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

Investigation of structural dynamics and Allosteric mechanisms of SAMHD1 protein complex via Molecular Dynamics studies

By: **Kajwal Kumar Patra**

Supervisors: Dr. Swati Bhattacharya, Prof. Saurabh Basu
Jan, 2018.

Investigation of structural dynamics and Allosteric mechanisms of SAMHD1 protein complex via Molecular Dynamics studies

A thesis submitted by

Kajwal Kumar Patra

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

Statement

The work contained in the thesis entitled "*Investigation of structural dynamics and Allosteric mechanisms of SAMHD1 protein complex via Molecular Dynamics studies*" has been carried out by me under the supervision of Dr. Swati Bhattacharya, Assistant Professor and Prof. Saurabh Basu, Professor, Department of Physics, Indian Institute of Technology Guwahati. This work has not been submitted elsewhere for the award of any degree.

(Kajwal Kumar Patra)
Department of Physics
Indian Institute of Technology Guwahati
Guwahati - 781039

January 10, 2018

Disclaimer

The bibliography included in this thesis is, by no means complete but contains the ones which are consulted thoroughly 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 the figures in this thesis are taken from other sources and properly cited.

(Kajwal Kumar Patra)
Department of Physics
Indian Institute of Technology Guwahati
Guwahati - 781039

January 10, 2018

Certificate

It is certified that the work contained in the thesis entitled "*Investigation of structural dynamics and Allosteric mechanisms of SAMHD1 protein complex via Molecular Dynamics studies*" by Mr. Kajwal Kumar Patra (Roll no. 136121025), a Ph.D. student of the Department of Physics, Indian Institute of Technology Guwahati for the award of Doctor of Philosophy has been carried out under our supervision. This work has not been submitted elsewhere for the award of any degree.

(Dr. Swati Bhattacharya)
Department of Physics
Indian Institute of Technology Guwahati
Guwahati - 781039

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

To my Family...

Acknowledgements

I express my sincere gratitude to all those people who made this dissertation possible. Foremost, I am very much indebted to my supervisor Dr. Swati Bhattachrya for her motivation, patient guidance, invaluable suggestions, constant encouragement and support throughout my research work. I am also thankful to her for introducing me to many advanced subjects in physics. My humble submission to Prof. Saurabh Basu, my co-supervisor, for being stood by me through thick and thin. I owe him a debt of gratitude for all his help and support when I needed it the most.

I convey my sincere thanks to my doctoral committee members, Dr. P. Kumar Padmanabhan, Dr. Tapan Mishra and Dr. S. Vimal Katiyar for constant encouragement, support, valuable suggestions and critical comments during the course of my work and specially during the annual review. I am thankful to all other faculty members of Physics department for being helpful in all regards. My special thanks to Dr. T. N. Day, Prof. P. Poulose for their timely help and support. I am also thankful to my teachers Debananda Sir (RJHS), Amar Sir, Dr. B. K. Dash (Ravenshaw), Dr. Sarmistha Ma'am (Ravenshaw) and others who inspired me to motivate towards Physics.

I wish to thank Department of Physics, Indian Institute of Technology Guwahati (IITG), to provide me with necessary computational facilities. I wish to thank Indian Institute of Technology Guwahati for providing a great library facility and the computer center with outstanding network connectivity and various computational resources. I thank all the technical assistant of the department. I would specially like to thank Mr. B. B. Purkayastha and H. Medhi in this regard. Susmita, my colleague and co-worker, deserves a special thanks for being there, in every situations of this PhD adventure. My heartily thanks to my seniors like Saurabh da, Dr. Kartik Sau, Dr. Debasish Das, Dr. Bappaditya ray, Dr. Sudin Ganguly, Bibhuti bhai, Shiba bhai and many more for their help and encouragements in research related

discussions. Special thanks to Pratap, Sangkha, Pankaj, Srikrishna, Bijita, Srimoy, Krishnanjan, Arun for their help and trust on me as well. I will be surely missing the entertaining socio-geo-political discussions over evening tea sessions.

I am grateful to IITG for the financial support. I am also grateful to National PARAM Supercomputing facility, Pune and PARAM-ISHAN supercomputing facility, IIT Guwahati for the high speed computational resources.

I am deeply indebted to my parents, family, Dipu bhai, Bada bhai, Nini and Kanhu for their unconditional love, constant support and encouragement towards my thesis work and beyond. Their true love and strong trust made me always to keep my faith on-board. They always stood behind me during the ups and downs of my life. To the end, I am deeply blessed to be a son of my parents. Lots of love to Bapi (my mother).

Kajwal

Abstract

The human sterile alpha motif and HD domain-containing protein 1 (SAMHD1) is a retroviral restriction factor in myeloid cells and non-cycling CD4+ T cells, a feature imputed to its phosphohydrolase activity since the enzyme depletes the cellular dNTP levels inhibiting reverse transcription. The catalytically active form of the protein is an allosterically triggered tetramer, whose HIV-1 restriction properties are attributed to its dNTP - triphosphohydrolase activity. The tetramer itself is assembled by a GTP/dNTP combination. This enzyme uses the strategy of deoxynucleotide starvation, which is thought to prevent effective reverse transcription of the retroviral genome hence restricting HIV-1 propagation. HIV-2 and SIV have evolved defences against SAMHD1, underscoring its role in restriction. It utilizes GTP-Mg+2-dNTP cross bridges to link and stabilize its adjacent monomers. Previous studies have provided high-resolution structures of GTP/dNTP bound enzyme complexes, but have not been able to provide information on dynamics. Very little is known about how it assembles into a tetramer and how long the tetramer stays intact. In this computational study, we provide a molecular dynamics based analysis of the structural stability and allosteric site dynamics in SAMHD1. We have investigated the allosteric links which assemble and hold the tetramer together. we report the investigation of structural and allosteric site dynamics of SAMHD1 complex with different ligands bound to the allosteric sites. Our preliminary studies have focused on the effect of nucleotide depletion on the stability of the complex and markers of incipient instability. Our studies show that only GTP bound SAMHD1 is a much more dynamic entity than the GTP/dATP locked tetramer.The absence of dATP has a greater detrimental effect on the stability of the complex than that of GTP. Later on, going beyond the phenomenological exploration of nucleotide depletions. we have used correlation network analysis along with MD techniques to study the flow of allosteric information across the active complex. We have found evidence of a reciprocal allosteric “handshake” occurring across monomeric units. We have also uncovered a short linker region as the nexus for funnelling the regula-

tory signal from phosphorylation at T592 from the surface to the interior core of the protein. We have also extended this analysis to a regulatory mutant of SAMHD1. Experimental studies have indicated that phosphorylation of T592 downregulates HIV-1 restriction. A similar result is also achieved by a phosphomimetic mutation T592E. While a mechanistic understanding of the process is still elusive, the loss of structural integrity of the enzyme is conjectured to be the cause of the impaired dNTPase activity of the T592E mutant. MD simulations show that the T592E mutation causes slightly elevated local motions which remain confined to the short helix (residues 591-595), which contains the phosphorylation site and do not cause long-range destabilization of the SAMHD1 tetramer within the timeframe of the simulations. Thus, the regulatory mechanism of SAMHD1 is a more subtle mechanism than has been previously suspected. Next we perform the MD studies of the monomeric forms of the SAMHD1 *wt* and Cys mutants in order to explore the regulatory mechanisms of redox regulations. While understanding the regulatory mechanisms of SAMHD1 protein complex is still remains a challenge, our current study has revealed interesting information regarding the role of cysteine residues on the switching mechanism. However, a complete picture of the exact mechanism of operation of the putative redox switch remains opaque. While all-atom MD simulations can play a vital role in unveiling the true picture, our study sheds new light on dynamics and allosteric information flow in this complex protein.

Contents

List of Figures

List of Tables

Abbreviations

SAMHD1 : sterile alpha motif HD domain-containing protein 1
HIV-1 : human immunodeficiency virus-1
MD : molecular dynamics
MC : monte carlo
NAMD : nanoscale molecular dynamics
dNTP : deoxynucleoside triphosphate
GTP : guanosine triphosphate
Catsite : catalytic site
Allosite : allosteric site
XRD : X-Ray diffraction
NMR : nuclear magnetic resonance
PBC : periodic boundary conditions
wt : wild type
CTD : C-terminal domain
LMI : linear mutual information
CM : covariance matrix
DCCM : dynamical cross correlation map
RMSD : root mean square deviation
RMSF : root mean square fluctuation
PCA : principal component analysis
DS link : disulfide link

Chapter 1

Introduction

1.1 Introduction

Proteins are the most versatile macromolecules and hold the key to practically all living processes. Even after several decades of intense research into the enzymatic activity of proteins, the link between their structures and function and regulatory mechanisms that allow proteins to perform crucial cellular functions,[1] we are only just scratching the surface. The exquisite link between the structure, function and dynamics of proteins is what sets them apart from other major classes of bio-macromolecules such as lipids, nucleic acids and carbohydrates, although exceptions abound. Consider some examples of protein functions that provide a glimpse of their versatility. Some of the most important functions of proteins include providing structure, catalysis, movement, signalling and transport. Proteins such as epidermal keratin and collagen act as building blocks. Another important class of proteins include transporters that serve to transport different species of molecules or ions or even electrons. Prominent among these are membrane proteins that allow molecules to cross the cell membrane forming a conduit between the intra and extra-cellular environments. Proteins are responsible for inter and intra-cellular signalling,[2] [3] are an important component of the immune system and play and essential role in converting chemical energy into mechanical energy in the case of muscles. Apart from these, there are numerous accessory proteins that assist or control other proteins. One of the most important functions of proteins is the catalysis of numerous complex biochemical reactions in the body. Catalytic proteins or enzymes combine high specificity to substrates with remarkable efficiency. No wonder, these are often called molecular machines that enable life. As a corollary to their multifaceted roles in living processes, any defect or malfunction in protein structure, folding or

mechanism may lead to several pathologies directly correlated to the type of protein involved. Consider antithrombin, a protein belonging to the serpin class, as an example. The wild type antithrombin is capable of controlling the thrombin enzyme, which is involved in coagulation related reactions.[4] Mutations may destroy the controlling ability of antithrombin, leading to venous thrombosis. Many neurodegenerative disorders are the results of aggregation due to malfunctions or glitches in protein folding. The diseases like Huntington, Creutzfeld-Jacob (Mad Cow) or the infamous Alzheimer are examples of such malfunctions.[5] Proteins can also be used to attack pathogenic viruses such such as HIV, SARS, hepatitis or to block bacterial synthesis to fight against the unwanted infections in the field of drug design. Consequently, the quest to gain deeper insights into the functioning of these remarkable molecules and to harness their power has engendered a vast interest in the medicinal and biophysical fraternities.[6] [7] [8]

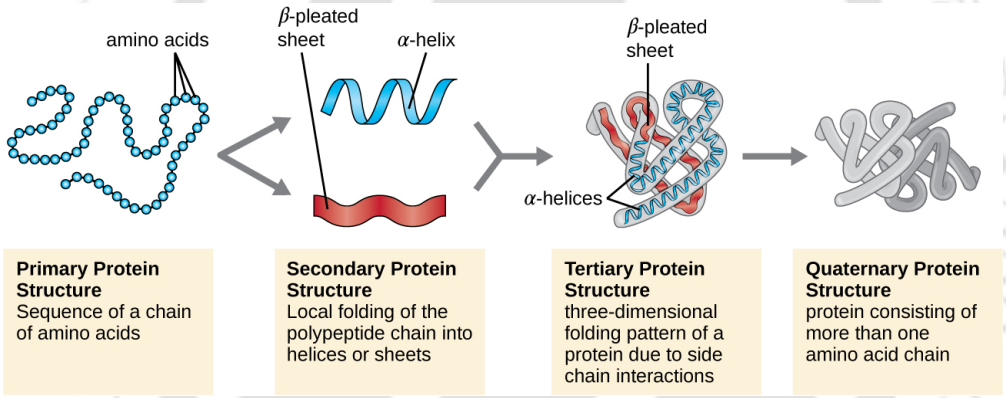

Figure 1.1: Time evolution of a protein from primary to quaternary structures. The image is adapted from the source of National Human Genome Research Institute (USA).

Proteins are synthesized as the polymeric sequence of amino acid residues. However, in order to be functional, most proteins need to acquire a specific three-dimensional structure.[9] The folding process from one-dimensional amino acid sequence to stable tertiary structures is usually extremely efficient and grants proteins their selective and diverse functions. The evolution of life has its origin in pre-existing peptide molecules formed from inorganic materials. The protoplasmic proteins differ abruptly from the proteins formed in the form of plant seeds of many species.[10] The synthesis of proteins is usually a two-step procedure.[11] The information encoded in the genetic material, DNA, is transcripted into mRNA. In the next step, called translation, the ribosomes in the cytoplasm or the ER synthesize proteins using the mRNA as a template. An essential role is played by tRNAs

that deliver the amino-acids to the site of synthesis and decipher mRNA codon. The initial structures, immediately after the synthesis of proteins, are unstructured polypeptide chains, made up of twenty different natural amino acids, linked together by peptide bonds.[12] Amino acids, the building block of proteins, are amphoteric compounds and made up of central carbon (C) and four groups that are attached to the central carbon : an amino group (-NH2), a carboxylic group (-COOH), a hydrogen atom (H) and a unique (-R) group that gives the amino acid its identity.[13] [14] [15] A peptide (or amide) bond links the -carboxyl group of one amino acid with the -amino group of another accompanied by the elimination of a water molecule.

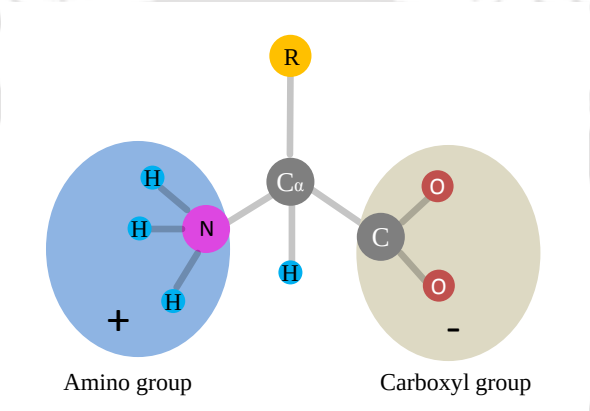

Figure 1.2: An amino acid structure with four basic group components .

The amino acids are often called residues. Sequential peptide linkages lead to the formation of long chains of amino acids, called polypeptide chains.[16] [17] Most natural polypeptide chains contain between 50 to 2000 amino acid residues. However, to be functional, most protein need to adopt a local or secondary and finally a tertiary or minimum energy stable conformational structure. When a primary structure finds a favorable atmosphere like aqueous environment, it spontaneously adopts its secondary and then tertiary structure to start functioning[18] [19]. Sometimes a protein may require the presence of a molecular chaperones or co-factors to find its tertiary structure. One of the grand challenges to emerge in the post Watson and Crick era was the "protein folding problem".[7] [8] Significant progress has been made towards understanding how the primary sequence, *i.e.* the sequence of amino acids in a polypeptide chain, encodes information regarding the native tertiary states and their functions. Currently, foldable proteins are routinely designed for specific functions. There exists a class of proteins that has lately garnered much attention, that are natively unstructured and yet functional. However, we will confine our discus-

sion to regular proteins that have to procure their well-folded structures in order to be functional. A functional protein has to undertake a journey from its primary sequences of amino acids to a stable minimal energy native state. This process highlights the dynamical nature of a protein. In fact, proteins are inherently dynamic entities traversing huge spatial and temporal orders of magnitude. An increasing number of studies attest to the importance of dynamics to the functions of proteins such as ligand binding, catalysis, signal transmission, etc. [20] [21] [22]

1.2 Landmarks in the study of proteins

The name "protein", coined from the Greek word "$\pi\varrho\omega\tau\varepsilon\iota\circ\zeta$", which means *"of the 1st importance"*. The importance of proteins was recognized already by Swedish chemist, Jacob Berzelius, recognized as one of the founders of modern chemistry who coined the name protein, in 1838. He also suggested that enzymes are cellular catalysts. In 1902 Fischer [15] and Hofmeister independently concluded that proteins are composed of a covalent chain of amino acids. [23] The X-ray crystallography studies of amino acids, peptides and fibrous proteins conducted by Pauling and Corey during the 1930s contributed substantially to the understanding of protein structure and led to the proposal of the secondary structure elements in 1951. [14] [13] [12] In 1953 Sanger presented the amino acid sequence of insulin. Five years later, a major breakthrough came when the first three dimensional structure of a globular protein, myoglobin, was determined. [18] With the molecular basis available, the development of biological science and the understanding of cellular processes accelerated considerably. Along with these findings the understanding of how proteins are coded on the DNA level has been of great importance. These mechanisms were elucidated mainly during the 1960s with the structural insights of Rosalind Franklin and Watson and Crick as starting point. In the 1970s the recombinant DNA technique emerged as a powerful tool with the main achievements being the recombinant expression of proteins in foreign hosts, rapid determination of protein sequences on the DNA level and the possibility of modification and engineering of proteins using more or less rational approaches. [24] [25] [26] Despite the fact that proteins have been investigated for more than 150 years and the rapid development of the field in recent years, the molecular mechanisms that determines how proteins adopt their three dimensional structures and how they recognize and bind to other biomolecules are not completely understood.

1.2.1 The chronological interpretations of protein nature

If we go back to the early studies of proteins or biomolecules, it is not that long time when people used to think of proteins as static entities. This can be traced back from the available reports of "key-lock" hypothesis proposed by Emil Fisher.[15] This was proposed at the end of nineteenth century to explain the enzymatic activities, commonly referred as the rigid docking mechanism in modern terminology. The concept of proteins as static entities persisted till the quarter of twentieth century. At that time the idea was so established that even the pioneer physicists of twentieth century like Erwin Schrödinger framed these molecules of life as some kind of "aperiodic crystals".[27] Later on, with the steady flow of experimental data especially from the field of enzymology, the idea of proteins as rigid bodies are found to be inadequate to explain the experiments. It was Linus Pauling who first came up with structural explanations for the function of the Haemoglobin protein in terms of ligand-mediated conformation changes[24] long before the X-Ray structures were solved. The "induced fit" theory proposed by Daniel Koshland which leads to the theory of "structural changes" of protein structures while binding with a foreign molecule or substrate has also provided the evidence of the relations between the structure, dynamics and function.[28] These links between "structure-dynamics-functions" became more evident when John Kendrew and Max Perutz[18] became the first to obtain the atomic resolution models. Later on, the "breathing-motion" analogy to a continuous movement of the protein structures was proposed by Linderstrom-Lang and Schellmann[29] and the idea was successfully probed by the experimentalists in the following decades. The role of flexibility in conformational changes for enzymatic mechanisms or in broader sense for the protein functions began to be accepted more preferably after the crystallographic apo and holo structures of Lysozyme in complex with inhibitors[26] [25] were solved successfully. Around the same time, the alternative views to induced fit theory of allostery and Manod-Wyman-Changeux models[30] were proposed.

After the first half of the twentieth century, there was a sharp drift in the field of biomolecular system studies. In the 70s a new era emerged when people started virtual experiments or simulation studies. The dynamic nature of the proteins were well established by the developements of molecular dynamics (MD) studies[31] and Nuclear magnetic resonance (NMR) studies[32]. Now one can probe the different scales of protein motions from atomic vibrations (in the range of femto seconds) to folding or misfolding of tertiary structures (in the range of micro seconds) with the help of

highly efficient computational facilities. Time-resolved crystallography[33], Förster resonance energy transfer (FRET)[34] [35] or neutron scattering methods[36] [37] [38] are some of the recent technological advancements in this area of research, that provide valuable information regarding macro biomolecular systems and their mechanisms.

1.3 Protein dynamics

In order to understand protein mechanisms and their corresponding functions, one needs to record the time evolution of the conformations in the form of a series of snapshots over a period of time. The smaller the time interval between two conformational states, the more detailed is the information regarding the mechanisms.[39] Depending upon the timescales, fluctuations can be categorized into two broad groups: the slow time scale process[40] (includes signal transduction, enzymatic catalysis, folding/unfolding)[41] [41] [42] and the fast time scale process (includes loop motion, hinge binding, side-chain rotations etc.).[43] [44] X-Ray crystallography is an experimental technique that provide high resolution images of the structures. However, the information is in the form of one or two small number of images showing a handful of conformational states. Hence, the method does not provide information about how the system transits between the two states.[45] [46] [47] [48]

Molecular Dynamics (MD) simulations[31], which can be used as a computational microscope to gain insights into the dynamics of the molecule is one of the most successful tools for the complex dynamical studies protein systems. In Chapter-2 we discuss the methodologies and theoretical aspects of MD simulations.[49] Although the recent developments both ion experimental and computational aspects has provided more pictures of the protein kinematics, the time and length scales involved in real biomolecular systems poses an enormous challenge in both computational and experimental studies. Nonetheless, the all-atom MD simulations of a large size protein is still limited to micro second time scales with a huge computational cost[50]. However, the enormous advances in computational efficiency and the development of special-purpose supercomputers and GPUs has no doubt aroused the curiosity of many researchers to explore hidden areas of protein mechanisms.

After the brief general introduction to the protein structure and dynamics, we can now begin the discussion of the main subject of my thesis.

1.4 Antiviral immunity

1.4.1 AIDS as a Pandemic

A rapid and uncontrolled spread of an infectious diseases to large number of people across continents or even worldwide is referred as a pandemic. Looking back at human history, our planet has been suffered from a number of devastating pandemics such as tuberculosis, smallpox, the Spanish flu (in 1918-1919), Black death (in 1350), plague of Athens (439 BC) etc. In recent times, HIV/AIDS has came across globally as a viral pandemic. According to the recent surveys, approximately 40 million people are suffering from HIV/AIDS worldwide. The agencies like Global Burden of disease study published a report in 2015, suggesting that the incidence of the HIV infection was at it's peak in the year 1997 with a rate of 3.3 million per year and then fell down to a certain level of about 2.6 million per year to the end of the year 2005 and became steady at these rates at current times .[51] [52]

1.4.2 HIV life cycle

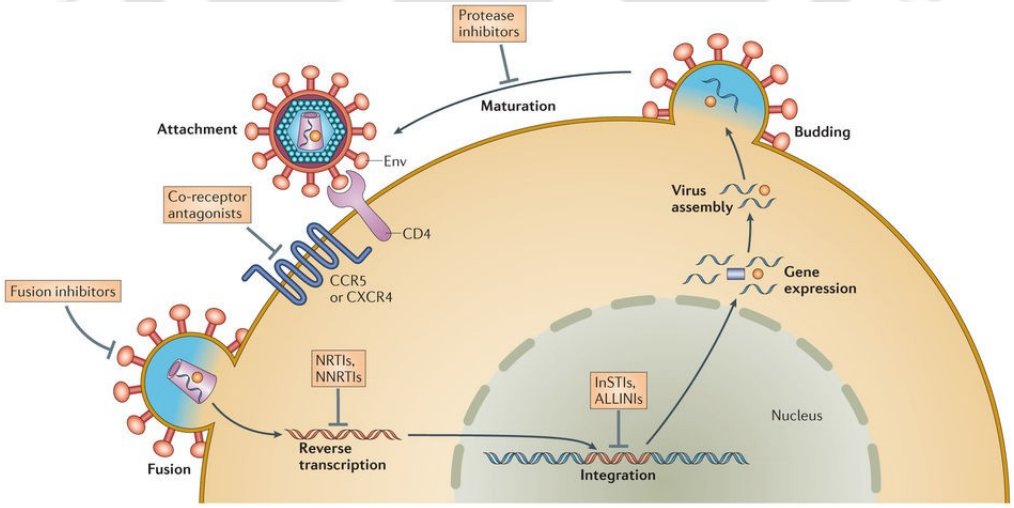

Figure 1.3: A schematic diagram of the life cycle of HIV-1 retrovirus inside the human host cell. This image is adapted from the Journal, *Nature Reviews Microbiology* **12**, 772–778 doi: 10.1038/mrmicro.3351[53]

We discuss briefly the method by which the HIV-1 virus infects the human cells and uses the host cell organelles to replicate and reproduce more HIV viruses that leads to syndrome known as AIDS.[54]

Human immunodeficiency virus (HIV) is retrovirus and constitutes a outer lipid-rich layer and a inner protein capsid. The inner capsid contains single stranded RNA molecules and also a specific kind of protein enzymes known as reverse transcriptase those act to transcribe RNA into DNA. The outer lipid rich layer contains protogen receptors those bind to the protein receptors present in the cell membrane of the human cell. In the first step, a matured HIV virus having all the necessary contents mentioned above, approaches to the human cell. The second step involves with the binding of the virus and the host cell. The protogens on the outer lipid layer of HIV virus binds with the protein receptors such as CD4 and CCR4 found on the cell membrane of the host cells. Once the binding takes place and HIV recognizes the specific human cells, the fusion process between the outer lipid layer of the virus and host cell membrane takes place. The lipid-rich envelop of virus fuses with the plasma membrane of the host cell injecting the contents including proteins and single stranded RNA molecules into the cytoplasm of the host cell. Inside the cytoplasm, the reverse transcriptase enzymes of the virus transcribes the RNA into DNA and forms the viral double stranded DNA. The viral dsDNA then travels into nucleus of the host cell. Inside the nucleus, a specific enzyme known as the retroviral integrase incorporates the viral dsDNA into the DNA genome of that human cell. The cell than transcribes the viral DNA back into the viral RNA as well as the viral mRNA to produce the viral proteins. Once the protein synthesis is done, the clusters of single stranded viral RNA as well as the viral proteins move towards the membrane of the cell and begin to push outward to come out from the cell and become matured into HIVs and ready to infect other host cells.

HIV, as a human pathogen was initially identified about 35 years ago.[55] [54] As a consequence of the inability of parent host to propagate a immuno response to the virus, it causes the chronic activation of the immune system that leads to a awful control of viral replication and immuno debilitation.[56] Initially HIV-1 was come to existence from Simian immunodeficiency virus in chimpanzees (SIV_{cpz}) which is responsible of the viral infection of its natural chimpanzee host.[55] While on the contrary the evolution of HIV-2 has happened from SIV_{sm} which has a greater replication rate in its natural simian host beyond causing any disease. Our studies are primarily focused on a protein that is a part of the intrinsic cellular defense system and helps in blocking HIV-1 infection.

1.4.3 Intrinsic cellular defenses

Intrinsic antiviral immunity[56], often called the first line of defense against retroviral infection, refers to form a innate immunity that directly renders a cell non-permissive to a specific class or species of virus by blocking viral replication and assembly. A subset of host cellular proteins, pre-existent in certain cell types, termed restriction factors, play a critical role in this defense by recognizing viral components and interfering with the viral life cycle. A thorough understanding of how these restriction factors defend against retro-viruses as well as viral counter-measures to intrinsic immunity effectors is key to developing new therapeutics to combat viral pandemics.

Both prokaryotes and eukaryotes have developed complex defense mechanisms to protect their cells from viral invasions. In addition to the conventional innate and adaptive immune responses, novel antiviral systems are emerging that are based on pre-existing, constitutively expressed, intracellular proteins (restriction factors) and that comprise an intrinsic immunity system. These factors can be considered as the front line of antiviral defense, as they act during the very first steps of virus-host interactions in an immunologically naive host, and they are often counterattacked by viral protective proteins.[57] Until now there are four anti-HIV restriction factors have been discovered. These are TRIM5 (triparitite motif)[58], BST-2 (bone marrow stromal cell antigen 2), A3G (apoliprotein B mRNA-editing, enzyme catalytic, polipeptide-like 3G (APOBEC3G)) and SAMHD1 (sterile alpha motif and HD Domain 1).[59] [60] [61]

1.4.4 "SAMHD1": As the HIV1 restriction factor

Amongst all the anti-HIV restriction factors discovered, SAMHD1 is the most recent one.[60] SAMHD1 is a protein that is actively expressed and functional in human myeloid cells and has been identified as an intrinsic restriction factor that blocks the viral replication of the HIV genome in the myeloid cells. The protein was initially identified as the human ortholog of the mouse gene Mg11 and induced through the interferon (IFN) treatment of dendritic cells and macrophages.[62] [63] [64] SAMHD1 forms a tetrameric complex with four monomeric segments each containing 626 amino acids. SAMHD1 includes a sterile alpha motif (SAM) domain and a Histidine Asparatate (HD) domain.[65] [66] [67] Both the domains are conserved in course of

time and play vital roles in signaling and nucleic acid metabolism. Furthermore, this protein complex is noticed to be associated with AGS (Aicardi-Goutières autoimmune syndrome) and acts as a negative regulator of the innate immune response.[68] SAMHD1 operates against the primary stages of the retroviral replications but its explicit mode of action is yet to be unveiled.

SAMHD1 as an exception

Unlike other discovered restriction factors, SAMHD1 is exceptional because of the fact that the retrovirus HIV-1 has no means to counteract SAMHD1 since it does not contain the auxiliary protein Vpx. Vpx is a Virion-associated protein encoded by certain strains of lentiviruses, like HIV-2, SIV, those can bypass the antiviral block in myeloid cells by employing the protein Vpx to target SAMHD1 for proteasomal degradation.[69] [62] [59] Pull-down experiments in which SAMHD1 has been co-precipitated with Vpx has indicated that it possesses the ability to restrict the retroviral replications.[70] Since no specific counter mechanism has yet been identified by which HIV-1 responds to SAMHD1, it appears that SAMHD1 may be a relevant and important solution to the HIV-1 pandemic.[71] [72] [73] [65]

1.5 SAMHD1: Origin and Structure

The Sterile Alpha Motif and HD containing protein 1 (SAMHD1) was initially discovered by cDNA screening of the peripheral blood monocyte derived dendritic cells.[74] [75] Human SAMHD1 is 72% identical to mouse MG11. Mutations in the gene are associated with Aicardi-Goutières autoimmune syndrome[68] [76] (AGS), an early onset neurodevelopmental disorder. SAMHD1 is implicated in restriction of retroviruses in terminally differentiated cells of myeloid lineage with non-cycling CD4+ T cells.[71] [77] [62] SAMHD1 is known to be an allosterically regulated deoxynucleotide- triphosphohydrolase , which decreases the cellular dNTP pool below the level required for virus reverse transcription. Crystallographic studies have provided structural insights into the operation of this enzyme.[74] The functional form of this protein is a tetramer. The N-terminal 148 residues of this protein are known to function as a nuclear localization signal. The catalytic core/HD domain of this protein (residue 114-626), which by itself forms a tetramer is necessary and sufficient for both phosphohydrolase activity (here after referred to as dNTPase

activity) as well as restriction of HIV-1.[78] [79] Each monomer of the HD domain contains two allosteric sites (Allosite 1 and Allosite 2), as well as one Catalytic site (Catsite).[66] The binding of GTP or dGTP to Allosite 1 and any dNTP to Allosite 2 trigers the formation of dimer of dimers. No other nucleobase is found to bind at these sites. Tetramerization is induced by GTP/dNTP combinations, with each tetramer possessing 4 $GTP - Mg^{+2} - dNTP$ cross bridges at the interface of three monomers. Crystal structures revealed that the HD domain consists of a major lobe (residues 115-373), minor lobe (residues 376-450) and finally a C-terminal domain (CTD) (residue 455-599) which features a prominent anti-parallel beta sheet. The major lobe contains the catalytic site. The CTD contains Thr592, phosphorylation of which by cdk1 has been suggested as an "on/off" switch for the enzyme.[67] [80] [81]

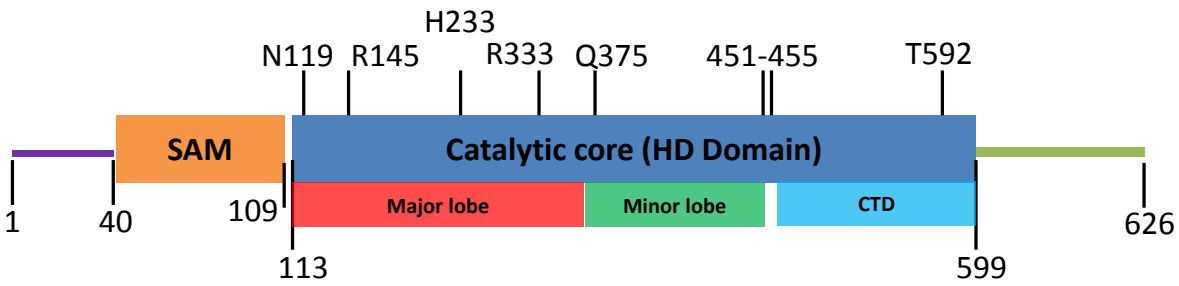

Figure 1.4: A qualitative bar representation of a monomeric segment of SAMHD1 protein complex, indicating the residue numbers from 1 to 626 with specified domain sections.

1.6 Motivations behind the study

Researchers have proposed that SAMHD1 acts as a negative regulator of innate immune response to interferon stimulating DNA.[82] [83] [84] Further studies attributed a exonuclease activity to the protein; suggesting that this might be a mechanism for antiviral activity. However, more recent experimental studies indicate that while SAMHD1 does display nucleic acid binding activity, it is not capable of functioning as a nuclease.

This does arise the question: how does SAMHD1 actually function as a restriction factor? The answer was found from the in-vivo studies which showed that SAMHD1 is highly expressed in resting CD4(+)T cells, which are resistant to HIV-1

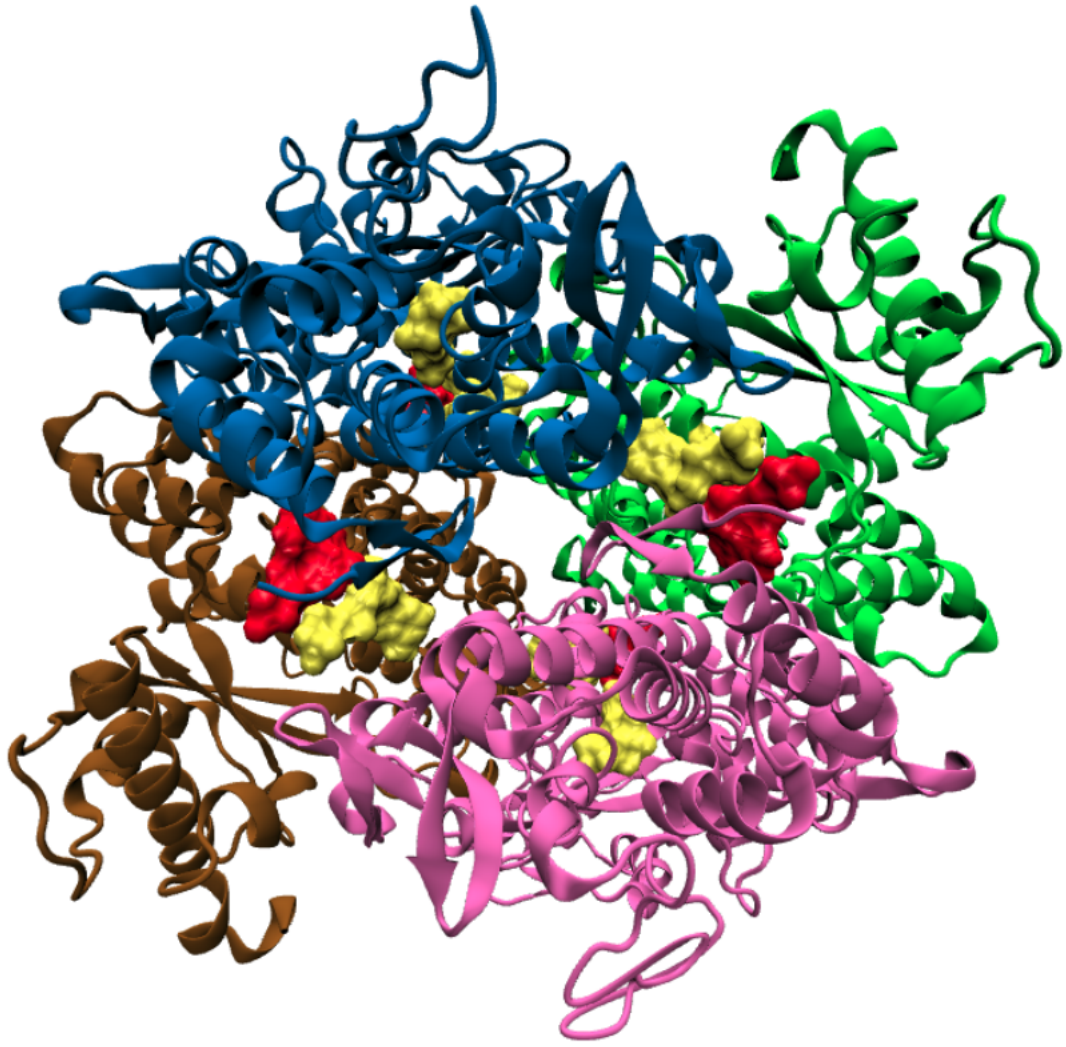

Figure 1.5: A cartoon representation of SAMHD1 protein complex with different chain colors. The dATP and GTP molecules are shown in red and yellow respectively.

infections. The strategy adopted by non cycling CD4(+)T cells circulating in peripheral blood and lymphoid organs is to use SAMHD1 to reduce the cellular dNTP levels. This prevents HIV-1 reverse transcriptase enzyme from producing DNA and hence stops infection. HIV-2/SIVmac/SIVsmm on the other hand, employ the protein Vpx to counter the restriction effect SAMHD1 by targeting it for proteosomal degradation.

The XRD derived structures[85] [18] have provided valuable insights into this enzyme. However, these insights have all taken the form of snapshots of various nucleotide bound states of the protein. The mechanistic details of how the triphosphohydrolase enzyme functions remains elusive. Simply put, we know what the end

states of the enzyme's functional cycle look like, but we do not know how the enzyme transits between these end states. Given that HIV-1 restriction by SAMHD1 is currently attributed to this very activity, and the concomitant ability to effectively starve the virus of dNTPs ; understanding the dynamics of this enzyme at a molecular level is of paramount importance . As this protein , especially in its tetrameric form is too large to carry out effective NMR experiments (even with the use of expensive methyl-labeling strategies), we have used molecular dynamics to study its behavior with particular focus on the dynamics at the allosteric sites.

The regulation of SAMHD1 based HIV restriction, is also an open question. Cellular levels of SAMHD1 are observed to be relatively constant during the cell division cycle. Interestingly, however, SAMHD1 contains a target sequence (592-TPQK-595) for cyclin dependent kinase 1 (cdk1). In vivo studies have led to a model where cdk1 phosphorylates SAMHD1 at T592, rendering it incapable of restricting HIV-1 during the S-phase of cycling cells. This loss of restriction capability does not correlate with a downregulation of the dNTPase activity of SAMHD1. The phosphomimetic mutations T592D and T592E have been shown to lose the ability to restrict HIV-1 in non-cycling cells, but were observed to retain their dNTPase activity through in-vitro studies. The unphosphorylable mutants T592V and T592A were able to restrict HIV-1 in non-cycling U937 cells.

A recent [86] [87] study has posited that structural changes in SAMHD1 associated with the presence of the negatively charged phosphate group on the side chain of T592 lead to structural changes in the neighborhood of the cdk1 recognition site, which in turn leads to impaired tetramer formation and hence a decrease in the dNTPase activity of the protein. Hence, we are interested in comparing the behavior of the *wt* with a phosphomimetic mutation. This aspect is considered in greater detail in Chapter 5.

Nevertheless, the importance of SAMHD1 in retroviral restriction is underscored by the fact that HIV-2/SIVmac/SIVsmm have evolved a defense against it: Therefore elucidating the enzymatic mechanism of SAMHD1 action is of great importance, both from the standpoint of understanding biophysical machinery, as well as due to the necessity of understanding HIV / host interaction.

The central enigma of retroviral restriction by SAMHD1 is how does an enzyme so ostensibly inefficient manage to lower dNTP levels to far below its Km (Michaelis constant)? The current state of the art involves well characterized assembled SAMHD1 tetramers, bound to different assembly and substrate nucleotide combinations. What is missing, however, is a picture of SAMHD1 dynamics at the

molecular level that could shed light on the enzymatic mechanism. In this thesis, we describe our computational studies using all-atom molecular dynamics simulations to systematically elucidate the mechanism of allosteric cross-talk and regulation of the protein.

Bibliography

[1] C. Mathews, K. Van Holde and K. Ahern, San Francisco, USA, edn **3** (2000).

[2] B. F. Volkman, D. Lipson, D. E. Wemmer and D. Kern, Science **291**, 2429 (2001).

[3] A. K. Gardino et al., Cell **139**, 1109 (2009).

[4] J. A. Huntington, Trends in biochemical sciences **31**, 427 (2006).

[5] C. M. Dobson, Nature **418**, 729 (2002).

[6] J. W. Kelly, Nature Structural & Molecular Biology **9**, 323 (2002).

[7] K. Huang, *Lectures on statistical physics and protein folding*, World Scientific, 2005.

[8] P. Echenique and J. L. Alonso, Molecular Physics **105**, 3057 (2007).

[9] C. B. Anfinsen, Biochemical Journal .

[10] K. Henzler-Wildman and D. Kern, Nature **450**, 964 (2007).

[11] F. A. Mulder, A. Mittermaier, B. Hon, F. W. Dahlquist and L. E. Kay, Nature structural & molecular biology **8**, 932 (2001).

[12] L. Pauling and R. B. Corey, Proceedings of the National Academy of Sciences **37**, 251 (1951).

[13] L. Pauling and R. B. Corey, Proceedings of the National Academy of Sciences **37**, 729 (1951).

[14] L. Pauling and R. B. Corey, Proceedings of the National Academy of Sciences **37**, 235 (1951).

[15] E. Fischer, European Journal of Inorganic Chemistry **27**, 2985 (1894).

[16] C. I. Branden et al., *Introduction to protein structure*, Garland Science, 1999.

[17] Y. Li and D. Zhao, Basics of molecular biology, in *Molecular Imaging*, pages 541–601, Springer, 2013.

[18] J. C. Kendrew et al., Nature **181**, 662 (1958).

[19] C. Ramakrishnan and G. Ramachandran, Biophysical Journal **5**, 909 (1965).

[20] K. A. Henzler-Wildman et al., Nature **450**, 913 (2007).

[21] P. McClean et al., Cell Biology Education **4**, 169 (2005).

[22] Y. Mu, P. H. Nguyen and G. Stock, Proteins: Structure, Function, and Bioinformatics **58**, 45 (2005).

[23] K. A. Dill, Biochemistry **29**, 7133 (1990).

[24] L. Pauling, Proceedings of the National Academy of Sciences **21**, 186 (1935).

[25] C. Blake et al., Nature **206**, 757 (1965).

[26] L. N. Johnson and D. Phillips, Nature **206**, 761 (1965).

[27] E. Schródinger, (1967).

[28] D. Koshald, Proceedings of the National Academy of Sciences of the United States of America **44**, 98 (1958).

