Summer 8-2017

Conformation of Transmembrane Segments of a Protein by a Coarse Grain Model

Sunita Subedi Paudel
University of Southern Mississippi

CONFORMATION OF TRANSMEMBRANE SEGMENTS OF A PROTEIN BY A COARSE GRAIN MODEL

by

Sunita Subedi Paudel

A Thesis
Submitted to the Graduate School,
the College of Science and Technology,
and the Department of Physics and Astronomy
at The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Master of Science

August 2017

CONFORMATION OF TRANSMEMBRANE SEGMENTS OF A

PROTEIN BY A COARSE GRAIN MODEL

by Sunita Subedi Paudel

August 2017

Approved by:

_____

Dr. Ras B. Pandey, Committee Chair
Professor, Physics and Astronomy

_____

Dr. Christopher Winstead, Committee Member
Professor, Physics and Astronomy

_____

Dr. Michael D. Vera, Committee Member
Associate Professor, Physics and Astronomy

_____

Dr. Parthapratim Biswas, Committee Member
Associate Professor, Physics and Astronomy

_____

Dr. Christopher Winstead
Chair, Department of Physics and Astronomy

_____

Dr. Karen S. Coats
Dean of the Graduate School

Published by the Graduate School

THE UNIVERSITY OF SOUTHERN MISSISSIPPI®

ABSTRACT

CONFORMATION OF TRANSMEMBRANE SEGMENTS OF A

PROTEIN BY A COARSE GRAIN MODEL

by Sunita Subedi Paudel

August 2017

The human voltage-gated proton channels (hH_V1) are critical in many physiological functions and control proton conduction in the cell. This process is governed by the cooperative response of different transmembrane segments of the protein. It is believed that the two subunits of the C-terminal dimer provide independent proton channel pathways through the membrane where the conformations of both monomers and dimer are key for selective proton transport. Conformational response of these transmembrane segments of the protein hH_V1 is studied by a coarse-grained model as a function of temperature where structural detail of a residue is ignored and its specificity is captured by its unique interaction. How residues of the protein hH_v1 assemble or disperse as the temperature varies is addressed using a coarse-grained Monte Carlo simulation where a knowledge-based residue-residue interaction matrix is used as input in the Metropolis algorithm. Contact maps, mobility, radius of gyration, and structure factors, are examined as functions of temperature due to the efficiency of this model. Thermal response of the radius of gyration of this protein in the low-temperature regime decreases on increasing temperature in which structure becomes more compact by reduced entropy while in the high-temperature regime, the radius of gyration increases with temperature before reaching a steady state value. The scaling of structure factor *S(q)* provides an estimate of the effective dimension (D) of the protein chain which becomes

globular conformation (D~3) with more connectedness in the low-temperature region and random coil (D~2) and then linear conformation (D~1) on increasing temperature further.

ACKNOWLEDGMENTS

First and foremost, a very special gratitude goes to my research advisor Dr. Ras B. Pandey without whom I could not do this project. I want to convey my sincere thanks to all my committee members; Dr. Christopher Winstead, Dr. Michael D. Vera and Dr. Parthapratim Biswas for their constant guidance and encouragement in my work. Also, I am thankful to the Department of Physics and Astronomy at the University of Southern Mississippi for providing the opportunity and support to me to pursue my Master's degree.

I am also grateful to our collaborators; Panisak Boonamnaj, Warin Jetsadawisut1, Dr. Sunan Kitjaruwankul, and Dr. Pornthep Sompornpisut for providing detailed results of the same protein segment through an all-atom model. I could not trust my result without the reinforcement from all-atom molecular dynamics simulation.Also, I would like to show appreciation to Ms. Jan Wilkinson for her boundless help from the beginning until the end for all paperwork and many more official works.

Last but not the least, my special thanks goes to my mentor as well as my best friend Mr. Durga P. Paudel who support me in my every step and always wants me to do the best. I am sure, I could not get my Master Degree in absence of your encouragement and guidelines.

DEDICATION

I am forever indebted to my parents- Narhari Subedi and Dhanu Maya Subedi, my in-laws- Nandala Paudel and Ranjati Paudel, my grandmother- Lila Paudel, my dear husband- Durga P. Paudel and my little boy- Denish Paudel, for their support and encouragement through the successful completion of my Master's degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF SYMBOLS

A	Any physical quantity
D	Diffusion Coefficient/Dimension
E	Internal Energy of the particle
F	Force applied
$\mathcal{H}$	Hamiltonian
K	Kinetic Energy of the particle
M_n	Mobility of the residue
N	Number of residues
N_n	Average number of residues
P_{eq}	Probability of configurational state at equilibrium
R_c	Root mean square displacement
R_g	Radius of gyration
$S(q)$	Structure Factor
T	Temperature
U	Potential Energy
V	Volume
X	Configurational state of system
Z	Partition function
a	Lattice constant
k_B	Boltzmann Constant
k_r	Spring constant
k_ϕ	Angle constant
k_θ	Multiplicative constant
m	Mass
q_i	Charge
r	Position of a residue
r_c	Equilibrium position of two residues
r_{cm}	Center of mass of the protein chain
t	Time
v	Velocity
x	Position of a residue in x-axis
y	Position of a residue in y-axis
z	Position of a residue in z-axis
ϵ_o	Absolute permittivity
ϵ_{ij}	Potential strength
σ_{ij}	Distance where potential equal zero
ϕ	Transition probability

LIST OF ABBREVIATIONS

AA	All Atom
ATP	Adenosine Triphosphate
CG	Coarse-Grain
CPU	Central Processing Unit
CTD	C-terminal Domain
DNA	Deoxyribonucleic acid
DOF	Degree of Freedom
FL	Full Length
hH_V1	Human Voltage gated proton channel
H_V1	Voltage gated proton channel
IMP	Integral Membrane Protein
M	Monomer
MC	Monte Carlo
MCS	Monte Carlo Time Steps
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NTD	N-terminal Domain
pH	Potential of hydrogen
PDB	Protein Data Bank
RMSD	Root Mean Square Displacement
tD	Tandem Dimer
TM	Transmembrane

CHAPTER I - INTRODUCTION

Protein folding is one of the most prominent fields of research in molecular biology. Several diseases, such as Alzheimer's disease, Parkinson's disease, and Huntington's disease (Dill & Maccallum, n.d.; Vianello, Domene, & Mavri, 2016) are identified with the misfolding of proteins. There are many experimental challenges in studying the structure and dynamics of proteins such as folding pathways, folding speed, energy landscapes, and misfolding in accelerating the understanding of protein structures and discovery of new drugs (Dill, Ozkan, Shell, & Weikl, n.d.). The experimental studies of proteins are hindered by low-resolution structural data so molecular dynamic simulation (MD) provides a complementary approach (Goloubinoff, 2014) with sufficiently high resolution spatial and temporal data on protein folding procedures. MD simulations enable the sampling of the states of proteins and calculation of the possible folding pathways. Classical all-atom MD simulation analyzes all the atomic detail and uses time steps in the femtosecond range which constrains the simulation to be quite slow and expensive. To overcome this issue, the atomistic detail of the protein is lowered by the use of the coarse-grain (CG) model (Scott et al., 2008) to simulate for large time scales by reducing the degrees of freedom (DOF) which is discussed in depth in the second chapter. In addition to the folding mechanism of the protein, simulation is significantly essential for the structural and functional analysis of the membrane proteins (Walker, 2010) which are found in a complex two-dimensional lipid bilayer environment where it is difficult to study details about membrane proteins experimentally. Therefore, computer simulation has become a useful tool in elucidating the complexity of macromolecular chemical architecture and geometrical structure.

I hereby present the study of the structure of a C-terminal domain of the human voltage-gated proton channel (hH_V1), using coarse grain Monte Carlo (MC) Simulation (Fritsche, Pandey, Farmer, Heermann, & Slocik, 2013; R B Pandey, Farmer, & Gerstman, 2015; Ras B Pandey & Farmer, 2012) as the main technique and comparing the result with the result from all-atom Molecular Dynamic (MD) Simulation obtained by our collaborators (Boonamnaj et al., 2017). We are obtaining an unexpected result on the conformation of the protein (hH_V1) with the change of temperature which is the main outcome of my work in this thesis.

First, I begin from the basic outline of protein, its types, and the detailed structural analysis of human the Voltage-Gated Proton Channel (hH_V1) protein. Second, I deliberate the theory that is applied while doing this project along with the definition of some technical terms employed while doing simulation. Third, I discuss the simulation methods with algorithms and brief statistical analysis of the local and global physical quantities that are obtained from the simulation. The result section follows where I have included all the results that I have obtained so far. Finally, I end with the conclusion and summary.

This introduction chapter starts with the basic details of proteins, motivation and the main objective of the thesis.

1.1 Proteins

The large complex molecules that underpin every aspect of biological activity are proteins (“Protein,” 2017). They are polymers in which amino acids act as monomers. There is a strong causal connection between the unique 3-dimensional structure of a protein and its specific biological function depending on the sequence of amino acids.

The human body is composed of 20 percent protein and protein makes up over 50 percent of the dry weight of the cell ("Amino Acids 1 | Peptide | Protein Structure," n.d.). Proteins have diverse biological functions such as: structural and mechanical (e.g. Keratin), hormones, enzymes, antibodies, acid-base balance, channels and pumps, transporters (hemoglobin), and regulators of gene expression. They can be classified based on form and main function into three groups: globular proteins like many enzymes, fibrous proteins for the structural role and membrane proteins which act as a channel or receptor for polar or charged molecules to pass through the cell membrane.

1.1.1 Amino Acids

Amino acids are the building blocks ("Amino Acids 1 | Peptide | Protein Structure," n.d.) or the structural units of proteins. They are organic compounds containing amine ($-NH_2$) and carboxyl (-COOH) functional groups, along with a side chain (R group) specific to each amino acid("Amino acid," 2017). The main elements of an amino acid are carbon, oxygen, hydrogen, and nitrogen with the exception of sulphur in some of them. There are more than 500 amino acids known, among them 20 play primary roles in the formation of proteins. They are called proteinogenic amino acids. In the form of proteins, amino acids form the second largest compound (water is the largest) of human muscles and other tissues.

Amino acids join together by the peptide bond and form polypeptides or proteins. Proteins are formed by step-by-step addition of amino acids with the formation of peptide bonds to a growing protein chain (as illustrated in the Figure 1.1) by a ribosome in a process called translation. Deoxyribonucleic acid (DNA) in the cell is responsible for the type and sequence of amino acids in a protein with the use of genetic code. The sequence

of amino acids plays a vital role for the structure and function of the protein. Proteinogenic amino acids can be broadly divided into three groups based on the propensity of the side chain to interact with polar solvents as; Hydrophobic (H) amino acids, Hydrophilic or Polar (P) amino acids and Electrostatic (E) amino acids.

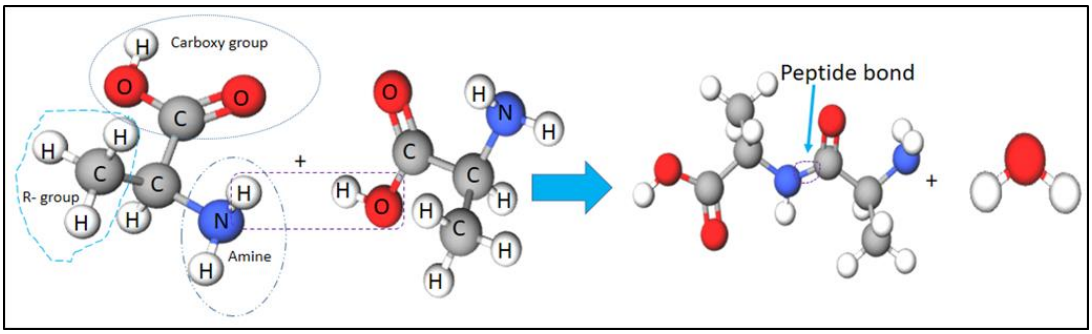


Figure 1.1 Two alanine amino acids forming the peptide bond by removing one water molecule from them.

1.1.2 Structure of Proteins

The structure of a protein provides insight into the function of the protein rather than its sequence of amino acids(Jha, Vishveshwara, & Banavar, 2010). In general, there are normally ten to several thousands of amino acids to form the protein with covalent bonds and different other non-covalent bonds. Proteins are grouped as nanoparticles as their size ranges from 1-100nm ("Protein structure," 2017). While performing biological functions, a protein structure undergoes changes in form which can be reversed back to the previous one. The different configurations of the same protein are known as conformational isomers or simply conformations and the transition phase from one form to the other is called a conformational change. There are four distinct structures of proteins: Primary Structure, Secondary Structure, Tertiary Structure, and Quaternary Structure.

The primary structure (as shown in Figure 1.2) of a protein is the linear sequence of amino acids in the polypeptide chain where amino acids are held together by covalent bonds named peptide bonds. The two ends of the protein chain are referred to as the N-terminal (amino terminal) and C-terminal (carboxyl terminal) based on the nature of the functional group on the terminal. Counting of the residues always starts at the N-terminal end (NH_2 -group) and ends at the C-terminal end (COOH-group) in the protein chain. It is assumed that all the information essential to determine the structure of a protein is present in its primary sequence(Kessel & Ben-Tal, 2011).

The secondary structure (as shown in Figure 1.2) of a protein is the regularly repeating local structure stabilized by hydrogen bonds between the main chain peptide groups. Proteins have regular geometry being constrained to specific values of the dihedral angles ψ and φ on the Ramachandran plot (i.e. plot to visualize dihedral angles ψ versus φ of amino acid residues in protein). There are two main types of secondary structure the α- helixes and the β- sheets. In this structure the covalent bond is a peptide bond and the non-covalent bond is an H-bond.

