential energy landscape with static pulling force for two model proteins. *J. Phys. Chem. B*, **116**, 8394-8411 (2012).

[237] J. A. V Butler. Studies in heterogeneous equilibria. Part II.-The kinetic interpretation of the Nernst theory of electromotive force. *Trans. Faraday Society*, **19**, 729–733 (1924).

[238] T. Erdey-Gruz and M. Volmer. Zur theorie der wasserstoffüberspannung. *Z. Phys. Chem.*, **150**, 203–213 (1930).

[239] A. Ahmed and H. Gohlke. Multiscale modeling of macromolecular conformational changes combining concepts from rigidity and elastic network theory. *Proteins*, **63**, 1038-1051 (2006).

[240] M. H. Kim and M. K. Kim. Review: elastic network model for protein structural dynamics. *JSM Enzymol Protein Sci.*, **1 (1)**, 1001 (2014).

[241] I. Putz and O. Brock. Elastic network model of learned maintained contacts to predict protein motion. *PLoS ONE*, **12 (8)**, e0183889 (2017).

[242] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78 (14)**, 2690 (1997).

[243] A. Chatterjee and D. G. Vlachos. An Overview of spatial microscopic and accelerated kinetic Monte Carlo methods. *J. Comput. Mater. Des.*, **14 (2)**, 253-308 (2007).

[244] L. C. Bock, L. C. Griffin, J. A. Latham, E. H. Vermass, and J. J. Toole. Selection of single stranded DNA molecules that bind and inhibit human thrombin. *Nature*, **355 (6360)**, 564-566 (1992).

[245] R. F. Macaya, P. Schultze, F. W. Smith, J. A. Roe, and J. Feigon. Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution. *Proc. Natl. Aca. Sci. U. S. A.*, **90 (8)**, 3745-3749 (1993).

[246] J. M. Burke and A. Berzal-Herranz. In vitro selection and evolution of RNA: applications for catalytic RNA, molecular recognition, and drug discovery. *FASEB J.*, **7 (1)**, 106–112 (1993).

[247] J. Zhou and J. Rossi. Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.*, **16**, 181-202 (2017).

[248] D. M. Tasset and M. F. Kubik, and W. Steiner. Oligonucleotide inhibitors of human thrombin that bind distinct epitopes. *J. Mol. Biol.*, **272 (5)**, 688–698 (1997).

[249] D. J. Patel, A. T. Phan, and V. Kuryavyi. Human telomere, oncogenic promoter and 5′- UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, **35 (22)**, 7429-7455 (2007).

[250] J. L. Huppert. Hunting G-quadruplexes. *Biochimie.*, **90 (8)**, 1140-1148 (2008).

[251] L. H. Hurley. Secondary DNA structures as molecular targets for cancer therapeutics. *Biochem. Soc. Trans.*, **29 (6)**, 692-696 (2001).

[252] B. Gatto, M. Palumbo, and C. Sissi. Nucleic acid aptamers based on the G-quadruplex structure: therapeutic and diagnostic potential. *Curr. Med. Chem.*, **16 (10)**, 1248-1265.

[253] M. Gellert, M. N. Lipsett, and D. R. Davies. Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. U. S. A.*, **48 (12)**, 2013-2018 (1962).

[254] A. N. Lane, J. B. Chaires, R. D. Gray, and J. O. Trent. Stability and kinetics of G-quadruplexes. *Nucleic Acids Res.*, **36 (17)**, 5482-5515, (2008).

[255] J. L. Mergny, A. De Cian, A. Ghelab, B. Saccà, and L. Lacroix. Kinetics of tetramolecular quadruplexes. *Nucleic Acids Res.*, **33 (1)**, 81-94 (2005).

[256] R. Krauss, I. et al. Thrombin-aptamer recognition: a revealed ambiguity. *Nucleic Acids Res.*, **39 (17)**, 7858–7867 (2011).

[257] R. V. Reshetnikov, A. V. Golovin, and A. M. Kopylov. Comparison of models of thrombin-binding 15-mer DNA aptamer by molecular dynamics simulation. *Biochemistry (Mosc.)*, 75 (8), 1017–1024 (2010).

[258] R. V. Reshetnikov, A. V. Golovin, V. Spiridonova, A. Kopylov, and J. Šponer. Structural dynamics of thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG) quadruplex DNA studied by large-scale explicit solvent simulations. *J. Chem. Theory Comput.*, **6 (10)**, 3003–3014 (2010).

[259] V. M. Marathias and P. H. Bolton. Structures of the potassium-saturated, 2:1, and intermediate, 1:1, forms of a quadruplex DNA. *Nucleic Acids Res.*, **28 (9)** , 1969–1977 (2000).