[29] K. Linderstrom-Lang and Schellmann, The Enzyme **1**, 443 (1959).
[30] J. Monod, J. Wyman and J.-P. Changeux, Journal of molecular biology **12**, 88 (1965).
[31] J. A. McCammon, B. R. Gelin and M. Karplus, Nature **267**, 585 (1977).
[32] K. Wuethrich, Accounts of chemical research **22**, 36 (1989).
[33] J. Hajdu et al., Nature Structural & Molecular Biology **7**, 1006 (2000).
[34] A. T. Brunger, P. Strop, M. Vrljic, S. Chu and K. R. Weninger, Journal of structural biology **173**, 497 (2011).
[35] T. Heyduk, Current opinion in biotechnology **13**, 292 (2002).
[36] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge and D. I. Svergun, Journal of the American Chemical Society **129**, 5656 (2007).
[37] F. Gabel et al., Quarterly reviews of biophysics **35**, 327 (2002).
[38] C. D. Putnam, M. Hammel, G. L. Hura and J. A. Tainer, Quarterly reviews of biophysics **40**, 191 (2007).
[39] Y. Shan, A. Arkhipov, E. T. Kim, A. C. Pan and D. E. Shaw, Proceedings of the National Academy of Sciences **110**, 7270 (2013).
[40] R. D. Schaeffer, A. Fersht and V. Daggett, Current opinion in structural biology **18**, 4 (2008).
[41] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).
[42] R. G. Smock and L. M. Gierasch, Science **324**, 198 (2009).
[43] V. Daggett and A. Fersht, Nature Reviews Molecular Cell Biology **4**, 497 (2003).
[44] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror and D. E. Shaw, Current opinion in structural biology **19**, 120 (2009).
[45] D. S. Glazer, R. J. Radmer and R. B. Altman, Structure **17**, 919 (2009).
[46] R. O. Dror, M. O. Jensen and D. E. Shaw, Elucidating membrane protein function through long-timescale molecular dynamics simulation, in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 2340–2342, IEEE, 2009.
[47] J. Hsin, J. Strümpfer, E. H. Lee and K. Schulten, Annual review of biophysics **40**, 187 (2011).
[48] D. Sellis, D. Vlachakis and M. Vlassi, Bioinformatics and Biology Insights **3**, 99 (2009).
[49] N. Tokuriki and D. S. Tawfik, Science **324**, 203 (2009).
[50] M. Akke, Current opinion in structural biology **12**, 642 (2002).
[51] M. S. Cohen, N. Hellmann, J. A. Levy, K. DeCock and J. Lange, The Journal of clinical investigation **118**, 1244 (2008).
[52] H. Wang et al., The lancet. HIV **3**, e361 (2016).
[53] S. B. Laskey and R. F. Siliciano, Nature Reviews Microbiology **12**, 772 (2014).

[54] F. Barre et al., Science **220**, 868 (1983).
[55] F. Kirchhoff, Nature Reviews Microbiology **7**, 467 (2009).
[56] P. D. Bieniasz, Nature immunology **5**, 1109 (2004).
[57] A. M. Sheehy, N. C. Gaddis, J. D. Choi and M. H. Malim, Nature **418**, 646 (2002).
[58] M. Stremlau et al., Nature **427**, 848 (2004).
[59] S. J. Neil, T. Zang and P. D. Bieniasz, Nature **451**, 425 (2008).
[60] N. Laguette et al., Nature **474**, 654 (2011).
[61] I. A. Oussenko, R. Sanchez and D. H. Bechhofer, Journal of bacteriology **184**, 6250 (2002).
[62] K. Hrecka et al., Nature **474**, 658 (2011).
[63] W. Liao, Z. Bao, C. Cheng, Y.-K. Mok and W. Wong, Proteomics **8**, 2640 (2008).
[64] N. Li, W. Zhang and X. Cao, Immunology letters **74**, 221 (2000).
[65] N. Beloglazova et al., The EMBO journal **30**, 4616 (2011).
[66] L. Aravind and E. V. Koonin, Trends in biochemical sciences **23**, 469 (1998).
[67] M. D. Zimmerman, M. Proudfoot, A. Yakunin and W. Minor, Journal of molecular biology **378**, 215 (2008).
[68] G. I. Rice et al., Nature genetics **41**, 829 (2009).
[69] N. Sharova et al., PLoS pathogens **4**, e1000057 (2008).
[70] C. Goujon et al., Journal of virology **82**, 12335 (2008).
[71] G. Doitsh et al., Cell **143**, 789 (2010).
[72] F. Kirchhoff, Cell host & microbe **8**, 55 (2010).
[73] L. D. Passerini, Z. Keckesova and G. J. Towers, Journal of virology **80**, 2100 (2006).
[74] A. Berger et al., PLoS pathogens **7**, e1002425 (2011).
[75] T. Sinkunas et al., The EMBO journal **30**, 1335 (2011).
[76] Y. J. Crow and J. Rehwinkel, Human molecular genetics **18**, R130 (2009).
[77] K. Strebel, J. Luban and K.-T. Jeang, BMC medicine **7**, 48 (2009).
[78] Z. C. Hartman et al., Journal of virology **81**, 1796 (2007).
[79] J. Schultz, P. Bork, C. P. Ponting and K. Hofmann, Protein Science **6**, 249 (1997).
[80] S. Kornberg, I. Lehman, M. J. Bessman, E. S. Simms and A. Kornberg, Journal of Biological Chemistry **233**, 159 (1958).
[81] C. A. Kim and J. U. Bowie, Trends in biochemical sciences **28**, 625 (2003).
[82] Y. J. Crow, Current opinion in immunology **32**, 7 (2015).
[83] M. N. Lee et al., Nature immunology **14**, 179 (2013).
[84] D. C. Goldstone et al., Nature **480**, 379 (2011).
[85] X. Ji, C. Tang, Q. Zhao, W. Wang and Y. Xiong, Proceedings of the National

Academy of Sciences **111**, E4305 (2014).
[86] Z. Wang, A. Bhattacharya, J. Villacorta, F. Diaz-Griffero and D. N. Ivanov, Journal of Biological Chemistry **291**, 21407 (2016).
[87] A. Bhattacharya et al., Scientific reports **6** (2016).

Chapter 2

Computational approaches and methodologies

2.1 Introduction

Computer simulation is a well developed and powerful tool to study the the microscopic details of macroscopic systems by implementing different models specified in terms of molecular structure and inter-atomic interactions.[1] The outcomes of the simulations are then compared to the experimental data or to the analytical predictions, if available, to verify the accuracy of the models and to predict the dynamical evolution of the systems.[2] Computer simulations are also quite useful to investigate atomistic details of complex macroscopic systems that are not accessible experimentally and hence help us understanding many complex phenomena.[3] [4] [5] [6]

Based on the simulations techniques, there are two main families in computer simulations, namely Molecular Dynamics (MD)[5] and Monte Carlo (MC) study[7] [8]. Along with these there is a whole range of hybrid techniques which combines features of both MD and MC. MC simulation was first performed by Metropolis in 1952[9] , on the MANIAC computer in Los. The Monte Carlo method includes a large variety of stochastic techniques based on the use of sampling (random numbers) and probability statistics to investigate problems from various fields. A simple Metropolis sampling method starts from a initial (random) configuration and involves computing the energy difference by displacing one or more particles and finally evaluating the transition probability which satisfies the detailed balance[9]. Subsequently, in 1957, Wood and Parker[7] performed other MC simulations on liquid Argon employing emperically designed Lennard-Jones (LJ) interaction potential. In 1962,

Rahman[5] proposed the algorithm of MD and applied it to the system of LJ particles. Since then, more complex systems, in terms of their extent and interactions between their constituent particles, such as polymers and biomolecules like proteins have been investigated via Molecular Dynamics[6]. Monte Carlo methods are generally used for calculation of thermodynamic properties. Although, Kinetic Monte Carlo methods (KMC) provide an opportunity to study the dynamics as well[10], the main advantage of MD over MC is that it gives a route to the dynamical properties of the system such as the time-dependant response to any change, transport coefficients and rheological properties etc. In this chapter we will be concentrating on the Molecular dynamics simulations and other trajectory analysis techniques.[11] [12] [13]

2.2 Molecular Dynamics

Simulations have brought about a conceptual change in the understanding of biomolecules. Instead of viewing protein functions such as enzymatic catalysis only in terms of structural data provided by high-resolution X-ray crystallography, one now recognizes the important role of the internal atomic motion.[14] [1] [15] MD simulation is a computational approach which predicts the behavior of an N-particle system in the course of time by solving equation of motion. MD simulation methods are further divided into two classes depending on the way the inter-atomic forces are calculated. The first kind is known as classical MD, involves iteratively solving Newton's equation of motion for a system of particles interacting via predefined forces. And the second kind is density functional theory (DFT)[11] [16] based *ab initio* MD (AIMD), such as Car-Parinello MD (CPMD)[12] or Born-Oppenheimer MD (BOMD)[17]. AIMD simulations are computationally expensive for large complex systems such as large proteins.[18]. Currently, a number of classical MD softwares are available for biomolecular systems such as CHARMM[19] [20], AMBER[21], GROMACS[22], NAMD[23] etc. We have used NAMD (NAnoscale Molecular Dynamics) to perform the MD simulations. Given below some key concepts for carrying out MD simulations under different ensembles.

In classical MD simulations, all possible conformations accessible to a particular system can be sampled by integrating Newton's equations of motion.[24] The force in each atom is obtained using appropriate potential energy functions. The time evolution of the equation of motion characterizes how the velocities and positions

varies with time. Newton's second law says:

$$\vec{F_i} = m_i \cdot \vec{a_i} \tag{2.1}$$

where $\vec{F_i}$ is the force applied to particle i of mass m_i, velocity $\vec{v_i}$, acceleration $\vec{a_i}$ at position r_i .

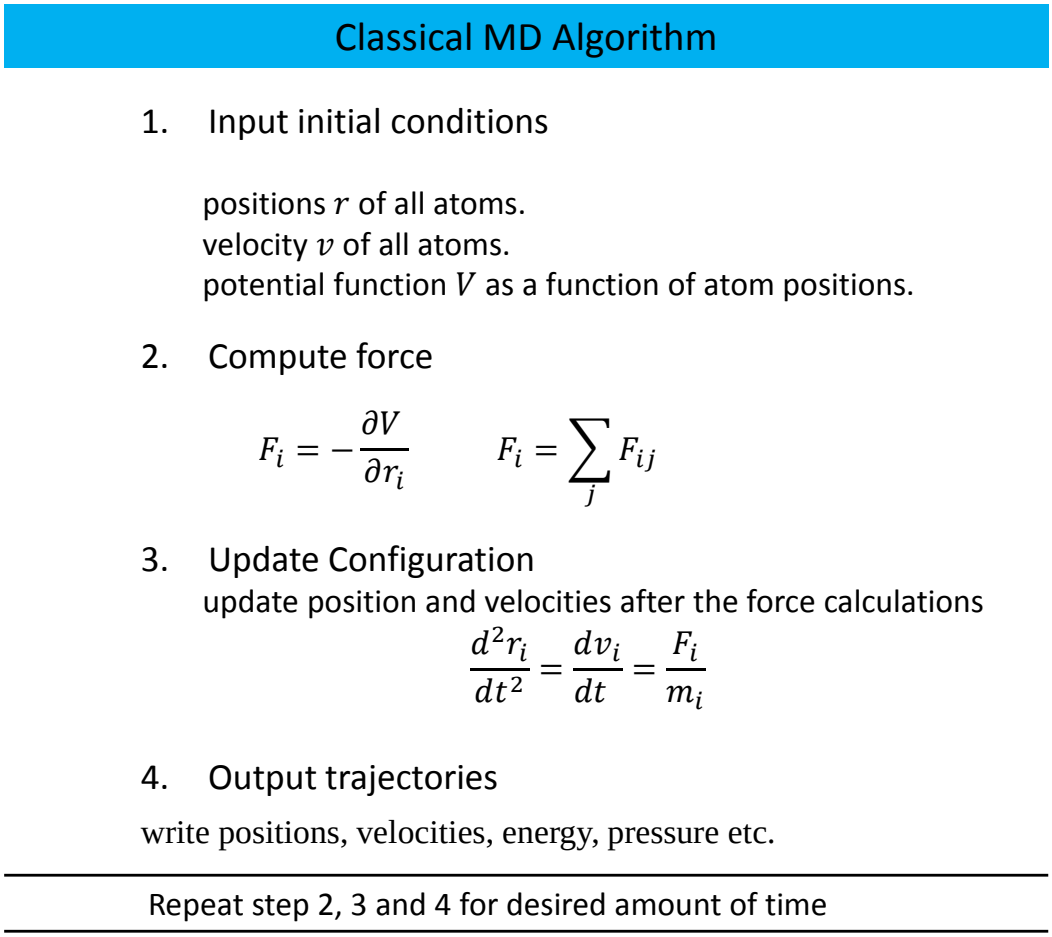

Figure 2.1: A schematic representation of the Classical MD algorithm

The initial coordinates required to start the biomolecular systems as protein simulations are obtained from the experimental methods like X-ray crystallography [25] [26] or NMR [27] [28]. The atoms in these systems are interconnected to each other producing a many-body problem. Hence, it is nearly impossible to solve the equations of motion analytically. Under such situations the equations have to be

solved using numerical methods. One can use finite difference methods to integrate equation of motion in MD simulations[6]. These methods generate MD trajectories with pairwise additive potential models. The integration involves many small steps each separated by a time step δt. The total force on a particle is calculated as the vector sum of its interaction with other particles in the system. The acceleration, obtained from the force, is used to update the position and velocities at time t, to determine the position and velocities at time $t + \delta t$, $t + \delta 2t$ and so on, approximated by Taylor series expansion. "The Velocity Verlet" algorithm[2] [14] is one of the most computationally efficient integration algorithms widely used in many MD codes including NAMD[29] [30] [23].

2.2.1 Velocity Verlet algorithm

In this method the velocity and position of the next time step is obtained from the current time step by following four ways[1] [14]

1. **"half-kick"**

$$\vec{v}\left(t + \frac{\delta t}{2}\right) = \vec{v}(t) + \frac{1}{2}\vec{a}(t)\delta t \tag{2.2}$$

Move the velocity at the half time step $\left(t + \frac{\delta t}{2}\right)$ using initial velocity and acceleration.

2. **"Drift"**

$$\vec{r}\,(t + \delta t) = \vec{r}(t) + \vec{v}\left(t + \frac{\delta t}{2}\right)\delta t \tag{2.3}$$

Move position of the atom to the next step using half moved velocity and previous position.

3. **"Compute force"**

$$\vec{F}_{t+\delta t} = \vec{F}(\vec{r}(t + \delta t)) \tag{2.4}$$

Acceleration at the next step $\vec{a}(t + \delta t)$ is calculated from the updated position; $\vec{r}(t + \delta t)$

4. **"Update velocity"**

$$\vec{v}\,(t + \delta t) = \vec{v}\left(t + \frac{\delta t}{2}\right) + \frac{1}{2}\vec{a}(t + \delta t)\delta t \tag{2.5}$$

The full move of velocity is completed by employing half moved velocity and updated the acceleration. This method is time reversible and conserves linear and angular momenta. It requires only one force evaluation for each time step, and for a fixed time, it exhibits a global error proportional to δt^2. Repeating these above four steps to all the atoms available in the system with desired number of steps will provide the trajectory of the simulation which can be further analyzed in more details to gain insights.

2.2.2 Potential energy functions

Force fields are the other important aspect of the MD simulation as these governs how the parts of the molecules interact with each other. For biomolecular systems, the potential function includes the summing over of a large number of bonded and non-bonded terms. [31] [19]

$$U(\vec{r}) = \sum U_{bonded}(\vec{r}) + \sum U_{nonbonded}(\vec{r}) \tag{2.6}$$

Covalent bonds with 2-,3- and 4-body interactions with $O(N)$ in the summation corresponds to the bonded potential terms. The interaction between all pairs of atoms (excluding the bonded term), with $O(N^2)$ terms in the summation, corresponds to nonbonded potential terms. we use the CHARMM force fields with NAMD to compute the biomolecular systems in aqueous environment. The bonded and nonbonded potential energy terms are as follows:

Potential energy functions for bonded terms:

The potential energy functions for bonded terms includes 2-, 3-, 4-body interactions of covalently bonded atoms as shown in Figure 2.2. The harmonic vibrational motion between an (i, j)-pair of covalently bonded atoms is described by the 2-body spring bond potentials given by,

$$U_{bond} = k(r_{ij} - r_0)^2 \tag{2.7}$$

where, the distance between atoms is given by $r_{ij} = \|\vec{r_j} - \vec{r_i}\|$, r_0 is the equilibrium distance, and k is the spring constant.

The angular vibrational motions occurring between the (i, j, k)-triples of covalently bonded atoms are described by the 3-body angular bond potentials, given by

$$U_{angle} = k_\theta(\theta - \theta_0)^2 + k_{ub}(r_{ik} - r_{ub})^2 \tag{2.8}$$

where, θ in the first term represents the angle in radians between vectors $r_{ij} = \vec{r_j} - \vec{r_i}$ and $r_{kj} = \vec{r_j} - \vec{r_k}$. k_θ represents the angle constant and θ_0 is the equilibrium angle. The second term, the Urey-Bradley term, describes (non covalent) spring between the outer i and k atoms which is active when constant $k_{ub} \neq 0$, where, $r_{ik} = \|\vec{r_k} - \vec{r_i}\|$ gives the distance between the pair of atoms and r_{ub} is the equilibrium distance.

The dihedral angle or 4-body torsion angle potential referred to Figure 2.2 describes the angular spring between the planes formed by the first three and last three atoms of a consecutively bonded (i, j, k, l)-quadruple of atoms,

$$U_{tors} = \begin{cases} k(1 + \cos(n\psi + \phi)) & \text{if } n > 0 \\ k(\psi - \phi)^2 & \text{if } n = 0 \end{cases} \tag{2.9}$$

where ϕ is the angle in radians between (i, j, k)-plane and (j, k, l)-plane. n, the integer constant, indicates the periodicity and is nonnegative. For $n > 0, \psi$ represents the phase shift angle and k represents the multiplicative constant. For $n = 0, \psi$ behaves as the equilibrium angle and the units of k changes to potential/rad^2.

Potential energy functions for Nonbonded terms:

The nonbonded potential terms involve interactions between all (i, j)-pairs of atoms, usually excluding pairs of atoms already involved in a bonded term. The Lennard-Jones (LJ) potential accounts for the weak dipolar attraction between distant atoms and the strong repulsion as atoms become close and is given by,

$$U_{LJ} = (-E_{min}) \left[\left(\frac{R_{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min}}{r_{ij}} \right)^{6} \right] \tag{2.10}$$

where $r_{ij} = \|\vec{r_j} - \vec{r_i}\|$ gives the distance between the pair of atoms. The parameter $E_{min} = U_{LJ}(R_{min})$ is the minimum of the potential term ($E_{min} < 0$, which means that $-E_{min}$ is the well-depth). Since the LJ potential approaches 0 rapidly as r_{ij} increases, it is usually truncated to 0 past a cutoff radius, requiring $O(N)$ compu-

tational cost.

The electrostatic potential is repulsive for atomic charges with the same sign and attractive for the atomic charges for the opposite signs,

$$U_{elec} = \epsilon_{14} \frac{C_{q_i q_j}}{\epsilon_0 r_{ij}} \tag{2.11}$$

where $r_{ij} = \|\vec{r_j} - \vec{r_i}\|$ gives the distance between the pair of atoms, and q_i and q_j are charges on the respective atoms. Coulomb's constant C and the dielectric constant ϵ_0 are fixed for all electrostatic interactions. The parameter ϵ_{14} is a unitless scaling factor whose value is 1, except for a modified 1-4 interaction, where the pair of atoms is separated by a sequence of three covalent bonds, in which case $\epsilon_{14} = \epsilon$, for a fixed constant $0 \leq \epsilon \leq 1$.

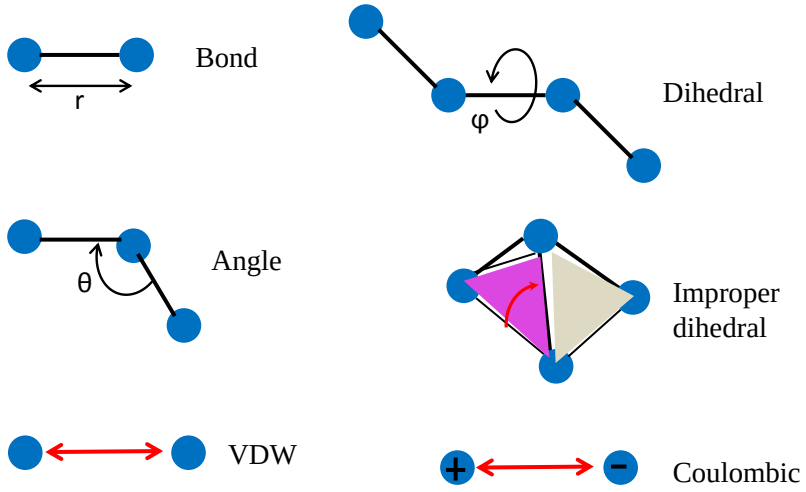

Figure 2.2: Empirical forcefields used as potential energy terms for biomolecular simulations

2.3 Periodic Boundary Conditions

In molecular dynamic simulations studies, Periodic boundary conditions (PBC) plays an important role in odder to avoid large time consumptions and minimize the surface effects. A bulk system usually contains a large number of atoms, of the order of 10^{23}, where as in a typical computer simulation the number of atoms are in

between the order of 10^3 to 10^5 (that is, the average size of of the cell is 20 to 50 Å). So there is always a higher probability that an ample amount of atoms on the surface of the simulation system. This creates an inhomogeneous atmosphere in between the surface atoms and the central atoms of the cell, which is not recommended for a bulk system simulations. To avoid such surface effects we employ periodic boundary conditions (PBS). The idea here is that, the original simulation cell is imagined to be encircled by the imaginary replicas of it's own so that all atoms of the original cell can find their infinite numbers of the images. For an orthogonal simulation cell (with sides L_x, L_y and L_z), the coordinates of the image atoms after employing PBC can be written as

$$\begin{aligned} x\prime &= x + n_x L_x \\ y\prime &= y + n_y L_y \\ z\prime &= z + n_z L_z \end{aligned} \tag{2.12}$$

where, x, y and z are the atoms the real simulation cell and n_x, n_y and n_z are the integers including both positive and negative numbers.

PBC can lead to infinite pairs of interactions which is not feasible. To avoid that a suitable cut off distance is required to limit the pair interaction. Typically the short range interaction goes down within half of the box length. Therefore the cutoff distance (R_c) is chosen such that it is less than or equal to $L/2$. It allows the atom to interact with other real atoms or it's image which ever is closest. This is also referred as minimum image convention.

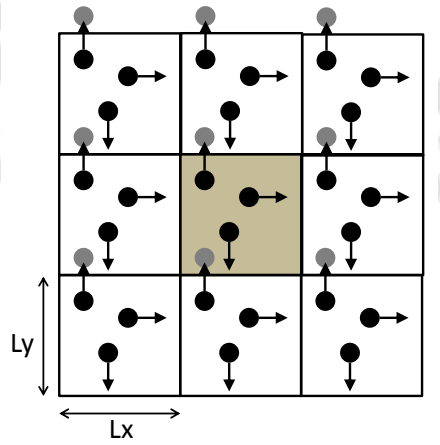

Figure 2.3: A schematic diagram of a 2-dimensional periodic boundary condition. The original cell box is in the center with gray background. while the other boxes are the replica images of the original box.

2.4 Energy minimization Techniques

In biomolecular simulations, we usually start our system preparation from a initial structure taken from the published data (for example, the protein initial structures are available in protein data bank (PDB)). But these initial structures do not necessarily correspond to stable conformations. Before starting the real simulations, it is necessary to bring the system into a stable or lowest energy conformation state. To achieve this, one need to perform the energy minimization on the initial structures. By employing these schemes, the geometry of the system changes step by step so that it can acquire a lowest potential energy conformation state. We discuss some common energy minimization methods.

2.4.1 Newton-Raphson method

A Taylor expansion of the potential energy surface (PES) is the basis of this method. The geometry (x) of the initial structure is updated in every step as

$$x' = x - \frac{E'(x)}{E''(x)} \tag{2.13}$$

where, $E'(x)$, $E''(x)$ are the first order and second order derivatives of the geometry of each points. This method is computationally expensive.[32]

2.4.2 Steepest descent Method

This method approximates the second derivatives as a constant and hence, does not required it to be calculated. The equation to update the geometry is

$$x' = x - \gamma E'(x) \tag{2.14}$$

where, γ is a constant. The minimization of the geometry takes place in the opposite direction of the steepest (or largest) gradient of the initial point. The process then shifts to next largest gradient for the minimization and continues until the minimization is done in all directions. This process is faster than Newton-Raphson method but requires more steps to achieve the minima.[32]

2.4.3 Conjugate gradient method

The Conjugate gradient (CG) method generates a set of directions those do not show the back and forth oscillating behavior like the steepest decent (SD) method during convergence. It makes use of the gradient history to decide a better direction of the next step. In SD method both the gradient and the direction of the successive steps are orthogonal. In CG method, the gradients at each points are orthogonal but the directions are conjugates. It is a property for a set of conjugates directions that, for a quadratic function of M variables, the minimum will be reached at M steps. The CG method moves in a direction $\mathbf{v}_n$ from a point $\mathbf{x}_n$, where $\mathbf{v}_n$ is computed from the gradient at the point and the previous direction vector $\mathbf{v}_{n-1}$ as given below,

$$\mathbf{v}_n = -\mathbf{g}_n + \gamma_n \mathbf{v}_{n-1} \tag{2.15}$$

γ_n is a scalar constant given by,

$$\gamma_n = \frac{\mathbf{g}_n \cdot \mathbf{g}_n}{\mathbf{g}_{n-1} \cdot \mathbf{g}_{n-1}} \tag{2.16}$$

The equation 2.15 can be used only after the first step and onwards. So the first step in the CG method is the same as the SD method (that is, in the direction of the gradient). There are many variants of the CG method, the equation 2.16 is the original Fletcher–Reeves algorithm. An alternative form for the scalar constant was proposed by Polak and Ribiere is given by

$$\gamma_n = \frac{\mathbf{g}_n - \mathbf{g}_{n-1} \cdot \mathbf{g}_n}{\mathbf{g}_{n-1} \cdot \mathbf{g}_{n-1}} \tag{2.17}$$

For purely quadratic functions the Polak–Ribiere method is identical to the Fletcher–Reeves method as all the gradients will be orthogonal. [32] [33]

2.5 Statistical ensembles and use of Thermostats

Integration of Newton's equation of motion only provides a constant-energy surface of a system. But to imitate the experimental conditions for the biological systems we may need the temperature and pressure to be constant in order to create a more real-like situation for the simulation. Based on which thermodynamical quantities to be fixed, several statistical ensembles can be generated. we briefly discuss some of the most commonly used statistical ensembles in molecular dynamics. For the

ensembles discussed here, the number of atoms is constant. [15]

NVE Ensemble (constant volume and energy): This is also referred as micro-canonical ensemble. The pressure and temperature are allowed to change but the energy is fixed. However due to the errors produced during integration process (like rounding and truncation errors) a minimal change in energy can be noticed.

NVT Ensemble (constant volume and temperature): This ensemble is also referred as the canonical ensemble. The temperature is controlled by direct temperature scaling or heat bath at desired temperature. [34] [1] [35]

NPT Ensemble (constant pressure and temperature): This is also referred as isothermal-isobaric ensemble and is a preferable choice when volume, pressure and densities are vital for the simulation. the pressure can be adjusted by coupling a pressure bath to the system [36] [37]

μVT Ensemble: This is also referred the grand canonical ensemble. Here the chemical potential of the system along with the volume and temperature is constant. [38]

2.5.1 Thermostats in MD simulations

In MD simulations thermostats are used to regulate and control the temperature of a system in canonical (NVT) ensemble. There are a number of widely used techniques available as thermostats to control the temperature. These include Anderson thermostat, Nosé-Hoover thermostat, Berendsen thermostat and Langevin thermostat. The idea to use the thermostat is to keep the mean temperature of the simulating system at a constant value through out the simulations. [9] [34] [39]

Anderson Thermostat

In this method, the system is coupled with a heat bath that imposes the constant temperature to the system. The coupling is represented by stochastic collision that randomly acts on arbitrarily chosen particles. The equation of motion include the Hamiltonian equations with an extra added term for stochastic collision in $\frac{dp_i}{dt}$,

$$H = \sum \frac{p_i^2}{2m_i} + \phi(q)$$
$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \tag{2.18}$$

Each instantaneous stochastic collision affects the momentum of the involved particle. The time interval distribution between the successive collision is given by,

$$P(t;\nu) = \nu e^{-\nu t} \tag{2.19}$$

where, $P(t,\nu)$ is probability that the next collision will take place in the interval $[t, t+\delta t]$. The Hamiltonian equation for the entire collection of particles are integrated over each stochastic collisions. The trajectory $\left(q^N(t), p^N(t)\right)$ for N particle and V volume can be used to find the average of any quantity such as,

$$\overline{F} = \lim_{T\to\infty} \int_0^T F\left(q^N(t), p^N(t); V(t)\right) dt \tag{2.20}$$

It shows the canonical distribution n phase space is invariant under repeated application of Anderson algorithm. The time average of any function F, calculated from an Anderson trajectory is equal to the ensemble average of F for the canonical average in which the temperature is T.

$$\overline{F} = \frac{1}{N!Q(N,V,T)} \int F\left(q^N(t), p^N(t); V(t)\right) e^{-\beta H(q^N(t), p^N(t); V)} dq^N dp^N \tag{2.21}$$

where, $Q(N,V,T) = \frac{1}{N!}\int e^{-\beta H(q^N(t), p^N(t); V)} dq^N dp^N$ is the partition function of the canonical ensemble.

Nosé-Hoover Thermostat

In stead of using stochastic collision in the simulated system, an extended Lagrangian with virtual co-ordinates and velocities is used in this thermostat.

Consider the system of N-particles with q_i' coordinates, mass (m_i), potential energy $\phi(q')$ and momenta (p_i'). An additional degree of freedom s is introduced such that

$$q_i' = q_i,\ p_i' = p_i/s,\ t' = \int_0^t dt/s$$

where, q_i, p_i and t are the virtual variables. The real velocity is given by,

$$\frac{dq_i'}{dt'} = s\frac{dq_i'}{dt'} = s\frac{dq}{dt}$$

The above transformation can be thought of as a time scaling by,

$$dt' = dt/s.$$

The Lagrangian of the extended system in terms of virtual variables is,

$$L_{Nosé} = \sum_{i=1}^{\infty} \frac{m_i}{2} s^2 \dot{q_i}^2 - \phi(q) + \frac{Q}{2}\dot{s}^2 - gkT \ln s \tag{2.22}$$

where, Q is the effective mass associated to s and g is the number of degrees of freedom of the system. The logarithmic dependence of the potential on variables is essential for producing canonical ensemble. The momenta conjugate to q_i and s are,

$$p_i = \frac{\partial L_{Nose}}{\dot{q_i} = m_i s^2 \dot{q_i}^2}$$
$$p_s = \frac{\partial L_{Nose}}{\dot{s}} = Q\dot{s}$$

The Hamiltonian for the extended system is,

$$H_{Nose} = \sum_{i=1}^{N} \frac{p_i^2}{2m_i s^2} + \phi(q) + \frac{p_s^2}{2Q} + gkT \ln s \tag{2.23}$$

Equation of motion using extending Hamiltonian are given by,

$$\begin{aligned}
\frac{dq_i}{dt} &= \frac{\partial H_{Nose}}{p_i} = \frac{p_i}{m_i s^2} \\
\frac{dp_i}{dt} &= \frac{\partial H_{Nose}}{q_i} = \frac{-\partial \phi}{\partial q_i} \\
\frac{ds}{dt} &= \frac{\partial H_{Nose}}{p_s} = \frac{ps}{Q} \\
\frac{dp_s}{dt} &= \frac{\partial H_{Nose}}{ds} = \frac{\sum \frac{p_i^2}{m_i s^2} - gkT}{s}
\end{aligned} \tag{2.24}$$

This Nosé-Hoover Thermostat method produces micro-canonical ensemble for the extended system.

Berendsen Thermostat

It is also referred as proportional time rescaling. Here a weak coupling is combined to a external temperature bath. This thermostat, also called proportional thermostat, tries to correct the deviation of actual temperature T from the prescribed temperature T_0. The temperature of the system is allowed to fluctuate in Berendsen thermostat. In this method, velocities are scaled at each time step such that the rate of change in temperature is proposal to the difference in temperature,

$$\frac{dT}{dt} = \frac{1}{\tau}(T_0 - T) \tag{2.25}$$

where, τ is the coupling parameter analogous to ν in Anderson Thermostat.

Berendson's method creates an exponential decay of the system towards the desired temperature.

$$T = T_0 - Ce^{-t/\tau} \tag{2.26}$$

From the above equations we can have,

$$\Delta T = \frac{\Delta T}{\tau}(T_0 - T) \tag{2.27}$$

This leads to the modification of momenta

$$p_i = \lambda p_i$$

where, λ is a scaling factor

$$\lambda^2 = 1 + \frac{\Delta T}{\tau T}\left(\frac{T_0}{T} - 1\right)$$

Langevin Thermostat

Motion of large particles through a continuum of smaller particles can be described by Langevin's equation,

$$\begin{aligned} \ddot{x} &= \nabla\phi - \gamma\dot{x} + \sigma\xi \\ or \qquad \frac{dq_i}{dt} &= \frac{p_i}{m_i}, \quad \frac{dp_i}{dt} = -\frac{\partial\phi(q)}{\partial q_i} - \gamma p_i + \sigma\xi_i \end{aligned} \tag{2.28}$$

The smaller particles create a damping force to the momenta, $-\gamma p_i$. The smaller particles also move in kinetic energy and give random kicks to larger particles.

σ and γ are connected by fluctuation-dissipation relation,

$$\sigma^2 = 2\gamma m_i kT$$

Langevin's equation can be used for molecular dynamics assuming that the atoms being simulated are surrounded in a pool of much smaller fictional particles (*eg.* proteins in aqueous solutions). In each time step, ΔT, the thermostat changes the momenta as

$$\Delta p_i = \left(\frac{\partial \phi(q)}{\partial q_i} - \gamma p_i + \delta j \right) \Delta t \tag{2.29}$$

where γp_i damp the momenta and δj is a Gaussian distributed random number with probability,

$$\rho(\delta j) = \frac{1}{\sqrt{2\Pi\sigma}} exp\left(-\frac{\|\delta p\|^2}{2\sigma^2} \right) \tag{2.30}$$

and standard deviation, $\sigma^2 = 2\gamma m_i kT$. Random fluctuating force represents the thermal kicks from small particles. The damping factor and random force combine to give the correct canonical ensemble.

2.5.2 Nosé-Hoover Langevin piston pressure control

This is one of the barostat schemes used to control pressure in NPT simulations. NAMD in particular, uses this method along with Langevin thermostat temperature controlling schemes to simulate the target systems.[40] [37] The equations of motion for the Langevin piston method are as follows:

$$\begin{aligned} r' &= p/m + e'r \\ p' &= F - e' - gp + R \\ V' &= 3Ve' \\ e'' &= 3V/W(P - P_0) - g_e e' + R_e/W \\ W &= 3N\tau^2 kT \\ < R^2 > &= 2mgkT/h \\ < R_e^2 > &= 2Wg_e kT/h \end{aligned} \tag{2.31}$$

where, W denotes the mass of the piston. R, being noise of atoms. τ is the oscillation period and R_e is noise of the piston. By specifying the pressure, oscillation, piston decay times and piston temperature, one can perform the NPT simulations by these schemes.

2.6 Constrained Dynamics

In MD simulations, Constrained Dynamics comes into the picture when we want to constrain some of the lesser significant dynamical process in order to get a larger benefit out of it[41]. The integration timestep for any simulations depends upon the fastest motion in the system. If we consider the biomolecular system, the vibration of the sidechain hydrogen–heavy atom bonds (X–H bond) show the fastest motion of the system. Usually these motions are assumed to be less significant in protein dynamics. So the X–H bond vibrations can be constrained to improve the speed of the integration process of the simulation. There are many algorithms like "SHAKE" algorithm[42] and "SETTLE" and "RATTLE" algorithm are available for constrained dynamics. In the "SHAKE" algorithms , the length of the hydrogen bonds are assumed to be constant and the unconstrained equation of motions are solved first. The "RATTLE" algorithm is a velocity version of the "SHAKE" algorithm. The "SETTLE" algorithm constrains the bonds in water in water molecules in the rigid water models. Both the algorithms are used by NAMD to perform the constrained dynamics which allows to extend the integration timesteps to 2 fs, else it would be limited to 1 fs for unconstrained dynamics.

2.7 Analysis of trajectories

After obtaining the trajectories with desired number of time steps from the MD simulations the next step is it confirm whether the simulation is fully equilibrated or not. Usually the analysis leads with the root mean square deviation (RMSD), root mean square fluctuation (RMSF) like calculations.

Root Mean Square Deviation (RMSD):
The deviation of a structure with respect to a particular conformation (or initial structure) is measured by RMSD. Here the average is taken over particles, with i, denoting the particle index

$$RMSD = \sqrt{\left(\frac{\sum_{i=1}^{N}(r_i(t_1) - r_i(t_2))^2}{N}\right)} \tag{2.32}$$

where, N is total number of atoms whose positions are being compared. $r_i(t_1)$ is the initial or reference position of particle i, $r_i(t_2)$ is the updated positions of particle i.

RMSD can be calculated in any segments of a protein system like the whole protein, protein backbone, C_α or any particular chain of the system.

Root Mean Square Fluctuation (RMSF):
RMSF is a measure of the particle-wise average deviation or fluctuation over a trajectory with respect to a reference, typically taken to be the average position of the particle *i.e.* $r_i^{ref} = \bar{r}_i$,

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T} |r_i(t) - r_i^{ref}|^2} \tag{2.33}$$

where T is total time frame of calculation. RMSF is averaged over time giving a particular value for each particle i.

Radius of Gyration (R_g):
It is the measure for the compactness of a structure.

$$R_g = \left(\frac{\sum_i |r_i|^2 . m_i}{\sum_i m_i} \right)^2 \tag{2.34}$$

where m_i is the mass of atom i and r_i is the position of atom i with respect to center of mass of the molecule.

2.8 Correlation network analysis

Correlation network analysis is a method to gain physical insights through an investigation of long-ranged correlations of complex systems. As a result, cross-correlation between each atoms/nodes with remaining others present in the system are assigned with numeric weights (between 0 to 1, some methods assign values -1 to +1, -1 indicating perfect anti-correlation), so that one correlation value C_{ij} (where i and j are the nodes or atoms or residues), represents the correlation of a node-pair. In this manner, a correlation matrix (C_{ij}) is created with all the residues or node wise cross-correlation values of the whole system. [43] [44] Once the correlation matrix is developed, then several other advanced techniques such as community network analysis, optimal and sub-optimal path analysis or principal component analysis can also be performed to gain deeper insights. [45] [46]

2.8.1 Correlation coefficient calculations

To characterize the interdependence among residues or atoms, one can quantify the correlation between pairs of components(i, j). Pearson's like approach [47] [48] is one of the conventional ways to determine correlation coefficient. The resulting correlation coefficient calculated by this method is also known as Pearson coefficient.

$$C_{ij} = \frac{\langle \Delta r_i(t) \cdot \Delta r_j(t) \rangle}{(\langle \Delta r_i(t)^2 \cdot \Delta r_j(t)^2 \rangle)^{1/2}} \qquad (2.35)$$
$$\Delta r_i(t) = r_i(t) - \langle r_i(t) \rangle$$

Where, i and j are indices corresponding to individual nodes or atoms, $r_i(t)$ is the location of node i at time t, and C_{ij} is the element of the correlation matrix at position (i, j). N is the total number of nodes present in the system. The absolute value of C_{ij} is larger when the motion of two nodes are highly correlated or anti-correlated. [44] [49]

But the disadvantage with Pearson like approach is that the coefficient suffers from two flaws; its inability to detect nonlinear correlations and its unwanted dependency on the relative orientation of the fluctuations. A more advanced technique, called the Linear Mutual information (LMI) [44] [50] approach avoids the second pitfall, in that it has no unwanted dependency on the relative orientation of the fluctuation. This LMI calculation returns a matrix of all atom-wise linear mutual information whose elements are denoted as C_{ij}. If $C_{ij} = 1$, the fluctuations are completely correlated and if $C_{ij} = 0$, the fluctuation of atom i and j are not correlated at all.

$$I_{lin}(x_i, x_j) = \frac{1}{2}\left[\ln \det CM_{(i)} + \ln \det CM_{(j)} - \ln \det CM_{(ij)}\right] \qquad (2.36)$$

Where, $CM_{(i)}$ is covariance matrix for displacement of node or atom i, $CM_{(j)}$ is covariance matrix for displacement of node or atom j and $CM_{(ij)}$ is the covariance matrix for atoms i and j respectively.

Correlation network construction

Network analysis of correlated motions can be employed to a correlated system to identify the segments with coupled dynamics. By applying the network analysis, we can construct a weighted network graph where each node represents an individual residue and the weight of the connection between nodes, i and j, represents

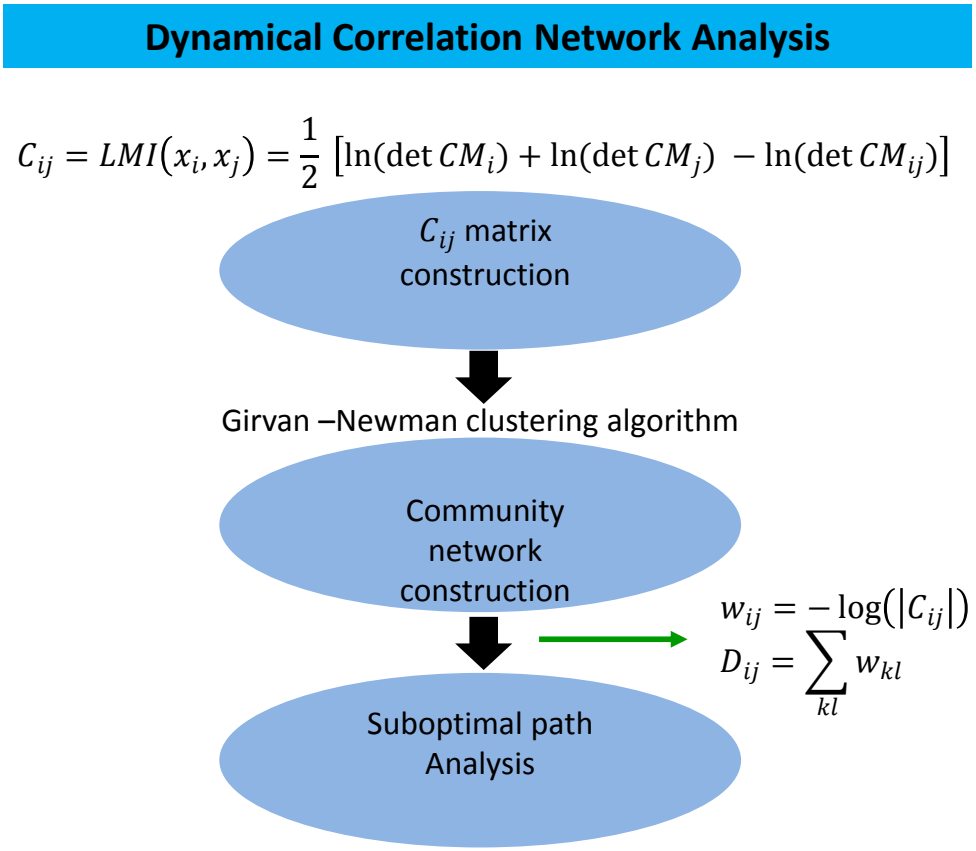

Figure 2.4: A schematic representation of step-wise Dynamical cross correlation analysis algorithm.

their respective cross correlation value, C_{ij}. The procedure of network construction widely follows the approach that is related to that introduced by *Sethi et al*[51], using multiple correlation matrices derived from input MD trajectories. The edges are then added for atom-pairs with $C_{ij} \geq C_0$, where C_0 (given cutoff C_{ij} value) is a constant. In addition, edges are added for atoms where $C_{ij} \geq C_0$ for atleast one of the structures and the $C_\alpha - C_\alpha$ distance, d_{ij}, satisfies $d_{ij} \leq 10$ Å for at least 75% of all conformations. Edge weights are then calculated with $-\log(< C_{ij} >)$, where $< \ >$ denotes the ensemble average. Girvan and Newmann betweenness clustering is then performed to generated aggregate nodal clusters or communities, that are highly intra-connected but loosely inter-connected.[52] [53]

The Girvan and Newman clustering method is based on the calculation of edge

betweenness of an edge. Edge betweenness of an edge is defined as the number of shortest paths between pairs of vertices that run along the edge. If there are more than one shortest path between a pair of vertices, each path is given equal weight so that total weight of all the paths is unity. The algorithm of this clustering method is as follows.:
step(1): calculate betweenness of all edges in the network
step(2): remove the edge with highest betweenness
step(3): recalculate the betweenness affected by the removal
step(4): repeat from step(2) until no edge remains.