The tertiary structure (as shown in Figure 1.2) of a protein is the overall three dimensional shape of the protein molecule which controls the basic function of the protein. Here, the α–helixes and β-sheets are folded into a compact globular structure. This structure of the protein also has peptide bonds as covalent interactions and H-bonds, salt bridges and disulphide bonds as non-covalent interactions. It is the most stable form of protein, often called the native state since it optimizes all sorts of interactions.

The quaternary structure (as shown in Figure 1.2) of protein is formed by several protein molecules which function as single protein complex. This structure is also

stabilized by the same covalent and non-covalent interactions as those involved in tertiary structure.

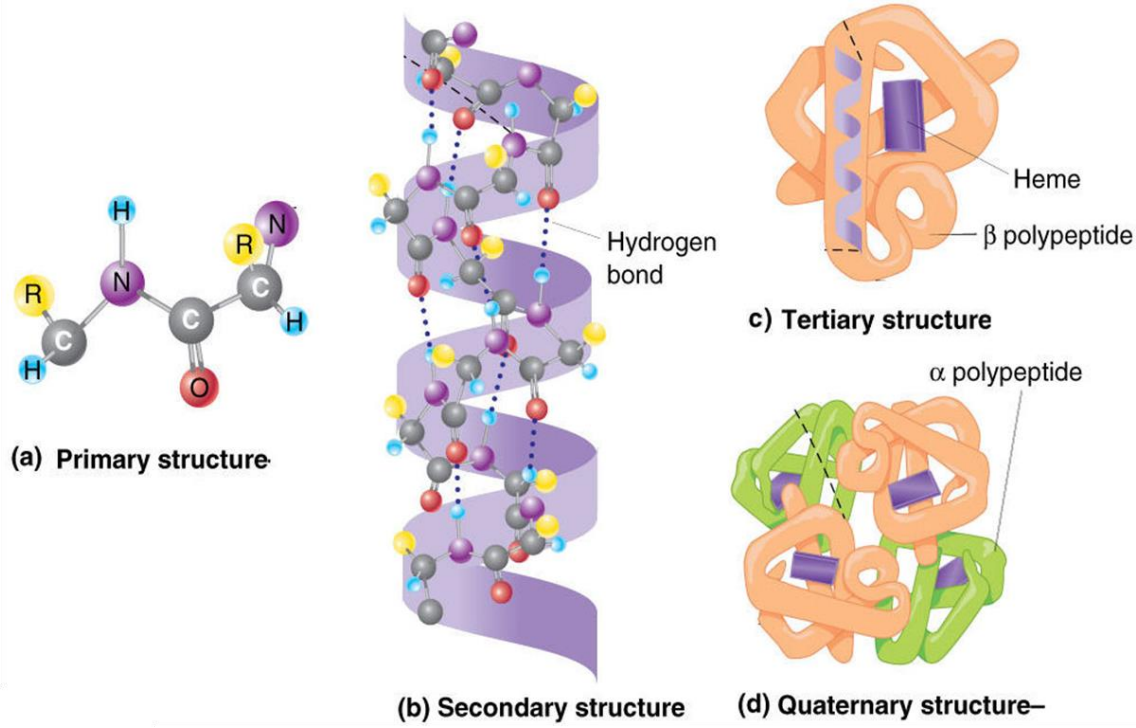


Figure 1.2 Different structures of the proteins

("Shape of proteins," n.d.)("Four levels of Protein Structure," n.d.)

1.2 Membrane Protein

Membrane proteins are important proteins for various biological activities such as cell-cell contact, surface recognition, cytoskeleton contact, signaling, enzymatic activity, transporting substance across the cell membrane and so on. They are the target of over 50% of all modern medicine and represent almost 40% of all proteins ("Membrane protein," 2017). Although membrane proteins are significant for their diverse function, they are still difficult to study at the molecular level. This is due to the complexity in producing and extracting them from their natural environment and purifying them in a

native conformation where they are naturally embedded in a hydrophobic environment made by phospholipids. They still represent a challenge to structural biologists as only 2% of their structure is known to date. There are different types of membrane proteins on the basis of their association with the lipid bilayer: integral membrane proteins, lipid-linked proteins and peripheral membrane proteins.

Integral membrane proteins (IMP) are permanently anchored with the biological membrane so they are challenging to detach from the membrane without the use of external agents. All transmembrane proteins (TM) are integral membrane proteins which span across the membrane. Besides TM proteins, there are integral monotopic proteins which are attached to only one side of the membrane and are integral membrane proteins. Nearly 20-30% of total proteins are IMP and most drugs target these proteins as they play a vital role in transporting materials, transmitting information, energy conservation and so on. Despite having great importance, IMP are difficult to study because of their instability outside their native membrane. To date, few structures are available.

Lipid-linked proteins are on the surface of the biological membrane and are embedded with lipids by covalent bonds while peripheral membrane proteins are momentarily attached to the lipid bilayer.

1.2.1 Ion Channels

Biological systems, from a simple unicellular organism like a bacterium to a complex multicellular organism like the human body, are reservoirs of charged particles called ions. Biological ions are either organic or inorganic and are present in the membranes of all cells.

Ion channels are the transmembrane proteins whose function includes establishing a resting membrane potential, and shaping action potentials and other electrical signals by gating the flow of ions across the cell membrane("Ion channel," 2017). They are the most important components in a range of biological processes that encompass quick changes in cells such as cardiac, skeletal and smooth muscle contraction, epithelial transport of nutrients and ions, T-cell activation and pancreatic beta-cell insulin release.

Ion channels are different from other ion transport proteins mainly in two basic ways(Eaton, 1985). First, the rate of transport of the ion through the channel is very high (i.e. nearly 10^6 ions per second or greater) and the next is that ions pass through the channels down their electrochemical gradient which is a function of ion concentration and membrane potential without the presence of metabolic energy (i.e. ATP).

Depending on the nature of the ion gating, species of ion passing through those gates and the number of gates, ion channels are classified into different types. Different channels can be categorized by gating (i.e. what opens and closes the channels). For an instant, voltage-gated ion channels open or close depending on the voltage gradient across the plasma membrane while ligand-gated ion channels open or close depending on the binding of ligands to the channel.

1.2.2 Voltage-Gated Proton Channel (H_V1)

The voltage-gated proton channel (H_V1) is an integral membrane protein, which has numerous physiological functions such as controlling the motility of sperm (Lishko, Botchkina, Fedorenko, & Kirichok, 2010), pH regulation (DeCoursey, 2013), reactive oxygen production (Rebolledo, Qiu, & Peter Larsson, 2012), activation of B-

cells(Decoursey, 2013), killing pathogens (DeCoursey, 2013), even it is thought to exacerbate the metastasis of breast cancer(Decoursey, 2013) and strokes and much more. H_V1 is unique among all voltage-gated channels in containing the pore and gate within its voltage sensing domain (VSD) and controlling proton conduction in cells. The H_V1 channel is found as a homodimer in the membrane and each subunit contains two essential parts: the transmembrane region (S1, S2, S3 and S4) and inner cytoplasmic region (N-terminal Domain, NTD and C-terminal Domain, CTD). The four transmembrane segments of each subunit are believed to constitute an independent proton permeation pathway and are responsible for sensing the potential of the so-called voltage sensing domain (VSD) for the purpose of gating while the CTD located downstream of the S4 helix is thought to be responsible for dimeric stabilization of the channel.

There are various characteristics and unique properties of H_V1 which include: specific proton conduction, gating in response to membrane depolarization, extremely small unitary conduction, intense temperature dependence of both conduction and gating, and the absence of inactivation. H_V1 is mainly responsible for the rapid movement of protons (H^+) across the cell membrane(Musset & Decoursey, 2012). When a single channel is open, it allows up to 10^5 protons (H^+) to cross the membrane per second but when it is close, it does not allow any ions to cross the membrane. The voltage dependence of these channels is sharply regulated by the pH gradient across the membrane although they open with membrane depolarization.

Voltage gated proton (H_V1) channels are a fascinating area for the researcher due to their expanding list of physiological importance in a variety of cells and species and they homologous to the Voltage Sensing Domain (VSD) of other voltage-gated ion

channels. Also, H_V1 violates many rules followed by other ion channels therefore they need separate study for all sorts of behavior. That is why H_V1 is unique and is an attractive subject for research.

Different species have different number of amino acids and they have their unique proposed name as listed in the table below:

Table 1.1

Proposed Name of the Voltage-Gated Proton Channel of few organism along with their length of amino acids.

Organism	Species	Proposed Name	Length of Amino acids
Human	*Homo sapiens*	hH_V1	273
Mouse	*Mus musculus*	mH_V1	269
Chicken	*Gallus gallus*	gGH_V1	235
Elephant	*Loxodonta africana*	LaH_V1	455
Green Algae	*Chlorella variabilis*	CvH_V1	480
Dog	*Canis lupus familiaris*	dH_V1	268
Sea Squirt	*Ciona intestinalis*	CiH_V1	342

Note: These informations illustrate the varying chain length of the Voltage-gated proton (H_V1) channel in different organism along with their different name (Decoursey, 2013).

1.2.3 Human Voltage-Gated Proton Channel

Human voltage-gated proton channel (hH_V1) is the voltage-gated proton channel in humans which is formed by 273 amino acid residues as mentioned in the above Table 1.1.

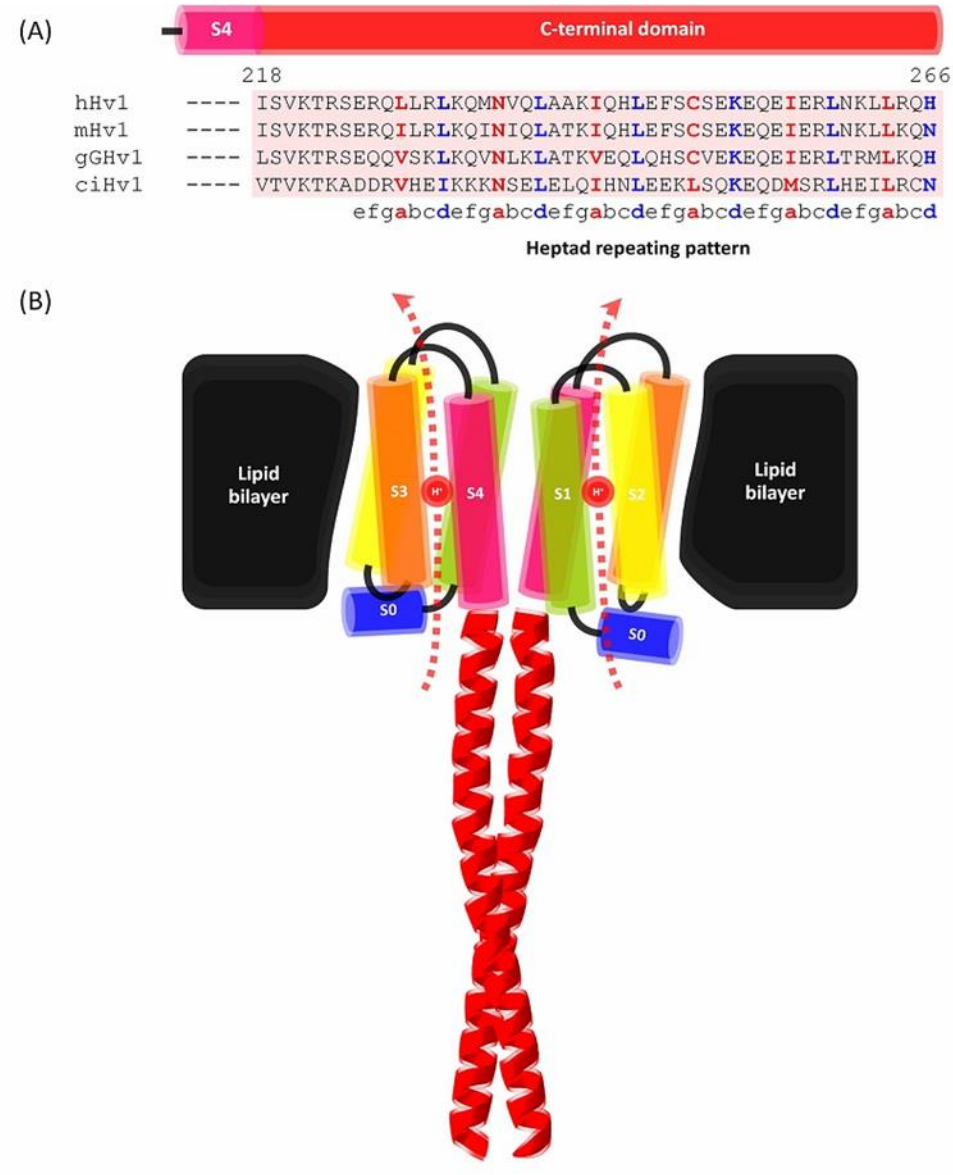


Figure 1.3 (A) Sequence alignment of the C-terminal domain of HV1 channels of various species (B) The coiled coil helix structure of a dimeric hHV1 channel

Note: (A) C-terminal domain of Hv1 channels form various species as mention in **Error! Reference source not found.** with a heptad r epeating $(abcdefg)_n$ pattern. Position 'a' and 'd' are hydrophobic positions highlighted in blue and red, respectively, and 'e' and 'g' are charged positions. (B)The above figure is based on experimental data and X-ray crystal structure is shown as red ribbons.

Figure 1.3 illustrates a dimeric assembly of hH_V1. The CTD of hH_V1 (PDB-3A2A) forms an α-helical coiled-coiled structure and provides the main contact between

the two subunits. The coiled-coil dimer consists of heptad repeating positions which contribute critically to intersubunit helical interactions. The different regions of this protein along with the residue number are mentioned in the table below.

Table 1.2

Different regions of hH_V1 protein

Regions of hH_V1	Position of Residues	Number of Residues
N-terminal Domain	1-100	100
Transmembrane Region (S1, S2, S3, S4)	101-220	120
C-terminal Domain	221-273	53
Residues taken for Simulation	218-266	49

Regions of hH_V1 protein with the position of residues and the number of residues (S. J. Li et al., 2010).