[260] V. B. Tsvetkov, I. et al. A Universal base in a specific role: tuning up a thrombin aptamer with 5-Nitroindole. *Sci. Rep.*, **5**, 16337 (2015).

[261] E. Kim, C. Yang, and Y. Pak. Free-energy landscape of a thrombin-binding DNA aptamer in aqueous environment. *J. Chem. Theory Comput.*, **8 (11)**, 4845–4851 (2012).

[262] P. Jayapal, G. Mayer, A. Heckel, and F. Wennmohs. Structure–activity relationships of a caged thrombin binding DNA aptamer: insight gained from molecular dynamics simulation studies. *J. Struct. Biol.*, **166 (3)**, 241–250 (2009).

[263] V. Limongelli. et al. The G-triplex DNA. *Angew. Chem.*, **52 (8)**, 2269–2273 (2013).

[264] C. Yang, S. Jang, and Y. Pak. Multiple stepwise pattern for potential of mean force in unfolding the thrombin binding aptamer in complex with Sr2+. *J. Chem. Phys.*, **135 (22)**, 225104 (2011).

[265] X. Mao, and W. H. Gmeiner. NMR study of the folding-unfolding mechanism for the thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG). *Biophys. Chem.*, **113 (2)**, 115-160 (2005).

[266] J. R. Williamson. G-quartet structures in telomeric DNA. *Annu. Rev. Bioph. Biom. Struct.*, **23**, 703–730 (1994).

[267] C. C. Hardin, A. G. Perry, and K. White. Thermodynamic and kinetic characterization of the dissociation and assembly of quadruplex nucleic acids. *Biopolymers*, **56 (3)**, 147–194 (2000).

[268] L. Sun, et al. Unfolding and conformational variations of thrombin-binding DNA aptamers: synthesis, circular dichroism and molecular dynamics simulations. *ChemMedChem.*, **9 (5)**, 993–1001 (2014).

[269] X. Zeng, L. Zhang, X. Xiao, Y. Jiang, Y. Guo, X. Yu, X. Pu, and M. Li. Unfolding mechanism of thrombin-binding aptamer revealed by molecular dynamics simulation and Markov state model. *Sci. Rep.*, **6**, 24065 (2016).

[270] P. Schultze, R. F. Macaya, and J. Feigon. Three-dimensional solution structure of the thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG). *J. Mol. Biol.*, **235 (5)**, 1532-1547 (1994).

[271] J. A. Doudna and T. R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418 (6894), 222-228 (2002).

[272] T. R. Cech, A. J. Zaug, and P. J. Grabowski. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, **27 (3)**, 487-496 (1981).

[273] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35 (3)**, 849-857 (1983).

[274] I. L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31 (13)**, 3429-3431, (2003).

[275] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9 (1)**, 133-148 (1981).

[276] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: a structure for Deoxyribose nucleic acid. *Nature*, **171**, 737-738 (1953).

[277] N. J. Baird, X. W. Fang, N. Srividya, T. Pan, and T. R. Sosnick. Folding of a universal ribozyme: the ribonuclease P RNA. *Q. Rev. Biophys.*, **40 (2)**, 113-161 (2007).

[278] G. Zemora and C. Waldsich. RNA folding in living cells. *RNA Biol.*, **7 (6)**, 634-641 (2010).

[279] R. Schroeder, A. Barta, and K. Semrad. Strategies for RNA folding and assembly. *Nat. Rev. Mol. Cell. Biol.*, **5 (11)**, 908-919 (2004).

[280] I. Shcherbakova, S. Mitra, A. Laederach, and M. Brenowitz. Energy barriers, pathways and dynamics during folding of large, multidomain RNAs. *Curr. Opin. Chem. Biol.*, **12 (6)**, 655-666 (2008).

[281] T. R. Sosnick and T. Pan. RNA folding: models and perspectives. *Curr Opin. Struct. Biol.*, **13 (3)**, 309-16 (2003).

[282] D. K. Treiber and J. R. Williamson. Exposing the kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.*, **9 (3)**, 339-45 (1999).

[283] D. K. Treiber and J. R. Williamson. Beyond kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.*, **11 (3)**, 309-314 (2001).

[284] S. A. Woodson. Recent insights on RNA folding mechanisms from catalytic RNA. *Cell. Mol. Life. Sci.*, **57 (5)**, 796-808 (2000).

[285] S. A. Woodson. Structure and assembly of group I introns. *Curr. Opin. Struct. Biol.*, **15 (3)**, 324-330 (2005).