2.8.2 Network path Analysis

We can perform optimal or sub-optimal path analysis between two nodes, treated as "source" and "sink" respectively. A fair amount of numbers of paths are sent from "source" to "sink". Tracking these paths and the intermediate nodes those lie on the paths can be useful to gain deep insights into the system dynamics and the mechanisms. Comparative path length distributions calculated from the path analysis, indicate the strength of the correlated motions under distinct conditions. In addition, normalized node degeneracy,that is, the fraction of the number of paths going through each node can also be calculated. Nodes or residues with high node degeneracy in any network are specified as "on-path" residues or nodes. [54] [55]

The interdependence among the nodes is represented as a connecting edge with an associated numeric value that reflects the strength. The absolute value of C_{ij} is larger when the motions of two nodes are highly correlated or anti-correlated. In order to compute signaling pathways, it is useful to construct a matrix where the opposite is true *i.e.* where small values indicate highly correlated or anti-correlated motions. Consequently, the correlation matrix is functionalized according to given equation,

$$\begin{aligned} w_{ij} &= \text{-log}(\|C_{ij}\|) \\ D_{ij} &= \sum_{kl} w_{kl} \end{aligned} \tag{2.37}$$

where, D_{ij} is the cumulative sum of the involved path-lengths between two distant nodes. As a point of clarification, each w_{ij} can be thought of as a "distance" in funtionalized correlation space.

2.9 Principal Component Analysis

Dimensionality reduction can be a very useful step for processing and visualizing high dimensional data-sets while retaining as much of the variance as possible. Principal Component Analysis (PCA) reduces dimension of data and often used to measure data in terms of it's principal component(orthogonal eigen vectors) [56] [57] [58].

From MD simulations one usually collects a large number of Boltzmann's ensemble of configurations of the simulated protein system from which the characteristic features of the system can be extracted. However, the trajectory derived from the MD simulation contains a large amount of noise evolved from the blend of short range atomic fluctuation and vital informations on large rearrangements. These inherent complications raised in MD simulations, makes it difficult to uncover the motions of interest or functional mechanisms. To get rid of such complexities, one option is to cluster conformations to find highly sampled regions in the conformational space, or else, one can apply Principal Component Analysis (PCA) to filter the main mode of collective motion from local noise. The motivation of employing PCA is to change the orthonormal basis of a complex trajectory so that it can be reduced to a lower-dimensional collection of the functional motions. Essential dynamics, a study of dynamics in the lower-dimensional sub-space, is designed by such enhanced sampling algorithms to find and analyze the conformational sub-spaces out of a raw and large MD trajectory [59] [60].

The schematic diagram of the PCA algorithm is given above in Figure 2.5. A ensemble of conformations from a MD trajectory is necessary to start the principal component analysis. All the conformations are to be superposed to a common reference structure by applying a least square fit. Then by using the fitted ensemble a Cartesian variance-covariance matrix (C) of positional fluctuations to be created in which each elements of the matrix represent the covariance between a pair of atoms by,

$$Cov(X,Y) = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{n} \tag{2.38}$$

The diagonal elements of the matrix represents the variance of each atom displacements where the off-diagonal elements represent the co-variance of the atomic fluctuation of each atom pair relative to their respective averages. Though PCA can be employed to any sub set of atoms, in protein dynamics the C_α atoms of each residues are taken into account for PCA calculations. Once the variance-covariance matrix is constructed, it is diagonalized to get a set of eigenvalues and their corresponding

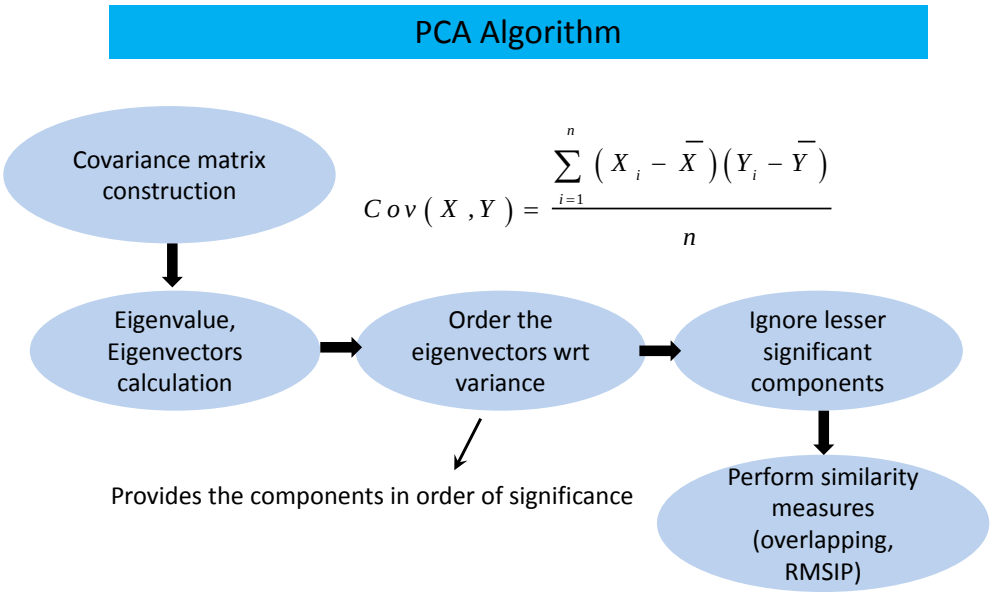

Figure 2.5: A schematic representation of step-wise Principal Component Analysis algorithm.

eigenvectors defining the principal components (PCs). If a protein system with N C_α atoms is taken into account, the matrix C is a $3N \times 3N$ matrix. After diagonalization, $3N-6$ eigenvectors with non-zero eigenvalues can be obtained. The eigenvalues corresponding to six eigenvectors describing the rotational and translational modes of the three axis of Cartesian spaces are zero. These 3N-6 eigenvectors indicate the directions of the collective modes in Cartesian space, with corresponding eigenvalues indicating the contributions of each components to the total fluctuation. Then the eigenvectors or principal components are sorted according to the decreasing order of variance so that PCs with higher variance can be referred as significant principal components and thus the less significant PCs are to be ignored to decrease the dimensionality of the system. [61] [62]

Bibliography

[1] M. Allen and D. Tildesley, Computer simulation of liquids 1987 (clarendon.

[2] M. Allen and D. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, 1989.

[3] A. Rahman, Physical Review **136**, A405 (1964).

[4] J.-P. Ryckaert and A. Bellemans, Chemical Physics Letters **30**, 123 (1975).

[5] P. Vashishta and A. Rahman, Physical Review Letters **40**, 1337 (1978).

[6] J. A. McCammon, B. R. Gelin and M. Karplus, Nature **267**, 585 (1977).

[7] W. Wood and F. Parker, The Journal of Chemical Physics **27**, 720 (1957).

[8] G. M. Torrie and J. P. Valleau, Journal of Computational Physics **23**, 187 (1977).

[9] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, The journal of chemical physics **21**, 1087 (1953).

[10] R. Car and M. Parrinello, Physical review letters **55**, 2471 (1985).

[11] P. Hohenberg and W. Kohn, Kohn W, Sham LJ (1965) Phys Rev **140**, A1133 (1964).

[12] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).

[13] X. Wu and S. Wang, The Journal of chemical physics **110**, 9401 (1999).

[14] D. Frenkel and B. Smit, Academic, San Diego .

[15] C. Kittel and D. F. Holcomb, American Journal of Physics **35**, 547 (1967).

[16] W. Kohn and L. J. Sham, Physical review **140**, A1133 (1965).

[17] D. Marx and J. Hutter, *Ab initio molecular dynamics: basic theory and advanced methods*, Cambridge University Press, 2009.

[18] V. A. Voelz, G. R. Bowman, K. Beauchamp and V. S. Pande, Journal of the American Chemical Society **132**, 1526 (2010).

[19] J. B. Klauda et al., The journal of physical chemistry B **114**, 7830 (2010).

[20] B. R. Brooks et al., Journal of computational chemistry **30**, 1545 (2009).

[21] R. Salomon-Ferrer, D. A. Case and R. C. Walker, Wiley Interdisciplinary Reviews: Computational Molecular Science **3**, 198 (2013).

[22] B. Hess, C. Kutzner, D. Van Der Spoel and E. Lindahl, Journal of chemical theory and computation **4**, 435 (2008).

[23] M. T. Nelson et al., The International Journal of Supercomputer Applications and High Performance Computing **10**, 251 (1996).

[24] B. J. Berne and J. E. Straub, Current Opinion in Structural Biology **7**, 181 (1997).

[25] C. D. Putnam, M. Hammel, G. L. Hura and J. A. Tainer, Quarterly reviews of biophysics **40**, 191 (2007).

[26] L. N. Johnson and D. Phillips, Nature **206**, 761 (1965).

[27] F. A. Mulder, A. Mittermaier, B. Hon, F. W. Dahlquist and L. E. Kay, Nature structural & molecular biology **8**, 932 (2001).
[28] M. Akke, Current opinion in structural biology **12**, 642 (2002).
[29] J. C. Phillips et al., Journal of computational chemistry **26**, 1781 (2005).
[30] L. Kalé et al., Journal of Computational Physics **151**, 283 (1999).
[31] A. D. Mackerell, Journal of computational chemistry **25**, 1584 (2004).
[32] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*, Cambridge unoversity press, 2007.
[33] R. L. Andrew, 2nd, editor.: Pearson Education Limited (2001).
[34] S. Nosé, The Journal of chemical physics **81**, 511 (1984).
[35] W. G. Hoover, Physical review A **31**, 1695 (1985).
[36] H. C. Andersen, The Journal of chemical physics **72**, 2384 (1980).
[37] G. J. Martyna, D. J. Tobias and M. L. Klein, The Journal of Chemical Physics **101**, 4177 (1994).
[38] D. Adams, Molecular Physics **29**, 307 (1975).
[39] H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola and J. Haak, The Journal of chemical physics **81**, 3684 (1984).
[40] S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, The Journal of chemical physics **103**, 4613 (1995).
[41] J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, Journal of Computational Physics **23**, 327 (1977).
[42] H. C. Andersen, Journal of Computational Physics **52**, 24 (1983).
[43] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon and L. S. Caves, Bioinformatics **22**, 2695 (2006).
[44] G. Scarabelli and B. J. Grant, Biophysical journal **107**, 2204 (2014).
[45] B. L. Kormos, A. M. Baranger and D. L. Beveridge, Journal of structural biology **157**, 500 (2007).
[46] B. L. Kormos, A. M. Baranger and D. L. Beveridge, Journal of the American Chemical Society **128**, 8992 (2006).
[47] T. Ichiye and M. Karplus, Proteins: Structure, Function, and Bioinformatics **11**, 205 (1991).
[48] P. Hünenberger, A. Mark and W. Van Gunsteren, Journal of molecular biology **252**, 492 (1995).
[49] I. Rivalta et al., Proceedings of the National Academy of Sciences **109**, E1428 (2012).
[50] O. F. Lange and H. Grubmüller, Proteins: Structure, Function, and Bioinformatics **62**, 1053 (2006).
[51] A. Sethi, J. Eargle, A. A. Black and Z. Luthey-Schulten, Proceedings of the National Acadamy of Sciences **106**, 6620 (2009).

[52] M. Girvan and M. E. Newman, Proc. Natl. Acad. Sci. USA **99**, 8271 (2001).
[53] M. E. Newman, Proceedings of the national academy of sciences **103**, 8577 (2006).
[54] A. T. Van Wart, J. Durrant, L. Votapka and R. E. Amaro, Journal of chemical theory and computation **10**, 511 (2014).
[55] A. Ghosh and S. Vishveshwara, Proceedings of the National Academy of Sciences **104**, 15711 (2007).
[56] K. Pearson, Philosophical Magazine **2**, 559 (1901).
[57] H. Hoteling, Journal of Educational Psychology **24**, 417 (1933).
[58] I. Jollife, (2002).
[59] S. d. G. B. Hayward, *Normal modes and essential dynamics*, Springer Publications, 2008.
[60] H. Abdi and L. J. Williams, Wiley Interdisciplinary Reviews: Computational Statistics **2**, 433 (2010).
[61] J. C. Liao et al., Proceedings of the National Academy of Sciences **100**, 15522 (2003).
[62] L. Skjaerven, A. Martinez and N. Reuter, Proteins: Structure, Function, and Bioinformatics **79**, 232 (2011).

Chapter 3

Structural rigidity and dynamics of SAMHD1 tetramer

3.1 Introduction

In the first study, we explored the effects of nucleotide depletion on the structural rigidity and dynamics of the tetrameric system.[1] [2] [3] [4] [5] A series of simulations were designed to probe the effect of the presence or absence of allosteric site bound GTP/dATP on the overall structure. Our study indicates that the Allosite 2 bound dATP is crucial to the stability to the tetramer.[6] [7] [8] [9] [10] [11] Although the time scale of the simulations were not sufficient to observe the dissociation of the tetramers in the absence of one or more of the allosteric site bound nucleotides, the study revealed large fluctuations in the surrounding residues in the absence of Allosite 2 dATPs.[12] [13] [14] [15] [7] [10]The simulations suggest that the dATP bound to Allosite 2 leads to an overall firming up of the protein, a long range effect that is perceptible even at the catlytic sites.

3.2 Methodology

3.2.1 Simulation System Setup

The crystal structure corresponding to the Protein Data Bank entry 4TNR[10] was used to generate the starting conformation for all atom MD simulations in an explicit water environment. The crystal structure 4TNR shows a tetrameric SAMHD1 with allosteric site 1 occupied by GTP and the allosteric site 2 occupied by dATP. Three of four catalytic sites are occupied by dATP while the fourth (in subunit A)

is vacant. The pdb structure was used to generate initial structures for four systems with allosite and catsite nucleotides selectively removed are listed in Table.1. In all four systems, the crystallographic waters were retained. The Mg^{+2} ions coordinated by allosteric site molecules and the catsite molecules were present in system 1 (wild type or *wt*) . However, the system 2 and 3 contained only the allosteric site 1 and allosteric site 2 molecules respectively and will be referred to as GTP-lost form and dATP-lost form respectively. The catsite ligand (dATP) is absent in both cases. In system 4 (the Catsite vacant form), only the catsite molecule is absent while both the allosteric sites are occupied. The missing residues in the loop 278-283 were inserted in the protein structures, where as the missing N and the missing C terminal residues were ignored. The four R206 and N207 residues were mutated back to histidine and aspartate in accord with the sequence of the *wt* SAMHD1 (Uniprot Q9Y3Z3-1). Disulfide bonds were introduced between residues 341 and 350. Each System was immersed in $\sim$ 59000 pre-equilibrated TIP3 water molecules. Na^+ and Cl^- ions were added at random positions to bring the net charge of the system to zero. Each system consists of $\sim 210,000$ atoms measured 13 nm $\times$ 12 nm $\times$14 nm.

3.2.2 General MD Methods

Molecular dynamics simulations were performed using NAMD 2.9 [16] [17]. All simulations employed periodic boundary conditions and multiple time stepping wherein local interactions were calculated every 2 femtosecond (fs) and full electrostatic eval-

System	GTP at Allosite-1	dATP at Allosite-2	dATP at catalytic site	NVT simlation-Set–1 length (ns) at 295 K	NVT simlation-Set–2 length (ns) at 295 K	NVT simlation length (ns) at 500 K
1	Present	Present	Present(3 out of 4 Catsites)	210	100	20
2	Absent	Present	Absent	210	100	20
3	Present	Absent	Absent	210	100	20
4	Present	Present	Absent	210	100	20

Table 3.1: List of MD simulations performed for all the systems created depending upon the presence and absence of the nucleotides in the SAMHD1 tetrameric complex.

uations were performed every two time steps. The Particle mesh Ewald [18] method was employed for long range electrostatic calculations. CHARMM31 [19] [20] force fields were employed along with the TIP3P water model. Covalent bonds involving hydrogen in water and other molecules were kept rigid using SETTLE [21] and RATTLE [22] algorithms. The cutoff and switching distance used for Van der waals and short ranged interactions were set to 12 Å and 10 Å respectively. Each system was minimized for 3000 steps using the conjugate gradient method and then equilibrated in the NPT ensemble using the Nosé-Hoover Langevin piston pressure control [23] at 295 K for at least 5 nanosecond (ns). Following equilibration, all NAMD simulations were performed in the NVT ensemble with the temperature maintained at 295 K using Langevin thermostat. The data was recorded at 10 picosecond (ps) intervals. [24] [25] All four systems were simulated for about 210 ns at 295 K. In addition 20 ns trajectories were also generated at elevated temperature of 500 K for all four systems. So the total simulation time exceeds 1 microsecond (μs).

3.3 Results and Discussions

To understand the role of Allosteric ligands in stabilizing the tetrameric structure of the SAMHD1, a series of all atom molecular dynamics simulations were performed and analyzed different aspects and properties of the trajectories.

3.3.1 Protein stability improved when both Allosteric site are occupied

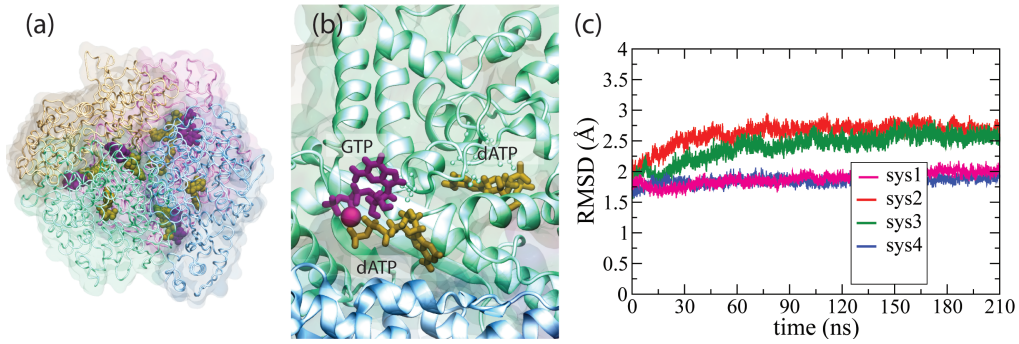

Figure 3.1: (a) Ribbon representation,(b) close view of allosteric site,(c) RMSD of prootein backbone

We first consider the stability of the *wt* SAMHD1. An analysis of the root mean

square deviation(RMSD) of the protein with respect to the crystal structure(4TNR) shows a high degree of stability when both the allosteric site are occupied (system 1 and 4) as shown in Figure 3.1(c). In contrast, the presence of the preferred ligand in only one of the allosteric site (system 2 and 3) does not impart the same level of stability to the tetramer as illustrated in Figure 3.1(c), which shows the time variation of the protein backbone RMSD with respect to the crystal structure. In addition the RMSD of the individual chains shown in Figure 3.2 are also plotted in order to identify any chain that shows greater deviations than others. System 1

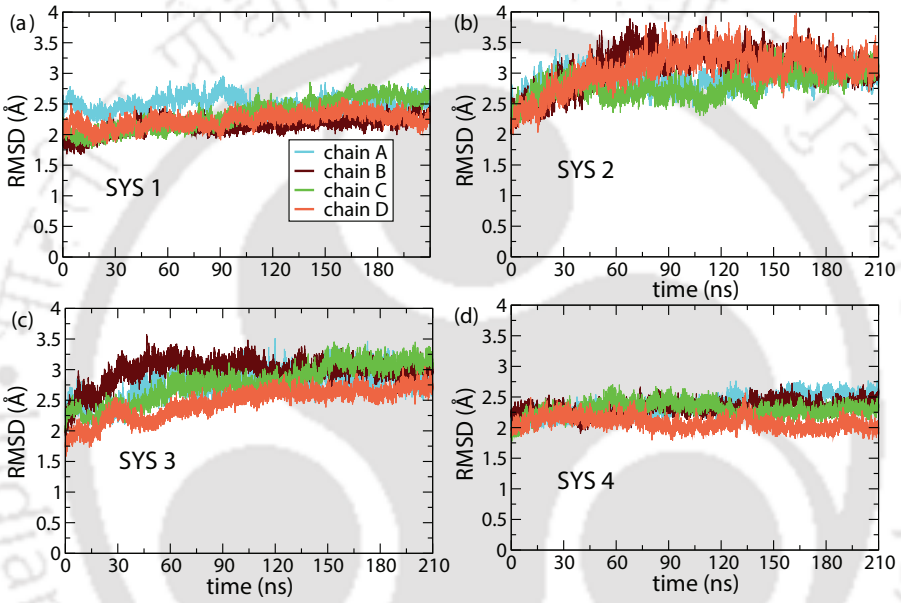

Figure 3.2: RMSD of protein backbone of individual chains in each of the systems (295 K)

and 4 with all allosteric site occupied show the smallest backbone deviation (average below 2 Å) in the 210 ns trajectory. In contrast, systems with only one allosteric site occupied, that is, system 2 (or GTP lost form in allostric site 1) and system 3 (or dATP lost form in Allosteric site 2) display a greater RMSD from the initial experimental structure. The individual chains also show considerable variation in the RMSD within the same time frame for the GTP-lost and dATP-lost form of systems. The RMSD obtained from the high temperature trajectories (see Figure 3.3 c) also showed the system 1 and 4 to have the smallest deviation from the crystal structure . System 3 in which the dATP is lost in Allosite 2, displayed the most rapid loss of stability at 500 K, with the large increase in RMSD ($\sim$ 7 Å) in 20 ns. The remaining systems showed an RMSD variation between 3 Å to 4 Å from the

original structure in the same duration at high temperatures.

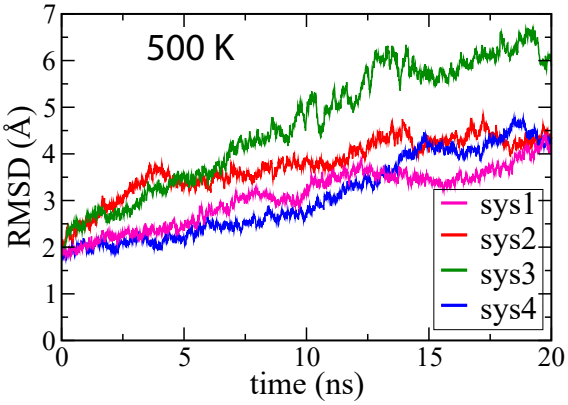

Figure 3.3: RMSD of protein backbone at high temperature (500 K) for the systems investigated.

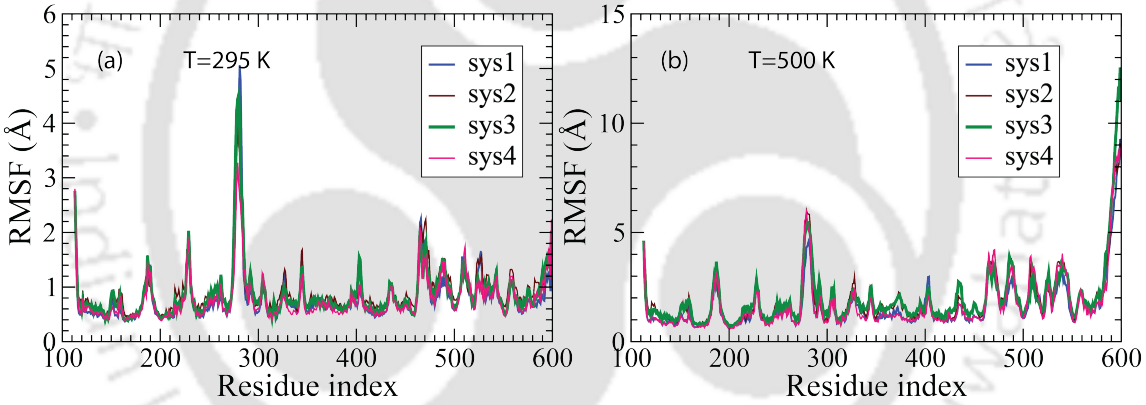

Figure 3.4: RMSF of protein backbone at (a) room temperature (295 K) and (b) at higher temperature (500 K) for the systems investigated.

Figure 3.4 shows a comparison of root mean square fluctuations (RMSF) of the protein backbone atoms of the of the individual systems both at room temperatures (295 K) and higher temperatures (500 K). Note that we performed the high temperature simulation only for 20 ns as a pilot-run to find any hint or interesting lead of the system dynamics that can be followed in real (295 K) simulations. Apart from the strong thermal vibrations in 500 K simulations, where the average RMSF of the systems lies in the range of 3-4 Å, the RMSF of the 295 K simulations shows the similar peaks with lesser average fluctuations. For both the cases, the dATP-lost form of the system (that is, system 3) has sharp peaks indicating comparatively large fluctuations with respect to other systems. The C-terminus domain (CTD)

region that belongs to the residue number 540 to 600 of the systems shows a sharp peak for the 500 K case that goes above 8 Å indicating the fluctuation is stronger compared to the core-HD domain.

3.3.2 Allosteric site 1 is stabilized when dATP is bound to Allosteric site 2

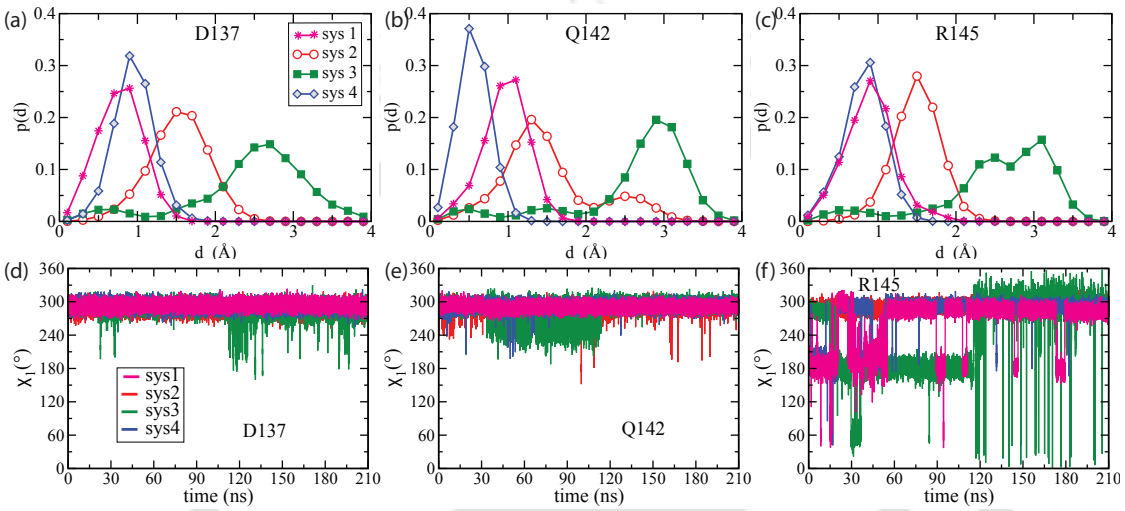

Figure 3.5: (a-c) Distribution of displacements of center of mass of allosteric site 1 residues for different systems investigated, (d-f) and their corresponding χ_1 dihedral variations with time, indicating the rotameric states of the residues.

To further analyze the effect of the ligand on the allosteric sites, we consider the deviation of the residues involved in hydrogen bond formation with the ligand molecules at the two pockets. D137, 142, R145 are the residues that are involved in stabilizing the GTP molecule in Allosite 1. The distribution of the distance of the center of mass of the three residues of chain A from the original position in the crystal structure is plotted in Figure 3.5(a, b, c). In system 1 and 4, where all the allosites were occupied, all three residues show little deviation from the original position with the peak of the distribution in each case below 1 Å . Surprisingly, system 3 with vacant Allosite 2, but with bound GTP in allosite 1 showed a greater deviation in the residue positions compared to system 2 in which Allosite 1 is empty but with Allosite 2 occupied. This is contrary to the expectations since the bound GTP at Allosite 1 is expected to influence the positional fluctuations of the nearby residues. However the average deviation of the D137, Q142, R145 ($\sim$ 3 Å) in system with dATP lost-form was almost twice that observed in system with GTP-lost form

($\sim$ 1.5 Å) indicating that the vacancy of Allosite 2 has a greater destabilizing effect on the allosteric pocket residues. The corresponding figures for the other three chains in Figure 3.6 depict similar behavior with peak of the displacements in all cases lying in the range 1-2 Å from the original position indicating minor fluctuations. System 3 (dATP-lost form) has a broader distribution with noticeable shift of the distribution peak in several residues (chain B D137, chain B R145, chain C Q142).

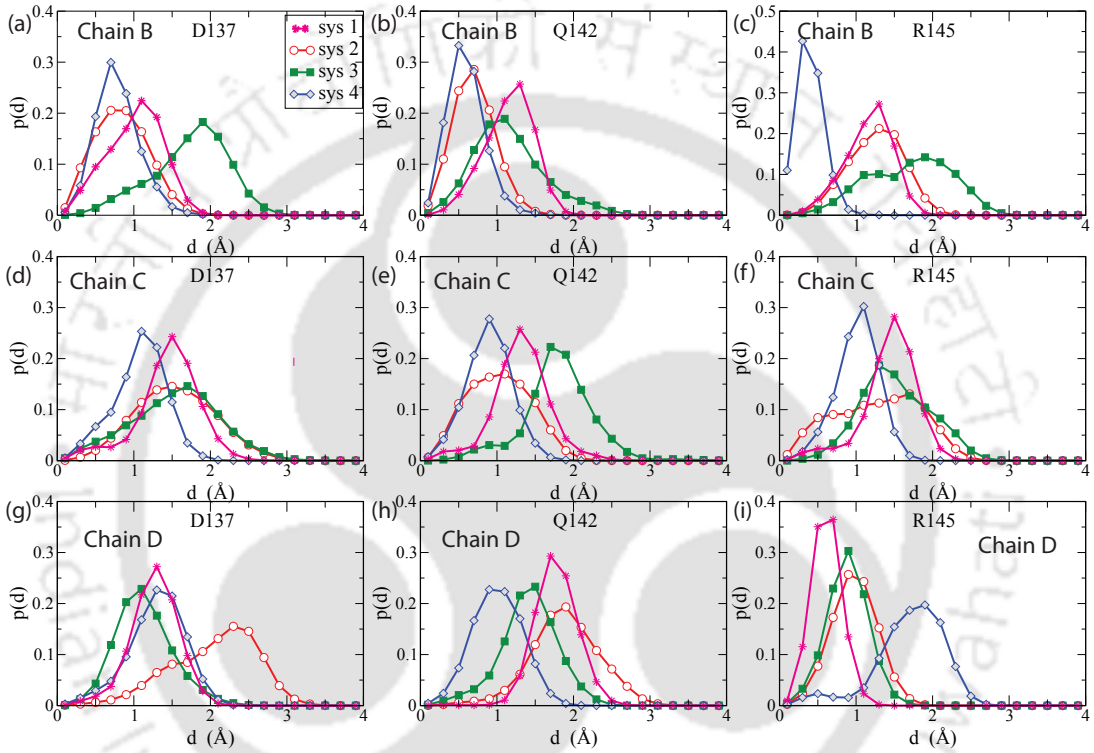

Figure 3.6: Distribution of displacements of center of mass of allosteric residues of chain B, C and D

3.3.3 The R145 sidechain in Allosteric site 1 is highly mobile, even when bound to GTP

To investigate the side chain dynamics, we analyze the χ_1 dihedral for these above three residues in *wt* SAMHD1. Figure 3.5(d, e) show the side chains of D137 and Q142 to be stiff and predominantly in the g^- state. Only the g^- state was observed in the systems with allosteric sites occupied (that is, system 1 and 4). In dATP-lost form (system 3), D137 shows intermittent transitions to the *trans* state. In case of Q142, the GTP and dATP-lost forms show occasional transitions to the *trans*

state. The degree of importance of allosteric communications in this protein can be gauged by the case study of residue R145. The residue is known to spontaneously mutate in AGS patients[26]the conservative mutation R145Q destroys the dTTP hydrolysis rate[13]. The non-conservative mutation R145A completely destroys the protein's ability to tetramerize. Based on this information, as well as published crystal structures[27], we would expect R145 to be relatively immobile when hydrogen bound to GTP. Contrary to expectations , R145 populates several rotomaric states with significant local transformations between g^- and *trans* states in all the cases except the GTP-lost form. The presence of both the Allosteric site ligands do not restrict the side chain fluctuations in R145 as shown in Figure 3.5(f). Transient excursions to the g^+ state was observed in system 3 and 4. The *trans* state of R145 χ_1 is significant in system 3 (43.8%) but total absent in system 2. The g^- state is predominant in system 1 (70%), 2 (100%), 4 (85.4%). Snapshots in Figure 3.7 shows the different orientations sampled by R145 in system 1. The long side-chain of R145 retains the substantial conformation entropy even when bound to GTP. In the dATP-lost form (system 3), the GTP drifts away from the initial positions resulting in enhanced fluctuations of proximal residues at Allosite 1.

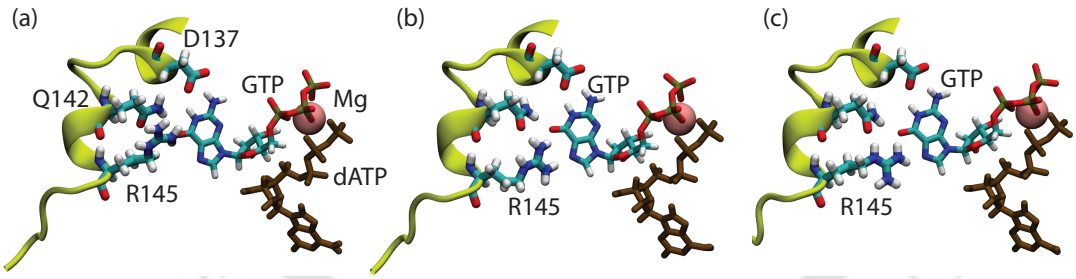

Figure 3.7: Snapshots of different conformational states adopted by R145 with respect to GTP. The protein backbone is shown in yellow ribbon. The residues along with the dATP and GTP molecules are in stick presentations. The Mg^{+2} ion is represented by pink sphere.

3.3.4 The Alloseric site 2 displays more motion when dATP is missing

To investigate how the Allosite 2 structure is affected by the presence/absence of ligands, we analyzed the displacement of the key residues surrounding the ligand, the rotameric state of the side-chain dihedrals and the distance between residues.

Figure 3.8 shows the distribution of the center of mass displacements of the residues N119, R145, V156 and R333 that interact with the dATP molecule at Allosite 2 in the *wt* SAMHD1. Here, we consider only one allosteric pocket (at the interface of chains A, D and B).

Overall, the residues were tightly bound in system 1 and 4 with displacement peaks $\sim$ 1 Å. The most pronounced deviations from the X-Ray structure is observed in dATP-lost form (system 3) in case of residues N119 (chain A) and V156 (chain D). The average displacement of N119 was about 4 Å in dATP-lost form. Large displacements ($>$ 3 Å) of the residues were also observed at the other three allosteric pockets of the dATP-lost form. The corresponding plots for other three Allosite 2 pockets are also provided in Figure 3.9, Figure 3.10 and Figure 3.11 . Each of these distribution of center of mass of the Allosite 2 residues plots show a common pattern that is the most sharp peaks for the residues correspond to the system 1 and system 4 where the nucleotide molecules (dATP and GTP) are occupied in the allosteric sites. In contrast, the dATP-lost form shows wider distribution of the displacement of allosite residues from the initial positions (see Figure 3.8, Figure 3.9, Figure 3.10 and Figure 3.11). The GTP-lost system (system 2) in which dATP molecules are present show comparatively sharper distributions with respect to dATP-lost form. Hence, it indicates that the allosteric site dynamics has a greater dependence on the presence or absence of the nucleotide molecules, specifically, presence of dATP in the active sites of SAMHD1 makes the terameric complex more subtle in the context of stability.

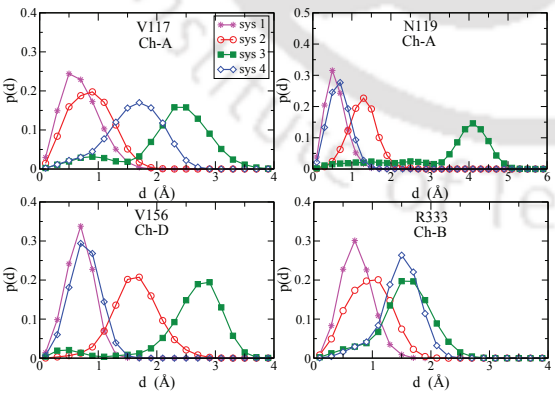

Figure 3.8: Distribution of displacements of centre of mass of Allosite 2 residues V117, N119, V156, R333 at the interface of chains A, B and D.

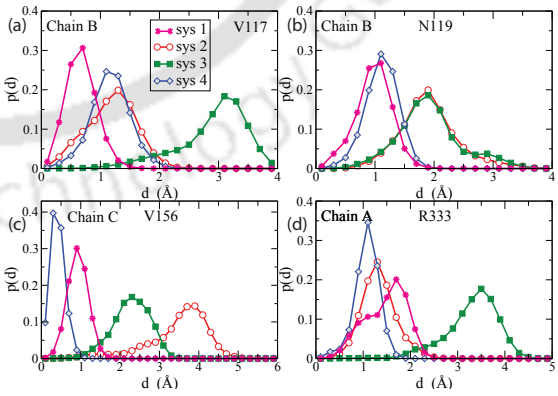

Figure 3.9: Distribution of displacements of centre of mass of Allosite 2 residues V117, N119, V156, R333 at the interface of chains B, A and C.

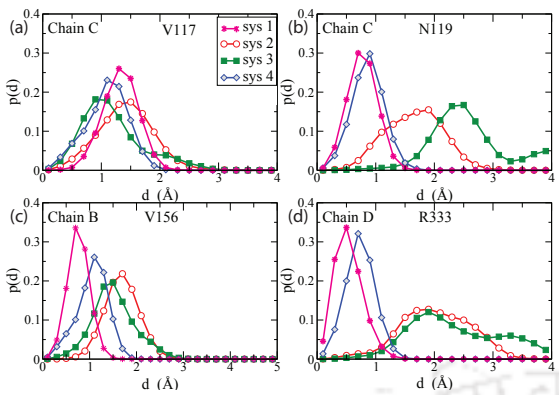

Figure 3.10: Distribution of displacements of centre of mass of Allosite 2 residues V117, N119, V156, R333 at the interface of chains C, D and B.

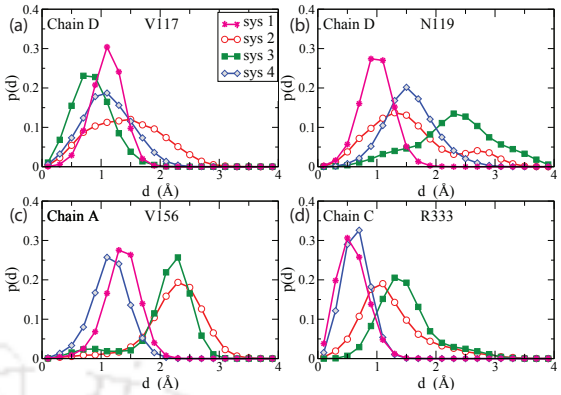

Figure 3.11: Distribution of displacements of centre of mass of Allosite 2 residues V117, N119, V156, R333 at the interface of chains D, C and A.

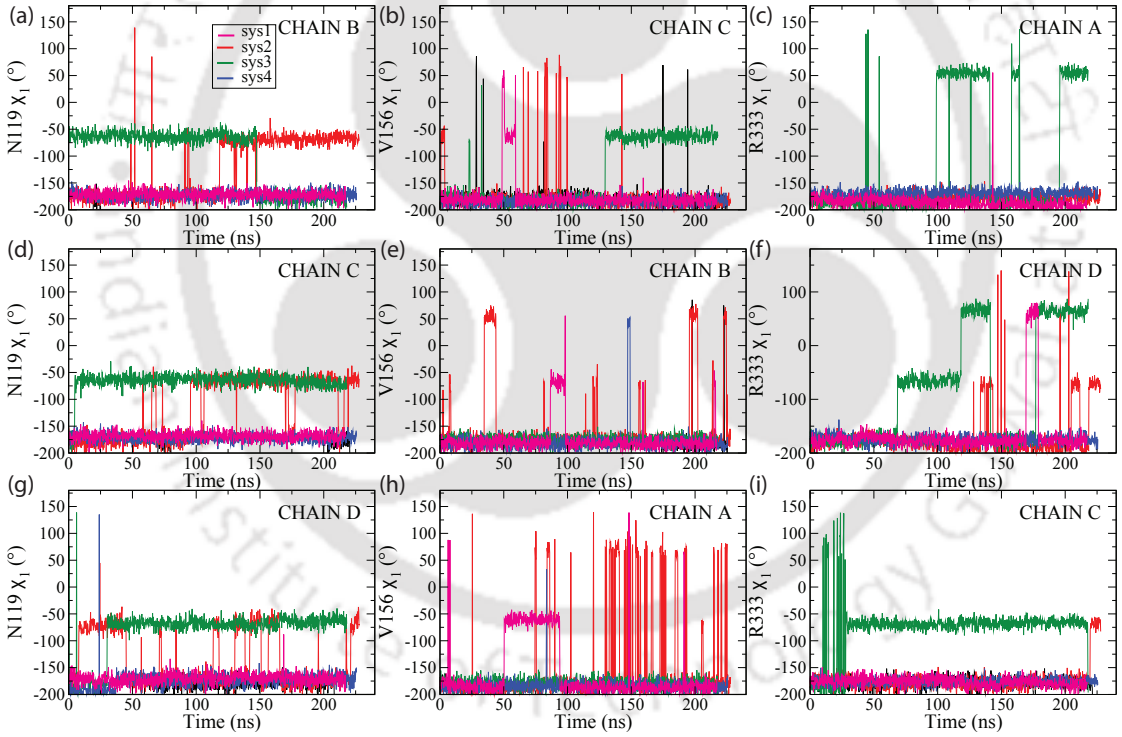

Figure 3.12: Side-chain χ_1 dihedral of residues N119, V156, R333 at Allosite 2 of different Allosteric pockets. Note that the plots in each row correspond to a particular allosteric pocket.

The side-chain χ_1 dihedral of N119, V156 and R333 of different Allosite 2 pockets are provided in Figure 3.12. Note that all the plots also show dihedrals to be stiff in system 1 and 4 (with nucleotides present in allosteric pockets) with a predominant

t state (that is average 180°). The only exception observed was an initial 50 ns excursion of V156 (chain A, see Figure 3.12 h) to g^- ($\sim -56°$) from the average t state. In the absence of dATP (system 3) the g^- and g^+ state are seen to be dominate. Figure 3.12, panels (c, f and i) indicate that the bound dATP pins down R333 to the t state thereby gluing adjacent subunits.