Several works reported that an H_v1 channel functions as a dimer with each subunit containing its own pore as mentioned above (Musset et al., 2010). The H_v1 channel exhibits temperature-dependent activation kinetics. At high temperature, the H_v1 channel showed faster activation kinetics, especially the monomeric H_v1 channel (truncated CTD, hH_v1-ΔC). The deletion of CTD makes it easier to activate than a dimeric H_v1 channel (full-length molecule, FL) (Fujiwara et al., 2012). This means that CTD may also be essential for regulation of activation processes in an H_v1 channel. Moreover, the gating mechanism of the proton transport is also relevant to the cooperative interaction between the two subunits (Musset et al., 2010). Each subunit undergoes the voltage-sensing conformational change before either pore can conduct a proton. The CTD also participates in the cooperative gating process via intersubunit

interactions. Thermal agitations were observed in elevating of the temperature, suggested by Fujiwara et al. (Fujiwara et al., 2012). This effect has reduced the residue-residue interactions and may lead to unstable dimerization. However, it remains unclear which exact part of the H_v1 channel is involved with the conformational change transmitted from one subunit to other during channel gating and how the global structure is stabilized by CTD. Obviously, this is a challenging problem with computational approaches involving all atomic details. Probing the structural evolution of the whole protein in a membrane environment is compute-intensive; even at a large-scale, all-atom MD simulations run for months on dedicated computer clusters and one can hardly see large-scale conformational responses(Kitjaruwankul, Khrutto, Sompornpisut, Farmer, & Pandey, 2016). Therefore, it would be interesting to explore the structural changes of the H_V1 channel in the absence of the transmembrane region via all-residue coarse-grained (CG) Monte Carlo (MC) methods as well as all-atom (AA) Molecular Dynamics (MD) simulations as attempted here. The conformational changes of the CTD of both the monomeric and dimeric structures have been monitored in this study.

The monomeric hH_V1-CTD [M-hH_V1-CTD: [^{1}H ^{2}Q ^{3}R … ^{48}S ^{49}I]] and its modified dimer [tD-hH_V1-CTD: [^{1}H … ^{49}I ^{50}A … ^{55}A ^{56}I … ^{104}H]] with six alanine (A) residues in a tandem linker, a connecting loop to maintain the dimer structure, (Figure 1.4 and APPENDIX B) are introduced for all-residue coarse-grained MC simulation and the obtained results are compared with all-atom MD simulation. The refined structure of a previous crystallographic report of hH_V1-CTD (Q. Li et al., 2015) is employed as a template of models for both computational approaches. The temperature-dependent manner of M-hH_V1-CTD monomer and tD-hH_V1-CTD with six-alanine linker has been

investigated. We study both local and global physical quantities of the monomer and its tandem dimer.

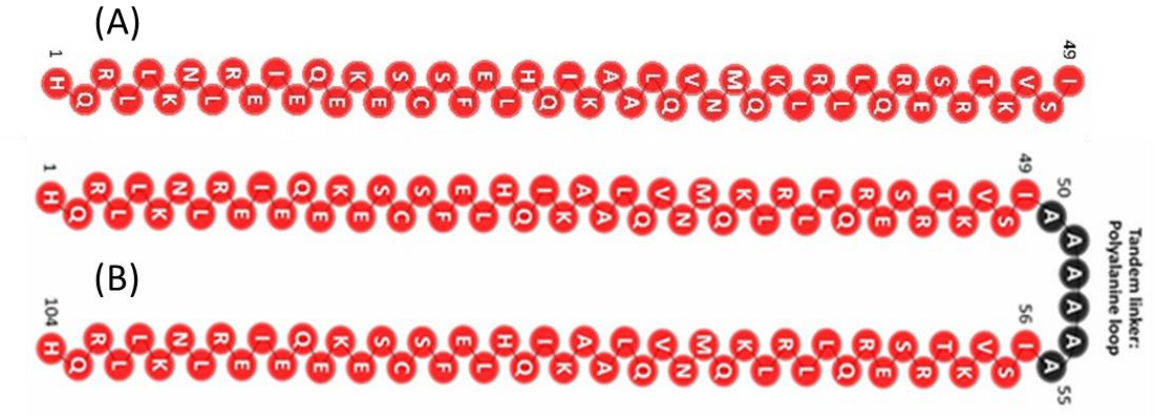


Figure 1.4 (A) Monomer chain of the protein hHV1-CTD from (B) Tandem dimer structure of the hHV1-CTD

Note: (A)Monomer chain of the protein hH_V1-CTD from the residue number 218-266 (49 residues) for the coarse-grain MC simulation.(B) Tandem dimer structure of the hHV1-CTD which is taken as input in our coarse-grain Monte Carlo simulation. More description of the letter symbols used is in APPENDIX B as supporting material.

CHAPTER II – MOLECULAR DYNAMICS AND MONTE CARLO SIMULATION

The application of computer simulation is an iterative process by which the identification of complex materials and processes can be substantially enhanced in a step by step process. In view of distinct conceptual and principal advantages, computer simulations in macromolecular materials have aroused great interest, and various complementary methods have been technologically advanced. In some cases experiments are limited and expensive so simulation can complement the experiment. There are two main classes of simulation technique considered here; one is Molecular Dynamics (MD) Simulation and the other is Monte Carlo (MC) Simulation. In addition, both of these techniques can be combined and form hybrid techniques in some problems.

2.1 Molecular Dynamics Simulation

Molecular Dynamics (MD) is a subject to study the interaction within many particle systems at an atomic resolution where the time evolution of a set of interrelating particles is followed by integrating their equation of motion. MD aims to build up the connection between microscopic information and macroscopic properties of the physical system. MD simulation delineates the atomistic detail of real atoms, assuming a given potential energy function and can solve complex many-body problems. It is classically exact with respect to a given interatomic potential as it gives unbiased dynamics of position and momentum in full atomistic detail(Adcock & McCammon, 2006; M. Allen, 2004; Binder, 2008). Although it is formally simple, it can handle complexity naturally as the system does the right things at right time and can capture emergent behavior.

With a good potential (U), MD simulation can be used to gain insight at the atomic level and make a prediction and interpret experiments and are supposed

themselves as experiments. It gives molecular level understanding even at a single molecule level. It is often exercised to refine 3D structures of proteins and other macromolecules based on experimental constraints from X-ray crystallography or NMR spectroscopy (M. Allen, 2004; Gelpi, Hospital, Goñi, & Orozco, 2015; Perlmutter et al., 2011).

MD simulation initializes position and momentum with the boundary conditions in space and time and determines the trajectory of atoms and molecules by numerically solving Newton's equation of motion.

$$F = ma \qquad (2.1)$$

where *F* is the force on an atom, *m* is the mass of the atom and *a* is the atom's acceleration. Also, force can be written in terms of the potential energy function as,

$$m\frac{d^2r}{dt^2} = F(r) = -\nabla \mathrm{U(r)} \qquad (2.2)$$

where *r* represents coordinates of all atoms and *U(r)* is the potential energy function where all sorts of interaction energies are involved that are discussed in the all-atom MD section below.

In MD simulations, analytical solution is impossible but numerical solution is straight forward with the use of the "Verlet algorithm" (M. . Allen & Tildesley, 1991) which uses the following equations;

$$r(t + \delta t) \approx 2r(t) - r(t - \delta t) + a(t)\delta t^2 \qquad (2.3)$$

where $r(t)$ is the positions, $a(t)$ accelerations, and $r(t - \delta t)$ is the position of the previous step. The equation (2.3) is obtained by eliminating the velocities $v(t)$ term from the following equations obtained by Taylor expansion about (t) :

$$r(t+\delta t) \approx r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots \qquad (2.4)$$

$$r(t-\delta t) \approx r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots \qquad (2.5)$$

2.1.1 All-atom Molecular Dynamics

The name itself gives a clear definition of all-atom MD simulation where every atomistic detail is considered for the simulation. All-atom MD simulation gives extensive knowledge of the system and if all trajectories of the system are known, everything can be computed. It easily accommodates non-equilibrium states and other complex situations beyond thermal equilibrium. Different interaction energies (i.e. non-bonded interaction and bonded interaction) are involved in this simulation from which we can obtain the force-field. The non-bonded interactions ($U_{non-bonded}$) involve the Lennard-Jones interaction (U_{VDW}) which describes dispersion and repulsion effects and Coulomb interaction or electrostatic interaction ($U_{Coulomb}$) which is given by:

$$U_{non-bonded}(r) = -\sum_{ij}\frac{q_i q_j}{4\pi\varepsilon_o r_{ij}} + \sum_{ij}\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad (2.6)$$

where q_i represents the charge of the particle i, ε_o is the absolute permittivity, r_{ij} is the distance between the particles, ϵ_{ij} is the depth of the potential well or the potential strength, σ_{ij} is the finite distance at which the inter-particle potential is zero which can be related to the distance at which potential reaches its minimum (r_{min}) by the relation $r_{min} = 2^{1/6}\sigma$. The bonded interactions (U_{bonded}) involve the bond length, bond angle and torsional angle. The bond-length (U_{bond}) refers to a two-body interaction which describes bond deformation. The bond-angle (U_{Angle}) refers to a three-body interaction that describes bond angle and geometry. The torsional angle ($U_{Torsion}$) refers to a four-

body interaction that describes rotation about certain dihedral angles of covalently bonded atoms (Kmiecik et al., 2016a). The total bonded potential is given by;

$$U_{bonded}(r) = \sum_{ij} k_r(r_{ij} - r_o)^2 + \sum_{ijk} k_\phi(\phi - \phi_o)^2 + \sum_{ijkl} k_\theta(1 + \cos(m\theta_{ijkl} - \gamma)) \quad (2.7)$$

where $k_r, k_\phi,$ and k_θ are the spring constant, angle constant and multiplicative constant respectively, $r_{ij} = \|\vec{r_j} - \vec{r_i}\|$ gives the distance between the atoms, r_o, and ϕ_o are the equilibrium distance and equilibrium angle. All of the physical forces can be determined using the potential energies (as mentioned above). Simulations use these force-fields for the study of the structure and dynamics of the protein which makes the problem complex.

Although all-atom MD can convey the atomistic detail and conformational flexibility, it is very difficult to get an accurate result as it cannot work on an integration time step of femtosecond (fs), rather the accessible timescale is about 10 nanoseconds (ns) in an advanced computer having high power. So, MD simulation is computationally demanding and expensive with limited run times where large-scale conformational changes are hard to model.

2.2 Monte Carlo Simulation

Monte Carlo (MC) Simulations are stochastic techniques where we use random sampling to learn about a system. MC simulation does not require a continuous energy function as an MD simulation does. It is an alternative method to discover low-energy regions of the space of atomic arrangements. In this simulation instead of using Newton's laws to move atoms, we consider random moves. By performing statistical sampling

experiments on a computer, MC provides approximate solutions for a variety of problems by replacing the actual motion of atoms by statistical sampling of their locations which is less computationally costly(Kessel & Ben-Tal, 2011). It samples configurations arbitrarily and decides whether to adopt or reject new conformations probabilistically. A basic feature of MC sampling is the use of an energy term which is not circumscribed to a differentiable function and uses any structural modification that does not violate basic assumptions. It is believed from many studies (Ingólfsson et al., 2014; Kmiecik et al., 2016b; Lässig & Valleriani, 2002) that, MC is faster than MD computationally as MC is not bound by complicated force-fields.

2.2.1 Static Quantities at Equilibrium

In statistical equilibrium mechanics (Kurt Binder, 2008), thermodynamic properties are calculated by the ensemble averages over all points in a high dimensional configuration space. In the canonical ensemble, the average of an unknown physical quantity A(X) is calculated using

$$\langle A \rangle_{N,V,T} = \frac{1}{Z} \int dX \, A(X) \, e^{-\frac{\mathcal{H}(X)}{k_B T}} \qquad (2.8)$$

where N is the number of particles, V is volume, T is temperature, Z is the partition function, X is a configurational state, $\mathcal{H}$ is the Hamiltonian, and k_B is the Boltzmann constant. Here N, V and T are held constant. With a number of samples states X_ν one may numerically compute the value of A as

$$\bar{A} = \frac{\sum_\nu A(X_\nu) e^{-\frac{\mathcal{H}(X_\nu)}{k_B T}}}{Z} \qquad (2.9)$$

The Boltzmann factor $P_{eq}(X) = e^{-\frac{\mathcal{H}(X)}{k_B T}}/Z$ is the probability the system will be in state X at any given point in time. In general, the integral cannot be solved analytically so MC simulations provide a numerical approach to this problem by generating a random sample of the configuration space points as in equation (2.7). However, the Boltzmann factor is so sharply peaked that this naïve implementation is rarely useful for physical systems. Approximately all of the randomly chosen states would have essentially zero probability, and the result would suffer from insufficient sampling. While it is true that the computed value $\bar{A}$ converges as the number of random states approaches infinity, it is clear that this naïve method is very inefficient. If, however, one chooses states according to the probability $P_{eq}(X)$ with the total number of states M, then at equilibrium Eq. (2.7) becomes

$$\bar{A} = \sum_{\nu=M_o+1}^{M} A(X_\nu) \qquad (2.10)$$

One can construct a Markov chain $X_1 \longrightarrow X_2 \longrightarrow X_3$ .... using a transition probability $\phi(X, X')$ to achieve the set of states. As we are considering the system in equilibrium, we apply the principle of detailed balance which is given below:

$$P_{eq}(X)\phi(X, X') = P_{eq}(X')\phi(X', X), \qquad (2.11)$$

From this we obtain

$$\phi(X, X') = \phi(X', X)e^{\frac{\mathcal{H}(X)-\mathcal{H}(X')}{k_B T}} \qquad (2.12)$$

Metropolis et al.(“Amino acid,” 2017) introduced the transition probability

$$\phi(X, X') = \begin{cases} e^{\frac{\mathcal{H}(X)-\mathcal{H}(X')}{k_B T}} & if\ \mathcal{H}(X')-\mathcal{H}(X)>0 \\ 1 & otherwise \end{cases} \qquad (2.13)$$

This is the most commonly used transition probability and is implemented in the Metropolis algorithm which is represented in the form of a flow chart in APPENDIX A.