[286] S. A. Woodson. Compact intermediates in RNA folding. *Annu. Rev. Biophys.*, **39**, 61-77 (2010).

[287] A. M. Mustoe, C. L. Brooks, and H. M. Al-Hashimi. Hierarchy of RNA functional dynamics. *Annu. Rev. Biochem.*, **83**, 441–466 (2014).

[288] P. C. Bevilacqua and J. M. Blose. Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annu. Rev. Phys. Chem.*, **59**, 79–103 (2008).

[289] W. Zhang and S.-J. Chen. RNA hairpin-folding kinetics. *Proc. Natl. Acad. Sci. U. S. A.*, **99 (4)**, 1931–1936 (2002).

[290] X. Xu, T. Yu, and S.-J. Chen. Understanding the kinetic mechanism of RNA single base pair formation. *Proc. Natl. Acad. Sci. U. S. A.*, **113 (1)**, 116-121 (2016).

[291] I. Jr. Tinoco, P. T. Li, S. B. Smith, and C. Bustanmante. Determination of thermodynamics and kinetics of RNA reactions by force. *Q. Rev. Biophys.*, **39 (4)**, 325-360 (2006).

[292] J. Vieregg, W. Cheng, C. Bustamante, and I. Jr. Tinoco. Measurement of the effect of monovalent cations on RNA hairpin stability. *J. Am. Chem. Soc.*, **129 (48)**, 14966-14973 (2007).

[293] M. T. Woodside, C. Garcia-Garcia, and S. M. Block. Folding and unfolding single RNA molecules under tension. *Curr. Opin. Chem. Biol.*, **12 (6)**, 640-646 (2008).

[294] X. Zhuang. Single-molecule RNA science. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 399-414 (2005).

[295] H. Xu, B. Plaut, X. Zhu, M. Chen, U. Mavinkurve, A. Maiti, G. Song, K. Murari, and M. Mandal. Direct observation of folding energy landscape of RNA hairpin at mechanical loading rates. *J. phys. Chem. B*, **121 (10)**, 2220, (2017).

[296] F. Colizzi and G. Bussi. RNA unwinding from reweighted pulling simulations. *J. Am. Chem. Soc.*, **134 (11)**, 5173-5179 (2012).

[297] N.-J. Deng and P. Cieplak. Free energy profile of RNA hairpins: a molecular dynamics simulation study. *Biophys. J.*, **98 (4)**, 627-636 (2010).

[298] D. R. Bell, S. Y. Cheng, H. Salazar, and P. Ren. Capturing RNA folding free energy with coarse-grained molecular dynamics simulations. *Sci. Rep.*, **7**, 45812 (2017).

[299] G. Hummer and A. Szabo. Free energy profiles from single-molecule pulling experiments. *Proc. Natl. Acad. Sci. U. S. A.*, **107 (50)**, 21441-21446 (2010).

[300] G. Colmenarejo and I. Jr. Tinoco. Structure and thermodynamics of metal binding in the P5 helix of a group I intron ribozyme. *J. Mol. Biol.*, **290 (1)**, 119-135 (1999).

[301] G.H. Vineyard. Frequency factors and isotope effects in solid state rate processes. *J. Phys. Chem. Solids*, **3 (1-2)**, 121-127 (1957).

[302] B. Alberts, A. Johnson, J. Lewis, M. Ra, K. Roberts, and P. Walter. Molecular biology of the cell. *Garland Science, New York, Fourth edition*, (2002).

[303] C. Branden and J. Tooze. Introduction to protein structure. *Garland Publishing, New York*, (1991).

[304] L. Stryer, J. M. Berg, and J. L. Tymoczko. Biochemistry. *W. H. Freeman and Co. Ltd. New York, Fifth edition*, (2002).

[305] WikiDoc. https://www.wikidoc.org/index.php/Amino_acid

[306] Wikipedia. https://en.wikipedia.org/wiki/Chirality_(chemistry)

[307] Biology Exams 4 U. http://www.biologyexams4u.com/2012/09/amino-acids-introduction.html#.WjjZD3lx3IU

[308] Wikipedia-User:LadyofHats. https://commons.wikimedia.org/wiki/File:Main_protein_structure_levels_en.svg.

[309] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7 (1)**, 95-99 (1963).

[310] UCSF Computer Graphics Lab. https://www.iop.vast.ac.vn/theor/conferences/smp/1st/kaminuma/UCSFComputerGraphicsLab/AAA.html

[311] Lumen Microbiology. https://courses.lumenlearning.com/microbiology/chapter/structure-and-function-of-rna/

[312] Atdbio. https://www.atdbio.com/content/5/Nucleic-acid-structure