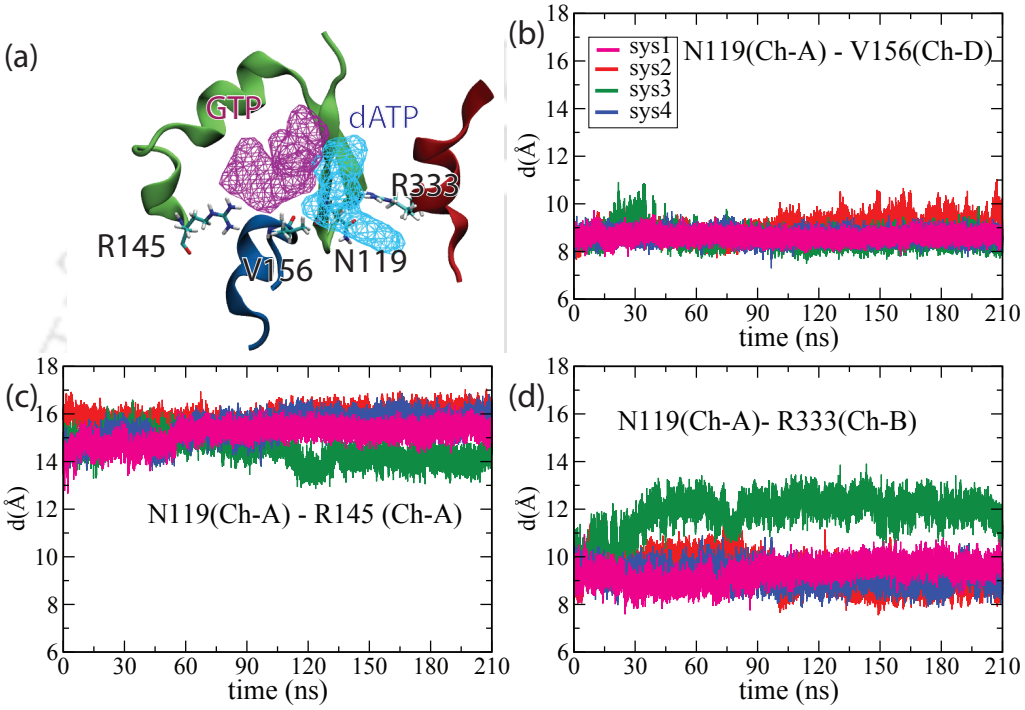

Figure 3.13: (a) The allosteric pocket indicating residues R145, V156, N119 and R333 involved in stabilizing interactions with GTP/dATP. The GTP/dATP molecules are represented by surface plots in violet and blue respectively. (b-d) The variation of the distances between the C_α atoms of N119 and other residues those interact with the GTP/dATP at the Allosite (1 and 2) with time. The Allosteric site lies at the junction of chain A, B and D.

Next we consider the separation between C_α atoms of certain residues (N119, R145, V156 and R333) involved in stabilizing the nucleotides (See Figure 3.13) bound to allosteric sites. The distance between N119 (chain A) and V156 (chain B) is steady (with less than 1 Å deviation from the mean) except for transients in the initial 50 ns in system 3 (dATP-lost form). The N119-R145 distances were found to be stable in all four systems with small fluctuations, although there is a difference between the mean distances in the various systems. The mean N119-R145 distance in dATP-lost form of system (omitting the initial 50 ns where the

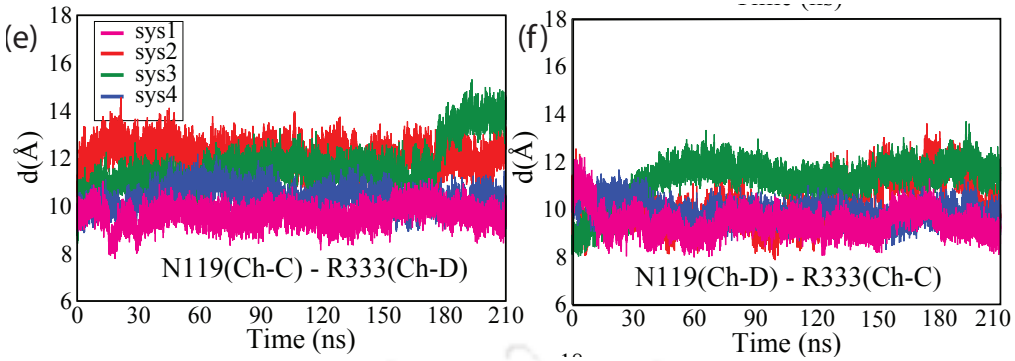

Figure 3.14: The variation of the distances between the C_α atoms of N119 of chain C and R333 of chain D and *vice-versa*.

fluctuation is large) is approximately 2 Å less than the corresponding mean distance in system 1. The difference is observed in all four chains (see Figure 3.14) indicating consistent deformation of the allosteric pocket in dATP-lost form. By far, the most conspicuous fluctuations are observed in the distance between chain A N119 and chain B R333 (see Figure 3.13 d). In the absence of the dATP molecule in Allosite 2, the distance between the residues increased by almost 4 Å in dATP-lost as shown in Figure 3.13(d). In system 1 and 4, the distances are stable with less than 1Å deviation about the mean distance. In the GTP-lost and dATP-lost form systems, the deviation of the mean N119-R333 distances from that of system 1 are found to be large (upto 6 Å) in Figure 3.14(e and f). Thus the structure of the allosteric pocket is substantially influenced by the presence or absence of the ligands. The presence of GTP alone (as in system 3) does not restrain the inter-chain distance in the vicinity of the allosteric sites as seen in the simulations. The N119-R333 distance shows the largest fluctuations and deviations from the initail structure indicating the likelihood of the separation of chains. A disturbance in the N119-R333 separation may be an early indicator of the loss of stability of the tetrameric complex.

3.3.5 Inter-helix (E355-A373) dynamics between- adjacent monomers

In all the tetrameric structures of SAMHD1 available in the protein databank, helices E355-A373 of adjacent monomers are coordinated at the both ends by allosteric site GTP/dATP molecules that interact to promote tetramer formations. Hence a disturbance of the allosteric site environment is expected to affect the helix structure. R372 has a highly mobile side chain that can interact with several proximal residues resulting in a (possibly) fluctuating salt bridge network. The inter-helix interactions

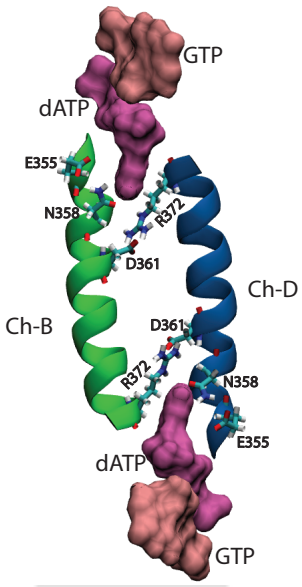

Figure 3.15: A snapshot of helix (E355-A373) of adjacent monomers (chain B in green and chain D in blue) coordinated by allosteric site dATP/GTP at the both ends. The probable inter-helix hydrogen bonding residues are shown in stick representations.

include critical hydrogen bonds between D361-R372, (possibly) E355-R372, and N358-R372 etc. In order to investigate the effect of the allosite occupancy on the feasibility and stability of the hydrogen bonds we calculated the distances between the relevant N and O atoms.

In Figure 3.16, the shortest of the four distances between the two carboxylate oxygen atoms (OD1/OD2) of D361 and the (NH1/NH2) nitrogen atoms of R372 of chain A-C, C-A, D-B and B-D are shown as a function of time. In most systems, the distance is steady (mean distance 2.7 Å of system 4 chains A-C) with occasional flipping of the arginine side-chain leading to an increase of the distance by upto 10 Å. The mean probability of hydrogen bond occurrence between D361 and R372 averaged over the four chains and using the cutoff 3.5 Å for system 1-4 are 0.78, 0.89, 0.88 and 0.76 respectively. The corresponding for hydrogen bond occurrence between N358-R372 is 0.25, 0.20, 0.26 and 0.45 for the four systems respectively (see Figure 3.17).

Therefore the D361-R372 hydrogen bond appears to be the most important with almost threefold greater probability of occurrence compared to the N358-R372 hydrogen bond in system 1. The distance between E355 and R372 is too large for hydrogen formation (see figure 3.18).

Figure 3.17 clearly shows the N358 oxygen (OD1/OD2)–R372 nitrogen (NH1/NH2)

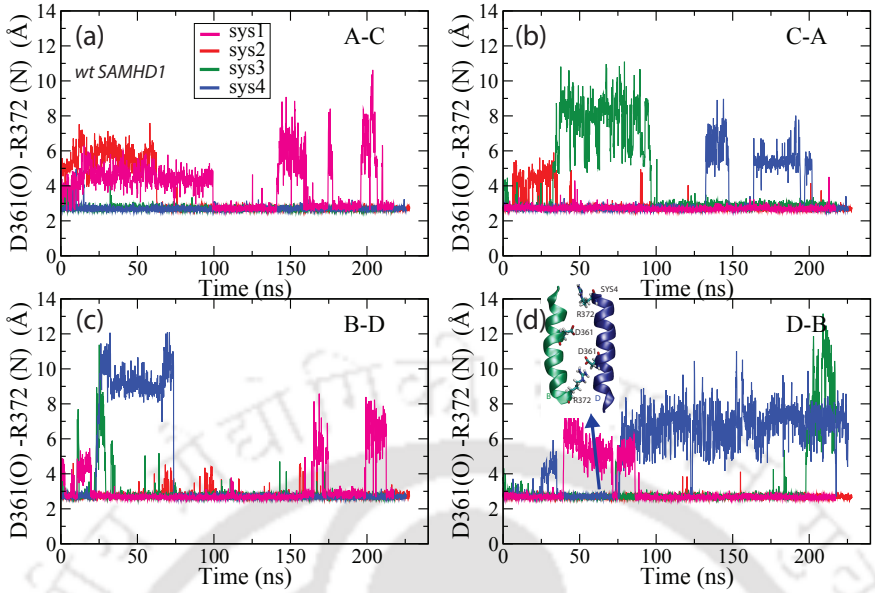

Figure 3.16: Shortest distances between two carboxylate oxygen atoms of D361 (OD1/OD2) and two nitrogen atoms (NH1/NH2) of R372 of chains (a) A and C, (b) C and A, (c) D and B and (d) B and D. The inset in (d) shows the R372 (chain D) side chain flipping away completely from D361 (chain B) in system 4.

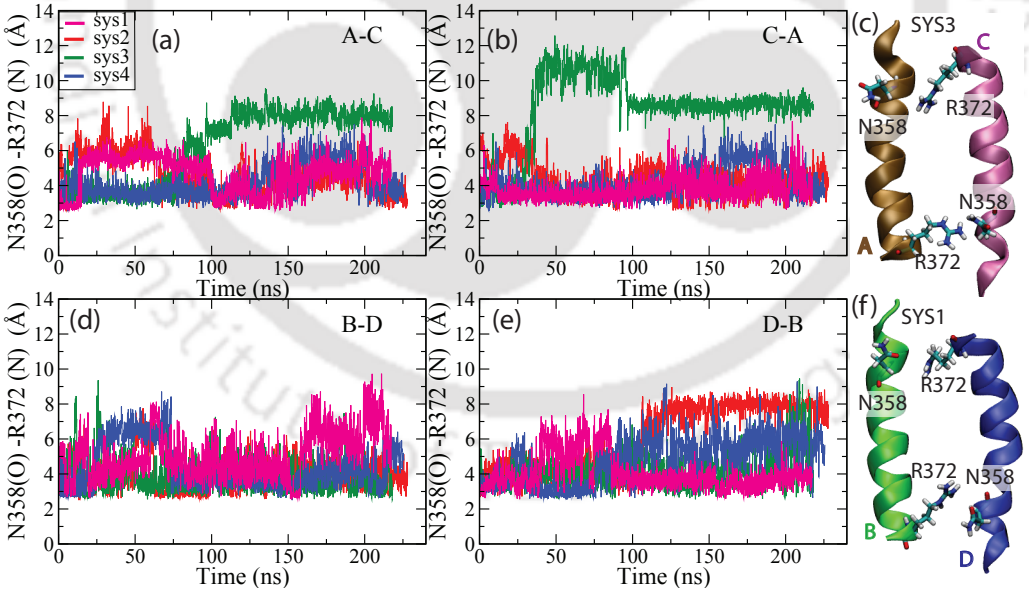

Figure 3.17: Shortest distances between (OD1/OD2) oxygen atoms of N358 and the (NH1/NH2) nitrogen atoms of R372 chains (a) A and C, (b) C and A, (d) D and B and (e) B and D. Panel (c) and (f) represent the snapshots the E355-A373 helices with residues N358 and R372 of adjacent chains A-C and B-D respectively.

distance in dATP-lost form to deviate significantly (upto 10 Å) from other systems.

To inspect the net effect of hydrogen bond formations on the interhelical distance, the distances between the center of mass of D361 and R372 of neighboring chains were measured (see Figure 3.19).

In contrast to Figure 3.16 where large fluctuations of the O-N distances were observed, the separation between the center of mass of the residues were steady with smaller fluctuations. The distance for system 1 (the complete system) between all pair of helices was found to be stable at ~ 8.9 Å in spite of intermittent breaking of hydrogen bonds. All other systems show occasional fluctuations (upto 2 Å) about the reference distance of system 1. In addition the distance measured in dATP-lost form (in Figure 3.19a, b) shows greater fluctuations about the mean (system 1). Overall, the interhelix distance was found to be steady. We find the separation between the center of mass of residues D361 and R372 a better indicator of the overall stability of the helices compared to the individual atom-to-atom distances which display greater fluctuations.

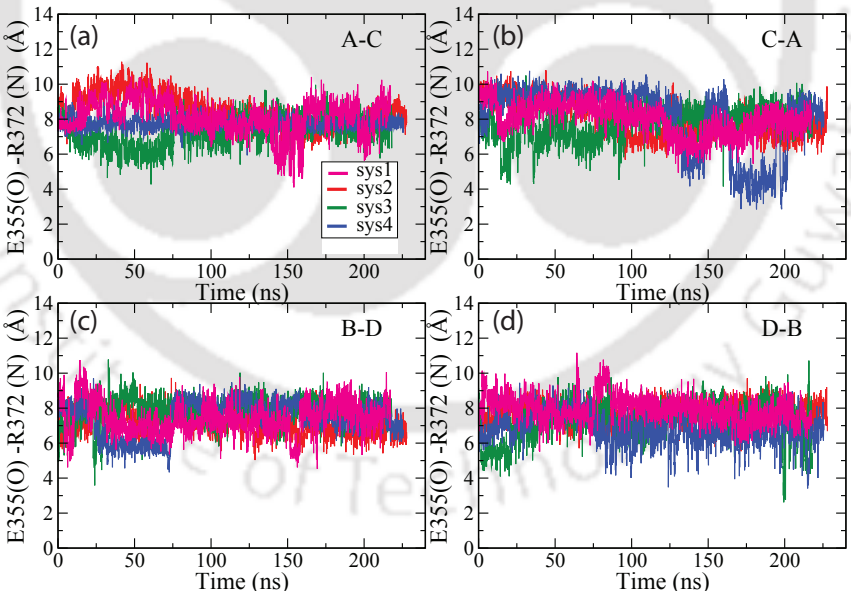

Figure 3.18: Shortest distances between (OD1/OD2) oxygen atoms of E355 and the (NH1/NH2) nitrogen atoms of R372 chains (a) A and C, (b) C and A, (c) D and B and (d) B and D. For all cases the distances are found to be longer than the cutoff H-bond distance.

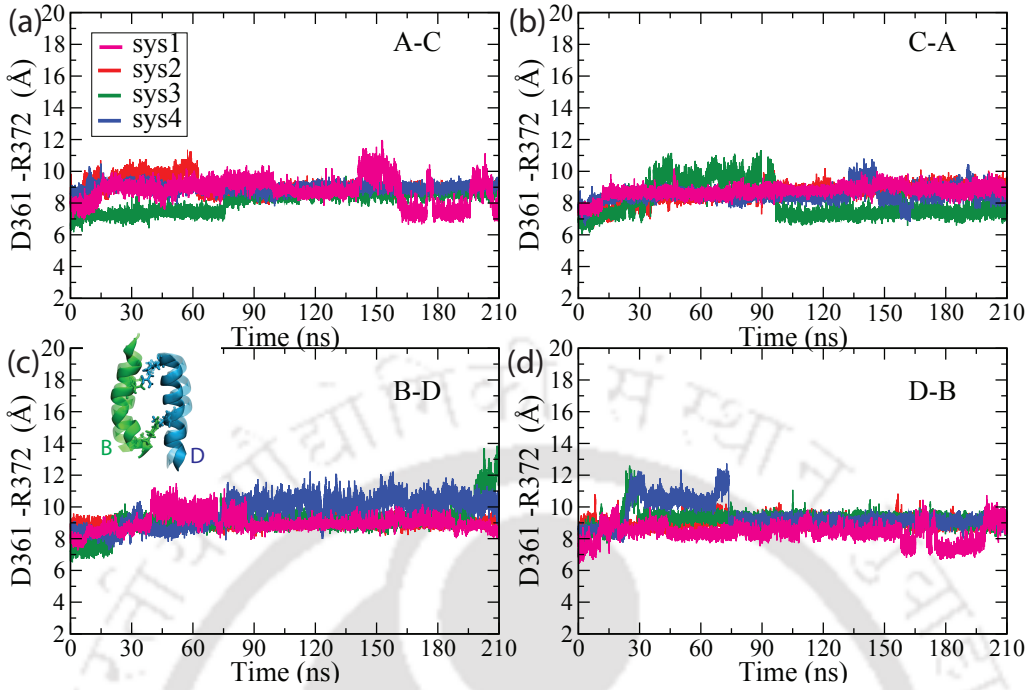

Figure 3.19: The distance between the center of mass of D361 and R372 of adjacent chains (a-b) A-C and C-A, (c-d) B-D and D-B with respect to time.

3.3.6 Increased fluctuations in catalytic site residues when Allosite 2 vacant

Figure 3.20 shows the displacement of the center of mass of the residues surrounding the catalytic site in chain C in the four systems. Note that the catalytic site in chain A in system 1 is vacant. The remaining three catsites are occupied by dATP in system 1 (corresponding plots are provided in the Figure 3.21, 3.22, 3.23). In most cases the distributions are unimodal with the peak within 2 Å of the initial position and small width (standard deviation less than 1 Å). However a few outliers observed in chain C include dATP-lost form (system 3) (H206, Y374, Q375) indicating that the destabilizing effect due to Allosite 2 vacancy extends to the catalytic site as well. The residues of the dATP-lost form exhibit significant deviation at the catsite in all chains compared to the other systems. The residues Y374 and Q375 exhibit large fluctuations in three of the four chains (chain A being the only exception) in dATP-lost system. No significant deviation is observed for the catalytic site residues in system 1 (except D218 in chain A) suggesting that the presence of the catalytic site substrate (in chains B, C and D) "firms" up the structure.

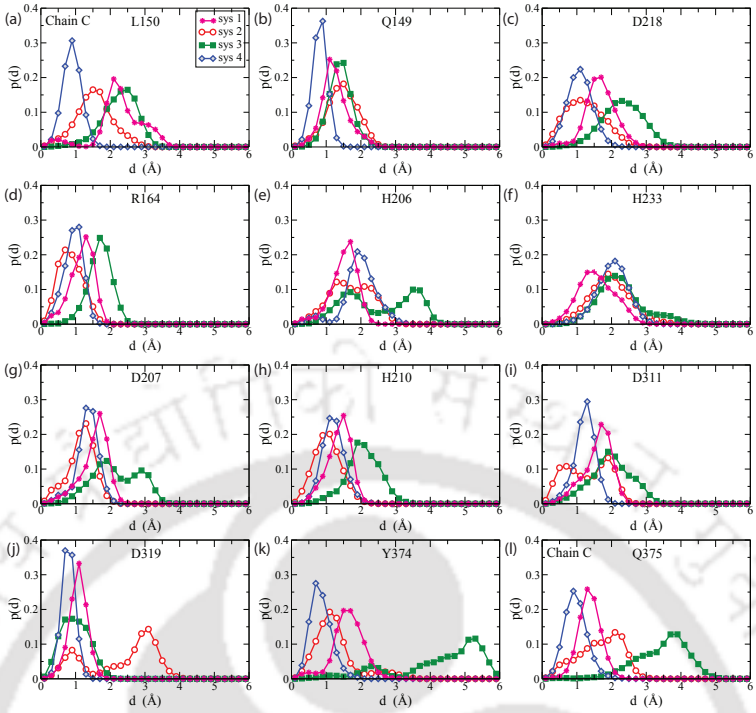

Figure 3.20: Catalytic site residue displacements: The distribution of the displacement of the center of mass of the residues flanking the catalytic site of the enzyme (chain C) from the initial positions in the crystal structure for the four investigated systems.

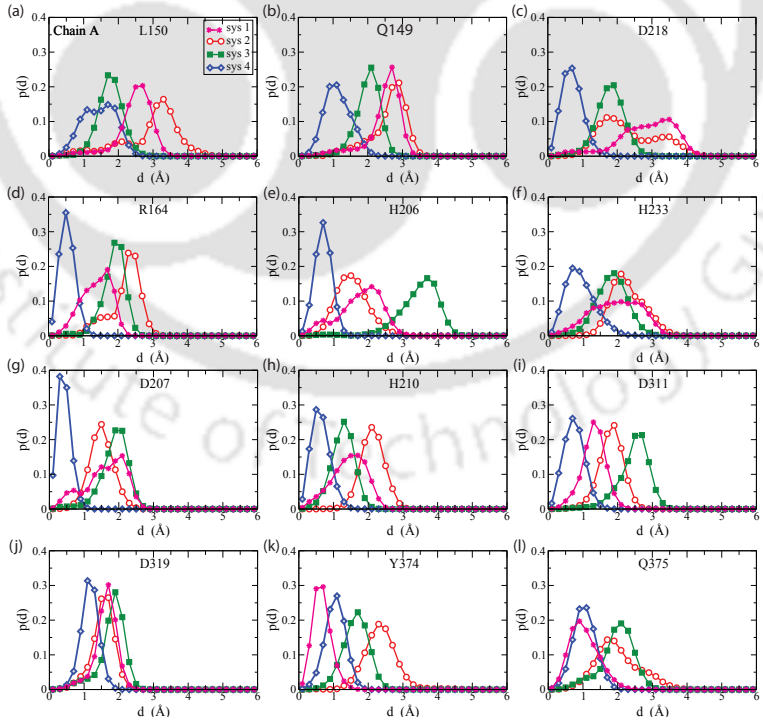

Figure 3.21: The auxiliary figure of Figure 3.20 correspond to chain A.

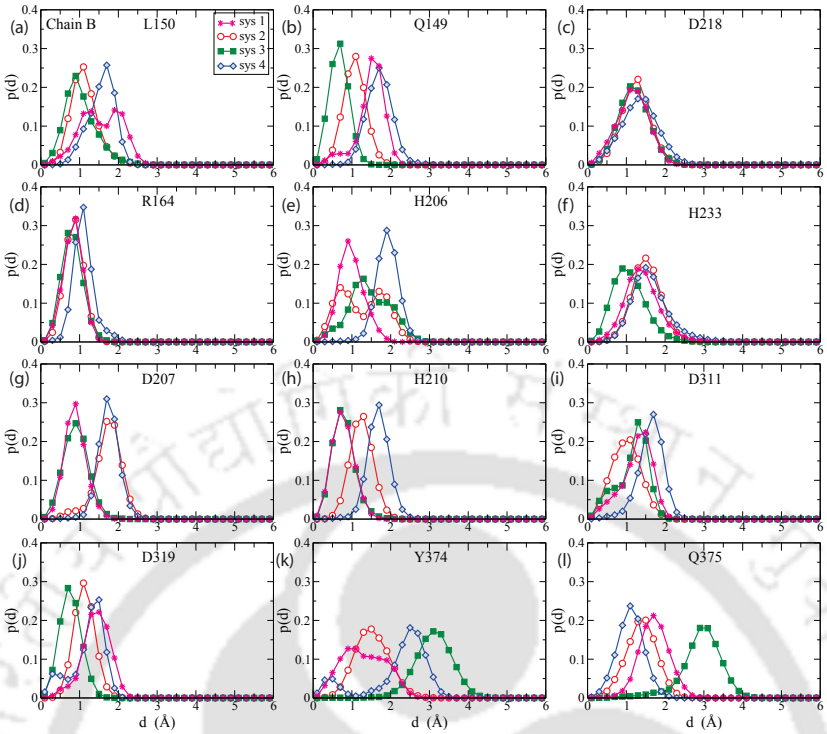

Figure 3.22: The auxiliary figure of Figure 3.20 correspond to chain B.

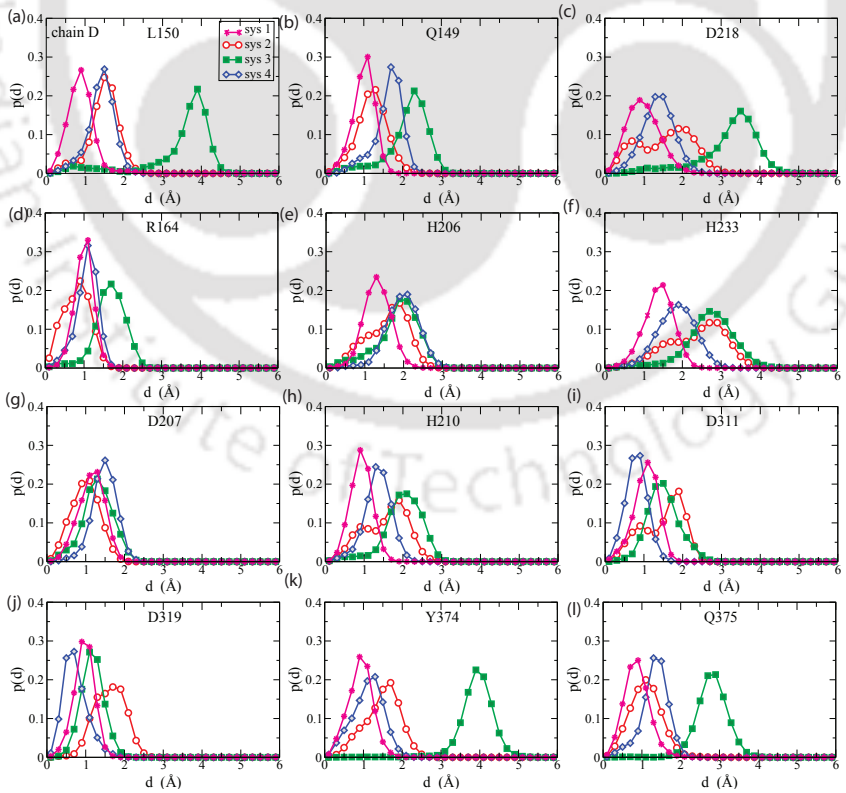

Figure 3.23: The auxiliary figure of Figure 3.20 correspond to chain D.

3.4 Conclusions

In this chapter, we report our first exploratory study where we designed a series of all-atom molecular dynamics simulations to examine the role of the nucleotides bound to the allosteric sites in imparting stability to the tetrameric complex. It is known that SAMHD1 can tetramerize in the presence of GTP and dNTP but not with the GTP alone. In congruence with existing experimental studies, we found that interactions between the dATP and the neighboring protein residues pin the chains together at the allosteric site. MD simulations indicate a "breathing motion" in dATP-lost form of the protein. In the absence of the dATP, the GTP drifts away from the initial positions leading to a increased fluctuations even in the residues flanking Allosite 1 (that is, residues D137, Q142, R145). The dATP, therefore, appears to "lock" the GTP in its place, imparting stability to the tetrameric structure. In addition a vacancy in Allosite 2 leads to large fluctuations of the residues (N119, V117, V156 and R333) directly involved in stabilizing the dATP. Since the proximal residues at Allosite 2 involve three subunits, the absence of dATP has a direct bearing on the stability of the tetramer. In particular, R333-N119 distance (the two residues belonging to different subunits) is found to be highly sensitive to the presence of the Allosite nucleotides with large fluctuations indicating a possible mode of disintegration of the complex. The removal of dATP has a greater detrimental effect on the stability of the complex than the removal of GTP. It should be noted that no X-Ray structure of the complex exists that stabilized by GTP alone and without the presence of dNTPs. There also exists no X-Ray structure of the complex, controlled only by dNTP and without GTP. However, such a situation is unlikely in physiological conditions since cellular GTP concentrations are 1000 times greater than dNTP concentrations.

The simulations suggest that the dATP bound to Allosite 2 leads to an overall firming up of the protein, a long range effect that is perceptible even at catalytic site. We may further conjecture that the GTP bound "loose" tetramer is "tighten up" and made active by the addition of dNTP to Allosite 2. Based on the estimated physiological concentrations of GTP($\sim 10\ \mu M$) and dNTPs (1000 fold less than GTP), this may well be a mechanism to keep SAMHD1 in a "ready" state and switch it on when necessary to drop dNTP levels. Indeed such a feedback-loop operation would be a strikingly elegant enzymatic solution to regulating dNTP pools. However, it is important to note that a only-GTP bound tetramer has not yet been observed via X-Ray crystallography. In the next chapter, we discuss our

investigation of allosteric mechanisms of SAMHD1 through analysis of cooperative dynamics.

Bibliography

[1] N. Li, W. Zhang and X. Cao, Immunology letters **74**, 221 (2000).
[2] G. I. Rice et al., Nature genetics **41**, 829 (2009).
[3] D. C. Goldstone et al., Nature **480**, 379 (2011).
[4] K. Hrecka et al., Nature **474**, 658 (2011).
[5] N. Laguette et al., Nature **474**, 654 (2011).
[6] T. L. Diamond et al., Journal of Biological Chemistry **279**, 51545 (2004).
[7] X. Ji et al., Nature structural & molecular biology **20**, 1304 (2013).
[8] J. Yan et al., Journal of Biological Chemistry **288**, 10406 (2013).
[9] C. Zhu et al., Nature communications **4**, 2722 (2013).
[10] X. Ji, C. Tang, Q. Zhao, W. Wang and Y. Xiong, Proceedings of the National Academy of Sciences **111**, E4305 (2014).
[11] O. Leavy, Nature Reviews Immunology **11**, 440 (2011).
[12] L. Wu, Retrovirology **9**, 88 (2012).
[13] J. A. Hollenbaugh et al., PLoS pathogens **9**, e1003481 (2013).
[14] B. Descours et al., Retrovirology **9**, 87 (2012).
[15] B. Kim, L. A. Nguyen, W. Daddacha and J. A. Hollenbaugh, Journal of Biological Chemistry **287**, 21570 (2012).
[16] L. Kalé et al., Journal of Computational Physics **151**, 283 (1999).
[17] J. C. Phillips et al., Journal of computational chemistry **26**, 1781 (2005).
[18] P. F. Batcho, D. A. Case and T. Schlick, The Journal of Chemical Physics **115**, 4003 (2001).
[19] J. B. Klauda et al., The journal of physical chemistry B **114**, 7830 (2010).
[20] A. D. Mackerell, Journal of computational chemistry **25**, 1584 (2004).
[21] S. Miyamoto and P. A. Kollman, Journal of computational chemistry **13**, 952 (1992).
[22] H. C. Andersen, Journal of Computational Physics **52**, 24 (1983).
[23] S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, The Journal of chemical physics **103**, 4613 (1995).
[24] H. C. Andersen, The Journal of chemical physics **72**, 2384 (1980).
[25] S. Nosé, The Journal of chemical physics **81**, 511 (1984).
[26] R. Behrendt et al., Cell reports **4**, 689 (2013).
[27] N. Yan and J. Lieberman, Nature medicine **18**, 1611 (2012).

Chapter 4

Allosteric signal transduction in SAMHD1

4.1 Introduction

Studying the enzymatic mechanism of SAMHD1 is of great importance, both from the stand point of understanding biophysical machinery, as well as due to the necessity of understanding HIV-host interactions [1] [2] [3] [4]. The central enigma of retroviral restriction by SAMHD1 [5] is how does an enzyme so ostensibly inefficient to manage lower dNTP level to far below its Km (Michaelis constant). [6] [7] Phosphorylation of SAMHD1 at Thr592 by cdk-1 has been suggested as an on/off switch for the enzyme [7] [8] [9]. However, the effect of phosphorylation is not well understood: it does not affect the kcat (Turnover rate) and Km of the enzyme. It was suggested that phosphorylation leads to structural collapse, but further studies demonstrated that it merely leads to faster tetramer disassembly upon nucleotide turn over [8] [9] [10]. Thus functional essays of SAMHD1 have yielded an incomplete picture, with the role of phosphorylation being especially opaque. Structural studies obtained from the X-Ray crystallography have been more successful. The current state of art involves well characterized assembled SAMHD1 tetramers, bound to different assembly and substrate nucleotide combinations [11] [12] [13] [14] [15]. What is missing, however, is a picture of SAMHD1 dynamics. Studies utilizing fluorescence spectroscopy and NMR have shed light on enzymatic mechanism [16] [17] [18] [19] [20] [12], but a molecular view of SAMHD1 dynamics remains elusive. While NMR based approaches that study the dynamics of large molecular weight structures do exist, they are very expensive [21]. In this chapter, the protein regulations like the allosteric information flow across the enzyme is highlighted via all atom MD simulations. In addition, we demon-

stration how the allosteric flow links with the on-path residues across the assembled SAMHD1.

Analysis of cooperative effects using molecular dynamics simulations [22] [23] has been well established. This technique relies upon cross correlating atomic fluctuations and can be used to establish hypernetworks that reflect the flow of allostery across a protein system. In this study, these ideas of cross correlating atomic fluctuations has been extended to develop the hierarchical model of interactions involving families and superfamilies of connected residues stretching across monomeric units of the assembled tetramer [24] [22] [23] [25] [26].

4.2 Materials and Methods

We have used the trajectories described in Chapter 3 for analysis. Hence the details of the system preparation and simulation protocols can be found in Section 3.2 of Chapter 3.

4.2.1 Correlation Network Analysis

The correlation network construction and analysis was performed with Bio3D package [27]. To identify and characterize the coupled dynamics between the different parts of the SAMHD1 machinery, first the C_α residue-wise linear mutual information(LMI) [25] [28] (as already mentioned in Chapter 2, Section 2.8.1, equation 2.36) was calculated as $I_{lin}(x_i, x_j) = \frac{1}{2}\left[\ln\det CM_{(i)} + \ln\det CM_{(j)} - \ln\det CM_{(ij)}\right]$, where $CM_{(i)}$ is the covariance matrix for the displacement of C_α atom of the ith residue and $CM_{(ij)}$ is the pair covariance matrix for residues i and j. A separate analysis for each of the three MD trajectories of the *wt* system resulted in three distinct matrices obtained as an ensemble average over multiple 50 ns windows along each 100 ns trajectory. Finally, three resultant distinct matrices (constructed from each 100ns trajectories) were again used to obtain a consensus matrix containing average LMI values. Then consensus matrix was pruned to cutoff 0.5 to create the correlation network out of the consensus matrix. The consensus matrix holds values from 0 to 1 whereas the correlation network created from the C_{ij} range 0.5 to 1. The network nodes represent the C_α atoms connected through edges weighted by the negative of the logarithm of the LMI values.

4.2.2 Network community Analysis

Network Path Analysis

To examine the origin of the distal site communications between specific residues, the optimal (shortest) and sub-optimal (close to but longer than optimal) path analysis are done through possible edges.[27] [29] [30] An ensemble of five hundred paths, representing the distribution of possible modes of communication between the nodes were collected for each pair of source-sink nodes. The path length (defined as the sum of the respective edge weights of the edges) distribution indicating the relative strength of correlations, was obtained in each case. Additionally, the normalized node degeneracy, indicating the fraction of total paths crossing each node, was calculated. Residues with high degeneracy were identified as important conduits to the communication network.[31] [26]

4.3 Results and Discussions

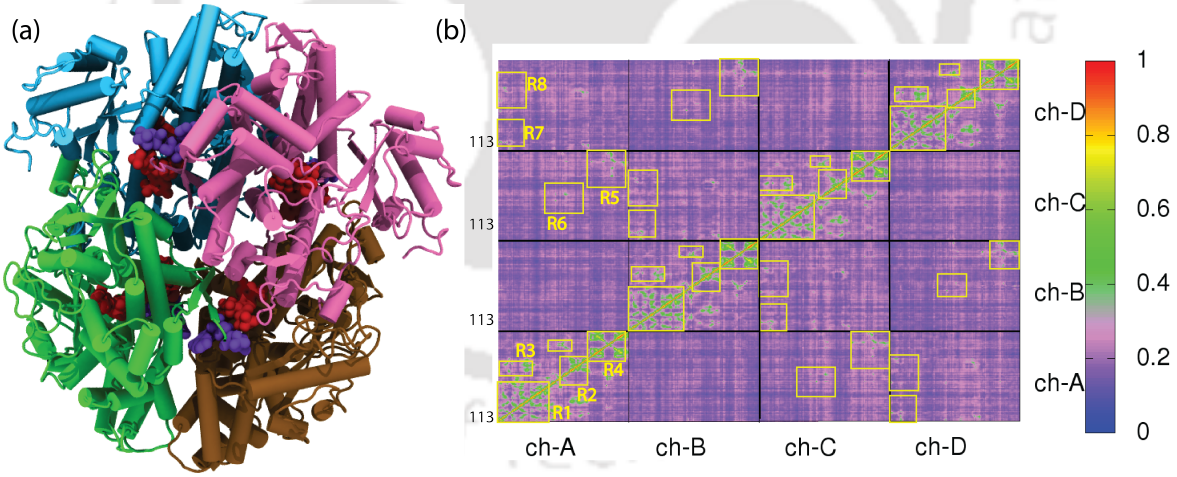

Figure 4.1: (a) SAMHD1 tetramer with chains in different colors. (b) Consensus cross correlation between C_α atoms of entire protein. Boxes highlight regions of significant correlations.

The main focus in this study was to identify "hot-spots" or residues that play a critical role either in the structural integrity of the tetramer complex or in allosteric signal transmission.[24] [32]

4.3.1 Correlation analysis of protein motions

The dynamical cross correlation map (DCCM) computed from the MD trajectories as described in chapter-2 and also in Materials and Method section are presented in Figure 4.1. Regions of significant correlations are highlighted by the yellow boxes. The main intra-chain correlations are indicated by the boxes R1, R2, R3, R4 for chain A. Similar correlations in the other chains, though present, are not highlighted in the figure. Amino acid residues at the two allosteric sites exhibit high correlations (region R1). Although most of these correlations are found in residues belonging to common structural elements (that is, near off-diagonal elements), highly coupled motions are also observed for non-contiguous residues such as R318 with I122 and D120, P130 with L197, H129 with Y257, Q140 with N248. Concerted motion is also observed between residues of two helices D440-Y450 and D383-F390 (region R2). Three proximal methionine residues, M239, M240 M416, exibit modorately coupled motion as highlighted in region R3. Several residues near the surface belonging to different structural elements are found to exhibit moderate correlations (region R4) such as W572 and E479.

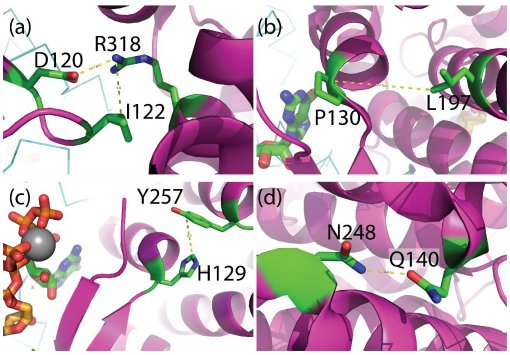

Figure 4.2: Snapshots showing (a) R318, D120 and I122 (b) P130 and L197 (c) Y257 and H129 (d) Q140 and N248

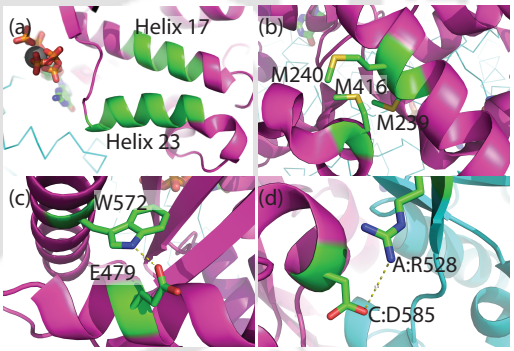

Figure 4.3: Snapshots showing (a) helices 17 and 23 (b) three proximal methionine residues M240, M416 and M239 (c) W572 and E479 (d) R528 (chain A) and D585 (chain C).

Region R5, R6, R7 and R8 indicate inter-chain correlations. only the correlations between residues of chain A with other chains have been highlighted by the boxes although similar correlations exist for other chains as well. Regions R5 and R6 represent correlation between chain A and chain C. Hydrogen bonds between Q539 (chain A) and E547 (chain C) (and *vice versa,* that is, Q539 of chain C with E547 of chain A) are associated with strong correlation between the two residues. Similar

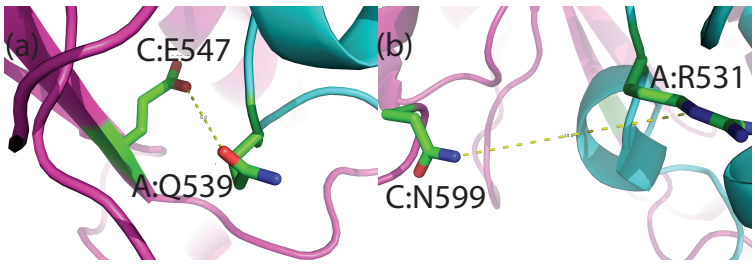

Figure 4.4: Snapshots showing (a) Q539 and E547 and (b) R531 and N599 of monomers A and C respectively.

coupling between chain D and chain B is also observed. Similarly, N328 and Q326 of chain A and chain C, respectively exhibit moderate correlations (region R6) that can be traced to intermittent hydrogen bonds between the two residues. Note that the reverse pair (that is, N328 of chain C and Q326 of chain A) also exhibit similar correlations. N119 (chain A) and F157 (chain D) abutting a common allosteric pocket, exhibit moderate coupling (region R7). Note that the features highlighted in Figure 4.1 arise from the consensus correlation matrix, obtained from multiple MD simulations as explained in previous sections.

Although the correlation analysis in Figure 4.1 provides evidence for distal site communications within the tetrameric complex and indicates residues that are important for the structural stability of the complex, a more detailed analysis was undertaken to elucidate allosteric pathways and residues crucial for signal propagation using network theory methods as described next.

4.3.2 Network and Community analysis of SAMHD1

A correlation network based on the LMI coefficients using the Bio3d package[27], was generated from the MD calculations as described in previous sections. For correlation coefficient calculation, the LMI approach was used instead of the standard Pearson like approaches to overcome the inherent underestimation in conventional covariance analysis. Subsequently, community network analysis was performed to partition the network into highly intracorrelated clusters based on the Girvan-Newman algorithm[33]. Table 1 presents the total community partitioning of the SAMHD1 complex. Fifteen consistently correlated sectors (or communities) emerged from the community analysis using the consensus correlation matrix.

The analysis elucidated a common pattern in the tetrameric complex; all residues of any given monomer (chain) can be grouped into five communities that display concerted motion. Four of these communities include nodes (residues) belonging to a single monomer, whereas the fifth community spans across two individual protein monomers that mirror each other's positions from adjacent monomeric units. The only exception was chain C, where the residues were partitioned into four communities instead of the expected five. The community partition using monomer B was illustrated here as a representative example. Figure 4.5(a) presents a snapshot of chain B of the complex in ribbon representation colored according to the communities. The name of the communities (C6 to C10) are indicated alongside. The allosteric pocket with the GTP and dATP molecule are indicated by red and purple sticks. The dATP molecule bound to the catalytic site is in the background.

Figure 4.5(b) represents the optimal community network of whole protein tetramer. The communities are labeled C1-C15. The edges of the communities of chains A and C (C1, C2, C3, C4, C5, C11, C12) are colored white while those of connecting

Family	Community	Monomer	Residues
F1	C1	A	113-145, 163-164, 166-209, 232-354, 516-530
	C6	B	113-214, 231-353, 355, 424, 519-527
	C11	C	113-145, 152-161, 164-208, 232-357, 515-533, 535
	C13	D	113-131, 158-159, 165-204, 251-360, 508-537
F2	C2	A	146-162, 165, 210-231, 376-453
	C7	B	215-230, 376, 378-423, 425-453
	C12	C	146-151, 162-163, 209-231, 358-452
	C14	D	132-157, 160-164, 205-250, 376-452
F3	C3	A	355-375
	C8	B	354, 356-375, 377
	C15	D	361-375
F4	C4	A	454-515, 544-599
	C4	C	534, 536-542, 544
	C5	C	453-514, 543, 545-599
	C5	A	531-543
	C9	B	454-518, 544-599
	C9	D	538-541
	C10	D	453-507, 542-599
	C10	B	528-543

Table 4.1: List of communities and their corresponding member residues obtained from the community network analysis of *wt* SAMHD1 tetramer complex

communities of chains B and D (C6, C7, C8, C9, C10, C13, C14, C15), are colored blue. Communities C1/6/11/13 -comprising family 1 are surrounded by a pink shaded box , they are the "floors" of the catalytic sites. Communities C3/8/15 comprising family 3 are the inside face of the catalytic site. Communities C2/7/12/14, comprising family 2 are shaded by triangles - they are the "roof" of the catalytic site. Communities C4/5/9/10, comprising family 4 are shaded by light green bars. The bar between C9 and C10 denote the allosteric handshake between them, as do the bars between C4 and C5.