2.2.2 Dynamic Quantities and Non-equilibrium System

For any non-equilibrium system, the probability of any particular microstate changes in time so that $P(X_\nu)$ becomes $P(X,t)$. In an infinitesimal period of time, the probability increases by the combined probabilities that the system transitions from any other microstate and decreases by the probability of transitions to any other microstate. This is represented mathematically by the master equation

$$\frac{dP(X,t)}{dt} = \sum_{X'} \Phi(X',X)P(X',t) - \sum_{X'} \Phi(\mathrm{X},X')P(X,t) \qquad (2.14)$$

where $\Phi\ (X,X')$ is the transition probability per unit time, and the Markov chain is now a sequence in time. We now seek an appropriate transition probability per unit time to produce a dynamic set of states for a system not necessarily at equilibrium. Here the transition probability can be written as:

$$\Phi(\mathrm{X},X') = \frac{1}{Z}\sum_{ii'} e^{\frac{-\mathcal{H}(Y_i)}{k_BT}\Phi(XY_i,X'Y_{i'}')} \qquad (2.15)$$

In the same way, the inverse transition probability is

$$\Phi(X',X) = \frac{1}{Z}\sum\nolimits_{ii'} e^{\frac{-\mathcal{H}\left(Y_{i'}'\right)}{k_BT}\Phi\left(X'Y_{i'}',XY_i\right)} \qquad (2.16)$$

From the conservation of energy and the transitional probability found in equation (2.10) which can again be obtained from eq. (2.13) and (2.14), the master equation can be written as:

$$\frac{dP(X,t)}{dt} = e^{\frac{-\mathcal{H}(X)}{k_BT}}\sum_{X'} \Phi\ (X,X')\left[P(X',t)e^{\frac{\mathcal{H}(X')}{k_BT}} - P(X,t)e^{\frac{\mathcal{H}(X)}{k_BT}}\right] \qquad (2.17)$$

Clearly, with the equilibrium probabilities, the expression inside the square brackets is zero. Therefore $dP(X,t)/dt = 0$ which is correct for equilibrium. Then we again use Metropolis algorithm to satisfy equation (2.9).

In MC simulation, a protein molecule is put into an initial configuration which is often a random configuration. There occur conformational changes over a time step using the Metropolis algorithm for the movement of the residues of the protein which use the Boltzmann probability distribution. Statistically, the residues of the protein move in such a way that every time it tends toward the state of lowest free energy called the native conformation or the state of equilibrium. This conformational procession of the protein from one state to another state of lower free energies is the main objective of the computational modeling.

2.2.3 Coarse-Grained Models

Although much detailed information at the atomistic level can be introduced by a classical all-atom model, pragmatic applications are still constrained by its algorithmic efficiency and the available computer power which is not enough even with special purpose supercomputers that are dedicated to atomistic molecular dynamics. To overcome such issues, coarse-grain (CG) models take a step which is computationally more effective and allows simulations of much longer time scales (Emperador, 2013). CG models are used widely in molecular biology for protein structure prediction, modeling of complex dynamics processes, depicting protein interactions with other proteins and peptides, understanding protein folding mechanisms, modeling of membrane proteins and so on (Kmiecik et al., 2016a).

In CG models, the atomic detail of the complex protein structure is lowered by the use of continuous representations of the geometry of the modeled structures which ultimately reduces the degrees of freedom (DOF) (Ingólfsson et al., 2014). Coarse-graining can be done either in MD or in MC simulation based on the demands of the problem. Here in our work, our problem is solved by coarse-grain MC simulation (Fritsche et al., 2013; Kitjaruwankul et al., 2016; R B Pandey et al., 2015; Ras B Pandey & Farmer, 2012) where we have analyzed lots of local and global physical quantities for the conformational analysis of the hH_V1 protein.

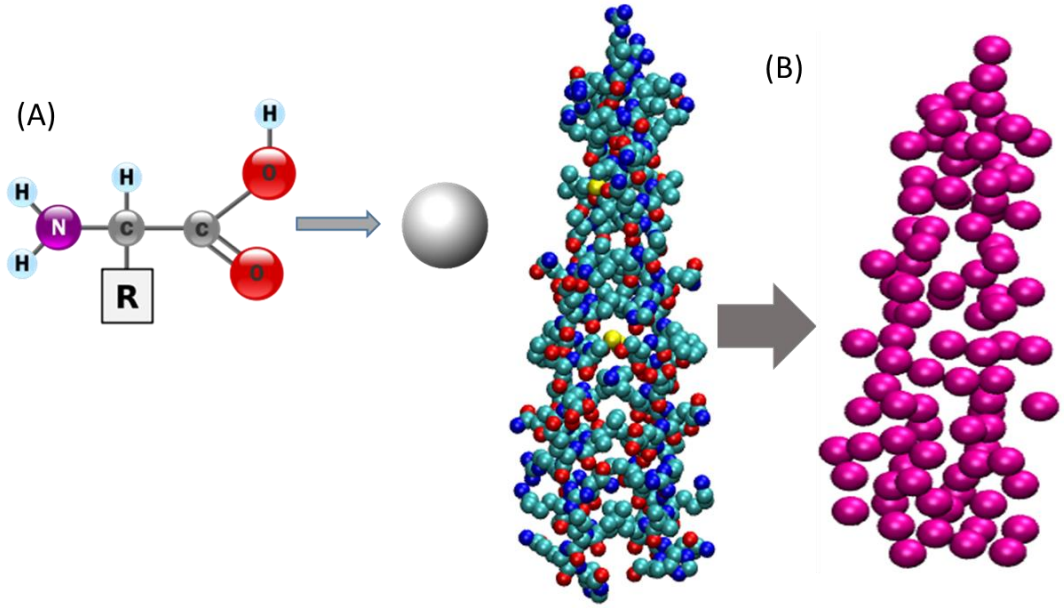


Figure 2.1 (A) One amino acid is equivalent to one residue in Coarse-grained Model. (B) All-atom representation of D-hHV1-CTD (PDB: 3A2A) and its corresponding coarse-grain model with 98 residues.

Note: This visualization is done by Visual Molecular Dynamics (VMD).

Numerous methods have been used for performing the coarse-graining of the protein chain. Among them we have chosen the one where we assign one bead for one amino acid as shown in the Figure 2.1 (A). The CG model of the protein from the all-

atom model can be done by different software packages. Figure 2.1(B) shows the CG model of the D-hH$_V$1-CTD from the all-atom model of if from the Visual Molecular Dynamics (VMD) software (Humphrey, W., Dalke, A. and Schulten, 1996). This simplified model is computationally less expensive and can be simulated for larger time steps. We have done up to ten million MC steps to get results.

2.2.4 Bond Fluctuation Model (BFM)

The Bond-Fluctuation Model (BFM) is a discrete lattice model in which bond length can vary to some extent. In this model, the smallest bond length is 2 (lattice units) and other allowed bond lengths are √5, √6, 3 and √10. These bond lengths occur on a cubic lattice. This model has some of the bond-length and bond-angle flexibility contained in the off-lattice model while maintaining some of the advantages of the discrete lattice model with rigid bonds. The bond-fluctuation approach used here is one of the most efficient and effective methods to investigate large-scale complex problems (Binder, 2008) in polymers, soft matter and proteins.

2.2.5 Excluded Volume Effect

The excluded volume effect is one of the constraints that we apply while moving the residues in the Metropolis algorithm in which two residues cannot occupy the same lattice site.

The simulations performed for this thesis include two chains: one monomer (M-hH$_V$1-CTD) of 49 residues and a tandem dimer (tD-hH$_V$1-CTD) of 104 residues that moves in a cubic box with the constraints of bond length and excluded volume.

CHAPTER III – SIMULATION METHODS

The main simulation employed in this work is a coarse-grained Monte Carlo simulation as discussed in the previous chapter. In this chapter, the main algorithm that is implemented is discussed.

3.1 Metropolis Algorithm

The Monte Carlo Simulation involves coarse-graining underlying host space and including structural details of the constituents and interactions as described in a recent study (Kitjaruwankul et al., 2016). The C-terminal domain of protein chain, M-hH$_V$1-CTD and tD-hH$_V$1-CTD, are represented by 49 and 104 residues as shown in Figure 1.4. They are tethered together by covalent (peptide) bonds on a cubic lattice in our coarse-grained approach. A residue occupies a cube of size $(2a)^3$ where a is the lattice constant; the excluded volume requires that two consecutive nodes along the backbone cannot be closer than two lattice constants. The covalent bond between successive nodes varies between 2 and √(10) in units of the lattice constant. The degrees of freedom (DOF) for each node to move and bonds to fluctuate are constrained to a cubic lattice in comparison to continuum space in an all-atom MD simulation as described above. There are, however, ample DOF for each node to move and covalent bond to fluctuate unlike the minimalist protein chain models on a discrete lattice with constant bond length. The degree of freedom also can be enhanced by fine-graining(Ras B. Pandey & Farmer, 2013).

The protein chain (M-hH$_V$1-CTD or tD-hH$_V$1-CTD) is initially placed in a random configuration. Each residue interacts (Kitjaruwankul et al., 2016) with its

neighboring residues within a range (r_c) of interaction with a generalized Lennard-Jones potential,

$$U_{ij} = \left[|\varepsilon_{ij}| \left(\frac{\sigma}{r_{ij}} \right)^{12} + \varepsilon_{ij} \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \quad , r_{ij} < r_c \qquad (3.1)$$

where r_{ij} is the distance between the residues at site i and j, $r_c = \sqrt{8}$ and $\sigma = 1$ in the unit of lattice constant. The potential strength, ε_{ij} is unique for each interaction pair with appropriate positive (repulsive) and negative (attractive) values used from knowledge-based contact interactions. A number of knowledge-based residue-residue interaction matrices (Betancourt & Thirumalai, 1999; Chen, Zhang, & Ding, 2004; Miyazawa & Jernigan, 1985; Ras B. Pandey & Farmer, 2013) have been developed over the years; we have extensively used some of these matrices including classic Miyazawa-Jernigan (MJ) interaction (Miyazawa & Jernigan, 1985). The Betancourt-Thirumalai (BT) interaction (Betancourt & Thirumalai, 1999) matrix appears to be improved over the classic MJ, therefore, we use it in this study.

The Metropolis algorithm is used to move each residue randomly at a temperature T with the following procedure. The supporting information flow chart of the procedure is in APPENDIX A. First, we randomly select a residue from the chain of a protein (i.e. the chain of 49 residues in M-hH$_V$1-CTD and the chain of 104 residues in tD-hH$_V$1-CTD) and one of its 26 neighboring lattice sites to move to. For instance, a residue at a site is selected randomly to move to a randomly selected neighboring lattice site, j. If site j is available, then, we check constraints (i.e. whether the change was permitted on the basis of the covalent bond length associated with the residue as a result of its move and by the excluded volume effect). Only if these constraints (bond fluctuation effect and excluded

volume) are satisfied, the residue is moved from site i to site j with Boltzmann probability exp(-ΔE_{ij}/T), where $\Delta E_{ij} = E_j - E_i$ is the change in energy between its new (E_j) and old (E_i)configurations; T is the temperature in the reduced units of the Boltzmann constant. An attempt to move each residue once defines the unit Monte Carlo Step (MCS). As the residues execute their stochastic motion, we keep track of a number of local and global physical quantities such as mean square displacement of the center of mass of the protein, the energy of each residue, its mobility, radius of gyration, and its structure factor. Simulations are performed for a sufficiently long time typically ten million time steps (10^7 steps) at each temperature with many independent samples (typically 100 and 200 samples) to estimate the average values of these quantities.

3.1.1 Lattice Size

In this simulation we use a cubic box of size 150^3 for the monomer (M-hH_V1-CTD) and the cubic box of size 340^3 for the tandem dimer (tD-hH_V1-CTD). But for the verification of our findings we have used different lattice sizes at the beginning and came to the conclusion that our findings are independent of the lattice size. We have used two different sizes of the cubic boxes for the monomer and the tandem dimer for providing sufficient space to move the residues.

3.2 All-atom Molecular Dynamics

For the all-atom MD simulation the following procedure is implemented. A 3D structure of hH_v1-CTD was obtained from Li et al (S. J. Li et al., 2010). The homology model of hH_v1 was generated using the structure refinement PaDSAR method (Sompornpisut, P.; Roux, B.; Perozo, 2008) that incorporates spin-labeling EPR solvent accessibilities as interaction restraints. Simulation systems were constructed with VMD

(Humphrey, W., Dalke, A. and Schulten, 1996). The all-atom dimer hH_v1-CTD model (D-hH_v1-CTD^{all}) was embedded in 150 mM NaCl solution. The simulation system had periodic box dimensions of ~73 × 73 × 103 Å. Side-chain ionization states were assigned based on pK_a calculations using the PROPKA(Olsson, M. H.; Sondergaard, C. R.; Rostkowski, M.; Jensen, 2011) with respect to pH 7. All molecular dynamics simulations were performed using NAMD(Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, 2005) with CHARMM22 force field parameters (MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; R, 1998) at fifteen different temperatures, ranging from *298 K-713 K*. Langevin dynamics was applied to keep the desired target temperature. Energy minimization used 6000 steps for each to relax any steric conflicts generated during system setup. NVT ensemble was applied to the systems which is appropriate for high temperature simulation. Each MD simulation was run for 45ns with a time step of 2 fs. A distance cut-off of 12 Å was used for calculating non-bonded interactions in both electrostatic interactions with particle mesh Ewald method and van der Waals interactions. A switching distance was set at 10Å(Darden, T.; York, D.; Pedersen, 1993). The TIP3P water was employed in simulations(Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, 1983). The structure analysis of all MD trajectories including the root-mean square deviation (RMSD) and radius of gyration (R_g) were analyzed.

3.3 Statistical Analysis

Proteins are analogous to all other molecules in the universe, which are subjected to the physical forces that empower the universe (Kessel & Ben-Tal, 2011). With the evaluation of one of the statistical physical quantity, entropy, a quantity related to the possible configuration and disorder of the system, the direction of the natural process i.e. folding and unfolding of the protein was predicted initially(Kessel & Ben-Tal, 2011). This concept came from the second law of thermodynamics which states that, "The entropy of an isolated system increases during any spontaneous process and approach to the equilibrium". Biological systems are in the state of constant temperature and pressure with fluctuating energy and volume where entropy cannot determine the direction of process rather Gibb's free energy is used in this circumstance. According to Kessel & Ben-Tal, "Spontaneous processes always proceed toward equilibrium by decreasing the free energy of the system to a minimum"(Kessel & Ben-Tal, 2011) . At equilibrium, the free energy change is zero, and the change in the standard free energy (ΔG^o) is related to the equilibrium constant (K_{eq}).

$$\Delta G^o = -RTlnK_{eq} \qquad (3.2)$$

where R the universal gas constant, and T is the temperature. This standard condition may not be practical in every case so the more general free energy change (ΔG), in term of ΔG^o along with the concentrations of the initial and final states of the process is,

$$\Delta G = \Delta G^o + RTln\frac{p[product]}{r[reactants]} \qquad (3.3)$$

where function p and r depend on actual product and reactant concentration. Hence by calculating the free energy change (ΔG), one is able to determine whether the process is spontaneous or not.

In addition to this spontaneous process, comprehending the free energy (ΔG) of a system is very important in estimating the relative stability of states and the direction of processes like protein folding, protein-ligand binding, and enzyme-catalyzed processes (Kessel & Ben-Tal, 2011) and can be written in term of enthalpy (ΔH) and entropy (ΔS),

$$\Delta G = \Delta H - T\Delta S \qquad (3.4)$$

From equation (3.4), enthalpy (ΔH) and entropy (ΔS) depicts the general tend to minimize energy and maximize disorder. Enthalpy (H) is the thermodynamic quantity which gives the quantitative information about the total heat content of the system. It is the sum of internal energy (E) of the system plus the product of pressure (P) and volume (V) as,

$$\Delta H = \Delta E + PV \qquad (3.5)$$

But biological systems usually do not experience a change in pressure and volume as they are either solid or liquid. Therefore the enthalpy change of such systems is mainly from the change in internal energy alone. Therefore,

$$\Delta H \approx \Delta E \qquad (3.6)$$

Internal energy (E) of the system is the sum of potential energy (U) and kinetic energy (K) as,

$$E = U + K \qquad (3.7)$$

Here, the potential energy (U) is the type of energy resulting from all covalent and most non-covalent interactions in the molecular system which is explained in detail

in Section 2.1.1 and the kinetic energy (K) is the result of thermally induced atomic motions in the molecule. In brief, the enthalpy change (ΔH) component of free energy change (ΔG) results from the formation or breaking of covalent bonds, changes in electrostatic or Van-der Waals interactions, and changes in thermally induced atomic motions(Kessel & Ben-Tal, 2011). Also, under constant pressure there occur transfer of heat ($Q_{(p)}$) between the system and its environment so can be written as,

$$Q_{(p)} = \Delta H \approx \Delta E \qquad (3.8)$$

This equation (3.8) is very helpful for the structural biophysicist to determine the change in enthalpy in the system by measuring the heat transfer to or from the system. The formation of an energetically favorable bond or non-covalent interaction, which is common in biological systems leads to decrease of ΔH by releasing heat from the system to its surroundings.

$$Q_{(p)}^{rev} = \Delta H \qquad (3.9)$$

The heat capacity (C_p) can be obtained by the relation;

$$C_p = \frac{Q_{(p)}^{rev}}{\Delta T} = \frac{\Delta H}{\Delta T} \qquad (3.10)$$

The next important part of the free energy change (G) is entropy (S) which is the natural logarithm of the number of possible configurational states of the system (Ω) multiplied by the Boltzmann constant (k_B) .

$$S = k_B ln\Omega \qquad (3.11)$$

Also, the change in entropy (ΔS) can be written in terms of the change in enthalpy (ΔH), heat transfer ($Q_{(p)}^{rev}$) and heat capacity (C_p) as,

$$\Delta S = \frac{\Delta H}{T} = \frac{Q_{(p)}^{rev}}{T} = \left(\frac{C_p}{T}\right)\Delta T \qquad (3.11)$$

The equation (3.11) allows the structural biophysicist to obtain change in entropy (ΔS) experimentally by simply measuring the simple heat capacity of the system(Kessel & Ben-Tal, 2011).

The next important measure for the structural similarity of the protein is root mean square displacement (RMSD), which measures the structural similarity between different proteins and is informative for the biophysicist to determine its structure. Basically there are two main criteria for protein folding as stated in Kessel and Ben-Tal, "The First is tight packing of atoms, needed to maintain compactness and optimize interaction stability and the next is efficient pairing of backbone amide and carboxyl groups in hydrogen bonds which is required to reduce the destabilizing effect of their de-solvation" (Kessel & Ben-Tal, 2011). For this the two structures to be compared are superimposed and the RMSD is calculated as follows;

$$RMSD = \left(\frac{\sum d_i^2}{n}\right)^{1/2} \qquad (3.12)$$

where n is the number of atoms compared between the two structures and d_i is the distance between each pair of atoms.

The RMSD value increases with the degree of structural dissimilarity and tends to depend on secondary structure to a greater extent than loops. This process of evaluating RMSD is done in all-atom MD while the way to determine RMSD is different in coarse-grain MC simulation. Root mean square displacement (R_C) as used in this case is a measure of the distance between the position of a particle (x) and some the reference point (x_o) and is the measure of the spatial extent of random motion. In the context of

biophysics, it is measured overtime to determine the diffusion behavior of the particles. Mathematically it is defined as;

$$R_c^2 = \langle (x - x_o)^2 \rangle = \frac{1}{N} \sum_{n=1}^{N} (x_n(t) - x_n(0))^2 \qquad (3.2)$$

where N is the number of residues to be averaged, $x_n(0) = x_o$ is the reference position of each residue, $x_n(t)$ is the position of each particle at time *t*. The linear relationship between the mean square displacement with time allows for the determination of a diffusivity constant (D) which can be written as;

$$\langle R_c^2 \rangle = d_i D t \qquad (3.3)$$

where d_i is the numerical constant which depends on dimensionality, d_i = 2,4 and 6 for 1D, 2D and 3D respectively, D is the diffusion coefficient and t is the time.

The radius of gyration (R_g) of a protein is the overall spread of the residues and is a measure of a protein's compactness. It is defined as the root mean square distance of the collection of residues from their common center of mass.

$$R_g = \sqrt{\sum_{i=1}^{N} \frac{(r_i - R_{cm})^2}{N}} \qquad (3.4)$$

where r_i is the position of the residue, R_{cm} is the center of mass of the protein chain with mass M and m_i is the mass of each residue. The statistical error of R_g can be calculated from the standard deviation. The center of mass is mathematically defined as,

$$R_{cm} = \frac{1}{M} \sum_{i=1}^{N} m_i r_i \qquad (3.5)$$

Furthermore, the spatial distribution of the residues of a protein chain, as the protein structure evolve can be examined by analyzing the structure factor S(q),

$$S(q) = \langle \frac{1}{N} \left| \sum_{j=1}^{N} e^{-iq.\vec{r_j}} \right|^2 \rangle_{|\vec{q}|} \qquad (3.6)$$

Where r_j is the position of each residue of the protein chain and $|q| = 2\pi/\lambda$ is the wave vector of wavelength λ

Different local and global physical quantities are analyzed by coarse-grain MC simulation where there is stochastic motion of the residues of the protein M-hH$_V$1-CTD and tD-hH$_V$1-CTD in a cubic lattice of size 150^3 and 340^3 respectively. While moving the residues, we keep track of the mean square displacement of the center of mass of the protein, the energy of each residue, its mobility, and contact map, radius of gyration and structure factors.

CHAPTER IV – RESULTS

All the results obtained from the simulation of M-hH$_V$1-CTD and tD-hH$_V$1-CTD for the local and global physical quantities are included in this section along with visual analysis of the snapshots. As it is easier to keep track of some of these physical quantities in a coarse-grained model, a large fraction of the results involve the CG approach and only a few of the results of all-atom approach are introduced to verify our result. However, the main result of all-atom MD simulation on the response of radius of gyration of the protein to temperature is compared with the results from the CG simulation. This enhances our understanding of the structural evolution of the protein and pathways for proton transport.

4.1 Local Physical Quantities

A number of local physical quantities of M-hH$_V$1-CDT and tD-hH$_V$1-CDT are studied to investigate the structural evolution of the protein chain.

Table 4.1

Different Analytical and Comparative term of monomer and tandem dimer

Protein Chain / Different Terms	M-hH$_V$1-CTD	tD-hH$_V$1-CTD
Low Temperature Regime (Reduced Scale)	0.010 – 0.020	0.016 -0.024
High Temperature Regime (Reduced Scale)	0.020 – 0.038	0.024 – 0.040
Lattice Size (Lattice constant)	150^3	340^3
Independent Samples	200	100

Note: These are the range of temperature and other sizes that are used in coarse-grained model.

The range of low and high temperatures depends on the size of protein chain so it is different for M-hH$_V$1-CTD and tD-hH$_V$1-CTD which can be seen in the Table 4.1. All of the quantities have arbitrary units in CG model therefore we need some other simulation to compare our result and to calibrate our data to get a realistic data. For this purpose most of our result in CG MC is compared with AA MD simulation obtained by our collaborators.

4.1.1 Snapshots

The very first visual analysis that gives a clearer understanding of the structure of the protein is its snapshot and contact map. Figure 4.1 shows a set of snapshots at low temperatures (T = 0.010, 0.014, 0.018, 0.020) of the conformation of M-hH$_V$1-CTD with the coarse-grain approach.

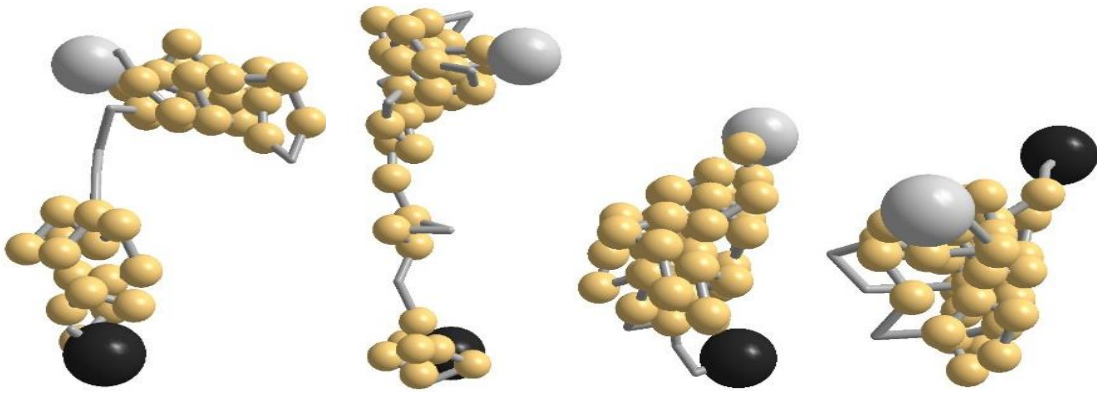


Figure 4.1 Snapshots of the protein M-hHV1-CTD conformation at the end of 10^7 MC steps with $T = 0.010, 0.014, 0.018, 0.020$ from left to right.

Note: Large spheres represent the end residue (residue 1- black and residue 49 – gray) and small spheres represent those within the range of interaction.

Although a snapshot is not enough to identify the complete trend, it does provide a glimpse of the structure. One can immediately see that the protein (M-hH$_V$1-CTD)

becomes more compact on increasing the temperature in the low- temperature regime. This may seem contrary to expectations from general physics. However, the protein chain expands on increasing the temperature further as shown in the snapshots below. Similarly, visual analysis of the snapshots of tD-hH$_V$1-CTD shows a similar trend in the low temperature regime (i.e. it contracts on increasing the temperature). The overall size of tD-hH$_V$1-CTD seems to contract on raising the temperature in the range $T = 0.016 - 0.024$ and expand over the temperature range $T = 0.024 - 0.040$.

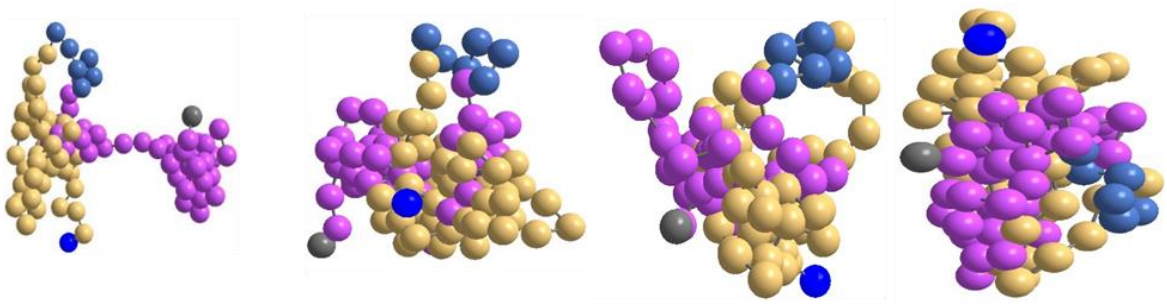


Figure 4.2 Snapshots of the protein tD-hHV1-CTD conformation at the end of 10^7 MC steps with temperature $T = 0.018, 0.020, 0.022, 0.024$ from left to right.