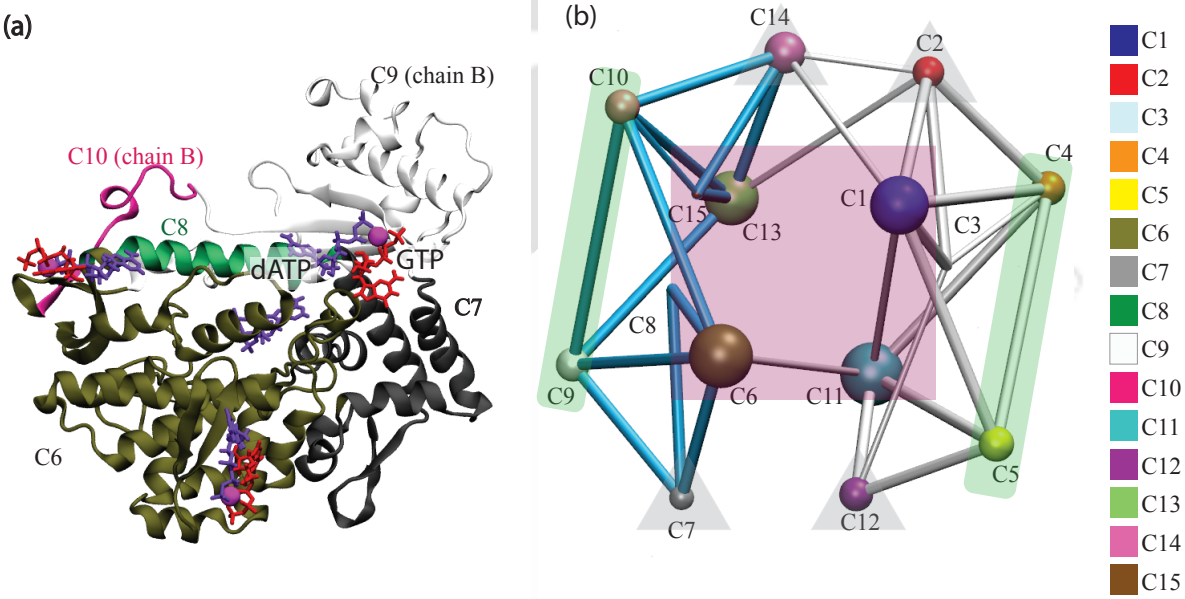

Figure 4.5: (a) Representative snapshot showing communities in chain B, the colors match the community partitioning in (b). (b) optimal community network of whole tetramer where each colore coded sphere represents one comminity. The communities are labeled from C1 to C15.

Based on the similarity of the community structure in the four chains, the communities were further grouped into four families. The first family, labelled F1, contains the communities, C1 (chain A), C6 (chain B), C11 (chain C) and C13 (chain D). The community, C6, belonging to chain B (Figure 4.5 a) is represented by tan colored ribbons, encompasses the allosteric pockets and also extends to the surface. C6 has three non-contiguous sectors, resdues 113-214, 231-353 and 519-527. The structural basis for the coupling of these sectors is discussed later. The next family, F2, includes the communities C2, C7, C12, and C14 of the chains A-D respectively. The C7 community, shown in grey in Figure 4.5(a), has two sectors;

a strip of residues, 215 to 230 sandwiched between two parts of C6, and a larger segment comprising of the residues 378 to 453. Several of the catalytic residues, for example, D311, R312, belong to the communities in the family F2. The third family, F3, includes the communities C3, C8 and C15 which includes the helix E355-A373. Notably, the community was absent in chain C, where the residues were included in the community C12 (part of the F2 family). The α-helix E355-A373 forms an independent community, C8, depicted by a dark green ribbon in Figure 4.5(a). Note that the E355-A373 helices from adjacent monomers (chain B and chain D) are coordinated at both ends by GTP molecules and interact with each other via critical hydrogen bonds (E355-R372 and N358-R372 of adjacent chains) that are essential for tetramer formation. The fourth family, F4, includes communities C4, C5, C9 and C10, each of which include residues from two chains. For example, C10 includes residues 528-543 from chain B and two segments from chain D, 453-507 and 542-599. Similarly, C9 includes residues 538-541 from chain D and the segments 454-518 and 544 to 599 from chain B. Thus the communities C9 and C10 mirror each other in chain B and D. Note that the communities C9 and C10 include residues from both chains (B and D) as listed in Table 4.1. The corresponding communities of chain A and C, that is, C4 and C5 also have a similar structure. Figure 4.6 shows the reciprocal allosteric handshake between the communities C9 and C10 that bridge chains B and D. The communities of the F4 family primarily includes the CTD residues.

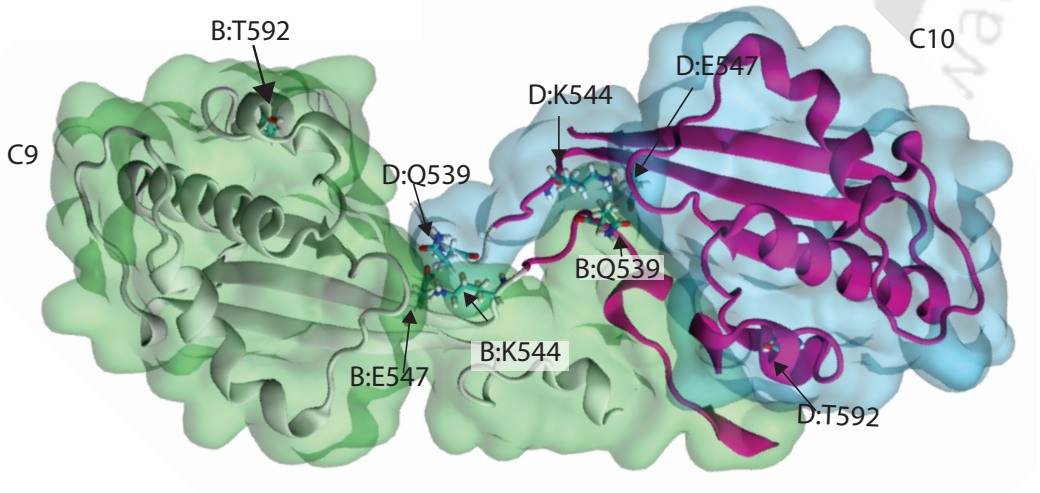

Figure 4.6: The communities C9 (white) and C10 (magneta) shown in ribbon representation enclosed by a transparent molecular surface colored according to the monomer (green for chain B and blue for chain D).

The communities of F4 family primarily includes surface exposed residues. The

community structure is further illuminated in Figure 4.7 that represents the RMSF (root mean square fluctuation) profile of four chains. The RMSF profiles in four chains in Figure 4.15 indicates the community structure by the background shade.

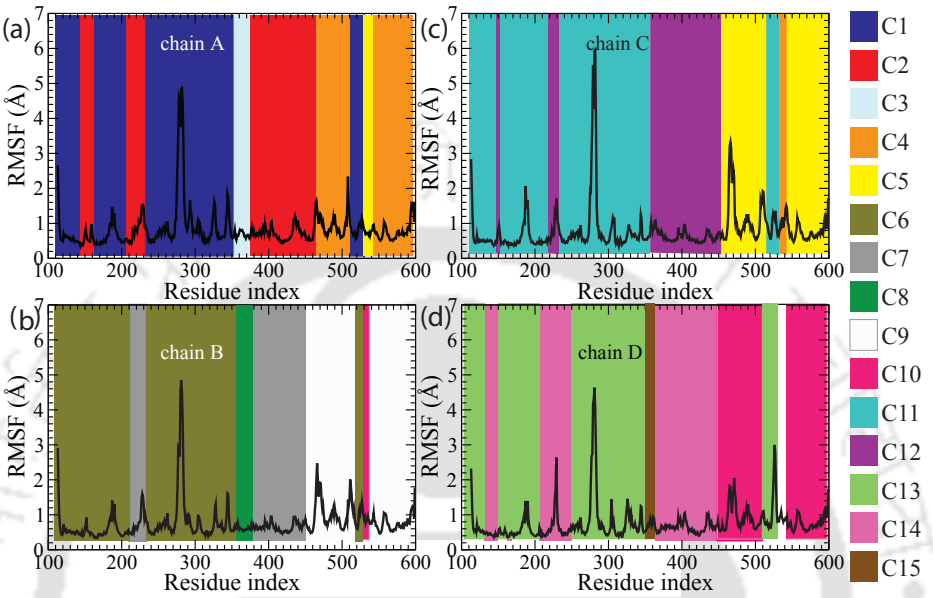

Figure 4.7: The RMSF of residues of four chains. The background is colored according to the Communities that include the corresponding residues.

4.3.3 Inter-chain Correlations

Next, the interactions between the monomers that are essential to the integrity of the complex are examined. Moderate correlations between residues of different chains are highlighted by boxes labeled R5, R6, R7 and R8 in Figure 4.1(b). To avoid cluttering, only the correlations between chain A with other monomers are labeled in that figure, although similar correlations are present between other pairs of monomers. Figure 4.8 shows representative images that illustrate how the couplings arise. Figure 4.8(a and b) correspond to region R5. The hydrogen bond between E547 and Q539 of the chain pairs A-C, C-A, B-D and D-B produce moderately strong correlations between the chains. These correlations lie at the heart of family F4 communities(C4, C5, C9, C10). The correlation of region R7 are illustrated in Figure 4.8(c and e). Residues V156 and R451 of chain D which interact with the allosteric site bound dATP and GTP, respectively are responsible for the dynamic coordination between chain A and D. The region R6 correlations (Figure

4.1 (b) are due to the residues N328 and Q326 of the chain pairs as shown in Figure 4.8(d). Figure 4.8(f) shows the helices E355-A373 from adjacent monomers, A and C coordinated at both ends by dATP molecules. Interaction between R372 of one chain and N358/D361 of the adjacent chain are instrumental in holding the helices together. Stacking interactions between H364 of the adjacent helices also contribute to the stability of the adjacent helices. In chain A, B and D, the major part of the helix E355-A373 forms an independent community (C3 in case of chain A). This is not the case in chain C, where the helix is apportioned between the larger C12 and C11 communities.

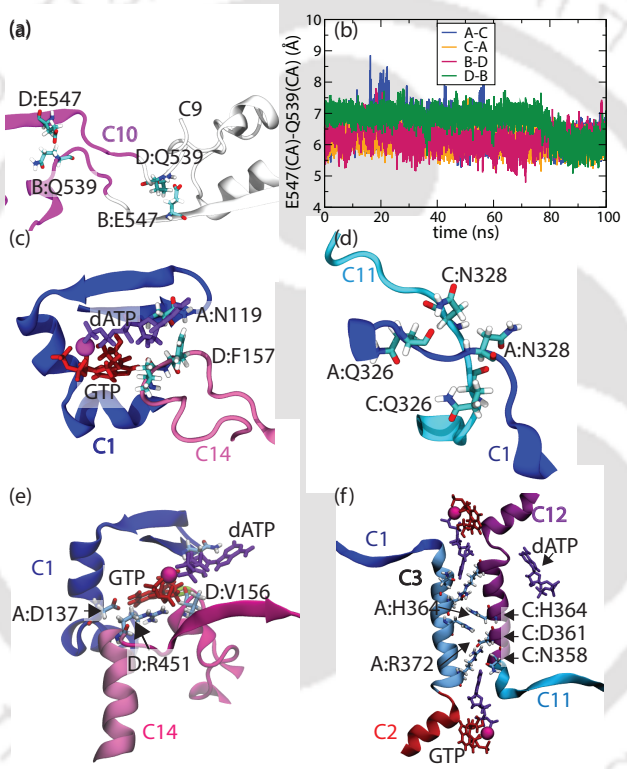

Figure 4.8: Inter-chain correlations. (a) Portion of the communities C9 and C10 represented as white and magenta ribbons. E547(chain B) and Q539(chain D) show direct interactions. (b) The $C_\alpha - C_\alpha$ distance between E547 and Q539 of the pairs of chains A-C,C-A,B-D and D-B obtained from the MD simulations. (c) The segment of chain A between V117 and R145 is represented in blue ribbon and portion of chain D in pink ribbon. (d) Interaction between the A:N328 side chain and the C:Q326 backbone of adjacent monomers result in moderate correlation. (e) Another view of allosteric pocket: The residues D:V156 and D:R451 are the two fingers of the pincer formation by which chain D encircles the allosteric site of chain A. (f) Adjacent anti-parallel helices E355-A373 of chain A and C with key residues forming hydrogen bonds. The residues are colored according to the communities.

4.3.4 Allosteric information flow via Network Path Analysis

Optimal and sub optimal paths of several pair of residues at key functional sites are being calculated to explore the connectivity and modes of signal transmission in the network. The method of calculation are already discussed in previous sections and also in Chapter 2. The resulting path ensemble reveals the diverse ways in which information can flow between the specific node pairs and also aids in the identification of key residues those are playing important role in passing the allosteric information across the system.

Pathways between surface site to catalytic core

We first examine communication channels between the phosphorylation target (T592) and the catalytic core. Consider the allosteric signal transmission between T592 (surface site residue) and H206 (catalytic core site residue) within same monomer (chain D). Since the phosphorylation of T592 by cdk-1 has been suggested as a means of regulation of protein, it is relevant to study the influence of T592 on the dynamics of the complex. The path ensemble mainly involved the track $T592 \rightarrow I591$ (of helix I591-K595) $\rightarrow A588 \rightarrow I587$ (of helix D583-A588) $\rightarrow Q567 \rightarrow A565$ (of helix D558-N577) $\rightarrow V457$ (of sheet K455-T460) $\rightarrow Y456 \rightarrow K455 \rightarrow F454 \rightarrow I381$ (of helix H376-D394) $\rightarrow T384 \rightarrow M216$ (of helix S214-R220) $\rightarrow S214 \rightarrow D207$ (of helix S192-R206) $\rightarrow H206$.

Figure 4.9(a) illustrates the pathways computed for chain D, whereas Figure 4.10 represents the node degeneracies of the intermediate residues/nodes those are taking part in transmitting the information flow for the corresponding pathways presented in Figure 4.9. Node degeneracy represents the fraction of the total computed paths crossing a given node or residue. Residues with high degeneracies are instrumental in controlling the flow of the allosteric signal. The T592–H206 pathways for other three monomers (chain A, B and C) along with their node degeneracy plots are also computed and illustrated in Figure 4.11 and Figure 4.12 respectively. While the pathways in all four chains show remarkable similarities, there are a few differences. In case of pathways between T592-H206 in all four chains, a multitude of pathways were found near the surface that later converged to a single bundle towards the core. The residues N452-K455 were found to play a critical role in funneling the signal from the branched pathways near the surface towards the core.

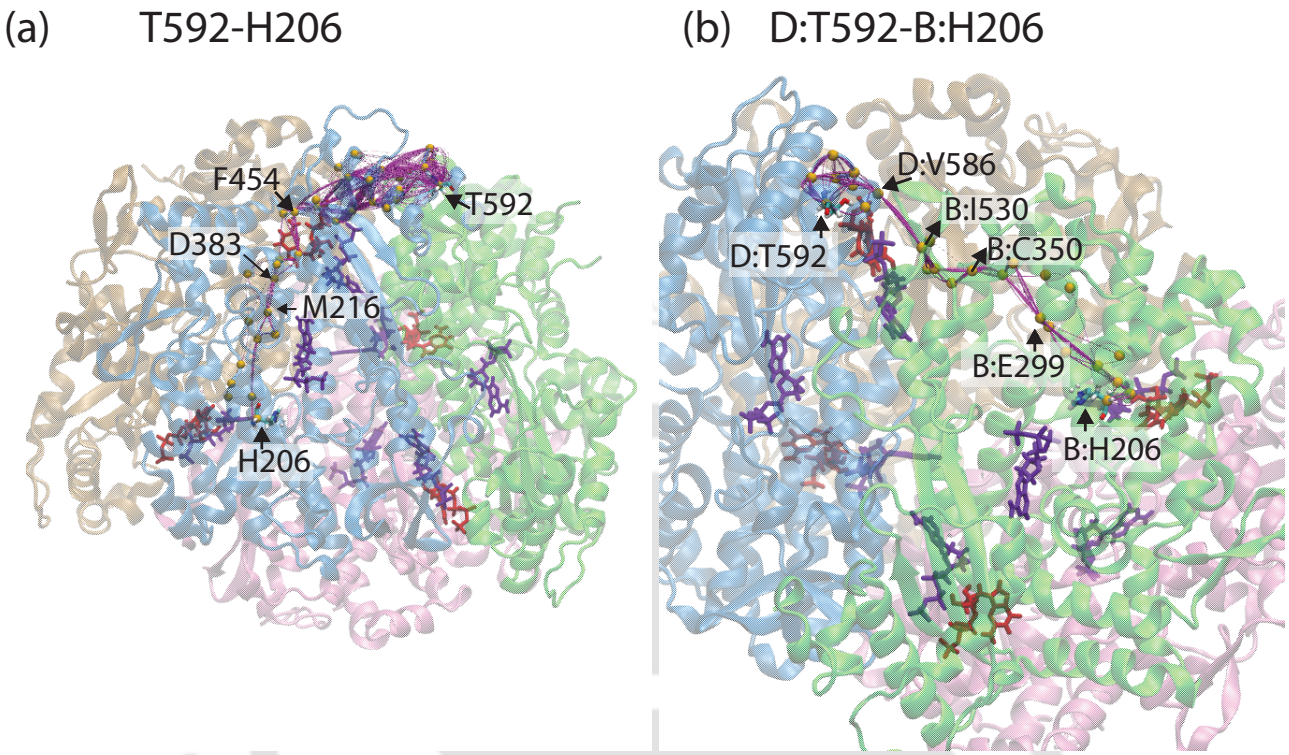

Figure 4.9: The optimal and sub-optimal pathways between key functional sites: (a) T592 and H206 of chain D. (b) chain D:T592 and chain B:H206. The intermediate important residues assisting in propagating the allosteric signal between two residues of interest are depicted as orange spheres along the suboptimal signaling pathways, depicted as magenta lines.

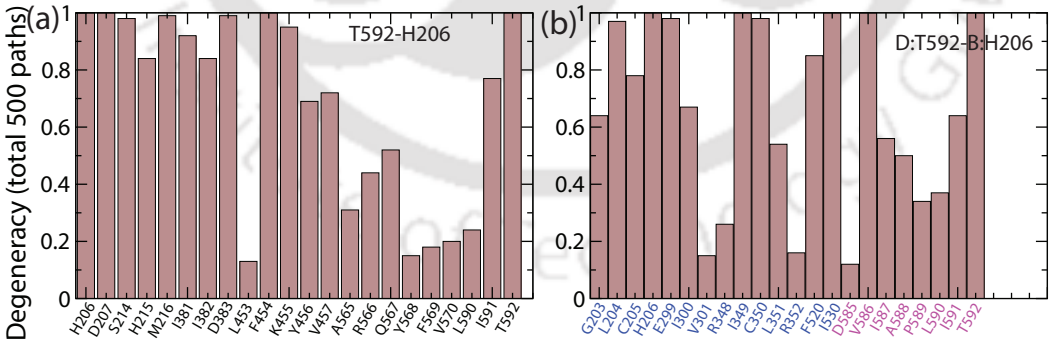

Figure 4.10: Node degeneracies corresponding to the paths in Figure 4.9 . In panel (b), since the pathways cover two monomers (chain D and B), the residue names and indices for chain B and D are coloured in blue and magenta respectively.

As an example of inter-chain signal transduction, consider the allosteric signal transmission between T592 (surface site residue, chain D) and H206 (catalytic core

site residue, chain B). Figure 4.9(b) illustrates the pathways. The must crucial inter-chain connection involved the residues V586 (chain D) and I530 (chain B) which correspond to the community C10. Note that these residues are adjacent to D585 of chain D and R528 of chain B which are intermittently connected by hydrogen bonds. The main connections involved the residues are $D:T592 \rightarrow D:I591 \rightarrow D:A585 \rightarrow D:V586 \rightarrow B:I530 \rightarrow B:F520 \rightarrow B:C350 \rightarrow B:I349 \rightarrow B:E299 \rightarrow B:I300 \rightarrow B:V301 \rightarrow B:G203 \rightarrow B:L204 \rightarrow B:H206$. The correlation between F520 and C350 that couple sequentially distant parts of the monomer may be traced to the interaction between proximal residues C522 and C350 (which are not connected by disulfide bonds in the simulation). Some of the important connections between sequentially distant residues included I349 (chain B, sheet E346-R352) and E299 (chain B, helix P291-I300) and G203 (chain B, helix S192-R206).

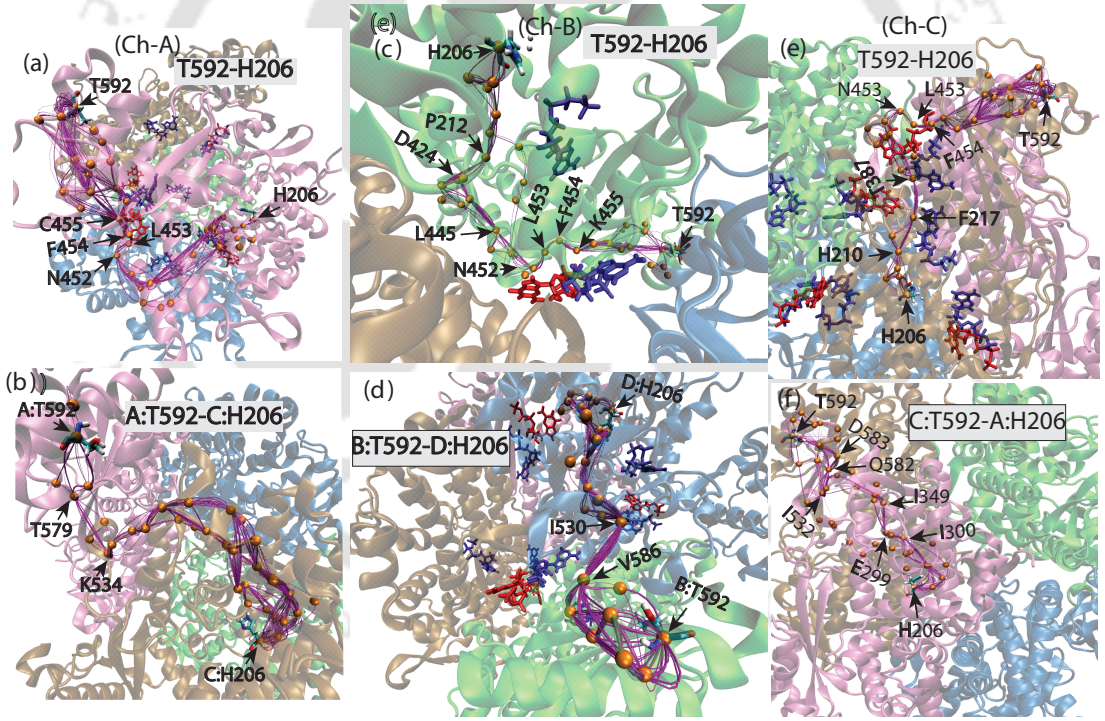

Figure 4.11: Auxiliary figure for Figure 4.9. The optimal and sub-optimal pathways between key functional sites: (a) T592 and H206 of chain A. (b) chain A:T592 and chain C:H206. (c) T592 and H206 of chain B. (d) chain B:T592 and chain D:H206. (e) T592 and H206 of chain C. (f) chain C:T592 and chain A:H206. The intermediate important residues assisting in propagating the allosteric signal between two residues of interest are depicted as orange spheres along the suboptimal signaling pathways, depicted as magenta lines.

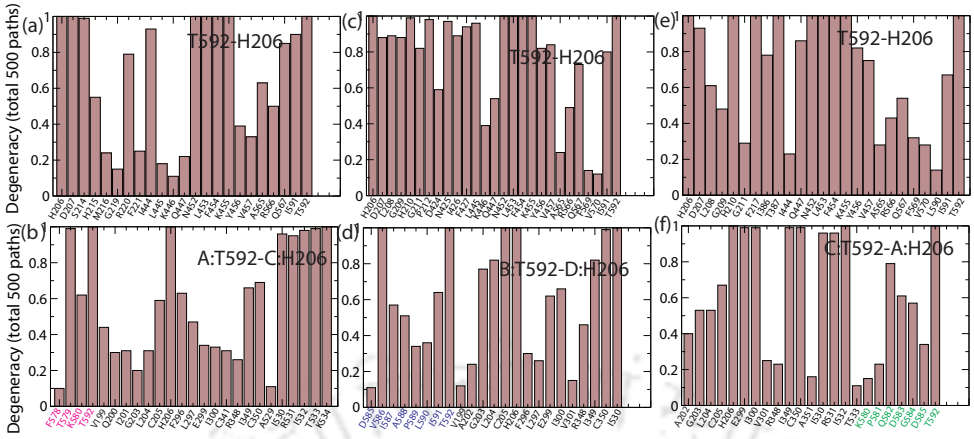

Figure 4.12: Node degeneracies corresponding to the paths in Figure 4.11 . Note that the pathways in panels (b), (d) and (f) connect different monomers. The residue names and indices for chain A in panel (b), chain B in panel (d) and chain C in panel (f) are coloured in magenta, blue and green respectively.

Pathways connecting the allosteric and catalytic sites of same monomer

An analysis of the communication pathways between the Allosite 2 residue N119 and the catalytic site residue Y374, both belonging to same monomer (chain D) (see Figure 4.13 a) reveals two distinct major groups of pathways. The minor signalling route crossed the helices D309-G324, N163-Q186, S214-R220 and H376-D394. A large amount of information flux between the allosite beta hairpin (strands K116-D120 and G124-L128) and helix D309-G324 passes through interactions between I122 and R318 (see Figure 4.13 b). The major signalling route between N119 and Y374 crossed helices H129-D137, T232-G249, D415-T420 and H376-D394. Information flux between the source, N119 and helix H129-D137 is transmitted along the backbone. Connection between I135 (of helix H129-D137) and L244 (of helix T232-G249), L244 and M416 (helix D415-T420), D415/D414 and D394 (helix H376-D394) formed the communication pathway between Allosite 2 and catsite (Y374).

A similar analysis of the communication pathways between N119 and the catalytic site residue D311 revealed D120, I122, R318 and D314 to be the critical links (see Figure 4.13 b). Again, the correlation between I122 and D318 (helix D309-G324) was found to be vital to the flow of information. Figure 4.13(c) shows the communication between the Allosite 1 residue, D137 and the catalytic site residue R164 to pass primarily along backbone. However, strong correlation between R145 and R164 are crutial for the information flow.

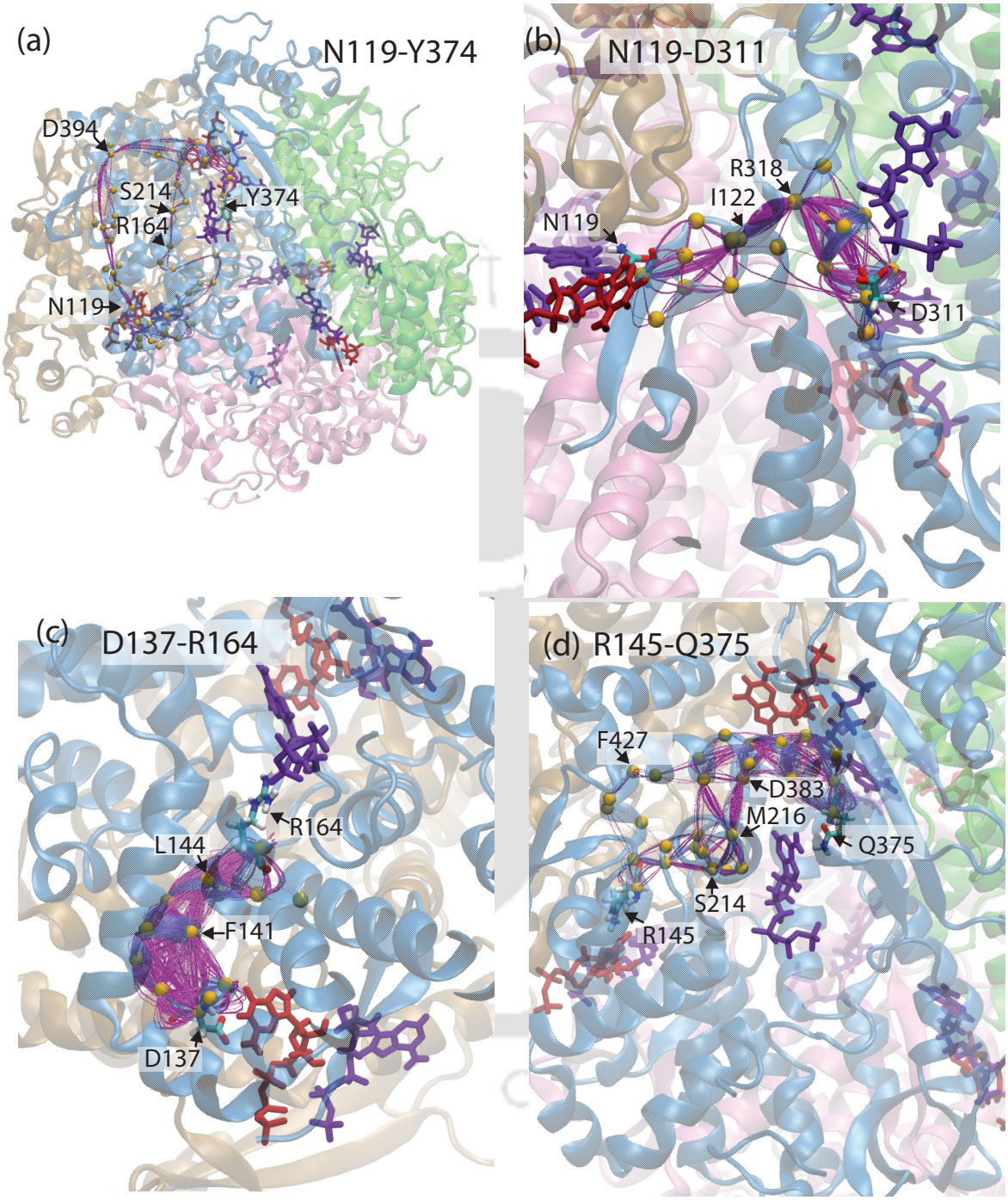

Figure 4.13: The optimal and sub-optimal pathways between key functional sites in chain D: (a) N119 and Y374, (b) N119 and D311, (c) D137 and R164, (d) R145 and Q375. The residues predicted to assist in propagating the allosteric signal between two residues of interest are depicted as orange spheres along suboptimal signaling pathways, depicted in magenta lines.

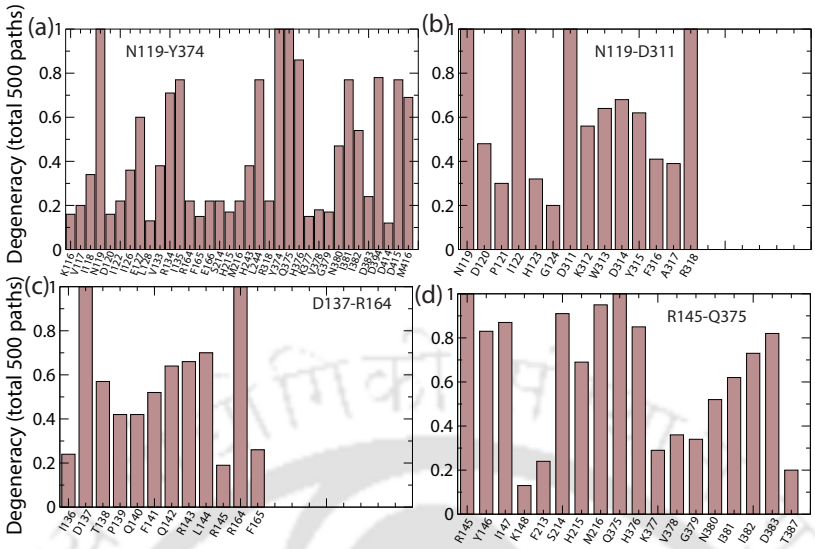

Figure 4.14: Nod degeneracies to the corresponding paths in Figure 4.13 .

As a second example of a connection between Allosite 1 and the catalytic site, we consider the pair, R145 and Q375 (see Figure 4.13 d). Correlation between I147 and F213 provides a clear link between helix L144-I147 and helix S214-R220. Connection between M216 and D383 (also I386) bridge the α -helices (S214-R220) and (H376-D394). The same residues were also identified as "hot-spots" in other pathways (see Figure 4.13 a).

Pathways between Allosteric site to catalytic site across monomers

Given the fact that tetramerization is essential to the enzymatic activity of SAMHD1 (the monomeric form is not known to possess triphosphohydrolase activity), it is necessary to understand the allosteric linkage beween the chains particularly between the allosteric and catalytic pockets. Figure 4.15 shows the communication channels between D137 of chain B, a key residue at allosteric site 1 that forms hydrogen bond with the GTP with the catalytic site residue Q375 of the other chains, C, D and A respectively.

Unsurprisingly, V156 of chain C is the key residue that bridges the two monomers (B and C). In contrast, the linkage between D137 (chain B) and Q375 (chain D) involve residues I350 (chain B) and V586 (chain D), both belonging to C10 community(see Table 4.1). The reciprocal path ways between Q375 of chain B and allosteric site D137 of chain A, C and D are also calculated and illustrated in Figure 4.17 along

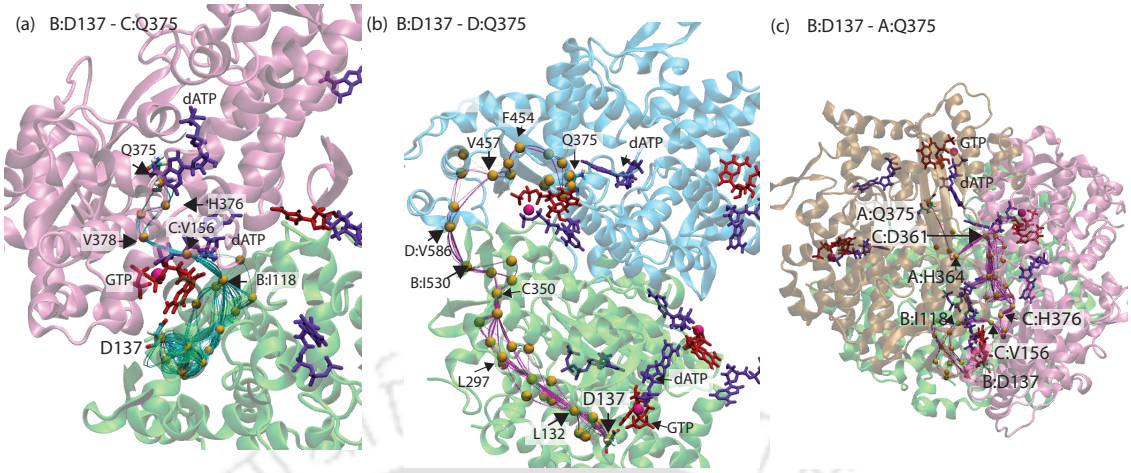

Figure 4.15: Inter-chain optimal and sub-optimal pathways between key functional sites: (a) D137 (chain B) and Q375 (chain C), (b) D137 (chain B) and (Q375 of chain D) and (c) D137 (chain B) and Q375 (chain A). The residues predicted to assist in propagating the allosteric signal between two residues of interest are depicted as orange spheres along suboptimal signaling pathways, depicted as cyan (for panel a) or magenta (for panel b and c) lines.

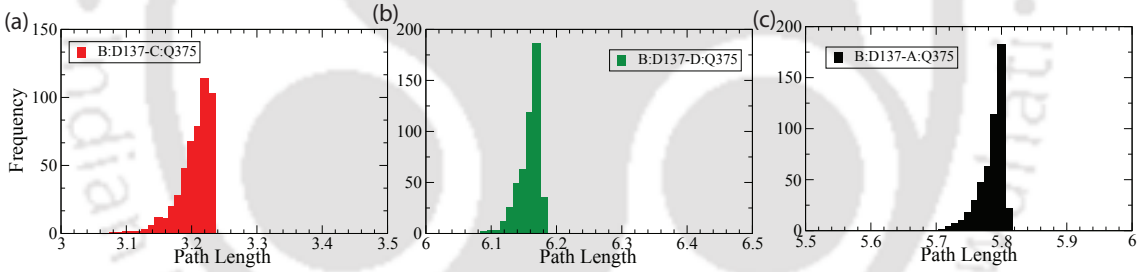

Figure 4.16: The path length distribution for the pair of residues (a) B:D137-C:Q375, (b) B:D137-D:Q375, and (c) B:D137-A:Q375 corresponding to optimal and sub-optimal path.

with their degeneracy plots shown in Figure 4.18. The same key residues such as V586, I530, V156, R451, interlinking the different chains are found in the paths even though, intra-chain communication proceeds through alternate pathways. Note that the correlation between V586 and I530 of neighbouring chains was also observed to be crucial in the communication between the surface exposed T592 residue of one chain and the H206 of the catalytic core of the adjacent monomer (Figure4.9 b). An important revelation in the path analysis is that the passage of the allosteric signals between the pairs of chain A-C and B-D primarily involve the communities of the family F4 (that is, C4, C5, C9 and C10) rather than the E355-A373 helix (family

F3). The communication between D137 of chain B and Q375 of chain A is tortuous, spanning three monomers(Figure 4.15 c) indicating tenuous correlation. The key inter-chain connections include D137 of chain B with V156 of chain C and D361 of chain C with H364 of chain A. The distribution of path lengths corresponding to the "source-sink" pairs in Figure 4.15 is presented in Figure4.16. The longer the path lengths, the more tenuous the connection between the sites. Hence, it is most likely that the D137 residue in chain B exerts maximum influence on the catalytic site in chain C. The longest path lengths are observed between the residues B:D137-A:Q375 while the shortest lengths are found between B:D137-C:Q375.

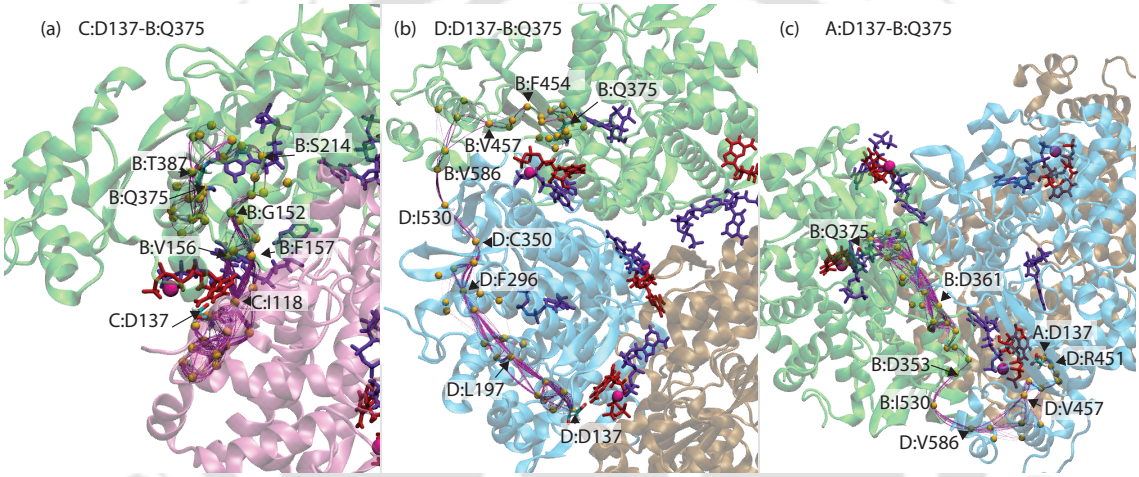

Figure 4.17: Auxiliary figure for Figure 4.15. The optimal and sub-optimal pathways between key functional sites: (a) D137 (chain C) and Q375 (chain B), (b) D137 (chain D) and (Q375 of chain B) and (c) D137 (chain A) and Q375 (chain B).

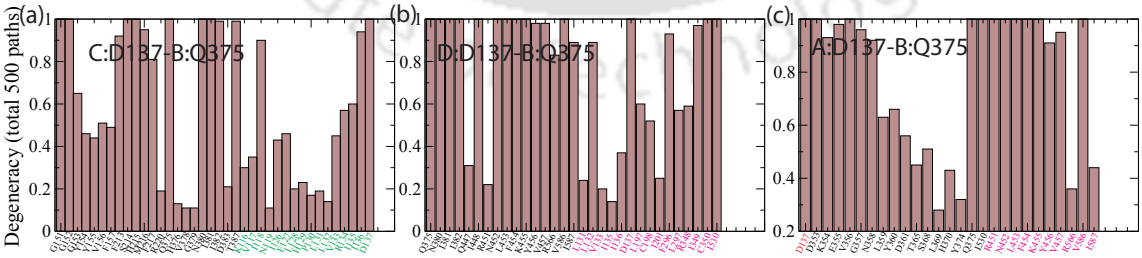

Figure 4.18: Node degeneracies corresponding to the paths in Figure 4.17. The residue names and indices are coloured according to the monomers. Residues of chain A, B, C and D are indicated by red, black, green and magenta respectively.

4.3.5 Node-Centrality calculations

The node centrality or betweenness centality, which gives the number of unique shortest paths crossing a node, is calculated and plotted for all the residues in Figure 4.19. The node centrality indicates the importance of the node or residue in controlling the flow of information.

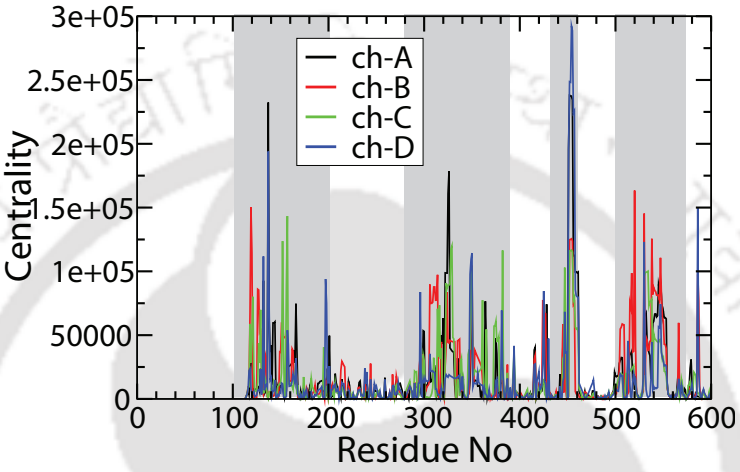

Figure 4.19: Residue-wise betweenness centrality for SAMHD1 network. A gray background is used to highlight residues with high centrality.

Although some variation between the centrality of the nodes belonging to the four monomers are observed, there is an overall consistency between the centrality values of residues of all four chains. Table.2 reports the residues with betweenness centrality more than 100000. In all four chains, N452, L453, F454, and K455 were found to possess high centrality values. These residues are found to be crucial in channeling the signal pathways from the surface towards the catalytic core (see Figure 4.9 a). In addition, the neighboring residues, R451 flanking the allosteric site, pins different chains together (D-A, A-D, B-C and C-B) as shown in Figure 4.8(e). A high centrality value was also observed for C350 in three of four chains. Of the three proximal cysteine residues C350, C341, C522, two are connected by a disulfide bond (C341 and C350). However, the path analysis reveals that the correlation between C350 and C522 play a significant role in transmitting information from the surface to core residues. Note that C350 was found to be an important link in the communication pathways between T592 and H206 of the neighboring chains (Figure 4.9 b).