Note: Here, the grey sphere represents residue 1, bright blue sphere represents residue 104, dark blue spheres represent tandem linker residues 50-55, pink spheres represent residues 2-49 and gold spheres represent residues 56-103.

In the low temperature regime, the residue –residue interaction dominates the thermal interaction which results in the contraction of the protein chain. This contraction increases the connectivity among the residues which further helps in the transport of protons in the channel. While in the high-temperature regime, the size of both protein chains M-hH$_V$1-CTD and tD-hH$_V$1-CTD increase on increasing temperature and reach a steady state which is shown in Figure 4.3 and Figure 4.4 respectively.

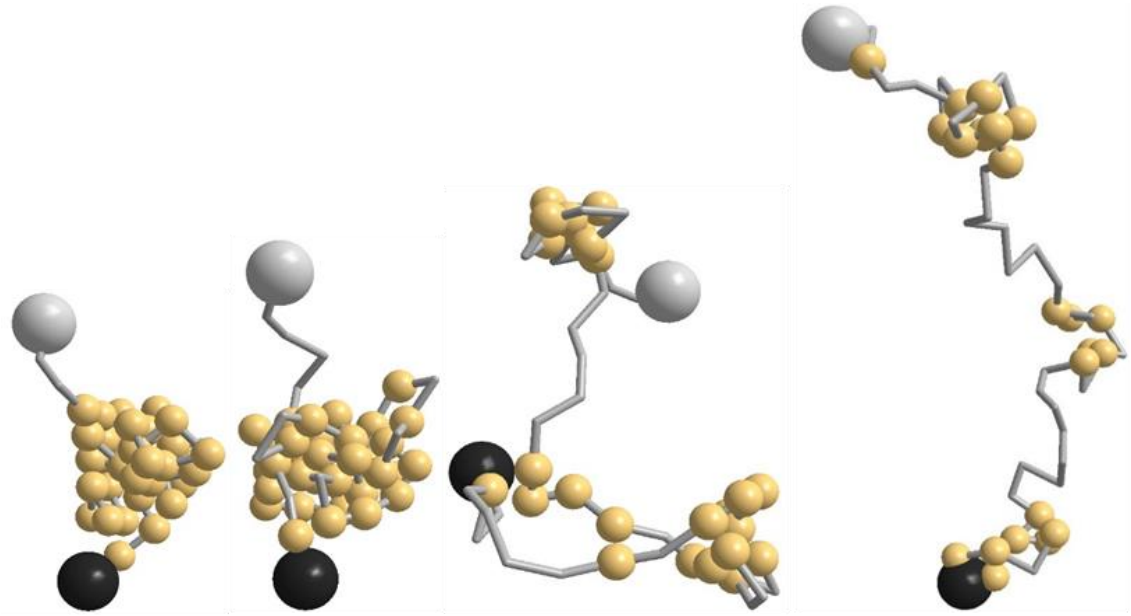


Figure 4.3 Snapshots of the protein M-hHV1-CTD conformation at the end of 10^7 steps with $T = 0.024, 0.028, 0.030, 0.034$ from left to right.

Note: Large spheres represent end residues (residue 1-black and residue 49-grey). Small spheres represent residues within the range of interaction.

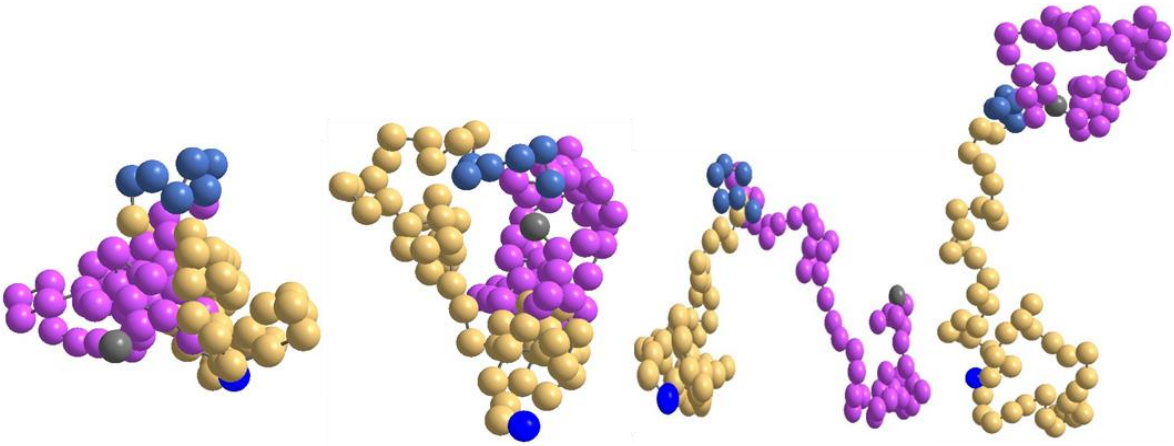


Figure 4.4 Snapshots of the protein tD-hHV1-CTD conformation at the end of 10^7 MC steps with $T = 0.026, 0.030, 0.034, 0.038$ from left to right.

Note: Here, the grey sphere represents residue 1, bright blue sphere represents residue 104, dark blue spheres represent tandem linker residues 50-55, pink spheres represent residues 2-49 and gold spheres represent residues 56-103.

4.1.2 Contact Maps

A contact map represents the distance between all possible amino acids pairs of a 3D protein structure using a two dimensional structure. It provides a glimpse of residues in a local proximity (i.e. within the range of interaction) which may change by changing the temperature as seen in the following figures. A few representative contact maps of the tandem dimer tD-hH$_V$1-CTD at low and high temperature are presented in Figure 4.5 and Figure 4.6.

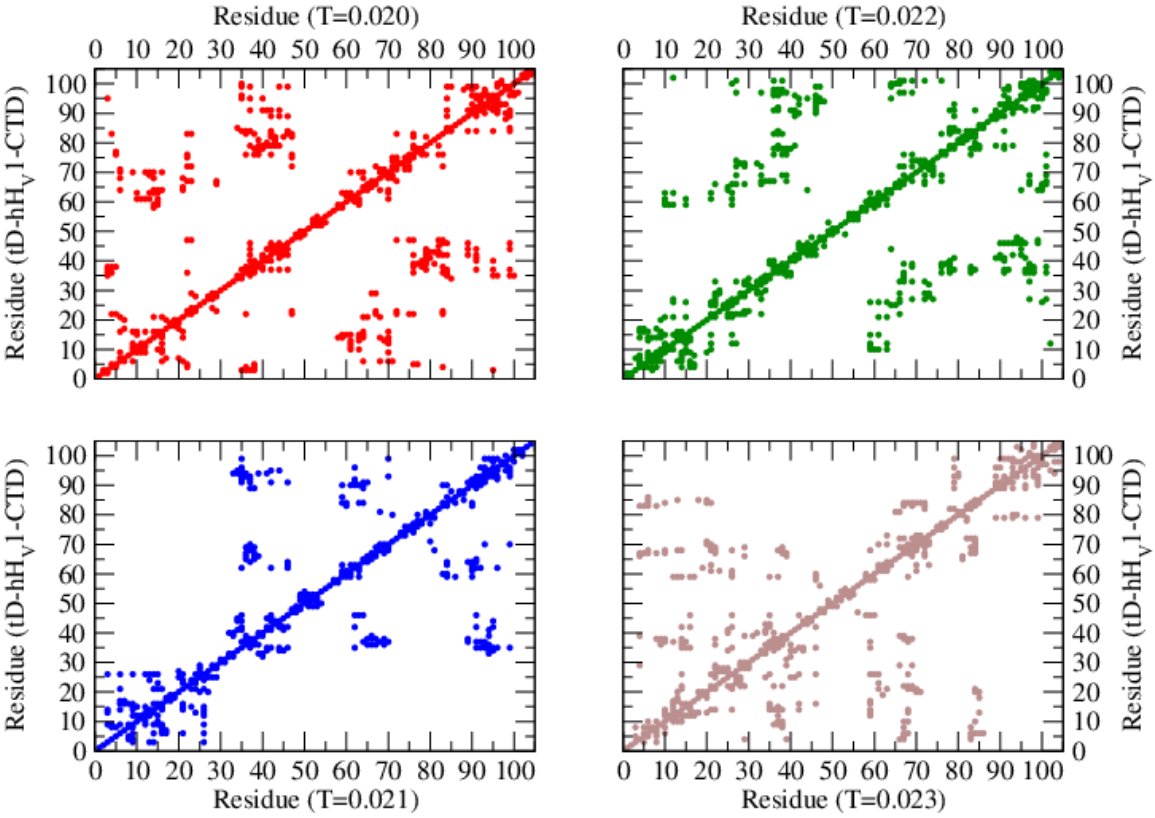


Figure 4.5 Contact maps of protein tD-hHV1-CTD in the low temperature regime from $T = 0.020 - 0.023$ for 104 residues.

To understand the figures, it would be beneficial to remember the sequence of residues in the two monomers i.e. ^{1}H -^{49}I and ^{56}I -^{104}H. At the temperature $T = 0.020$

(Figure 4.5), we see some localized assembly almost around the same segments ^{1}H-^{22}L and ^{84}E-^{104}H in both monomers almost like a mirror image. This map changes on changing the temperature. At low temperature we can see the connectivity of the residues increase on raising the temperature which is illustrated in Figure 4.5.

While in the high-temperature regime, the connectivity of the residues reduces on increasing the temperature further. However, the mirror symmetry around the tandem linker ^{50}A-^{55}A seems to persist at $T = 0.020 - 0.022$ with some fluctuations at $T = 0.023$. At high temperature ($T = 0.028, 0.029$) larger loops appear which eventually vanish on raising the temperature further ($T = 0.030, 0.031$) which is seen in Figure 4.6

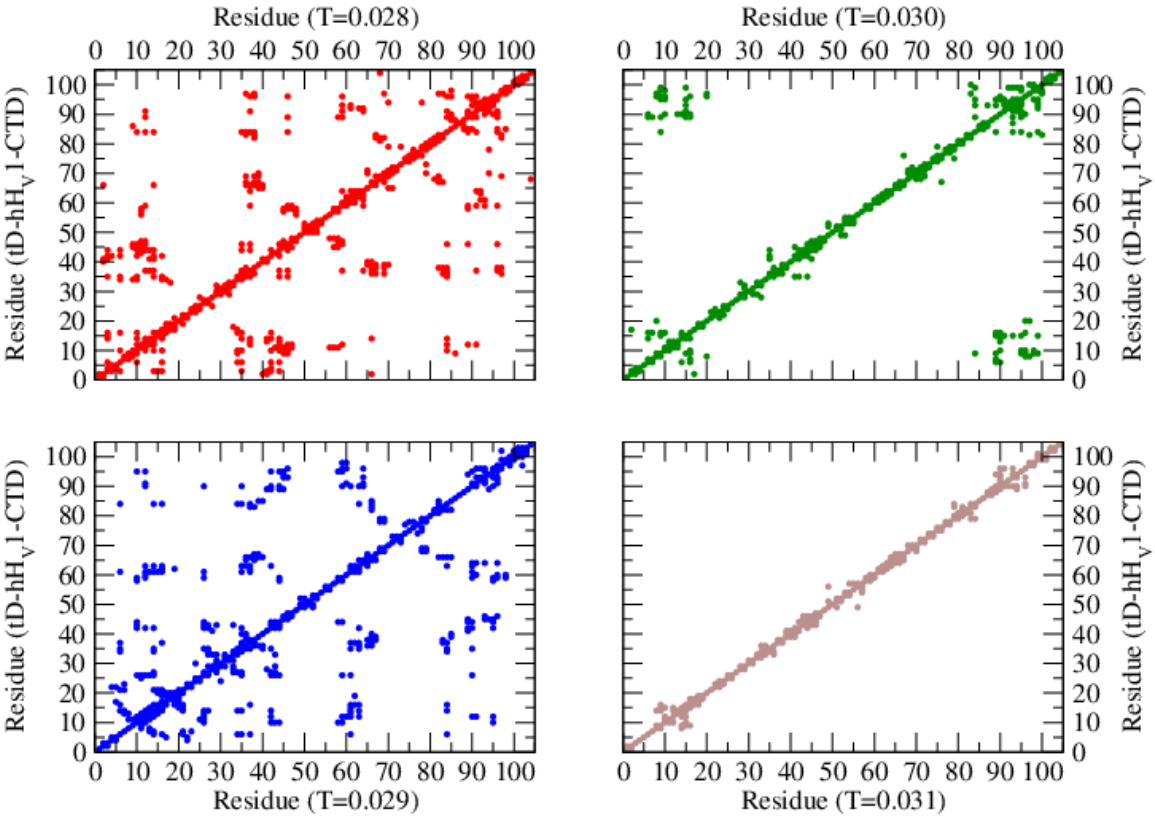


Figure 4.6 Contact maps of the protein tD-hHV1-CTD in the high temperature regime from $T = 0.028 - 0.031$ for the 104 residues.

4.1.3 Mobility Profiles

The mobility (M_n) of a residue is the probability of its successful moves per unit time step in our course-grain MC simulation. So, to understand the transport of protons via the connecting pathways along the structure of the protein, it would be informative to examine the local mobility and structure profile. Since the structural evolution of M-hH$_V$1-CTD appears to be preserved in its tandem dimer as seen above, we would like to concentrate now on tD-hH$_V$1-CTD. The mobility profiles of the residues of tD-hH$_V$1-CTD at low temperatures are presented in Figure 4.7.

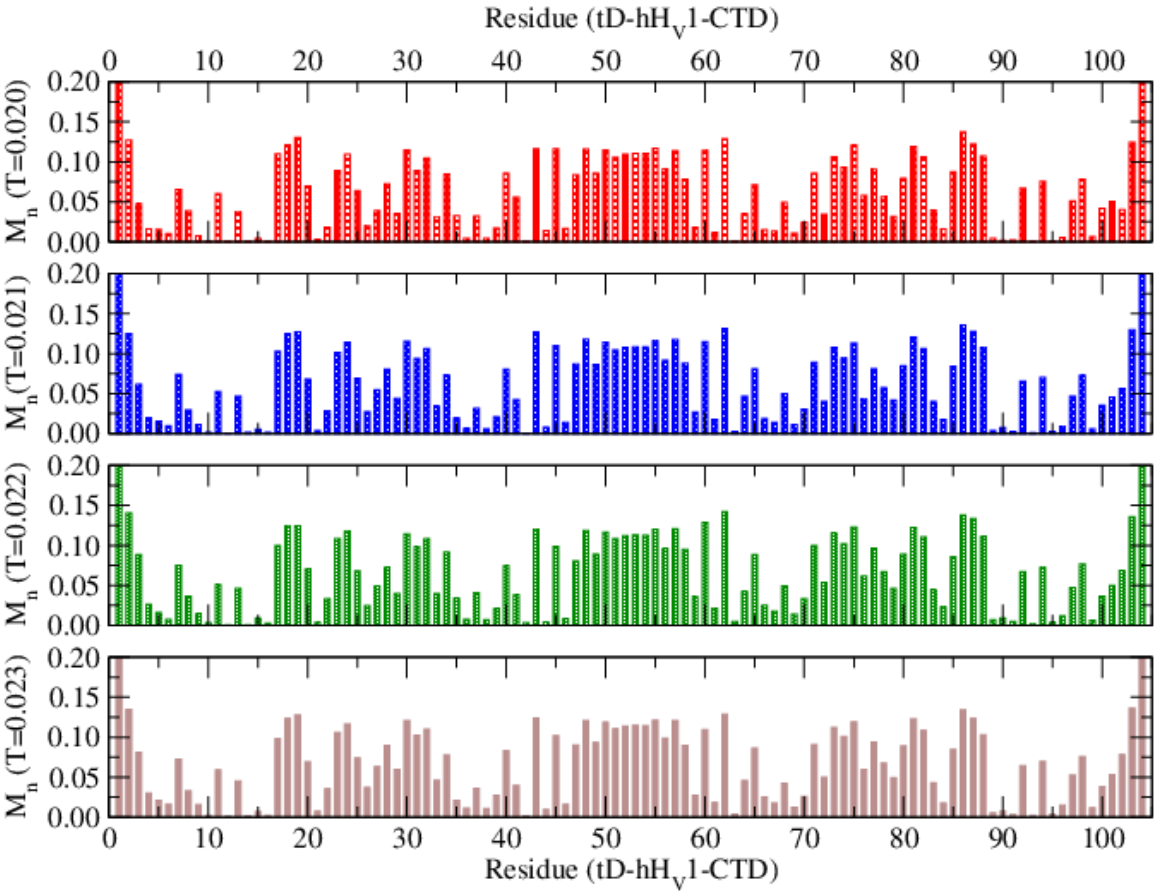


Figure 4.7 Mobility (M_n) - successful hops per unit MCS of the protein tD-hHV1-CTD in the low temperature regime from $T = 0.020 - 0.023$ for the 104 residues.

Apart from the residues towards the end, the high mobility of residues (^{50}A-^{55}A) of the tandem link and its surrounding residues is persistent even in the low temperature regime as shown in Figure 4.7. In addition, there are localized segments with relatively high mobility (i.e. ^{17}S-^{20}F, ^{85}F-^{88}S) and low mobility (i.e. ^{7}N-^{15}K, ^{90}K-^{98}N) which have mirror symmetry around the tandem link. The mobility profiles of the residues of tD-hHv1-CTD at high temperatures are presented in Figure 4.8.

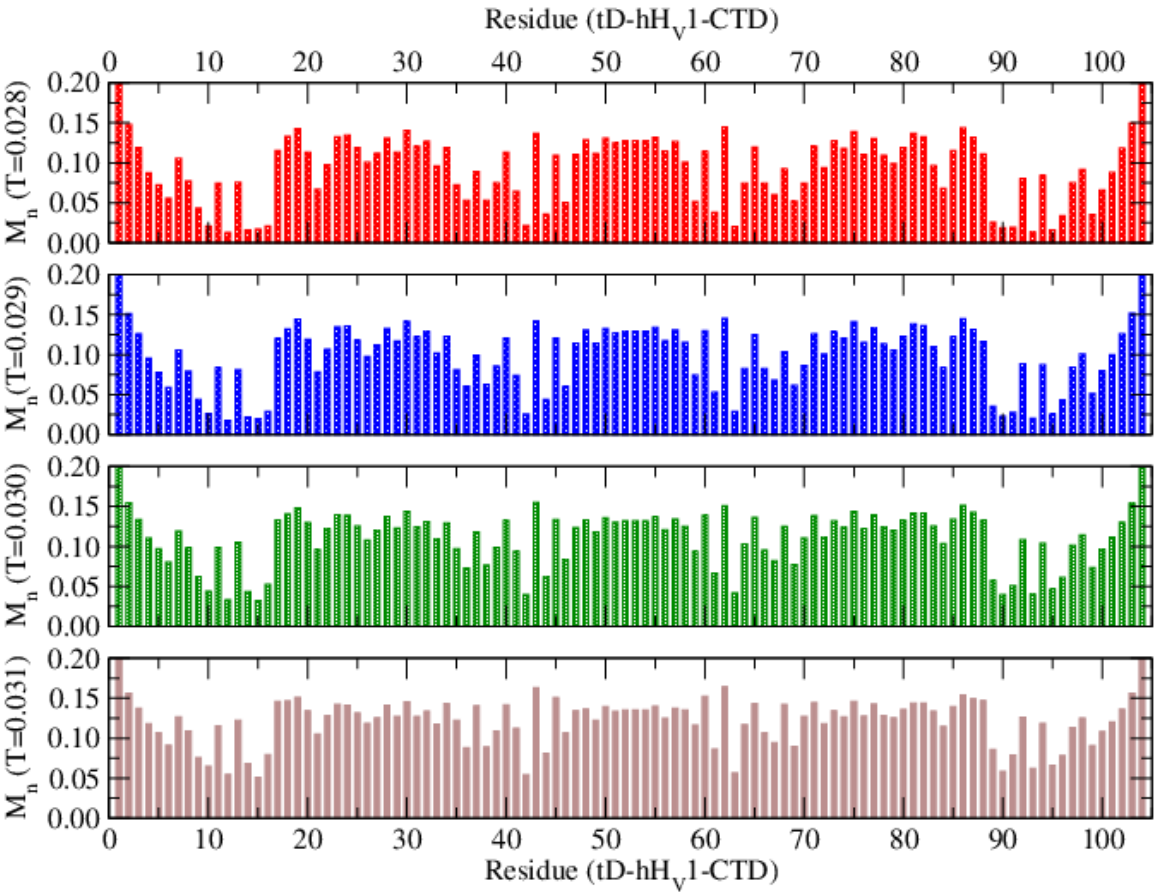


Figure 4.8 Mobility (Mn)- successful hops per unit MCS of the protein tD-hHV1-CTD for the high temperature regime $T = 0.028 - 0.031$ for the 104 residues.

Symmetry in mobility profiles seem to be consistent with the experimental observations that both monomers provide coordinated, similar pathways for proton

transport. Raising the temperature leads to higher mobility (Figure 4.8) while preserving the characteristic profile; the distinction among the local segments is however enhanced at lower temperatures (Figure 4.7).

4.1.4 Average Number of Residues

The average number of residues within the range of interaction is a measure of the interacting contact density which in the steady state provides some insight into the segmental morphology of the protein. The segmental structural profiles of the proteins at low and high temperatures are presented in Figure 4.9 and Figure 4.10 respectively.

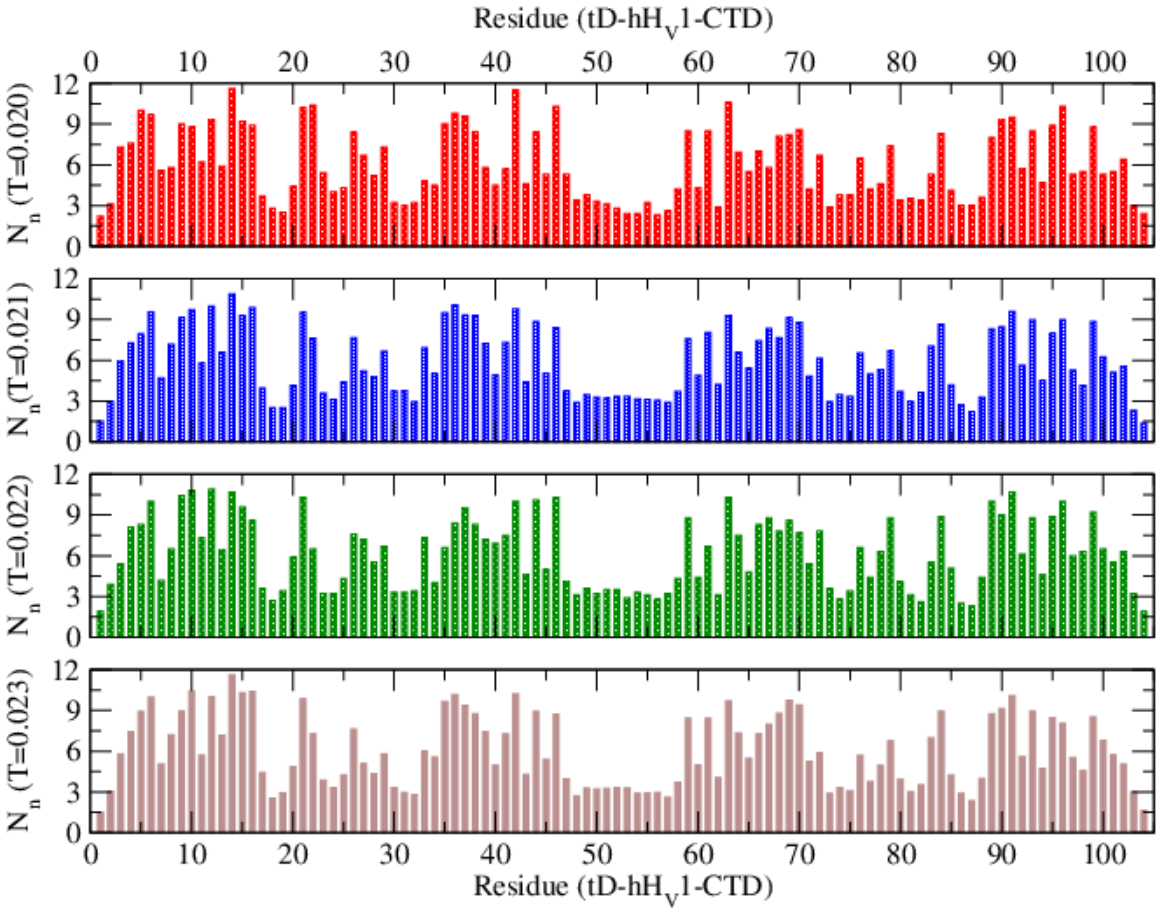


Figure 4.9 Average number N_n of residues within the range of interaction (a measure of the contact density profile) at temperature $T = 0.020 - 0.023$.

The mobility profile and self-assembled segmental residue profiles, look complementary to each other. Segments that have a larger number (N_n) of interacting residues have lower mobility. The segment containing the tandem link has a lower number (N_n) of surrounding residues and high mobility; it is highly flexible. Segments with higher values of N_n have more connecting pathways for proton transport. The distribution of segmental connectivity pathways is symmetric about the tandem link.

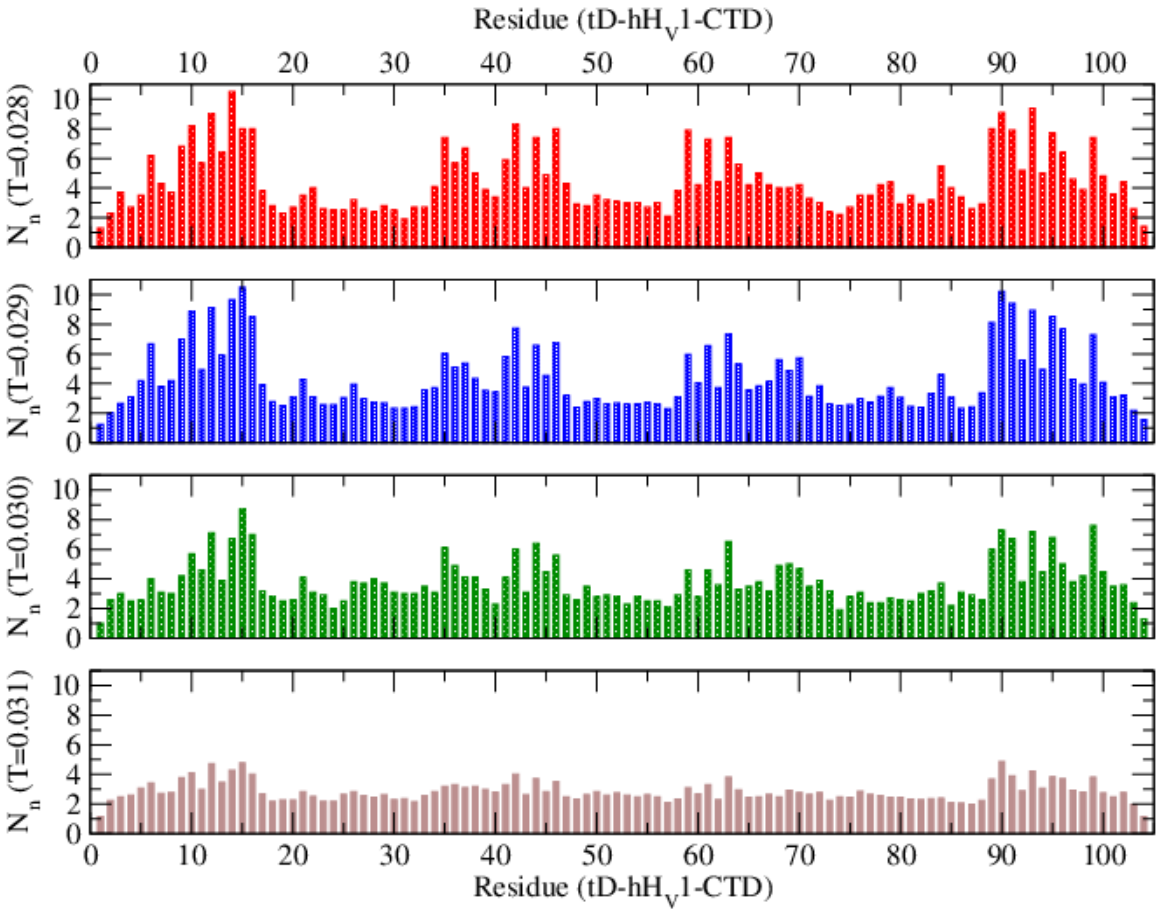


Figure 4.10 Average number N_n of residues within the range of interaction (a measure of the contact density profile) at temperature $T = 0.028 - 0.031$.