Residue Centrality							
Chain A		Chain B		Chain C		Chain D	
137	232400	119	150295	152	123735	132	111753
325	160847	120	114617	157	143354	137	193977
326	178452	**350**	104482	328	117487	349	109510
451	196060	**452**	124790	329	120010	**350**	114358
452	237978	**453**	125120	**350**	100091	451	195953
453	237630	**454**	125490	382	116455	**452**	248611
454	237372	**455**	123920	447	102934	**453**	248308
455	234356	520	163366	**452**	116042	**454**	292309
456	112407	530	145348	**453**	116344	**455**	288803
457	112466	538	125572	**454**	116716	456	224243
		547	110666	**455**	115138	457	225866
				534	100163	530	123044
						586	149846

Table 4.2: Residue-wise betweenness centrality greater than 100000 are listed for four chains. A high value of centrality is found in the residues N452, L453, F454 and K455 in all four chains shown in bold font.

4.3.6 Crucial hydrogen bonds for catalytic pocket

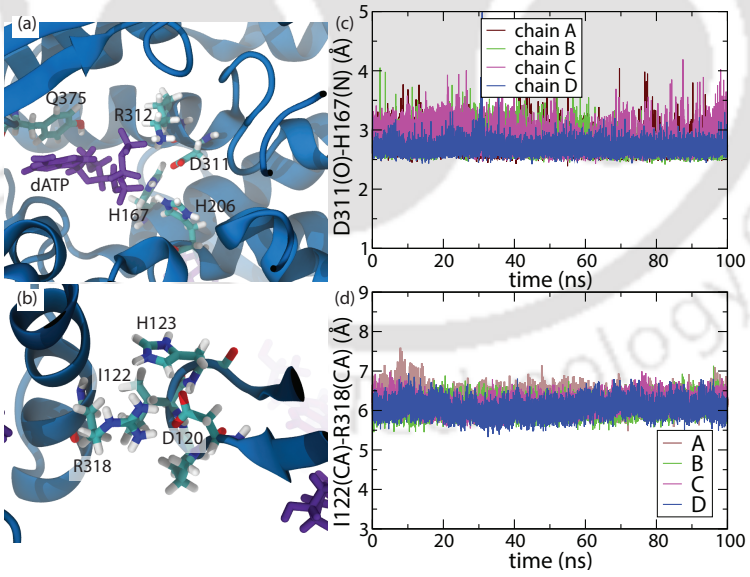

Figure 4.20: (a)Snapshot of dATP (purple sticks) bound to catalytic site in chain D (blue ribbon) (b) proximal residues R311,H123,I122 and D120 are presented as sticks. A double hydrogen bond between D120 and R318 pins the α-helix (D309-G324) to the beta hairpin. (c) Distance between D311 (side-chain O) and H206 (N) obtained from MD simulations. (d) The $C_\alpha - C_\alpha$ distance between I122 and R318 of four chains from Set 1 MD simulations.

Figure 4.20(a) represents a snapshot showing the dATP bound to the catalytic site in one of the chains. A closer inspection shows that the interaction between D311 and H167 is instrumental in maintaining the shape of the catalytic pocket. The shortest distance between the OD1/OD2 atoms of D311 and the NH1 atom of H167. Figure 4.20(b) shows the variation of the distance with time as measured from Set 1 MD simulations. The distance was found to be steady with an average of 2.8 Å in three of the four chains. The corresponding occurrence of the hydrogen bond was greater than 50% in three of the four monomers.

4.3.7 Perturbations to the community network

To examine the role of the "hot-spots" in signal transmission, the community network (Figure 4.5 b) was modified by selectively deleting certain edges. Specifically, the role of the linker residues 452-455, C350 (part of the putative redox switch) [34] and the edge between V586 and I530/R531 of adjacent chains were probed. The residues D137 of chain B and Q375 of chain D (and *vice versa*), belonging to the allosteric and the catalytic sites of neighboring chains were selected for the analysis, the paths are shown in Figure 4.15(b). As shown in Figure 4.21 a,b), the deletion of the linker 452-455, that is, elimination of all edges connected to the specified residues, leads to weaker communication between the allosteric and catalytic site residues in neighboring chains as indicated by longer path lengths.

Figure 4.21(c-d) shows that the deletion of edges connecting the C350 (part of the redox switch) also lead to longer pathways, that is, diminished communication between the active sites of adjacent chains. Finally, the deletion of the edges between B:530/531 and D:V586 was found to weaken the communication between the specified active sites of the neighboring chains B and D. Figure 4.21(e-f) show a significant shift in the path length histograms towards longer paths, supporting our hypothesis regarding inter-chain signal transmission. The shift in the path length distributions were more pronounced in the last example (Figure 4.21 e-f) as compared to the first two (Figure 4.21(a-d), highlighting the influence of the interaction V586-I530/R531 links in signal transmission between the allosteric and catalytic sites of adjacent chains.

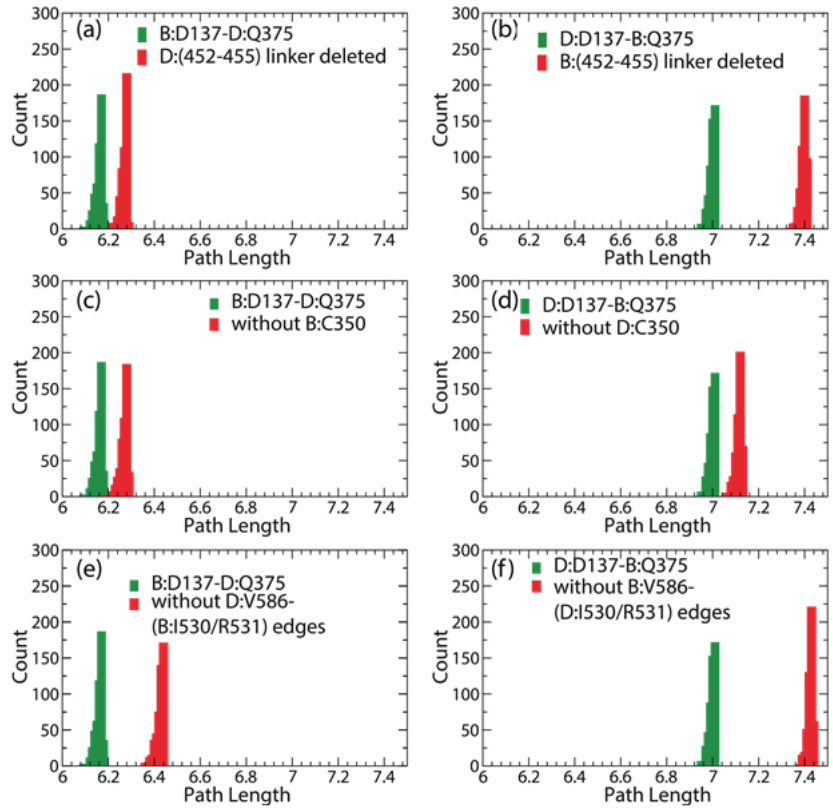

Figure 4.21: Path length distributions between allosteric site residue D137 and catalytic site residue Q375 of adjacent chains. Green bar denotes original network and red bar represents the perturbed network where specific edges were deleted.

4.3.8 Principal Component Analysis of Reciprocal Allosteric Handshake

To elucidate the overall patterns of motion in the SAMHD1 complex, Principal Component Analysis (PCA) was employed on the C_α atoms of the structure. The analysis was performed in two ways. First, in order to focus on the "allosteric handshake" between the monomers, we performed PCA using only C terminal domain of adjacent chains (that is, considering A-C and B-D separately). Next, we carried out a similar analysis for the entire complex. In each case three separate calculations were performed using three independent trajectories. Figure 4.22 illustrates the results of the PCA performed using CTD residues 455-599 of pairs of adjacent monomers (*viz.* A-C and B-D in separate calculations) on trajectory 3. Although all the PCs are involved in the collective motion, the contribution to the total variance diminishes rapidly after first few eigenvectors as shown in Figure 4.22(e). The first

four eigenvectors accounted for almost half of the total variance in all cases.

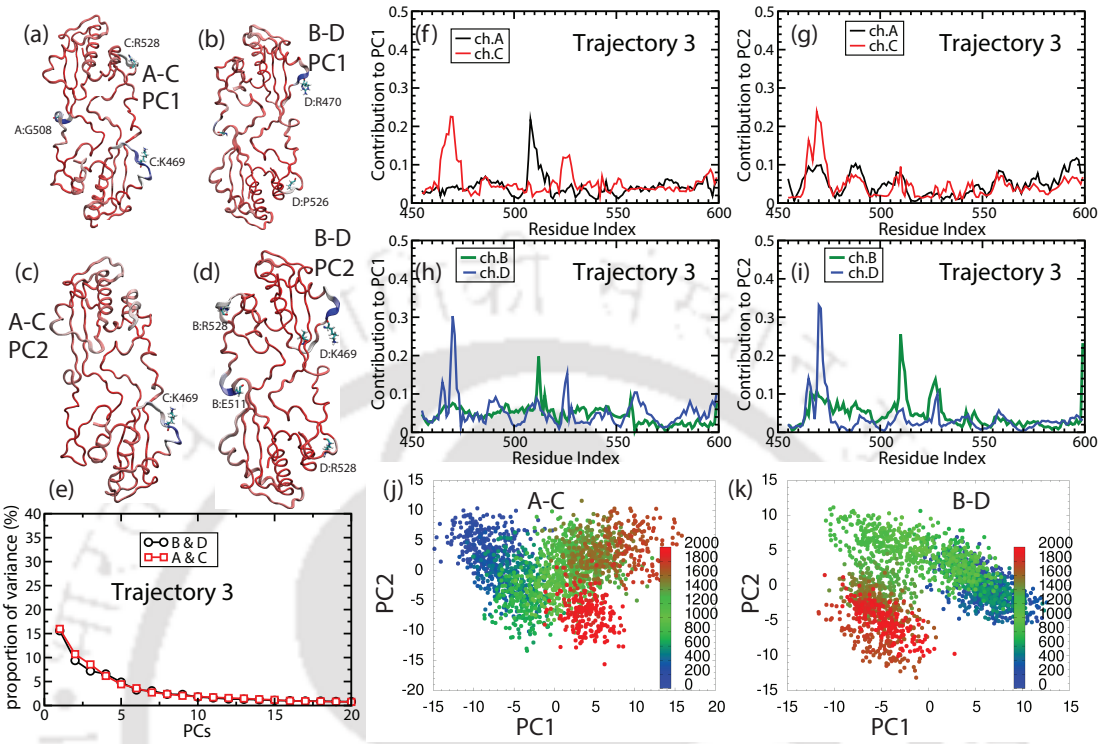

Figure 4.22: PCA of Trajectory-3 MD data performed using only C terminal domain (CTD) of pairs of adjacent chains: A-C and B-D. The interpolated structures representing motion along principal component 1 of chains A-C and B-D are shown in panels (a) and (b) respectively. The principal component 2 of chains A-C and B-D are shown in panels (c) and (d). The backbone protein in blue color, indicating high contribution to the PCs and red indicating invariant structure. (e) represents contribution of all calculated PCs to the total variance as percentage. Panels (f-g) and (h-i) present the residue-wise contribution to the first and second principal components performed on chains A-C and B-D respectively. Panels (j-k) present the cross plots of the first two principal components, the color code represents the time evolutions from early (blue) to final (red) frames.

An indication of collective motion of adjacent monomers is obtained from the simultaneous peaks in both monomers in the plot showing the residue-wise contribution to the first two PCs (Figure 4.22 f-g). In case of PC1, almost all sharp peaks are observed in segments G464-I466. D506-E511, and the loop F520-I530.

Two of the most dynamic parts of the CTD that contribute to the top two PCs are the segments 505-511 and 520-530. The flapping of the segment F520-I530 is particularly interesting because of two reasons. First, it closes the gap between C522 and C340/C350. Recent experimental studies have indicated a redox switch operating between the three cysteine residues that can have important consequences

for regulation of the system[34]. Second, the residue I530 is involved in inter-chain communication as seen previously from the path analysis. The scatter plots in panels (j) and (k) of Figure 4.22 essentially give a two-dimensional representation of the conformational space occupied by the system. The gradual migration of the points in the PC1-PC2 scatter plots (Figure 4.22 j,k) indicate that the protein is undergoing breathing motions while the tetramer stays intact. Figure 4.23 shows the contribution of PC3 for the contribution of all the three trajectories.

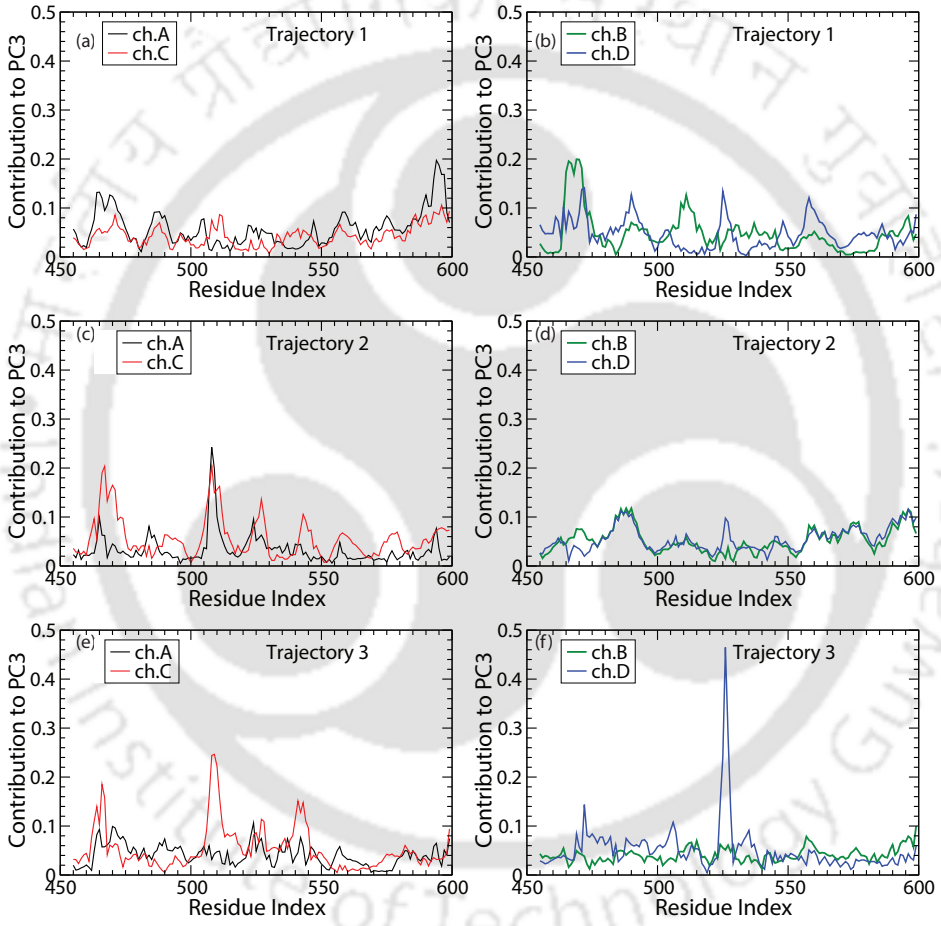

Figure 4.23: PCA of CTD for pairs of adjacent monomers. The contribution of each residue of the CTD (residue 455-599) to PC3 for the three trajectories are presented.

Figure 4.24 and 4.25 present the results of PCA performed using entire tetrameric complex. As in the case of PCA performed using the CTD of pairs of monomers, the residue-wise contributions to the top three PCs (Figure 4.24) shows the same segments in the CTD to be involved in the motion (G464-I466, D506-E511, and F520-I530). In addition, other segments in the core with high contributions can be

identified: 186-190, 228-230 and smaller peaks between 300 and 350. The peaks were found in similar locations in all three trajectories in case of the top three PCs. The loop 278-283 has been ignored in all these PCA calculations since the segment was not resolved in the XRD structure (4TNR.pdb) and was found to have high fluctuations that dominated analysis. However, sharp peaks at the residues flanking the segment, that is, 277 and 284, were observed. Figure 4.25 shows the cross plot projecting the dynamics of the system on the top two Principal Components. The migration of the points in the PC1-PC2 scatter plot along semi-circular or U-shaped paths, characteristic of diffusive behavior in a shallow basin, is indicative of collective motions in the protein complex, which nevertheless retains the quaternary structure that is, breathing motions.

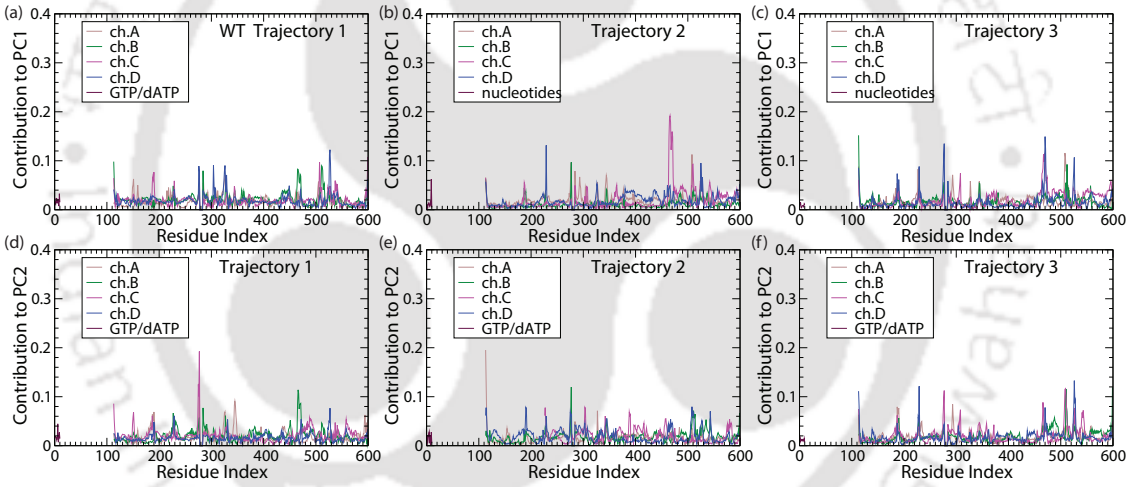

Figure 4.24: PCA for the whole tetrameric complex. Residue-wise contribution to the total variance. The two rows denote the top two PCs while the three columns indicate three trajectories.

PCA of Trajectory-1 and Trajectory-2 MD data performed using only C terminal domain (CTD) of pairs of adjacent chains: A-C and B-D. The interpolated structures representing motion along principal component 1 and principal component 2 of chains A-C and B-D are plotted. A similar trend of characteristics as obtained for the PCA of Trajectory-3, are also observed for these two trajectories. The plots are shown in Figure 4.26 and Figure 4.27 respectively.

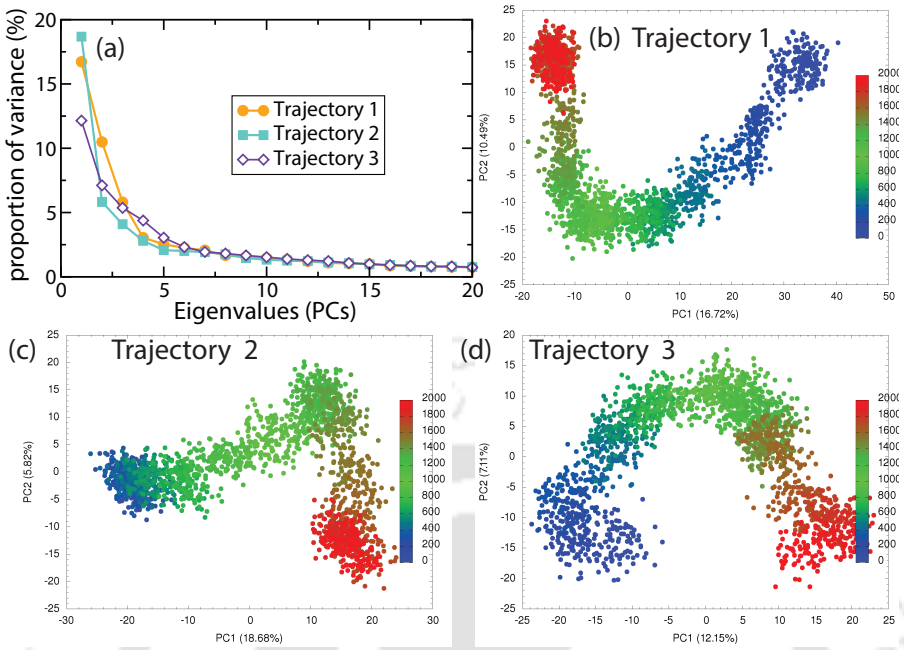

Figure 4.25: PCA of total tetrameric complex. (a) The contribution of all calculated principal components to the total variance as percentage. (b)-(d) Scatter plots presenting the projection of dynamice of the complex onto the first two principal components for the three trajectories respectively. The color code represents the time-evolution of the system starting from blue.

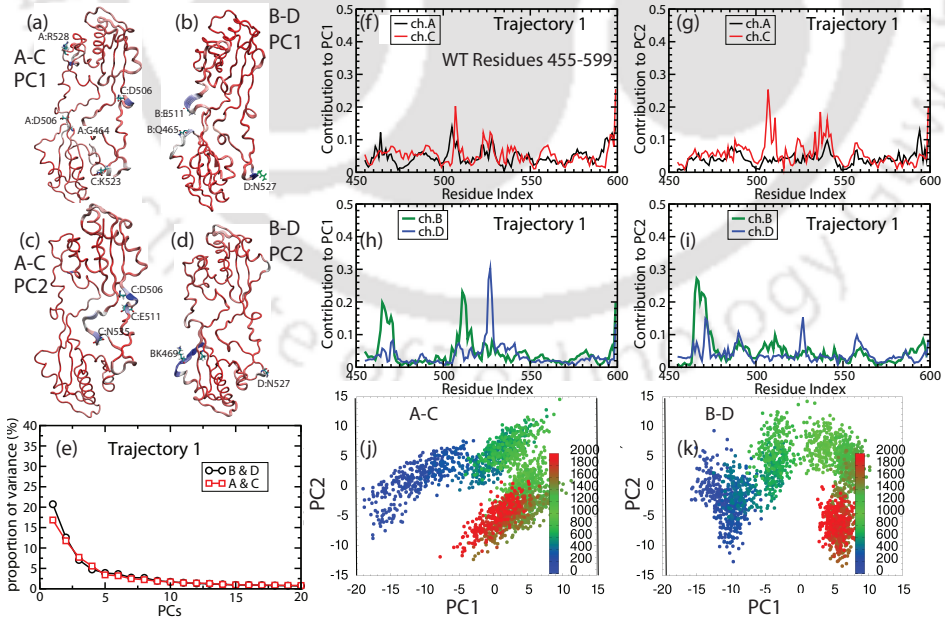

Figure 4.26: Auxiliary figure to Figure 4.22, corresponding to Trajectory-1.

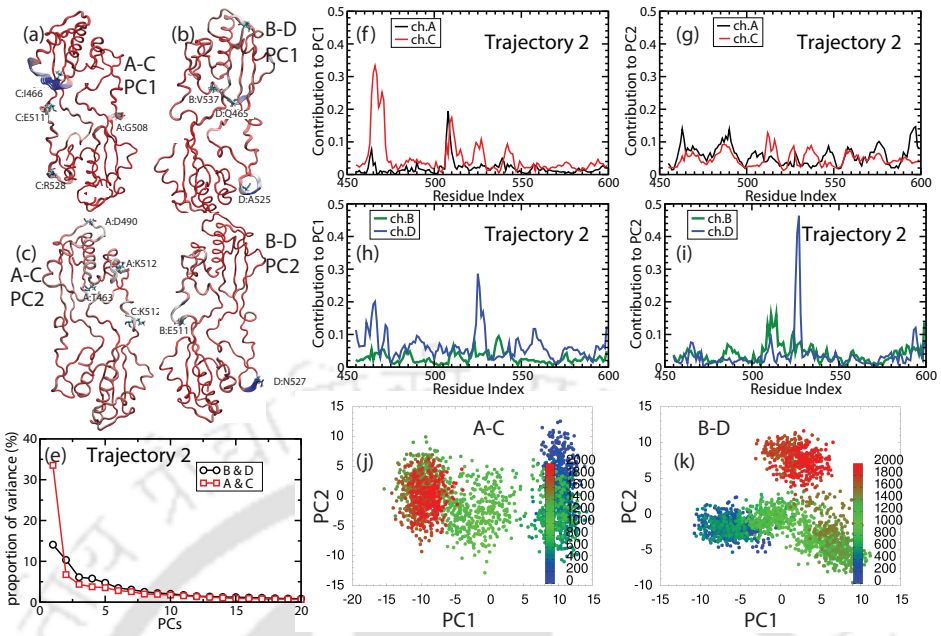

Figure 4.27: Auxiliary figure to Figure 4.22, corresponding to Trajectory-2 .

4.4 Discussions

The existence of concerted motions between adjacent, inter-chain helices in SAMHD1 monomers, as detected by correlation analysis is unexpected. Indeed, the allosteric dATP is stabilized by beta strands 1 and 2, which are anchored at either end to helices 1 (residue 129-137) and helix 12 (residue 309-324). Helix 1, in particular is "double-anchored" at both ends to helix 9 (residue 248-257). Thus the allosteric dATP, by its presence, confers a rigidity to the tertiary structure of the monomer which extends from the Allosite, all the way to surface. Thus the Allosteric sites are heavily "scaffolded" by a network of correlated residues. This is in line with conclusions from the recent study in chapter-3 [32]. However, it is the presence of the correlation groups (R5, R6, R7, R8) which span across monomeric units which yields new insights into how the different monomers "talk" to one another. Surface residues E547, for instance, belonging to beta strand 11 (residue 546-554) is anchored to residue Q539 of the adjacent monomer. However, more information can be extracted when one can look beyond individual pairs of residues and start looking at the co-moving groups of residue. Hence an analysis of the community partitioning of residues begins to shed more light on how this protein functions.

Taking as an example of chain B, the bulk of the monomer is one co-moving bundle of helices which are identified as community C6. This community forms one

wall of the catalytic crevice which accommodates dNTPs, and indeed forms many of the side-chain stabilizing interactions which select and hold the dNTP in place, including the catalytically critical residues D311, H206 and D207. However, C6 only makes sparing contacts with the two allosteric sites. The "inside" face of the catalytic pocket is formed by community C8, which is essentially helix 14 (residue E355-A373). This community makes, not just close contacts with the substrate dATP (by Y374), but also makes direct contacts with both allosteric pckets, with one end of the helix helping to stabilize the dNTP at one end (R372) and the other end stabilizing the dATP in that pocket (N358). Further, helix 14 (E355-A373) that forms the community makes two hydrogen bonds (D361-R372 and N358-R372, as discussed in chapter-3) at either end with the equivalent helix from the adjacent monomer. Thus, community C8 is not just responsible for inward allosteric communication from the allosteric pocket to the catalytic core, but also for outward communication to the adjacent monomer. Community C7 forms the "roof" of the catalytic pocket and community C9 forms the "outer wall". Viewed in isolation, community C9 can be readily seen to compose the bulk of the CTD of SAMHD1 monomer (chain B). It is not surprising that this domain would be a co-moving entity. Looking closer, however, C9 actually comprises two large segments of the CTD, but does not include the connecting 25 linker residues. Surprisingly, it does reach across the tetramer interface and includes the corresponding linker residues from the adjacent monomer (in this case, chain D). Thus the two CTDs of the adjacent monomers are involved in an exquisitely reciprocal allosteric handshake. It is also worth noticing that the allosteric communication, as well as the hydrogen bonding between communities C3-C12 and C8-C15 (helix E355-A373) of opposing monomers spans the chain A-C and chain B-D "tetramer" interface and not the chain A-D and chain B-C "dimer" iterface. This might yield an elegant explanation for why the SAMHD1 dimer is catalytically inactive: the allostric breathing motions that this protein undergoes are only truely seen in the tetrameric state. [9] [8]

The next phase of anlysis involved drilling down from identifying co-moving groups of residues to identifying specific pathways and nodes through which allostreic information travels. As expected it was found that several pathways for information flow from the catalytic site to the allosteric sites within a monomer. However, path analysis proved to be truely revelatory when information flow across monomeric units are made. It was already found that the communities C9/C10 are involved in allosteric handshakes across monomers. Path analysis revealed even

more fine-grained results: residues I350 (chain B) and V586 (chain D) are vital nodes for information flow across monomers, which is largely transduced across surface pathways. These are located close to R528 (chain B) and D585 (chain D) which are involved in intermittent hydrogen bonding. Finally, the question of SAMHD1 regulation by phosphorylation remains. Previous studies report that this leads to structural collapse of the CTD have been disputed. Present analysis indicate that the transmission of allosteric signals from the surface exposed cdk-1 phosphorylation site - T592 to the catalytic site are seen to converge and funnel through a critical bottleneck: N452-K455. Apropos the community partitioning table, this is the link between communities C7 and C9 (chain B). Structurally, this is the connector between the minor lobe and the CTD of the SAMHD1 monomer. Clearly this short linker is of significant regulatory importance, as it transmits the effect of cdk-1 phosphorylation from the CTD to the entirety of the protein tertiary structure. this is also refected in high "centality" of these residues. Mutagenesis studies on the linker residues (residue 452-455) are suggested- both to study restriction defects as well as the in-vitro dNTPase activity. However, the consequences of this regulatory signal are yet to be uncovered. It is known that the phosphomimetic mutants of SAMHD1 are restriction incompetent, but they are dNTPase capable. Thus this regulatory signal potentially triggers some other, hitherto undiscovered property of this protein, which may be an alternate HIV-1 restriction mechanism.

4.5 Conclusion

In the previous empirical study described in chapter-3[32], the effect of GTP/dATP occupancy and vacancy was explored in the Allosites of the SAMHD1 tetramer. In this study, the mechanistic basis for the phenomenological observations are undertaken. In other words, the structural linkages and the communication channels are revealed that allows SAMHD1 to function as a molecular engine. SAMHD1 is an enormously complex protein system which makes NMR studies difficult. This is where the all atom MD study comes. Starting from the high resolution X-Ray structures, we have uncovered the pathways of allosteric communications which show how the monomeric units of the active tetramer communicates via a reciprocal "handshake".

Bibliography

[1] G. I. Rice et al., Nature genetics **41**, 829 (2009).
[2] D. C. Goldstone et al., Nature **480**, 379 (2011).
[3] K. Hrecka et al., Nature **474**, 658 (2011).
[4] N. Laguette et al., Nature **474**, 654 (2011).
[5] O. Leavy, Nature Reviews Immunology **11**, 440 (2011).
[6] E. C. Hansen, K. J. Seamon, S. L. Cravens and J. T. Stivers, Proceedings of the National Academy of Sciences **111**, E1843 (2014).
[7] L. H. Arnold et al., PLoS pathogens **11**, e1005194 (2015).
[8] A. Bhattacharya et al., Scientific reports **6** (2016).
[9] Z. Wang, A. Bhattacharya, J. Villacorta, F. Diaz-Griffero and D. N. Ivanov, Journal of Biological Chemistry **291**, 21407 (2016).
[10] L. M. Koharudin et al., Journal of Biological Chemistry **289**, 32617 (2014).
[11] T. L. Diamond et al., Journal of Biological Chemistry **279**, 51545 (2004).
[12] X. Ji et al., Nature structural & molecular biology **20**, 1304 (2013).
[13] J. Yan et al., Journal of Biological Chemistry **288**, 10406 (2013).
[14] C. Zhu et al., Nature communications **4**, 2722 (2013).
[15] X. Ji, C. Tang, Q. Zhao, W. Wang and Y. Xiong, Proceedings of the National Academy of Sciences **111**, E4305 (2014).
[16] H.-M. Baldauf et al., Nature medicine **18**, 1682 (2012).
[17] B. Descours et al., Retrovirology **9**, 87 (2012).
[18] B. Kim, L. A. Nguyen, W. Daddacha and J. A. Hollenbaugh, Journal of Biological Chemistry **287**, 21570 (2012).
[19] L. Wu, Retrovirology **9**, 88 (2012).
[20] J. A. Hollenbaugh et al., PLoS pathogens **9**, e1003481 (2013).
[21] V. Tugarinov and L. E. Kay, Journal of biomolecular NMR **28**, 165 (2004).
[22] B. L. Kormos, A. M. Baranger and D. L. Beveridge, Journal of structural biology **157**, 500 (2007).
[23] B. L. Kormos, A. M. Baranger and D. L. Beveridge, Journal of the American Chemical Society **128**, 8992 (2006).
[24] A. Ghosh and S. Vishveshwara, Proceedings of the National Academy of Sciences **104**, 15711 (2007).
[25] G. Scarabelli and B. J. Grant, Biophysical journal **107**, 2204 (2014).
[26] I. Rivalta et al., Proceedings of the National Academy of Sciences **109**, E1428 (2012).
[27] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon and L. S. Caves, Bioinformatics **22**, 2695 (2006).
[28] O. F. Lange and H. Grubmüller, Proteins: Structure, Function, and Bioinfor-

matics **62**, 1053 (2006).
[29] M. Girvan and M. E. Newman, Proc. Natl. Acad. Sci. USA **99**, 8271 (2001).
[30] M. E. Newman, Proceedings of the national academy of sciences **103**, 8577 (2006).
[31] A. T. Van Wart, J. Durrant, L. Votapka and R. E. Amaro, Journal of chemical theory and computation **10**, 511 (2014).
[32] K. K. Patra, A. Bhattacharya and S. Bhattacharya, Proteins: Structure, Function, and Bioinformatics (2017).
[33] M. Girvan and M. E. Newman, Proceedings of the national academy of sciences **99**, 7821 (2002).
[34] C. H. Mauney et al., Antioxidants & Redox Signaling (2017).

Chapter 5

Phosphomimetic mutation T592E of SAMHD1: Strucural Stability and Dynamics

5.1 Introduction

Phosphorylation is a process of coupling or addition of phosphoryl groups to amino acids in proteins. Phosphorylation and its counterpart, dephosphorylation, are very critical for many cellular processes. It is especially important for protein functions as this modification activates (or deactivates) almost half of enzymes, thereby regulating their function. [1] [2] [3] Reversible phosphorylation of proteins is an important regulatory mechanism that occurs in both prokaryotic and eukaryotic organisms. [4] [5] [6] [7] Kinases phosphorylate proteins and phosphatases dephosphorylate proteins. Many enzymes and receptors are switched "on" or "off" by phosphorylation and dephosphorylation. Reversible phosphorylation results in a conformational change in the structure in many enzymes and receptors, causing them to become activated or deactivated. Phosphorylation usually occurs on serine, threonine, tyrosine and histidine residues in eukaryotic proteins. In prokaryotic proteins phosphorylation occurs on the serine, threonine, tyrosine, histidine, arginine or lysine residues. [8] [9] Histidine phosphorylation of eukaryotic proteins appears to be much more frequent than tyrosine phosphorylation. The addition of a phosphate ($PO^{+}4$) molecule to a polar R group of an amino acid residue can turn a hydrophobic portion of a protein into a polar and extremely hydrophilic portion of a molecule. In this way protein dynamics can induce a conformational change in the structure of the protein via long-range allostery with other hydrophobic and hydrophilic residues in the protein.

As we discussed in earlier chapters (see motivation section in Chapter 1) SAMHD1 contains a surface exposed target sequence (592TPQK595) for phosphorylation by cyclin-dependant kinase-1(cdk1)[10] [11] [12]. In–vivo studies led to a model which suggests that cycling cells show phosphorylation of SAMHD1 at T592, and reduced ability to restrict HIV-1.[13] [14] [15] This loss of restricting capability does not correlate with down-regulation of dNTPase activity of SAMHD1. While in–vitro studies have shown that the phosphomimetic mutations, T592D and T592E lose their ability to restrict HIV-1 in non-cycling cells but retain their dNTPase ativity, the unphosphorylable mutants T592V and T592A were able to restrict HIV-1 in non-cycling U937 cells. These contradictions in the results suggest that SAMHD1 employs a mode of restriction independent of lowering dNTP levels. On the contrary, the recent studies suggested that structural changes in SAMHD1 associated with the presence of negatively charged phosphate group on the side chain of T592 lead to structural changes in the neighborhood of the cdk1-regulation site, which, in turn leads to impaired tetramer formation and hence a decrease in dNTPase activity in proteins. Whether phosphorylation or phosphomimetic mutations (T592E or T592D) cause an ensemble wide collapse of the tetrameric structure of the protein is still an open question.[16] [17] [18] In this chapter we investigate the effect of the T592E mutation on the structural stability of the tetramer in comparision with the wild type (wt) tetrameric system. In addition, a comparative study of correlation network analysis and allosteric information flow between the T592E mutated protein and wt protein is also carried out.[19] [20] [21] [22] [23]

5.2 Methods

5.2.1 System preparation

The starting conformations of the explicit solvent simulations were based on the high resolution crystallographic structures (PDB Code 4TNR[24]) of the tetrameric SAMHD1 complex with Allosite 1 occupied by GTP and the Allosite 2 occupied by dATP. Three of the four catalytic sites in the crystal structure are occupied by dATP while the fourth (in subunit A)is vacant. The initial structure was generated from the original crystal structure. The crystallographic waters were retained along with the Mg^{+2} ions coordinated by allosteric site ligands. The unresolved portions

in the loop (278-283) were inserted in the protein structures whereas the missing N terminal and C terminal residues were ignored. The four R206 and N207 residues were mutated back to histidine and aspartate in accord with the sequence of the *wt* SAMHD1(Uniprot Q9y3Z3-1). Disulfide bonds were introduced between residues 341 and 350. The assembled system was immersed in $\sim$ 59000 pre-equilibrated water molecules. Na^+ and Cl^- ions were added at random positions to bring the net charge of the system to zero. The system consisted of $\sim 210,000$ atoms measured 13 nm $\times$12 nm $\times$14 nm. The T592E mutant was created by replacement of the threonine residue by glutamic acid at the residue position 592 on all four chains. Additional ions were added to restore the electrical neutrality of the T592E mutated system. This mutated system corresponded to the *wt* protein, that is, all the allosteric sites were occupied by GTP/dATP and three of the catalytic sites were occupied by dATP.

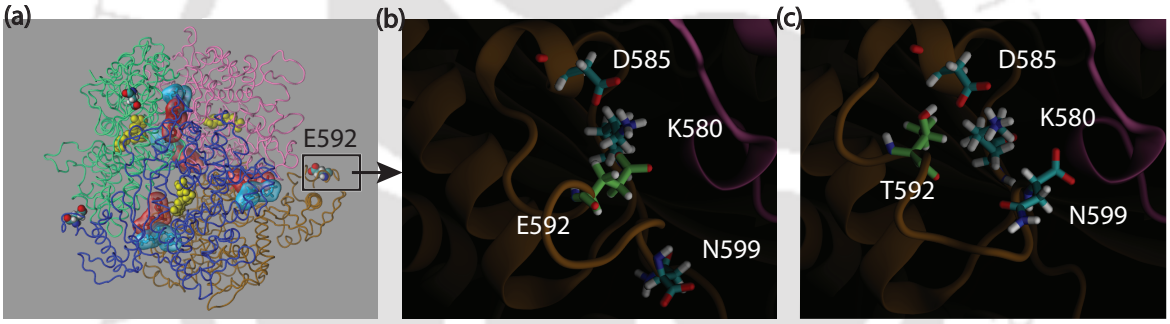

Figure 5.1: Ribbon representation of the T592E mutant with the location of the E592 residue indicated by a box. The chains A,B,C and D are colored in brown, pink, blue and green. The allosite GTP and dATP are represented by blue and green surfaces respectively, the dATP bound to catalytic site is represented by yellow spheres. The mutated residue E592 is in stick representation. (b) Snapshot showing the local environment of E592 in chain A. (c) snapshot of environment of T592 residue in the *wt* along with the charged residues in its vicinity.

5.2.2 General MD methods

All simulations were performed using the NAMD 2.9[25] [26] package with the CHARMM31 force fields[27] [28]. Analysis was performed with BIO3D package[29] and VMD.[30] All MD simulations were carried out using periodic boundary conditions and particle-mesh Ewald electrostatics[31] for long range electrostatic calculations. The SETTLE[32] and RATTLE[33] algorithms were employed to constrain the covalent bonds

involving the hydrogen atoms. The operational parameters included a 2 fs timestep while the cutoff and switching distances were set to be 12 and 10 Å respectively. The T592E mutated system was minimized for 3000 steps using the conjugate gradient method and then equilibrated in the NPT ensemble using Nosé-Hoover Langevin piston pressure control[34] at 295K for at least 5 ns. Following equilibration, two independent sets of MD calculation of the T592E system, with trajectory length of 100ns each were performed in the NVT ensemble with the temperature maintained at 295 K using the Langevin thermostat and then the comparisons were preformed with the *wt* trajectories already present (previous chapters).[35] [36]

5.2.3 Correlation Network Analysis

The correlation network construction and analysis were performed with Bio3D package[29]. To identify and characterize the coupled dynamics between the different parts of the *wt* and the T592 mutated SAMHD1 machinery, first the C_α residue-wise linear mutual information (LMI) were calculated as in equation 2.36. A separate analysis for each of the two MD trajectories of the T592E system resulted in two distinct matrices obtained as an ensemble average over multiple 50 ns windows along each 100 ns trajectory. Then the averaged consensus matrix for the T592E system containing LMI values was obtained following the same method used in Chapter 4 for the *wt* system. Later on, the consensus matrix was pruned to a cutoff 0.5 in order to create the correlation networks out of the consensus matrices. The network nodes represent the C_α atoms connected through edges weighted by the negative of the logarithm of the LMI values.

5.3 Effect of the T592E mutation on the structural stability and the dynamics at the active sites

5.3.1 T592E mutation has negligible effect on overall protein stability

The protein backbone RMSD calculated from the two trajectories of T592E variant shows a small increase from upto 2.2 Å (Figure 5.2 b) which is comparable to the RMSD for the *wt* system. An analysis of RMSD computed for the individual chains shows a noticeable increase of fluctuations ($\sim$ 3 Å RMSD) in some chains (B and C) in trajectory 1 of the T592E mutant (Figure 5.2 c) compared to the *wt* system.

However the overall change in the RMSD over the last 50ns of both trajectories is small.

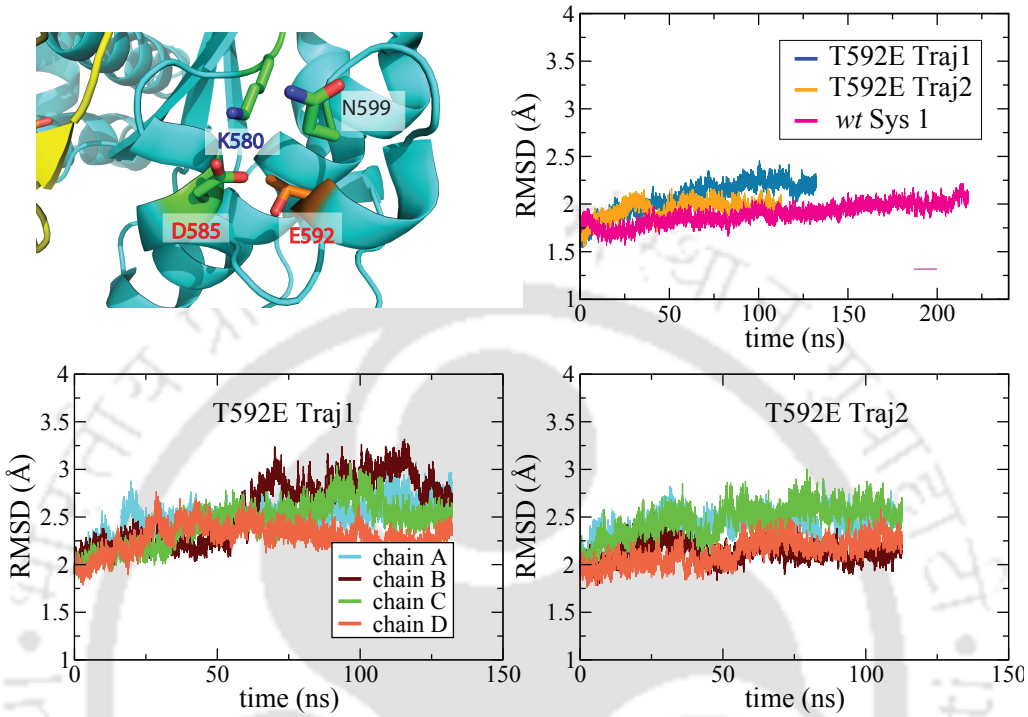

Figure 5.2: Snapshot of the mutation region showing the relative orientation of the E592 (mutated residue) and nearby charged/polar residues from the T592E variant. Variation of the protein backbone RMSD vs time with respect to the initial structure for (b) the two trajectories of T592E mutant (blue and orange) compared to *wt* system 1 (pink), (c-d) the four chains of trajectory 1 and 2 of the T592E mutant respectively.

5.3.2 Minor local fluctuation confined to short helix containing the mutation

The root mean square deviation of the protein backbone of the helical segments $560 - 574$ and $583 - 588$ near the mutation site in the T592E system exhibit little variation from the *wt* system. (see Figure 5.3 a and b). However a sharp increase in the RMSD of the $591 - 595$ segment (Figure 5.3 c) indicates an elevated motion in the residues. An analysis of the side-chain (Figure 5.3 f) is observed in both the *wt* as well as in the T592E mutant. All results in Figure 5.3 correspond to chain A of the T592E and the *wt* systems. Additional plots for B, C and D are also given in Figure 5.4.

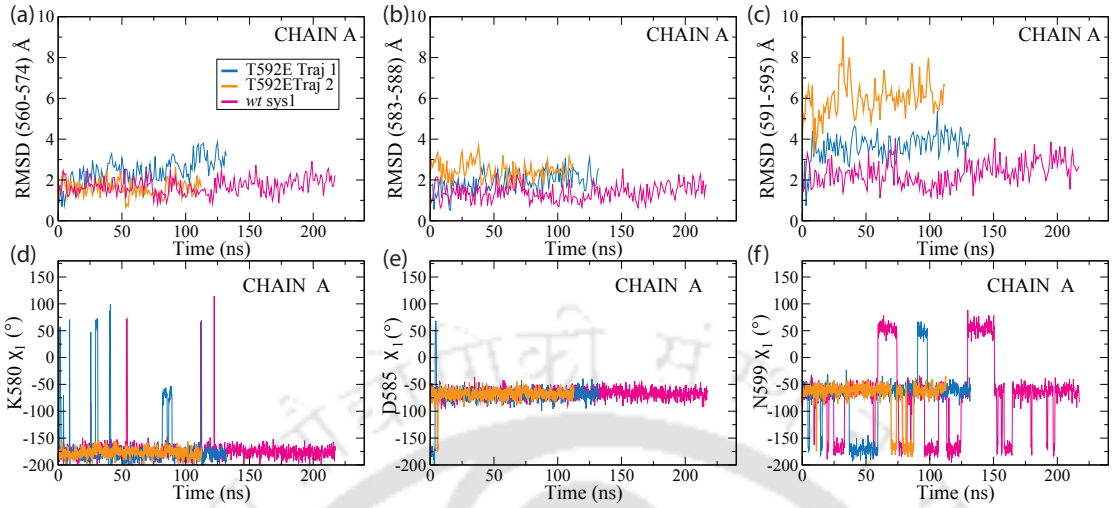

Figure 5.3: RMSD of helical segments (a)560-574 (b) 583-588 (c) 591-595, near the mutation site of the T592E trajectories and compared with the *wt*(magenta). the plots here are taken from the chain A. Rotameric state of the χ dihedral of charged/polar residues near the mutation site: (d) K580, (e) D585, (E) N599.

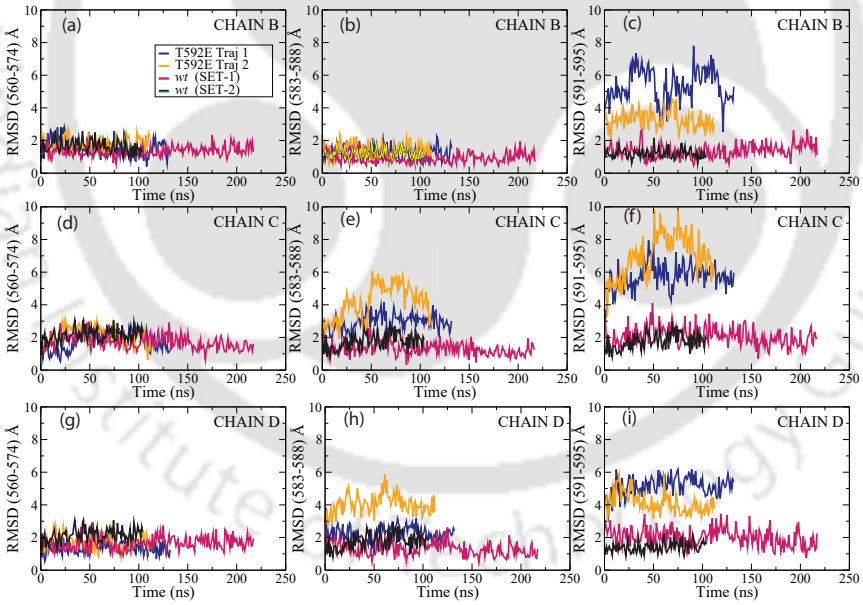

Figure 5.4: RMSD of helical segments (a)560-574 (b) 583-588 (c) 591-595, near the mutation site of the T592E trajectories and compared with the *wt* (magenta). The plots here are taken from the chain B, C and D.

5.3.3 Dynamics in allosteric site environment

An analysis of the allosteric site dynamics of the T592E variant carried out reveals little difference compared to the reference *wt* system. Figure 5.5 shows the compar-

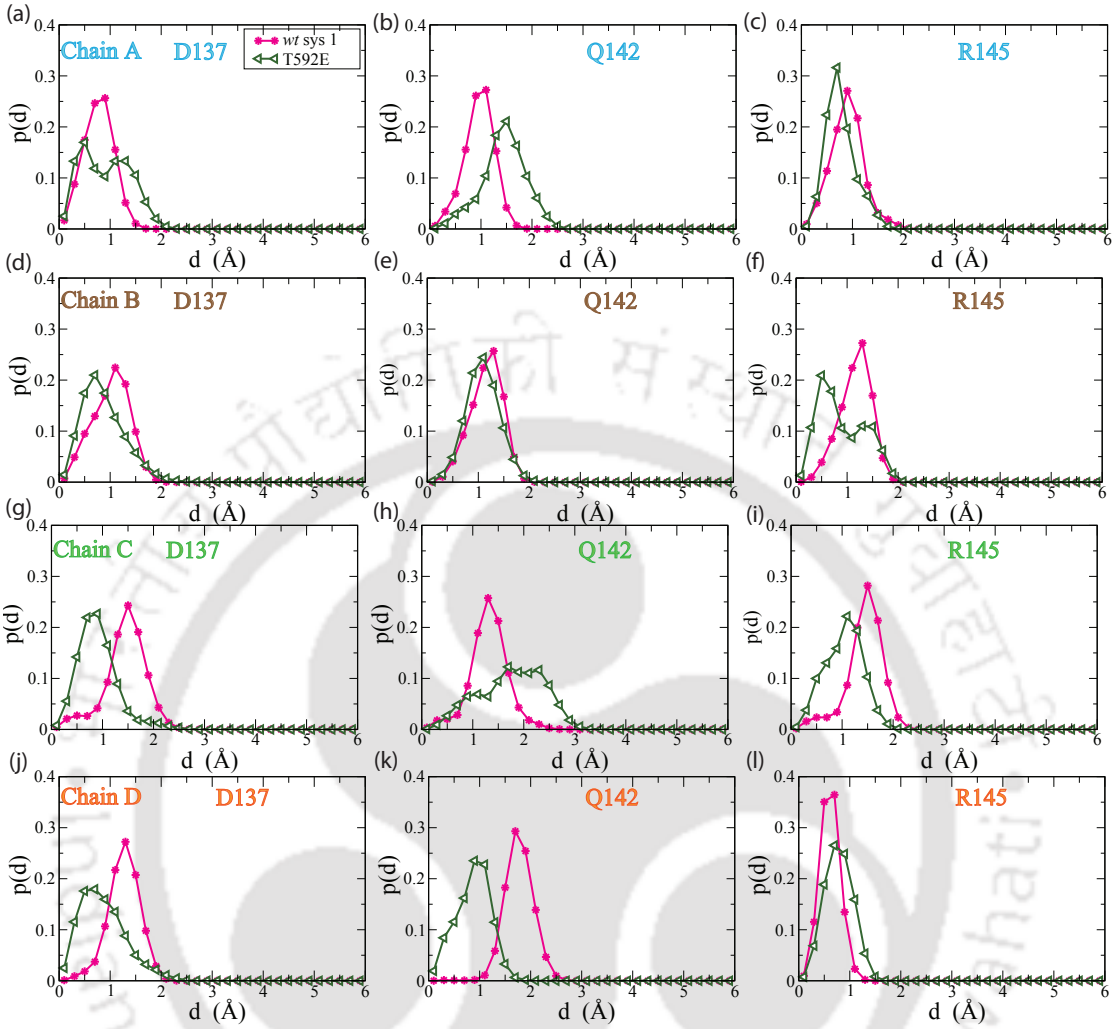

Figure 5.5: The distribution of the displacement of center of mass of residues D137, Q142, R145 near allosite 1 for the T592E and *wt* system. Panels (a-c) correspond to chain A, (d-f) to chain B, (g-i) to chain C and (j-i) to chain D respectively.

ison of the distribution of the displacements of center of mass of the residues D137, Q142 and R145. These residues are proximal to Allosite 1 region. Although minor shifts in the peak positions are noted in both systems, with the deviations being below 2 Å.

Figure 5.6 shows the distribution of the displacements of the residues flanking Allosite 2 (at the junction of chains A, D and B). Corresponding figures for the remaining three allosites are also calculated provided in Figure 5.7, Figure 5.8 and Figure 5.9 . In all cases, the differences between the T592E variant and the *wt*

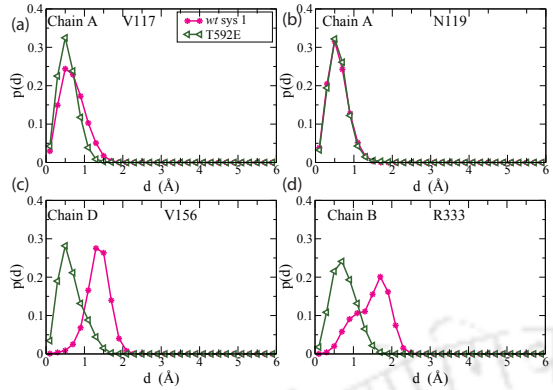

Figure 5.6: Distribution of the displacement of the center of mass of residues V117, N119, V156 and R333 surrounding allosite 2 at the interface of chain A, D and B for the T592E variant and the *wt* system.

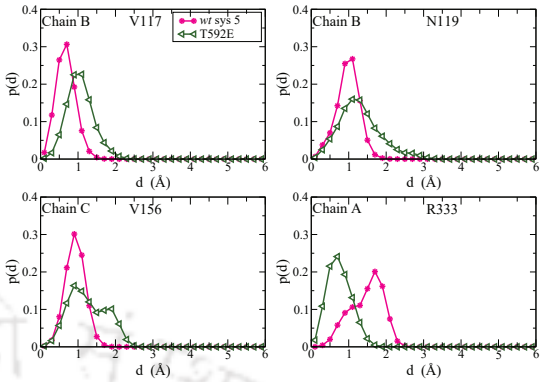

Figure 5.7: Distribution of displacements of centre of mass of residues V117, N119, V156, R333 surrounding allosite 2 at the interface of chains B, A and C for the T592E variant and the *wt* system.

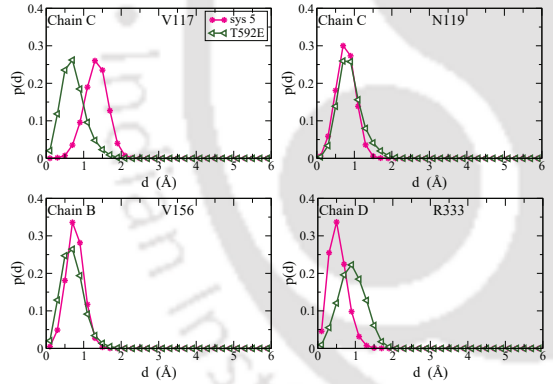

Figure 5.8: Distribution of displacements of centre of mass of residues V117, N119, V156, R333 surrounding allosite 2 at the interface of chains C, D and B for the T592E variant and the *wt* system.

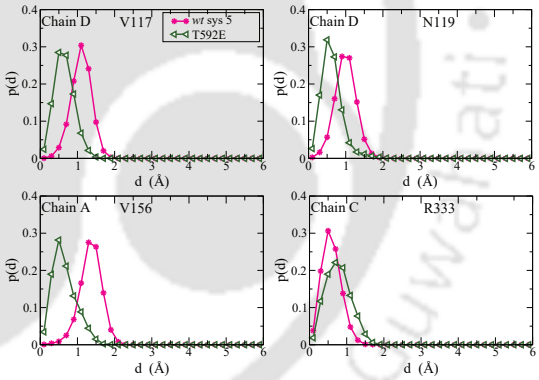

Figure 5.9: Distribution of displacements of cetre of mass of residues V117, N119, V156, R333 surrounding allosite 2 at the interface of chains D, C and A for the T592E variant and the *wt* system.

system are minor. The inter-residue distances in the allosteric pockets of the T592E variant (in Figure 5.10) are stable with small fluctuations ($\sim$ 1 Å) about the mean and exhibit little difference with respect to the *wt* system. Similarly, the rotameric states of the Allosite residues N119, V156 and R333 in the T592E mutant are in congruence with the corresponding states in the *wt* system.

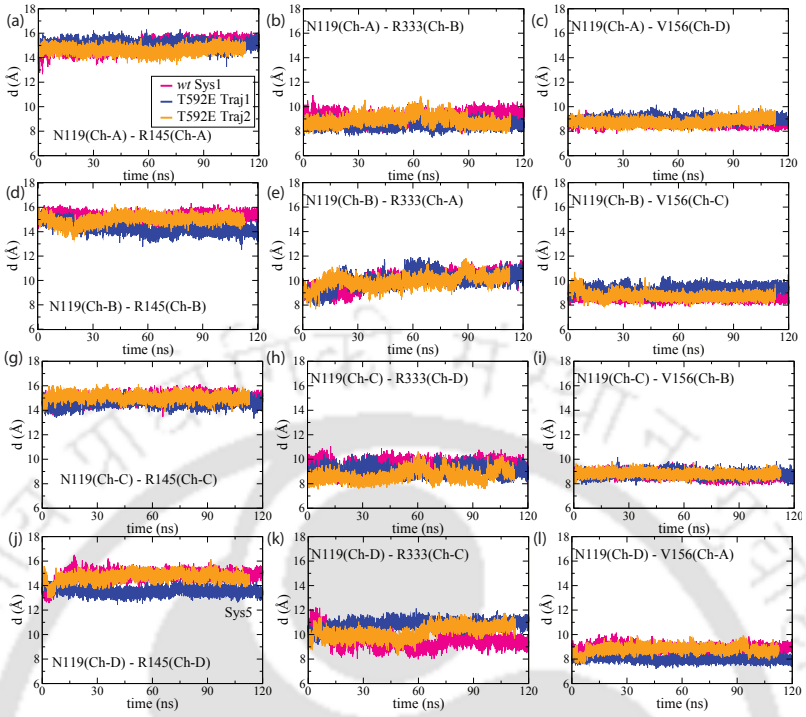

Figure 5.10: A comparison of inter-residue distances in the allosteric pockets between the trajectories of T592E mutant and the *wt* system. The plots indicate minimal fluctuations due to T592E mutations with respect to the *wt* system.

5.3.4 Inter-helix distance shows a little deviation from *wt*

In Chapter 3, we have investigated the inter-helix (E355-A373) distance between two adjacent monomers (see Chapter-3, Section 3.3.5, Figure 3.19). Changes in interhelix distances could be a precursor to the dissociation of the complex. For that reason, here also similar calculations were performed. The results are then compared to the *wt* system. The shortest distance between the (OD1/OD2) atoms of D361 and (NH1/NH2) atoms of R372 from the T592E trajectories in Figure 5.11 show that there is little difference form the *wt*. Apart from the occasional fluctuations that increase the separation of the atoms, the mean distance is steady. The center of mass separation of D361 and R372 are observed to be steady (Figure 5.12) in trajectory 1 of the T592E mutant. In trajectory 2 of the T592E, the distance between the D361 (chain D) and R372 (chain B) shows intermittent fluctuations of $\sim$ 3 Å about the mean. The measured distances in the other chains show little fluctuations indicating a stable inter-helix separation. The mean distance calculated from the T592E trajectories are close to the *wt* system confirming that the mutation at residue 592 does not affect the dynamics of the helices E355-A373 in the time scale simulated.

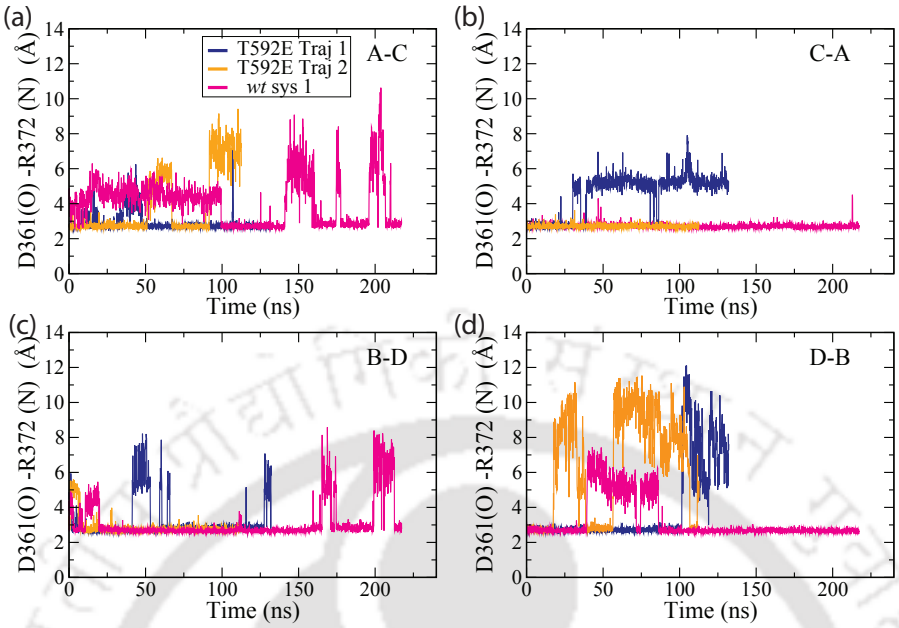

Figure 5.11: Shortest distances between the two carboxylate oxygen atoms of D361(OD1/OD2) and two nitrogen atoms (NH1/NH2) of R372 of chains (a) A and C, (b) C and A, (c) D and B and (D) B and D for *wt* (magenta) system compared to the two trajectories of the T592E mutant(blue and orange).

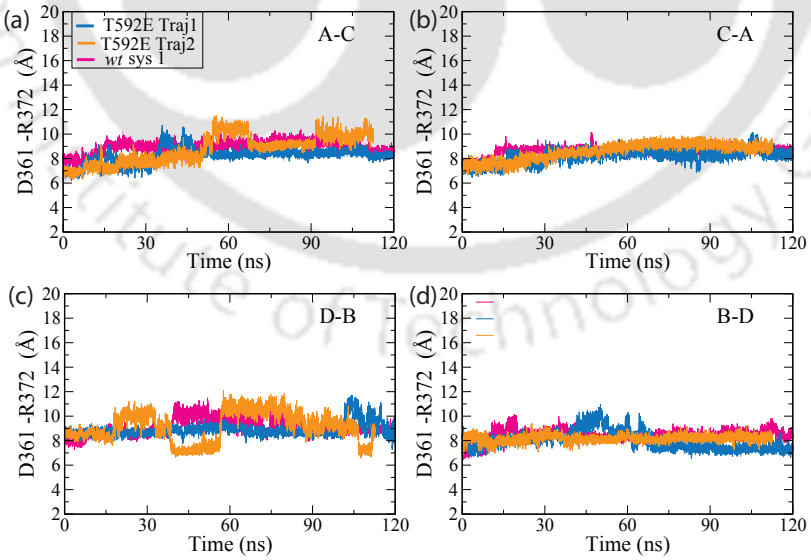

Figure 5.12: Distance between center of mass of D361 and R372 of chains (a) A and C, (b) C and A, (c) D and B and (D) B and D for *wt* (magenta) system compared to the two trajectories of the T592E mutant(blue and orange)

5.4 Alterations in allosteric communications in T592E variant

To understand the regulatory the role of CTD residue, Thr592, a community network analysis of the phosphomimetic variant T592E was performed.[29] [37] [38] In the previous study, MD simulations of the T592E mutant suggested local fluctuations near the site of the mutation that did not extend to the core of the complex. The two independent 100 ns trajectories are subjected to network analysis in order to examine if more subtle effects on the correlated dynamics are present. Figure 5.13 pictorially depicts the community partitioning in the *wt* and T592E mutant system.[39] [23]

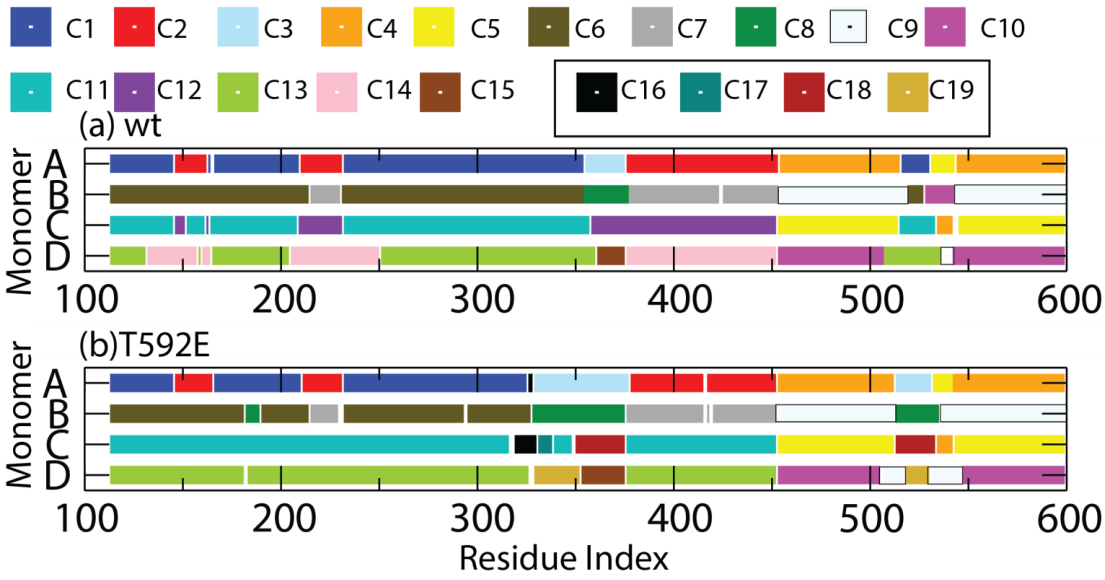

Figure 5.13: The community partitioning of the *wt* and T592E systems indicated by color coded horizontal bar. Each monomer (chain) is depicted by a bar that is colored according to the community. The color code of the extra communities not present in the *wt* system, but identified in the T592E variant are indicated by the box at the top.

In the *wt* system (Figure 5.13 a), a clear link between the CTD and the core is observed. The communities of Family F1 (that is, C1, C6, C11 and C13) each had a section in the CTD (approx. residues 516-530), sandwiched between the communities of the family F4, that provided a connection between the surface and the core. The connection is eliminated in the T592E variant, that is, the communication

between the core (Allosite) and the surface is significantly weakened in the T592E mutant. Instead, the members of the F3 Family (C3, C8, C15 and the new community C18) are connected to the CTD. Now the E355-A373 helices, which form the F3 family, were effectively dynamically decoupled from the other parts of the complex in the *wt* system. The isolation is concomitant to their role as structural anchors as described previously. In contrast, in the T592E variant, the communities of the F3 family have a segment extending to the CTD, indicating stronger dynamic coupling between the helices and the surface, which may be a precursor to the dissociation or loosening of the complex. [19] [40]

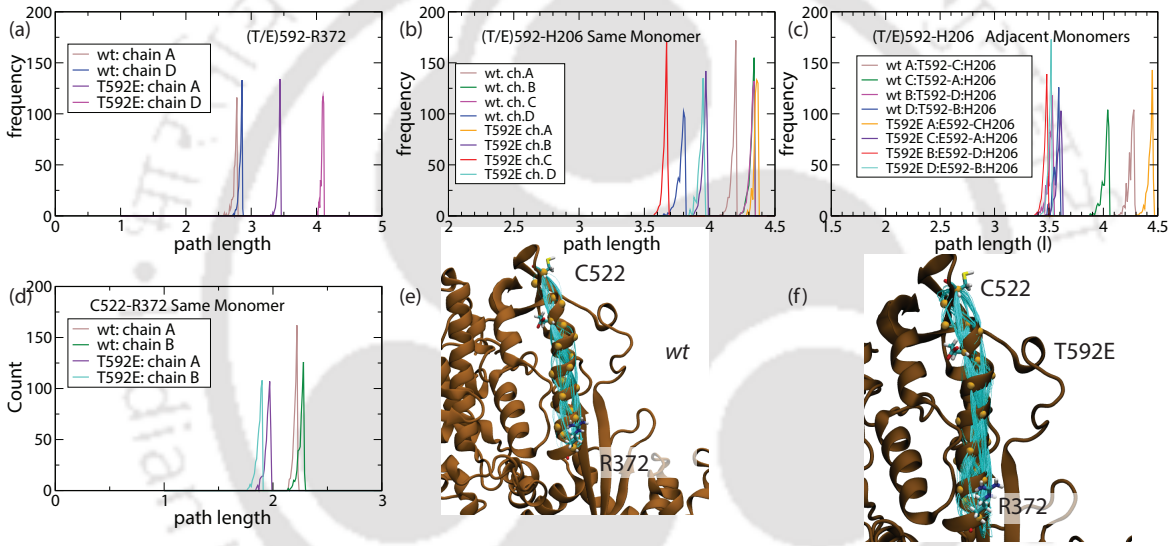

Figure 5.14: The path length distributions calculated for the *wt* system compared to those of the T592E variant. Source-sink pairs considered include (a) the regulatory site residue 592 and anchor-helix residue R372, (b) regulatory site residue 592 and the catsite residue H206 of the same monomer, (c) residue 592 and H206 of adjacent monomers (A-C and B-D) and (d) the surface site C522 and the anchoring helix residue R372. The calculated sub-optimal pathways between C522 and R372 in the (e) *wt* and (f) the T592E mutant.

To examine the link between anchoring helices and the CTD further, the direct information flow between the surface residue 592 and R372 on the helix in the *wt* and the T592E mutant was first considered. In each of the cases considered (Figure 5.14 a), the path lengths has increased in T592E mutant, indicating weaker correlation between E592 and the helix sites. Since the network topology in the T592E variant (Figure 5.13) shows the helix E355-A373 to be connected to the surface strip in three of four chains. We next consider the communication pathways between C522 and

the helix. In this case (Figure 5.14 d), the path lengths were found to be consistently smaller in case of T592E mutant indicating a tighter communication between the two sites. The snapshots of Figure 5.14(e,f) show that the paths involve the same nodes in both the *wt* and the T592E mutant. Thus the correlation between C522 and the helix terminus R372 is stronger in case of the T592E mutant even though the pathway itself is not changed.

To investigate other effects of the T592E mutation, the communication between the surface site 592 and the catalytic site H206 in the same monomer and in adjacent monomer are also considered in figure 5.14(b,c). This corresponds to paths computed for the *wt* in Figure 4.9 of previous chapter (Chapter 4). The results here are more ambiguous, although a shift towards shorter paths is observed in case of inter-chain communication in T592E mutant, suggesting stronger correlation between distant paths of the complex. It appears that the effect of T592E on the anchoring helices may trigger a loosening of the complex rather than a direct effect on the catalytic site. This is in line with results obtained by *Ivanov and coworkers* [17] [18] who have shown that the T592D mutation does not lead to collapse of secondary structure. The present studies indicates that the phosphomimetic mutations affects signal transmission to the protein core, but it is a subtle effect, and may be potentially connected to the redox-regulation implicated residue C522. [41] [42] [43]

Figure 5.16–Figure 5.15, presenting the PCA of the T592E variant, show the first three PCs to be dominated by the fluctuations of the C terminal residues 590-599. This is again, in line with the result by *Bhattacharya and Ivanov* [17] where the effect of the phosphomimetic mutation T592D was not found to propagate.

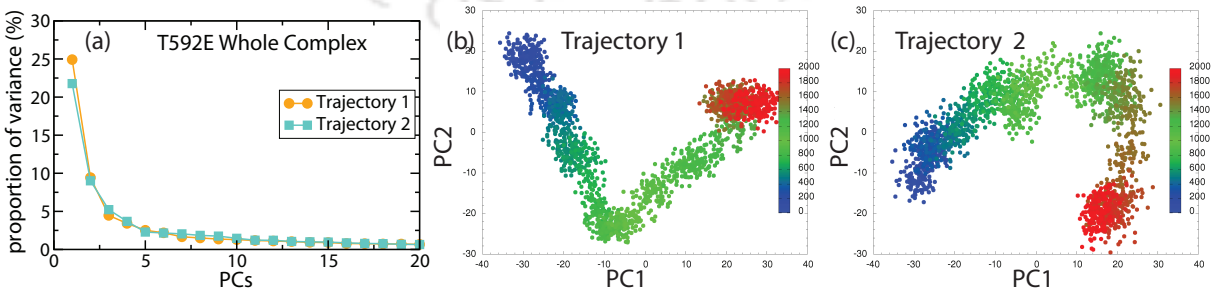

Figure 5.15: PCA of the T592E complex. (a) Proportion of variance contributed by PCs computed from the two trajectories. (b)-(c) Cross plots showing the dynamics of the complex projected on the top two PCs in case of the two trajectories.

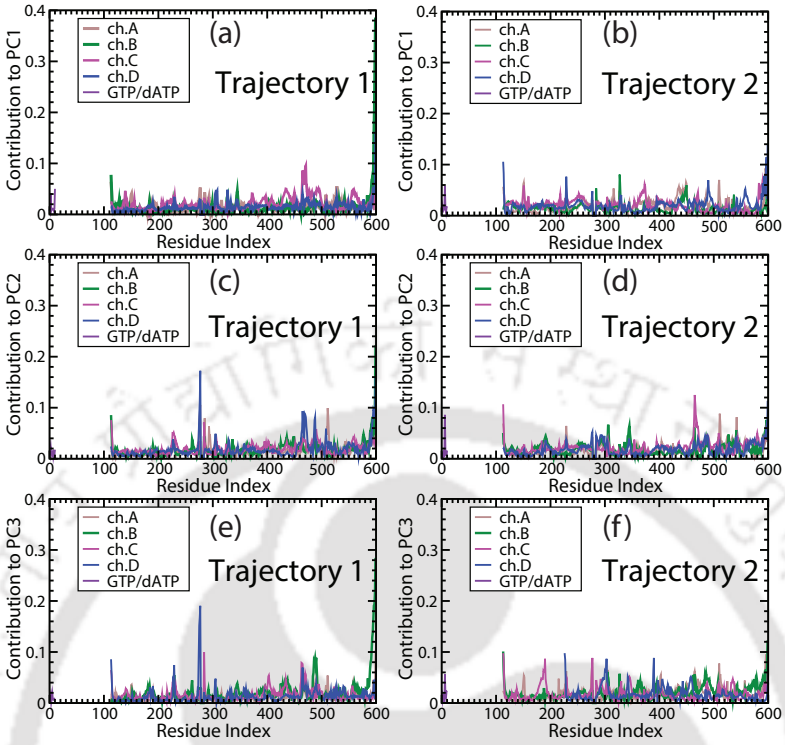

Figure 5.16: PCA of T592E complex. Residue-wise contribution to the total variance in case of first three PCs are presented in three rows. The columns indicate the two trajectories.

5.5 Conclusions

In contrast to conjectures about disruption of the folding at the tetramer interface region by the introduction of the additional negative charge at T592 by phosphorylation or phosphomimetic mutation, MD simulations showed that the T592E mutation does not alter the overall stability of the protein in the duration of the MD simulations. The mutation leads to enhanced fluctuation of the proximal residues which, however, does not extend to the entire chain. However, given the numerous caveats associated with molecular dynamics simulations and their ability to sample the conformational space of a large protein, it is can not be posited that the T592E mutation may not disrupt the structure of the tetramer at long time scales. Also, from the network analysis we find an altered dynamics in correlation space of T592E variant system. The communication networks between the CTD and the allosteric core are significantly weaker in comparison to the wt system. This signifies the incompe-

tent allosteric communications between surface and core of the mutant variant. In addition, we see the anchoring helices which compromises the communities of F3 family are no longer dynamically decoupled from the other parts of the complex (as we see this in wt system). There is a clear indication that for the T592E variant, the anchoring helices are dynamically coupled to the surface residues and are no longer isolated. As the dynamically isolated communities of the anti-parallel helices are necessary for the helices to act as anchor points for the protein, this results weakening in anchoring the complex and can be a instigator to the dissociation of the protein complex.

Bibliography

[1] A. P. Oliveira and U. Sauer, FEMS Yeast Research **12**, 104 (2012).

[2] F. Tripodi, R. Nicastro, V. Reghellin and P. Coccetti, Biochimica et Biophysica Acta (BBA) - General Subjects **1850**, 620 (2015).

[3] P. Vlastaridis et al., GigaScience **6**, 1 (2017).

[4] A. J. Cozzone, Annual Review of Microbiology **42**, 97 (1988), PMID: 2849375.

[5] J. Stock, A. Ninfa and A. Stock, Microbiological reviews **53**, 450 (1989).

[6] C. Chang and R. C. Stewart, Plant Physiology **117**, 723 (1998).

[7] D. Barford, A. K. Das, and M.-P. Egloff, Annual Review of Biophysics and Biomolecular Structure **27**, 133 (1998), PMID: 9646865.

[8] J. Cieśla, T. Fraczyk and W. Rode, Acta Biochimica Polonica **58**, 137 (2011).

[9] H. Saier Jr. M, J Mol Microbiol Biotechnol **9**, 125 (2005).

[10] A. Berger et al., PLoS pathogens **7**, e1002425 (2011).

[11] Z. C. Hartman et al., Journal of virology **81**, 1796 (2007).

[12] J. Schultz, P. Bork, C. P. Ponting and K. Hofmann, Protein Science **6**, 249 (1997).

[13] M. D. Zimmerman, M. Proudfoot, A. Yakunin and W. Minor, Journal of molecular biology **378**, 215 (2008).

[14] S. Kornberg, I. Lehman, M. J. Bessman, E. S. Simms and A. Kornberg, Journal of Biological Chemistry **233**, 159 (1958).

[15] C. A. Kim and J. U. Bowie, Trends in biochemical sciences **28**, 625 (2003).

[16] L. M. Koharudin et al., Journal of Biological Chemistry **289**, 32617 (2014).

[17] A. Bhattacharya et al., Scientific reports **6** (2016).

[18] Z. Wang, A. Bhattacharya, J. Villacorta, F. Diaz-Griffero and D. N. Ivanov, Journal of Biological Chemistry **291**, 21407 (2016).

[19] A. Ghosh and S. Vishveshwara, Proceedings of the National Academy of Sciences **104**, 15711 (2007).

[20] B. L. Kormos, A. M. Baranger and D. L. Beveridge, Journal of structural biology **157**, 500 (2007).

[21] B. L. Kormos, A. M. Baranger and D. L. Beveridge, Journal of the American Chemical Society **128**, 8992 (2006).

[22] G. Scarabelli and B. J. Grant, Biophysical journal **107**, 2204 (2014).

[23] I. Rivalta et al., Proceedings of the National Academy of Sciences **109**, E1428 (2012).

[24] X. Ji, C. Tang, Q. Zhao, W. Wang and Y. Xiong, Proceedings of the National Academy of Sciences **111**, E4305 (2014).

[25] J. C. Phillips et al., Journal of computational chemistry **26**, 1781 (2005).

[26] L. Kalé et al., Journal of Computational Physics **151**, 283 (1999).

[27] A. D. Mackerell, Journal of computational chemistry **25**, 1584 (2004).
[28] J. B. Klauda et al., The journal of physical chemistry B **114**, 7830 (2010).
[29] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon and L. S. Caves, Bioinformatics **22**, 2695 (2006).
[30] W. Humphrey, A. Dalke and K. Schulten, Journal of molecular graphics **14**, 33 (1996).
[31] P. F. Batcho, D. A. Case and T. Schlick, The Journal of Chemical Physics **115**, 4003 (2001).
[32] S. Miyamoto and P. A. Kollman, Journal of computational chemistry **13**, 952 (1992).
[33] H. C. Andersen, Journal of Computational Physics **52**, 24 (1983).
[34] S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, The Journal of chemical physics **103**, 4613 (1995).
[35] S. Nosé, The Journal of chemical physics **81**, 511 (1984).
[36] H. C. Andersen, The Journal of chemical physics **72**, 2384 (1980).
[37] M. Girvan and M. E. Newman, Proc. Natl. Acad. Sci. USA **99**, 8271 (2001).
[38] M. E. Newman, Proceedings of the national academy of sciences **103**, 8577 (2006).
[39] A. T. Van Wart, J. Durrant, L. Votapka and R. E. Amaro, Journal of chemical theory and computation **10**, 511 (2014).
[40] K. K. Patra, A. Bhattacharya and S. Bhattacharya, Proteins: Structure, Function, and Bioinformatics (2017).
[41] C. H. Mauney et al., Antioxidants & Redox Signaling (2017).
[42] J. K. Yao and M. S. Keshavan, Antioxidants & redox signaling **15**, 2011 (2011).
[43] S. M. Marino and V. N. Gladyshev, Antioxidants & redox signaling **15**, 135 (2011).

Chapter 6

MD studies of monomeric forms of the SAMHD1 wt and Cystine mutants

6.1 Introduction

In our previous MD work, we have studied the stability of SAMHD1 assembly and methods of allosteric communication within the assembled tetramer. We have also explored the consequences of phosphomimetic mutations of The592 principally because phosphorylation at T592 by cdk-1 has been postulated to be a regulatory switch for SAMHD1 dNTPase activity. All of these studies, however, have not yet touched upon other regulatory mechanisms of SAMHD1 and this remains a very important lacuna in the growing literature of this protein system. One of the most visible regulatory mechanisms is redox regulation. Glutathione (GSH) is a cellular tripeptide with free sulfahydryl groups which is found in the cytosol at a concentration of 5 mM. This has long been hypothesized as an agent to remove harmful reactive oxygen species (ROS) such as superoxides, etc . However, an ancillary role of GSH in redox regulation of proteins is increasingly becoming apparent.[1] [2]

A monomer of SAMHD1 complex contains 10 residues of cysteine (Cys), with the following residue numbers: 177, 198, 205, 266, 320, 341, 350, 522, 554 and 573. Out of these cysteine residues the residue numbers 177, 266, 341 are exposed to the surface and residues 341 and 350 could also be disulphide linked. However, we are mostly interested in the three cys residues C341, C350 and C522. This is because in the only dimeric structure of SAMHD1 solved via XRD (3u1n.pdb, by Goldstone, et al in 2011), C341 and C350 are linked by a disulphide bond[3]. This is not seen

in subsequent tetrameric structures. Further, C522 is a neighbouring Cys residue. Therefore, it is not implausible that redox modulation of the thiol sidechains of these residues could occur in cells - either in isolation, or in concert with each other. Secondly, significant gains in enzyme activity were seen in reducing conditions (such as incubation with TCEP and DTT) as reported in experimental assays of dNTPase activity by Wang et al[4] and Bhattacharya, et al[5]. In contrast, deactivation of the enzyme was observed after incubation with hydrogen peroxide (unpublished, Ivanov lab).

Extracellular, or secreted proteins frequently employ disulphide bonds to confer additional stability[6]. For cellular proteins such as SAMHD1, the presence of an intramolecular disulphide bond could also be explained away in terms of stability had it been formed across monomeric units. But this is not the case. Thus, the question is: why does the protein have exposed (or quasi exposed) sulfahydryl groups. Is there a regulatory role played by these Cys residues?

A full and exhaustive study of redox regulation of SAMHD1 will involve identifying the redox active Cys residues, assembling GSH groups on the sidechains of these residues and performing extensive MD simulations. However, the first step is to simply ask the question: are these Cys sidechains critical in protein assembly and stability. In order to answer this question, we have mutated these three Cys residues to Serine and have evaluated MD trajectories of the mutants.

6.2 System preparation

The starting conformations of the explicit solvent simulations were based on the high resolution crystallographic structures (PDB code 4TNR[7]) of the tetrameric SAMHD1 complex. Only one monomer has been picked out from the tetrameric complex and the allosteric and catalytic site molecules are excluded from the test monomer. The initial structures for six different systems with varying Cys mutants are created. Among these, three systems are prepared by selectively mutating Cys to Ser : C341S, C350S and C522S. In one system, C522 is glutathionylated (that is, bonded to GSH). The remaining two systems are *wt* (that is, without mutations), with one of them having no disulfide bonds between the cysteine residues ($wt1$) and the other with the residues 341 and 350 connected by a disulfide linkage ($wt2$). The details of the created monomeric systems with simulation details are shown in given Table 6.1 . The crystallographic waters were retained in all five systems. The unresolved portions in the loop (278-283) were inserted in the protein structures

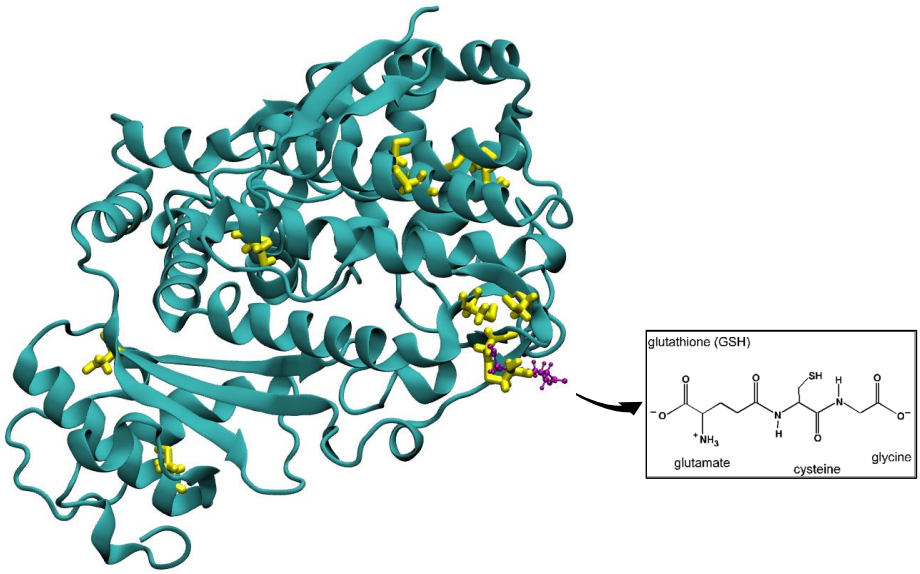

Figure 6.1: The protein segment of SAMHD1 monomer represented by cartoon representation, the yellow segment shows the Cys residues and the purple colored molecules shows the glutathione (GSH) which binds with residue C522. An elaborate structure of GSH molecule is shown in right side.

whereas the missing N terminal and C terminal residues were ignored. The four R206 and N207 residues were mutated back to histidine and aspartate in accord with the sequence of the *wt* SAMHD1 (Uniport Q9Y3Z3-1). Each system was immersed in pre-equilibrated TIP3 water molecules. Na^+ and Cl^- ions were added at random positions to bring the net charge of the system to zero. Each system consists of $\sim 70,000$ atoms measured 9 nm $\times$ 8 nm $\times$ 10 nm .

6.3 General MD methods

All simulations were performed with NAMD 2.9[8] [9] package with the CHARMM31 force fields[10] [11]. Analysis was performed with Bio3D package[12] and VMD[13]. All MD simulations were carried out using periodic boundary conditions and particle-mesh Edward electrostatic calculations[14]. The SETTLE[15] and RATTLE[16] algorithms were employed to constrain the covalent bonds involving hydrogen atoms. Operational parameters include a 2 fs time-step while the cutoff and the switching distances were set at to 12 and 10 Å respectively. Each system was minimized for 3000 steps using conjugate gradient method and then equilibrated in the NPT ensemble using the Nosé-Hoover Langevin piston pressure control[17] at 295 K for at least 5 ns. Following equilibration, two sets of MD calculations, with trajectory

length of 200 and 300 ns respectively were performed in the NVT ensembles with the temperature maintained at 295 K using the Langevin thermostat. The data was recorded at 10 ps intervals [18] [19].

System	Cys mutations	Allosite and Catsite molecules	Glutathione (GSH) attached to resid 522	NVT simlation-Set 1 length (ns) at $295K$	NVT simlation-Set 2 length (ns) at $295K$	NVT simlation length (ns) at $500K$
M1	C341S	Absent	Absent	200	300	20
M2	C350S	Absent	Absent	200	300	20
M3	C522S	Absent	Absent	200	300	20
GWT	No	Absent	Present	200	300	20
WT1 (with-out DS)	No	Absent	Absent	200	300	20
WT2 (341-350 DS linked)	No	Absent	Absent	200	300	20

Table 6.1: List of MD simulations performed for all the systems created from the monomeric SAMHD1 complex.

6.4 Results

To understand the role of cysteine (Cys) to serine (Ser) mutations on the monomeric form of SAMHD1, two sets of independent series of all atom molecular dynamics with 200 ns and 300 ns trajectory lengths of each systems were performed and then analyzed.

6.4.1 Fluctuation builds up at CTD region for particular mutations

We first analyzed the root mean square fluctuations (RMSF) of the C_α atoms of different monomeric mutants simulated in two independent sets. Figure 6.2 represents the comparisons of the average RMSF of two sets of simulation data for individual monomeric systems. The strong fluctuations were observed at the tail

of the CTD region (570-599) in C350S mutant (color blue). It indicates incipient structural instability of the monomeric variant (C350S) originating at CTD. C341S (green) also shows deviation in RMSF for residues in CTD (around 520-530). In no simulation did we connect 341 to 522 by disulfide bonds. Though residue number 522 is not sequentially close to either C341 or C350, it may approach close to the two cysteine residues geometrically, giving rise to the possibility of a redox switch operating between the three cysteine residues.

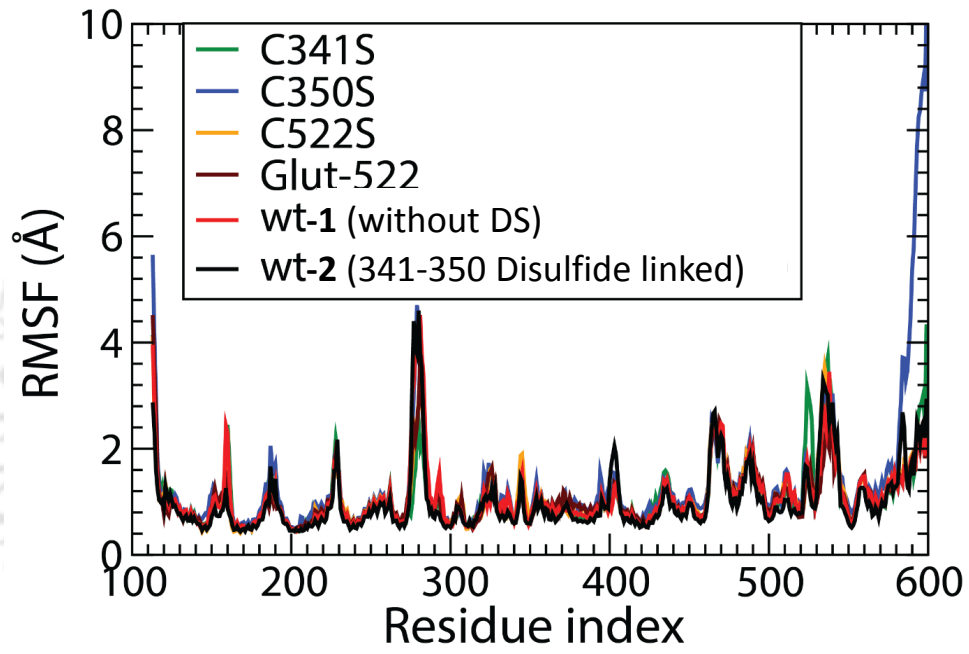

Figure 6.2: RMSF of monomeric units, averaged over two sets of independent simulations .

Disruption occurs in stability of C350S mutant after 200ns

We then, performed an analysis of the root mean square deviation (RMSD) of the proteins with respect to their initial crystal structures. The analysis was done in two ways: first by taking the entire monomer into considerations, and second, by considering only the CTD (residue 455-599). The RMSD calculation shows a sound degree of stability for both the monomeric CTD as well as the whole monomeric unit till 200 ns as shown in Figure 6.3, panel (a and c) for Set-1 simulations. In contrast, a significant disturbance in the RMSD is observed in the Set-2 simulations, after 200 ns for the C350S variant (See Figure 6.3 panel b, d) of the monomeric unit.

The RMSD increases from about $\sim$ 2.5 Å to upto $\sim$ 6.5 Å. Similarly for the CTD, the increase is from about $\sim$ 3 Å to $\sim$ 8 Å and in some instances upto $\sim$ 10 Å. The effect is most prominent between residues 580-599 (Figure 6.3 d).

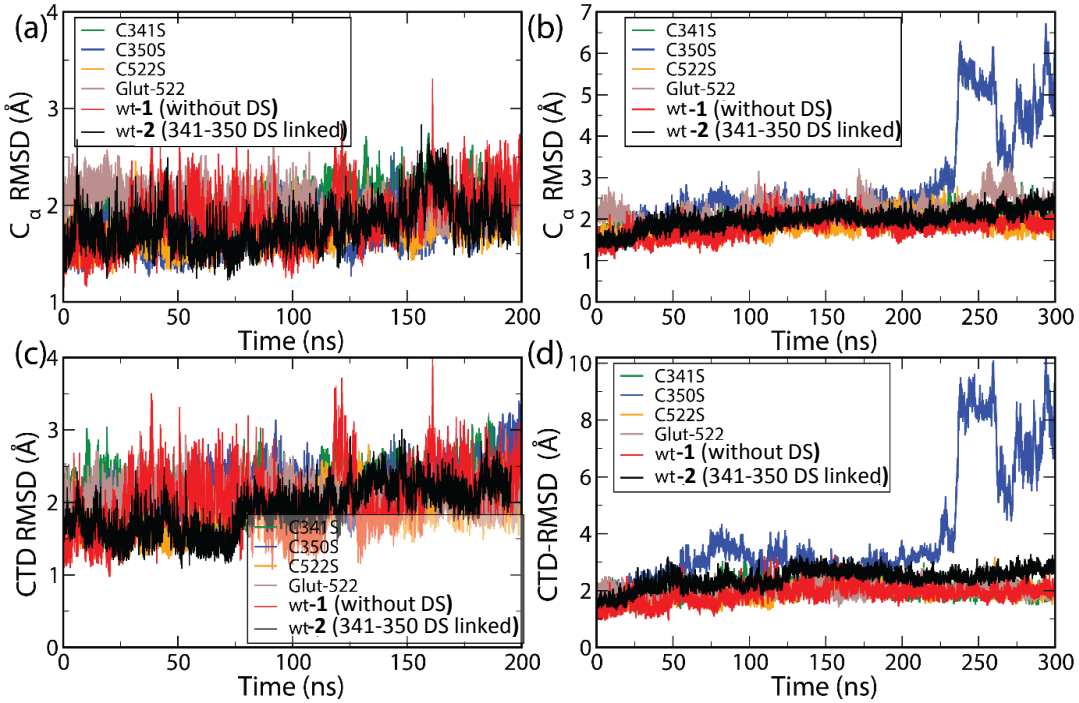

Figure 6.3: RMSD of the C_α atoms of the whole chain for the systems studied (a) Set-1 and (b) Set-2. RMSD of C_α atoms of only the CTD (residues 455-599) from (c) Set-1 and (d) Set-2 simulations.

6.4.2 Correlation Analysis of the SAMHD1 monomers

To further analyze the effect of $Cys \rightarrow Ser$ mutations on the correlated dynamics between different parts of the monomeric units we have performed dynamic correlation analysis by Bio3D package and constructed the correlation coefficient matrices for each systems from their corresponding simulated trajectories. We have followed the same process as we did in our previous chapters (Chapter 4).

First the C_α residue-wise linear mutual information (LMI) was calculated using, $I_{lin}(x_i, x_j) = \frac{1}{2}\left[\ln\det CM_{(i)} + \ln\det CM_{(j)} - \ln\det CM_{(ij)}\right]$ where CM_i is the covariance matrix for the displacement of C_α atom of the i^{th} residue and CM_{ij} is the pair covariance matrix for residues i and j. Here the averaged consensus correlation matrices for each systems were constructed over a total 500 ns of MD trajectories including Set-1 and Set-2 simulations.

Figure 6.4 represents the cross correlation coefficient matrix (Linear Mutual Information) of C_α atoms for (a) the $wt1$ (no Disulfide links) system, (b) C350S mutant system and (c) the difference between the matrices of (a) and (b). In panel (a), the regions of moderate to high correlations are indicated by boxes labelled R1-R8. Violet and orange patches in panel (c) indicate the regions of maximum difference between the wt and the C350S mutant. Note that the entire CTD domain, particularly residues have high correlation (R1, R2 and R3 in panel a). In particular, there is a moderate correlation between the strips of residues 460-480 and 570-599 (region R3 in panel a). This is weakened in the C350S variant. Additionally the correlations between the residues between 570 to 599 (region R2) are also found to be reduced in the C350S variant. However, the correlations in the core (major lobe) is not affected much by the C350S mutation.

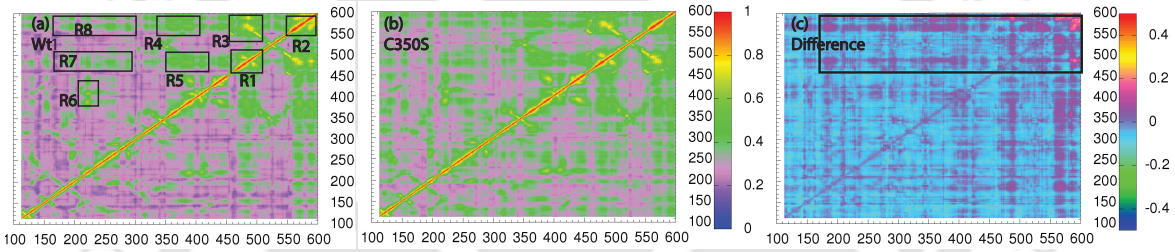

Figure 6.4: Cross correlation network matrices of (a)WT, (b) C350S system and (c) their difference matrix .

We have also extended the comparisons of correlation matrix of the $wt1$ system to the other $Cys \rightarrow Ser$ mutated systems. The Figure 6.5 represents the cross correlations matrices and their differences from the correlation matrix of $wt1$ system. The C341S mutation shows a general loss of correlations in the minor lobe as indicated by the violet patches in Figure 6.5(b) (the box labelled B1). In addition, there is a noticeable diminution of correlations between the strips 340-400 and the minor lobe (451-599) indicated by the box B2. Correlations within the strip 340-400 are also reduced. We find that C522S mutant shows very little difference (Figure 6.5 c and d). Interestingly, the glutationated C522 system shows reduced correlations between the minor lobe and the strip 340-370 (see the box B4 in Figure 6.5 f). Thus the correlation network is more severely disrupted in the glutathionated C522 system compared to the C522S variant.

To see the difference in the dynamics of correlation space between atoms of the SAMHD1 tetrameric system and monomeric system, we did a comparison of the

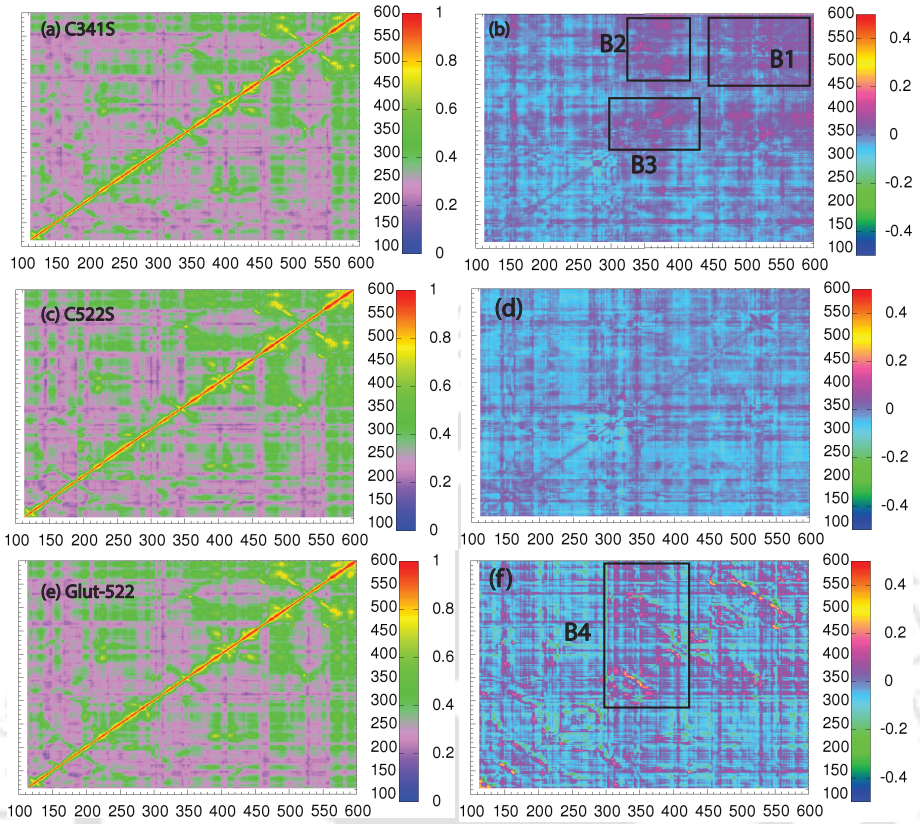

Figure 6.5: Cross correlations matrices of (a) C341S and (b) Difference between CC matrix of the *wt* (without Disulfide bonds, Figure 3a) and C341S, (c) C522S and (d) Difference between *wt*1 and C522S , (e) Glut-522 and (f) difference.

correlation coefficients within a monomer from the *wt* tetramer simulations (i.e, the other chains are present but not considered in the figure) to the *wt*1 without DS links monomeric system. Figure 6.6 represents difference between the *wt*1 monomer (without disulphide linkage between C341-C350) and the *wt* tetramer (having those linkages). What is interesting, is that apart from the specific anchors, most of the map is blue or close to blue (indicating lower correlations) in panel (a). This indicates that the correlations are suppressed or absent. In contrast, if we look at the monomeric systems, one can easily find lots of "spurious" cross correlations present. Thus, tetramerization appears to quench the spurious correlations and channels the correlations through a network of pathways.

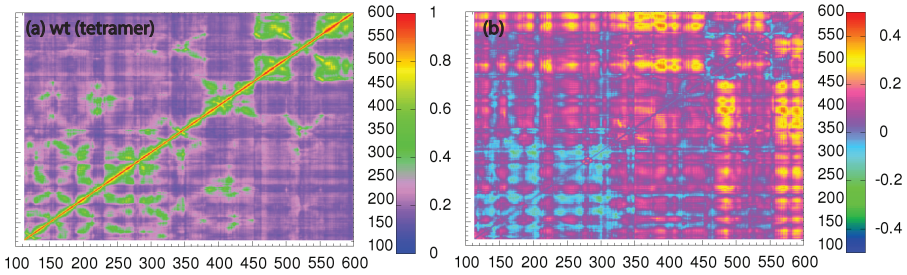

Figure 6.6: Cross correlation network map of (a) *wt* (tetramer), (b) difference map between *wt*1 (without DS link monomer) and the *wt* tetramer with DS links .

6.4.3 Scatter plots reveals potential switching mechanism of the system

Scatter plot referred to Figure 6.7 showing distances between C_α- atoms specified. The *wt*1 (without any disulfide linkages), *wt*2 (with C341 and C350 connected by disulfide bond), C341S, C350S, C522S, Glutathionated C522 systems are represented by red, black, green blue, orange and brown symbols respectively. Only those systems with disulfide bond between C341 and C350, that is, *wt*2 (black) and C522S (orange), have a shorter distance between C341 and C350 (about 5.6 Å). The other systems register a larger distance (about 6.4 Å).

Note that in panel (a), the *wt*2 system with disulfide linked C341-C350 (black circles), the distance between C341 and C522 shows a bimodal distribution. A similar distribution is seen in case of C341S (green diamonds) and the C350S (blue stars) variants, where the distance between the residue 341 and C522 has a bimodal distribution. The bimodal character of the distribution indicates a switch [20] [21]. The 522 modified systems (C522S and the glutathionylated C522) do not show such a distribution. The 350-522 distance does not show a bimodal distribution for the same systems (panel b). It is not surprising that the C341-C350 C_α distribution will be much the same when the C341-C350 disulfide bond is disrupted, as compared to *wt*1 (red). The disulfide bond is simply disallowed by mutating C341 or C350 to serine. However, it is not at all obvious how the attachment of a glutathione molecule to C522 (brown) manages to have the same effect as a disruption of the C341-C350 disulfide linkage. The bimodality of the C341-C522 distance is lost when the C522 residue is either glutathionated, or mutated to serine (Ser). This raises the question of whether the bimodality was the result of a transient interaction between C341 and C522? Furthermore, could this be connected to a potential C341-C350 disulfide linkage? Experimental investigations may clarify the questions raised by

the plot.

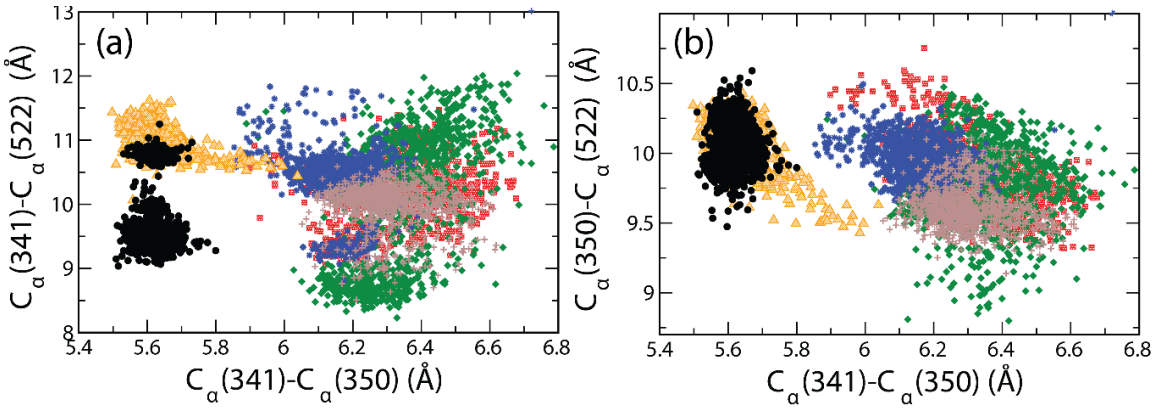

Figure 6.7: Scatter plot representing distances between C_α atoms specified. The wt1, wt2, C341S, C350S, C522S, Glutathionylated C522 systems are represented by red, black, green blue, orange and brown symbols respectively.

6.4.4 Dynamic correlation increases in $Cys \rightarrow Ser$ monomeric mutants

Figure 6.8 represents the cross correlation of the residue (a) 341, (b) 522 and (c) 350 and (d) 530 with all other residues (calculated using the C_α atoms). The $wt1$ (without disulphide bonds), $wt2$ (with 341-350 connected by disulphide bonds), C341S, C350S, C522S and C522-glutathionated systems are represented by red, black, green, blue, orange and brown lines respectively. In addition the the correlations in the wt tetramer (chain B only) is represented by purple lines. In all cases, the correlations in the wt tetrameric system are generally quenched except for specific correlations such as between C341 and C350. The baseline of the correlations in the tetrameric system is below 0.2. In contrast, for all the monomeric systems studied, the correlations between the specific residues (341, 350, 522 and 530) with the rest of the chain are generally higher. The C350S system (blue) shows a noticeably higher correlation of S350 with other residues (panel c) compared to all the other systems. Also note that we have calculated the correlations with residue 530. The importance of the residue 530 in the allosteric handshake is already described in the previous chapter (Chapter 4). Here we noted a peak in the correlations in panels (a and c) next to C522. The peak corresponded to the residue 530. Hence we also calculated the correlations with respect to 530 in panel (d) as it appears to play an important role in channeling any allosteric signal to the interior.

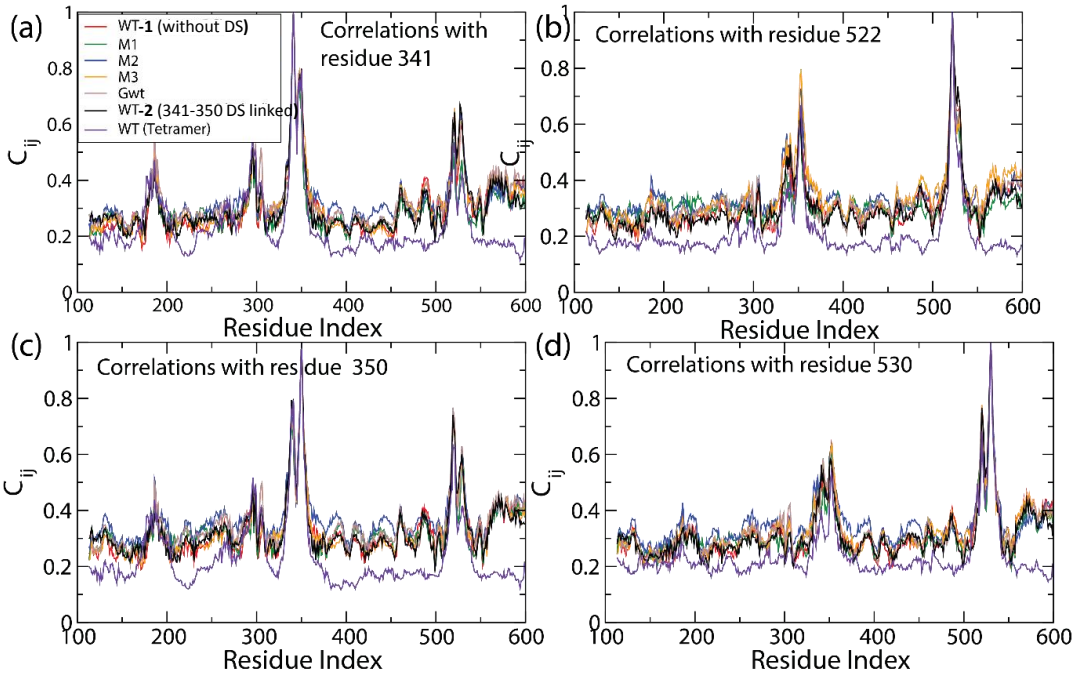

Figure 6.8: Cross correlation of the residue (a) 341, (b) 522 and (c) 350 and (d) 530 with all other residues (calculated using the C_α atoms). The wt1, wt2, C341S, C350S, C522S and C522-glutathionylated systems are represented by red, black, green, blue, orange and brown lines respectively.

6.4.5 Compactness of the monomeric structures

Next we have calculated the radius of gyration of the monomeric variants as a measure of the compactness of the structures. Note that in Figure 6.9, in case of the tetramer, the subunit (as a whole) has a distinctly smaller radius of gyration compared to all monomeric systems (panel a, purple line). The peaks are broad or shifted compared to the tetrameric system in case of the catsite residues (panel bottom, left). The *wt*2 monomer (with disulfide linked C350 and C341) has a broad distribution with two peaks suggesting that the shape of the catalytic pocket is different from the tetramer.

In all four panels of Figure 6.10, large deviations seen between the tetrameric system (purple) and the rest of the monomeric system including the *wt* ones. All indicate that the geometry of the allosteric site and the catsite is considerably altered in the monomeric systems. In panel (a), broader peaks with shifted centers show that the allosteric site is distorted in the monomeric systems. Hydrogen bonds between R318 and D120 were found to be crucial to intrachain communication network

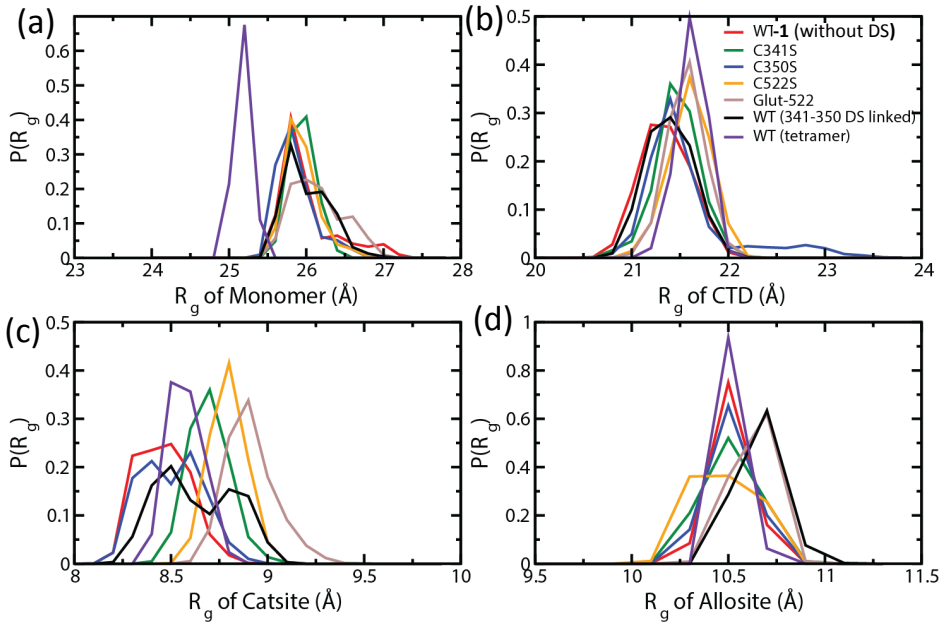

Figure 6.9: Radius of gyration of (a) entire monomer, (b) C terminal domain, (c) selection of catsite residues (374, 375, 206, 207, 311, 319, 218, 210, 149 and 150) and (d) allosteric site residues (117 to 145).

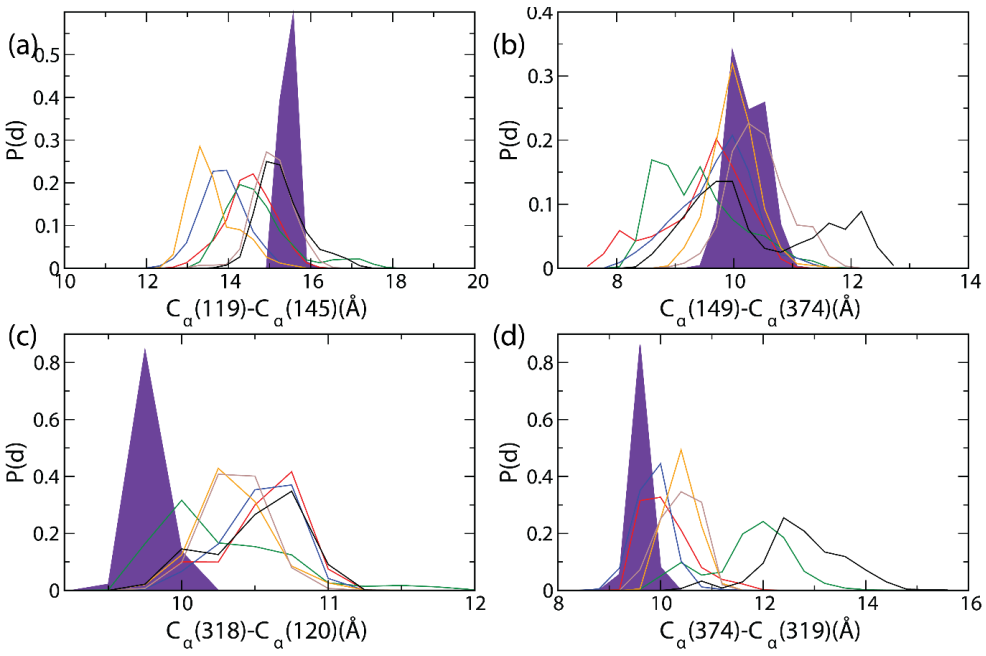

Figure 6.10: Distances between C atoms of select residues: (a) 119 and R145 at the allosteric site, (b) 149 and 374 at the catalytic site, (c) 318 and 120 and (d) 374 and 319.

between allosite and catsite residues in previous study. But the distance between the two residues is seen to undergo large fluctuations in the monomeric systems (see panel c).

6.5 Discussion

The correlation network analysis in Chapter 4 hinted the significance of the cysteine residues. The three cysteines, C341, C350, C522, appear to play an important role in the flow of allosteric information, particularly between the CTD and the major lobe. There exists the possibility of a redox switch operating between the three, which can lead to a change in the dynamics of the system. We have performed a series of all-atom MD simulations to examine how the dynamics of the monomeric forms of SAMHD1 differ from the tetrameric system and also to illuminate the role of cysteine residues in protein assembly and stability.

Our analysis revealed large fluctuations in the CTD of the C350S mutant. An analysis of the dynamic correlations in all the monomeric variants shows the C522S to be least affected compared to the other mutants and even the glutathaionylated C522 system. This is an interesting point because experimental evidence (C522S) does suggest that this mutant is fully capable of tetramerization, as opposed to the impaired tetramerization ability of C341S and C350S. Our results indicate a possible answer; the C522S mutant simply has fewer short and long range disruptions in its secondary structure.

Finally, we find C350S mutation has a significant effect on the dynamics of the whole momomeric system compared to other mutants. Strong correlation between residue C350 and C522, even in the absence of disulfide links is really interesting and in further studies may provide more clues on dynamical cross correlations and switching mechanisms of the SAMHD1 protein complex. Understanding the regulatory mechanisms of SAMHD1 still remains a challenge. While our current study has revealed interesting information regarding the role of the cysteine residues, the exact mechanism of operation of the putative redox switch remains opaque. In addition, the complete picture may be accessible only if one takes into account the tetrameric system. Molecular simulations can play a vital role in unraveling this puzzle.

Bibliography

[1] P. A. D. T. V. Pomella A, Visvikis A and C. AF., Biochem Pharmacol **66**, 1499 (2003).

[2] H. Z. H N Forman and A. Rinna, Mol Aspects Med **30**, 1 (2008).

[3] D. C. Goldstone et al., Nature **480**, 379 (2011).

[4] Z. Wang, A. Bhattacharya, J. Villacorta, F. Diaz-Griffero and D. N. Ivanov, Journal of Biological Chemistry **291**, 21407 (2016).

[5] A. Bhattacharya et al., Scientific reports **6** (2016).

[6] C. S. Sevier and C. A. Kaiser, Nature reviews Molecular cell biology **3**, 836 (2002).

[7] X. Ji, C. Tang, Q. Zhao, W. Wang and Y. Xiong, Proceedings of the National Academy of Sciences **111**, E4305 (2014).

[8] L. Kalé et al., Journal of Computational Physics **151**, 283 (1999).

[9] J. C. Phillips et al., Journal of computational chemistry **26**, 1781 (2005).

[10] J. B. Klauda et al., The journal of physical chemistry B **114**, 7830 (2010).

[11] A. D. Mackerell, Journal of computational chemistry **25**, 1584 (2004).

[12] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon and L. S. Caves, Bioinformatics **22**, 2695 (2006).

[13] W. Humphrey, A. Dalke and K. Schulten, Journal of molecular graphics **14**, 33 (1996).

[14] P. F. Batcho, D. A. Case and T. Schlick, The Journal of Chemical Physics **115**, 4003 (2001).

[15] S. Miyamoto and P. A. Kollman, Journal of computational chemistry **13**, 952 (1992).

[16] H. C. Andersen, Journal of Computational Physics **52**, 24 (1983).

[17] S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, The Journal of chemical physics **103**, 4613 (1995).

[18] H. C. Andersen, The Journal of chemical physics **72**, 2384 (1980).

[19] S. Nosé, The Journal of chemical physics **81**, 511 (1984).

[20] C. H. Mauney et al., Antioxidants & Redox Signaling (2017).

[21] C. Klomsiri, P. A. Karplus and L. B. Poole, Antioxidants & redox signaling **14**, 1065 (2011).

Chapter 7

Conclusions

The sterile alpha motif domain and histidine-aspartate domain-containing protein 1 (or SAMHD1), the most recently discovered anti-HIV restriction factor, has generated immense interest in its activity and regulation on account of its role in blocking HIV-1 infection in dendritic cells. SAMHD1 presents an avenue for restricting HIV-1, that however, operates only to provide immunity to terminally differentiated CD4+T cells and certain other myeloid cells. It is a human triphosphohydrolase protein.[1] [2] [3] [4] The functionally active form of SAMHD1 is an allosterically triggered tetramer which utilizes GTP-Mg+2-dNTP cross bridges to link and stabilize adjacent monomers which then cleave the triphosphate group of dNTPs.[5] [6] [7] [8] [9] This results in the lowering of dNTP levels to as low as a few 10s of nM obstructing efficient reverse transcription, and therefore propagation of HIV-1 is hindered. Studies have also suggested that SAMHD1 possesses an exonuclease activity, which may also be involved in HIV-1 restriction. However, the exonuclease activity has been disputed and is not considered a mechanism for antiviral resistance in myeloid cells. Nevertheless, the importance of SAMHD1 in retroviral restriction is underscored by the fact that HIV-2/SIVmac/SIVsmm have evolved a defense against it: by employing the protein Vpx to target SAMHD1 for proteasomal degradation. The enzyme has the potential to provide a solution to the HIV/AIDS pandemic that has already taken an enormous human and economic toll. However, the complex allosteric and regulatory mechanisms of the enzyme is yet to be unravelled.[10] [11] It is difficult to envision how an enzyme as ostensibly inefficient as SAMHD1 can possibly lower dNTP levels so dramatically in non-cycling cells. The effect of phosphorylation at residue T592 is of paramount importance. This residue lies in a semi-detached lobe, away from the bulk of the protein and is not involved in the dimerization interface. Phosphorylation of SAMHD1 at Thr592 by cdk-1 has been suggested as an

on/off switch for the enzyme.[11] [12] [13] However, the effect of phosphorylation is not well understood: it does not affect the kcat and Km of the enzyme. Bishop and co-workers[11] have conducted X-ray crytallography experiments where they were unable to see any electron density for the immediate region surrounding residue T592, when it is phosphorylated by CDK2/cyclin A. It was suggested that phosphorylation leads to structural collapse, but further studies demonstrated that it merely to faster tetramer disassembly upon nucleotide turn over. Thus functional essays of SAMHD1 have yielded an incomplete picture, with the role of phosphorylation being especially opaque.[12] [13] [14] The XRD derived structures have provided valuable insights into this enzyme. However, these insights have all taken the form of snapshots of various nucleotide bound states of the protein. The mechanistic details of how the triphosphohydrolase enzyme functions remains elusive. What is missing, is a picture of SAMHD1 dynamics at the molecular level that could shed light on the enzymatic mechanism.[15] [16] [17] [18] [6]

In our computational studies of structural dynamics and allosteric mechanisms of SAMHD1 complex, all-atom MD simulations were applied to investigate the structural characteristics, allosteric interactions, the effect of the phosphomimetic T592E mutation, allosteric information flow and monomeric dynamics. Our empirical studies[19] show that only GTP bound SAMHD1 is a more dynamic entity than the GTP/dATP locked tetramer. The absence of dATP has a greater detrimental effect on the stability of the complex than that of GTP. However, given that cellular GTP concentrations are 1000 times greater than that of dNTPs, it is unlikely that a GTP deleted structure could exist in physiological conditions. We found the dATP deleted structure is prone to "breathing motions". The binding of dATP to Allosite 2 leads to an overall firming up of the protein, a long range effect perceptible even at catalytic site. Indeed, this may well be a mechanism to keep SAMHD1 in a "ready state and switch it on when necessary to drop dNTP levels. Thus, our results corroborate experimentally suggested models where SAMHD1 can exist in cells as an inactivated GTP-bound monomer/dimer equilibrium, ready to be assembled into active and "rigidified" tetramers when induced by dNTPs. Indeed, the regulation of dNTP concentrations is likely to be due to an interplay of Ribonucleotide Reductase (RNR) and SAMHD1, with SAMHD1 being held in an "inactive dimer pool" until dNTP concentrations increase to the point where it is needed. This provides a very elegant buildup/breakdown feedback pathway for the control of dNTP levels.

Subsequently after our empirical studies, we have delved into the mechanistic basis for the phenomenological observations of our results. In other words, we

have explored the structural linkages and communication channels allow SAMHD1 to function as a molecular engine.[20] Starting from high-resolution X-ray structures, we have uncovered pathways of allosteric communication which show how the monomeric units of the active tetramer communicate via a reciprocal "handshake". We have found several pathways for allosteric information flow from catalytic core to the allosteric sites within and across the monomers. Path analysis has been instrumental to reveal even more explicit results. The residues from adjacent chains like I530 (chain B) and V586 (chain D) are found to be vital for information flow across monomers. The high "centrality" count of the short linker residues (452-455) has indicated their importance in signal transduction in SAMHD1. We have also demonstrated that mutations at the catalytic site affects the kinetics of tetramer assembly (but not its equilibrium thermodynamics). We have also found the avenue taken by a phospho-regulatory signal to the core of the protein. PCA has revealed that the protein undergoes breathing motions, while the quaternary structure stays intact.

Phosphorylation is proposed to be a control mechanism of SAMHD1 activity downregulation by disallowing tetramer formation. However, as Bhattacharya[13], and Wang[12] have demonstrated through solution phase enzymatics and kinetics studies, the phosphomimetic mutation T592D does not impair the formation of tetramer, nor does it affect the enzymatic activity: it does appear to reduce the stability of the tetramer, once formed. Our preliminary studies of dynamics induced by the T592 mutation demonstrate that there is no large scale breakdown of secondary and tertiary structure of the mutated system and are in line with the conclusions of Bhattacharya[12] and Wang.[13] However, phosphorylation of T592 undoubtedly plays a regulatory role-in vivo studies indicate that T592D and T592E mutants of SAMHD1 are incapable of restriction. Going beyond the preliminary analysis of the dynamics, when we applied correlation network analysis, we have witnessed more complicated effects. We found that the anchoring helices (E355-A373) became correlated to the rest of the chain indicating that the helices are no longer dynamically isolated as found in the case of *wt* tetrameric complexes. However, we are still limited to relatively small MD trajectory time scales. The dynamics of this protein at the millisecond time scale remain unexplained. The elucidation of such a phosphorylation based controlling mechanism awaits further experimental and computational investigation. In that context, there is a need for more solution state biophysics experimental data: such as from realtime FPLC, small angle X-ray scattering, which would reveal information about the average particle size.

Another aspect explored in this thesis was the redox regulation in SAMHD1 which involves redox active cysteine residues. To examine if the cysteine residues (C341, C350, C522) sidechains are critical in protein assembly and stability, we have performed the all-atom MD simulations of $Cys \rightarrow Ser$ mutated monomeric forms of SAMHD1 enzyme complex. An interesting result, that a strong correlation between residue C350 and C522 was observed even in the absence of the disulfide links. Amongst the mutated forms, the C350S system has shown a significant effect on the dynamics of the whole monomeric systems compared to other mutants. While the scatter plots has indicated the potential switching mechanisms of the system, a complete picture of the operation of the putative switch remains opaque.

To summarize, we find SAMHD1 as an enormously complex protein system. Since its discovery about 6 years ago, intense efforts have been made by immunologists, biochemists, and structural biologists to uncover its secrets. Various plausible models have been proposed, involving, for instance, the formation of a monomer dimer equilibrium, which can be driven further to a stable, activated tetramer by dNTPs. However, the best structural biology toolkits cannot overcome the fact this is a 245 kDa tetramer. NMR studies are extremely difficult, if not impossible at this molecular weight. X-ray crystallography only provides still images of protein "eigenstates, and it only registers dynamics as the absence of electron density data. Thus, there is a paucity of information to work with. The molecular dynamics simulations performed has yielded insights that were not anticipated from experimental studies alone. Much remains to be done, in terms of uncovering the regulatory mechanism and elucidating the role of nucleic acid binding and beyond. A proper understanding of the workings of this inscrutable enzyme can potentially enable us to test its druggability and switch its restriction activity "on" and "off" at will. That is the long-term goal of our research project - which has enormous implications, not just from the perspective of understanding a complex molecular machine, but also from the public health viewpoint.

Bibliography

[1] N. Laguette et al., Nature **474**, 654 (2011).
[2] D. C. Goldstone et al., Nature **480**, 379 (2011).
[3] K. Hrecka et al., Nature **474**, 658 (2011).
[4] G. I. Rice et al., Nature genetics **41**, 829 (2009).
[5] T. L. Diamond et al., Journal of Biological Chemistry **279**, 51545 (2004).
[6] X. Ji et al., Nature structural & molecular biology **20**, 1304 (2013).
[7] J. Yan et al., Journal of Biological Chemistry **288**, 10406 (2013).
[8] C. Zhu et al., Nature communications **4**, 2722 (2013).
[9] X. Ji, C. Tang, Q. Zhao, W. Wang and Y. Xiong, Proceedings of the National Academy of Sciences **111**, E4305 (2014).
[10] E. C. Hansen, K. J. Seamon, S. L. Cravens and J. T. Stivers, Proceedings of the National Academy of Sciences **111**, E1843 (2014).
[11] L. H. Arnold et al., PLoS pathogens **11**, e1005194 (2015).
[12] A. Bhattacharya et al., Scientific reports **6** (2016).
[13] Z. Wang, A. Bhattacharya, J. Villacorta, F. Diaz-Griffero and D. N. Ivanov, Journal of Biological Chemistry **291**, 21407 (2016).
[14] L. M. Koharudin et al., Journal of Biological Chemistry **289**, 32617 (2014).
[15] B. Descours et al., Retrovirology **9**, 87 (2012).
[16] B. Kim, L. A. Nguyen, W. Daddacha and J. A. Hollenbaugh, Journal of Biological Chemistry **287**, 21570 (2012).
[17] L. Wu, Retrovirology **9**, 88 (2012).
[18] J. A. Hollenbaugh et al., PLoS pathogens **9**, e1003481 (2013).
[19] K. K. Patra, A. Bhattacharya and S. Bhattacharya, Proteins: Structure, Function, and Bioinformatics (2017).
[20] K. K. Patra, A. Bhattacharya and S. Bhattacharya, Journal of Chemical Information and Modeling **57**, 2523 (2017).

List of publications

1. *Uncovering allostery and regulation in SAMHD1 through molecular dynamics simulations.*
   K K Patra, Akash Bhattacharya, S Bhattacharya., Proteins **85**, 1266-1275 (2017).

2. *Allosteric signal Transduction in HIV-1 Restriction Factor SAMHD1 proceeds via Reciprocal handshakes across Monomers.*
   K K Patra, Akash Bhattacharya, S Bhattacharya., J. Chem. Inf. Model. **57**, 2523–2538 (2017).

3. *Allosteric dynamics of SAMHD1 studied by molecular dynamics simulations simulations.*
   K K Patra, Akash Bhattacharya, S Bhattacharya., J. Phys:Conf.Ser. **759** 012026 (2016).

4. *Studies on the monomeric forms of the SAMHD1 wt and Cysteine mutants.*
   K K Patra, Akash Bhattacharya, S Bhattacharya., to appear (2018)