From the analysis of the segmental connectivity pathways, it can be seen that residues ^{3}R-^{16}E and ^{89}E-^{102}R have relatively high values of N_n (high contact density with local globular morphologies) and mirror symmetry around the tandem link. Also it is noticed that the distinction in distribution of segmental globular structures decreases on increasing the temperature (Figure 4.10) with low values of N_n at high temperatures where residue-residue interaction becomes irrelevant and the structure of the protein conforms to a random-coil conformation.

4.2 Global Physical Quantities

The main idea of coarse-grained model is to evaluate the local and global physical quantities effectively for large amount of time by ignoring all atomistic detail of the protein chain. Below are the results of the global physical quantities obtained from CG MC Simulation.

4.2.1 Root Mean Square Displacement

The variation of the root mean square displacement (R_c) of the center of mass of the protein (either M-hH_V1-CTD or tD-hH_V1-CTD) with Monte Carlo time Steps (t) reveals the structural change in the globular dynamics of the protein chain characterized by the power law $R_c \propto t^k$. Figure 4.11 depicts the log-log scale plot of the root mean square displacement (R_c) with the time, which shows the frozen state or no motion state ($k \to 0$) of the protein at low-temperatures (i.e. $T = 0.010, 0.013, 0.015, 0.018$) and then to sub-diffusion state ($\frac{1}{2} > k > 0$) where there is slow motion at the intermediate temperature ($T = 0.025$) in the transition regime. On further increasing the temperature ($T = 0.027, 0.028, 0.029$), highly mobile protein with diffusive motion ($k \approx 0.5$) occurs which is clearly illustrated in the Figure 4.11.

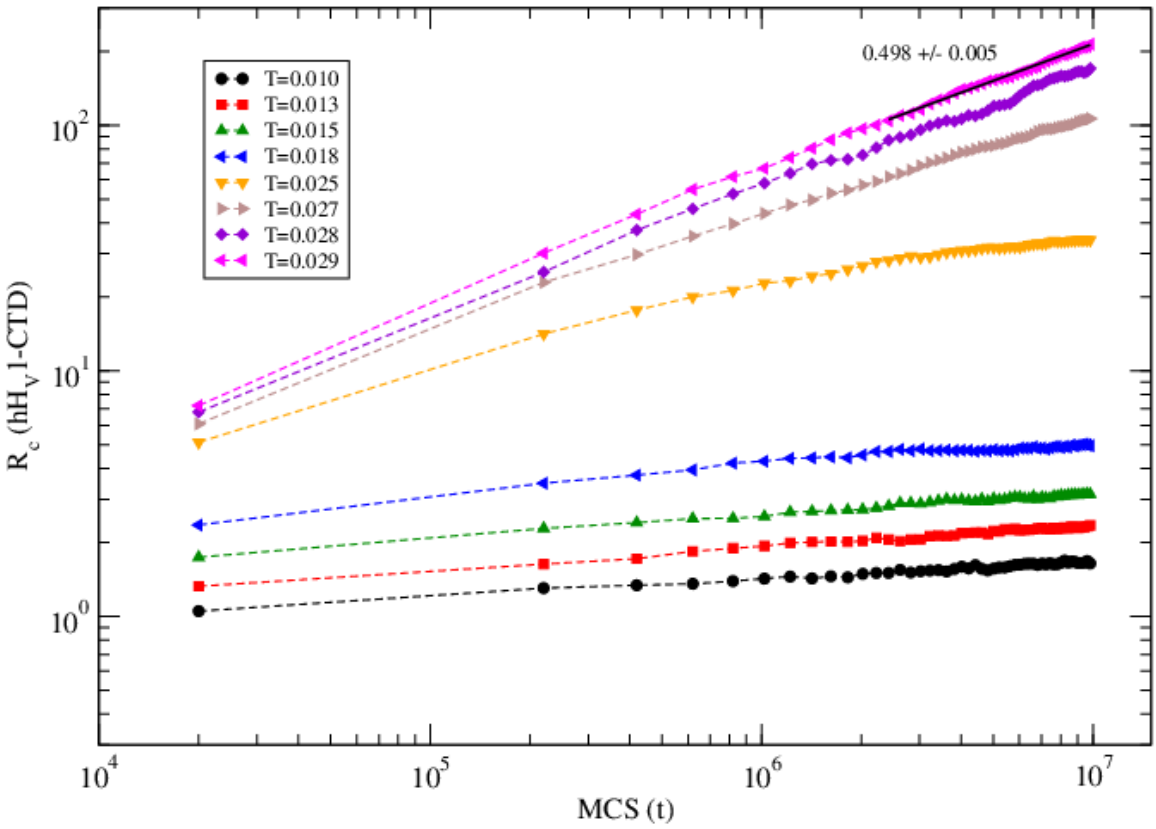


Figure 4.11 Log-log plot of Root Mean Square displacement (Rc) with Monte Carlo time Step (t) in the different temperature regimes.

4.2.2 Energy

The final folded conformation of the protein (either M-hHv1-CTD or tD-hHv1-CTD) is the one in which free energy is minimized. Figure 4.12 illustrates the variation of the energy with the Monte Carlo time Step (t) at different temperatures ($T = 0.010, 0.013, 0.015, 0.018$) in the low-temperature regimes.

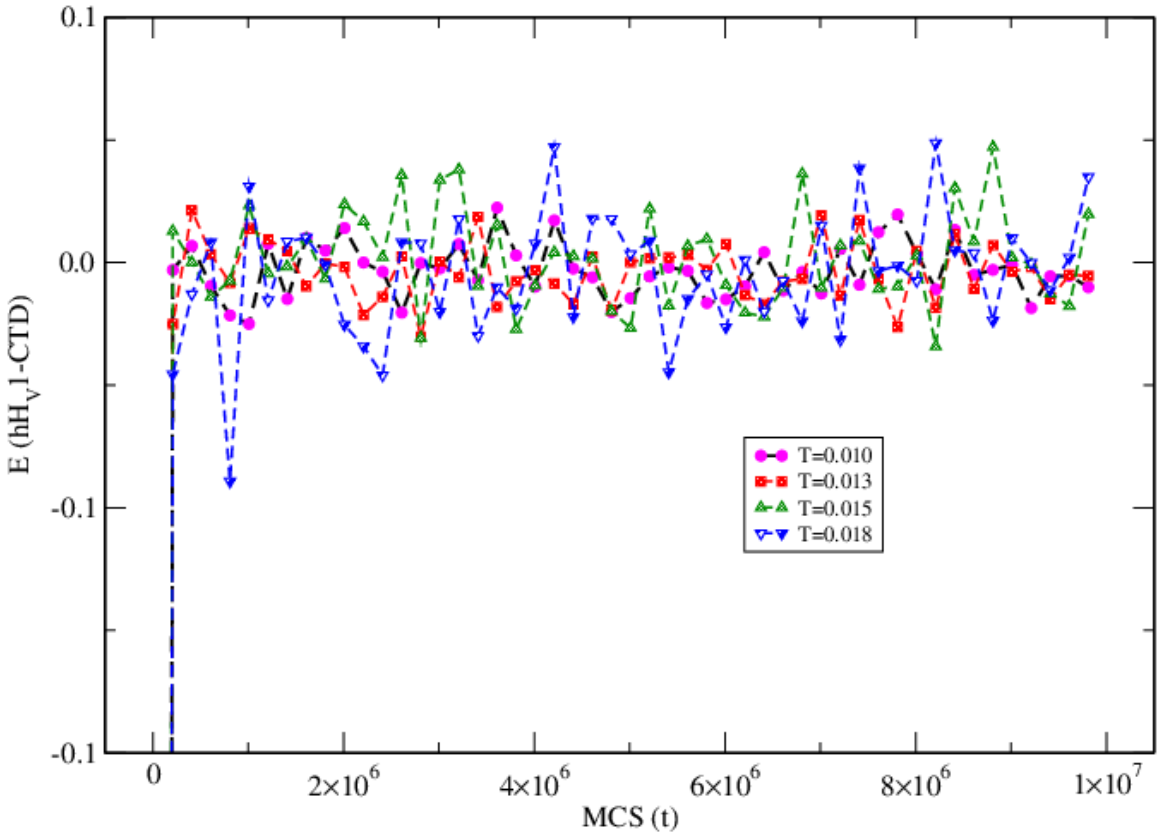


Figure 4.12 Variation of the Energy (E) with the Monte Carlo Time Steps (t) at different temperature regions in low-temperature regime.

4.2.3 Radius of Gyration (R_g)

The radius of gyration is an indicator of protein structure compactness in which it is concern about how can regular secondary structures are compactly packed into three-dimensional structure. Figure 4.13 illustrates the variation of the radius of gyration (R_g) with the Monte Carlo Time Step (t) in different temperature regions ($T = 0.010,\ 0.013,\ 0.015, 0.018$) at the low-temperature regime. It is believed that if a protein is stably folded, it will likely maintain relatively steady value of R_g while if a protein unfolds, its R_g changes over time which can be seen from Figure 4.13.

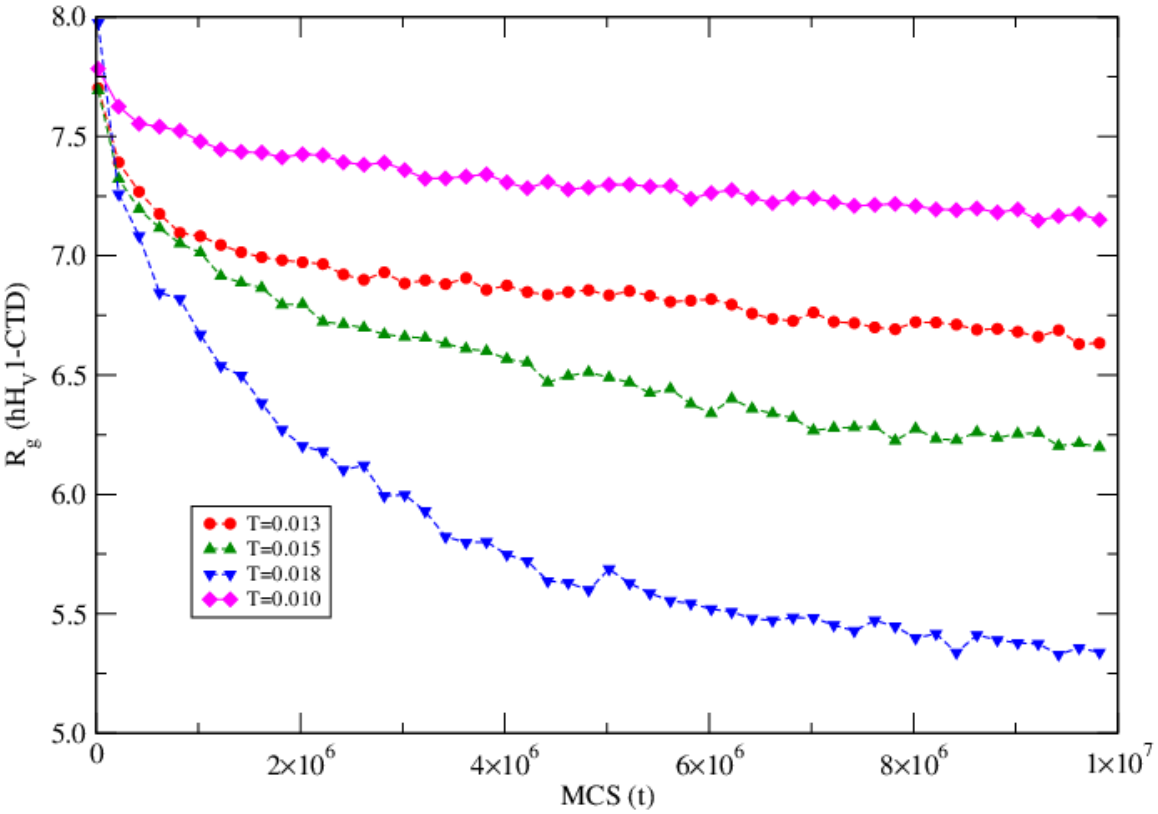


Figure 4.13 Variation of the radius of gyration (Rg) with the Monte Carlo Time Steps (t) in different temperature regions.

The size of the protein can be estimated by evaluating the radius of gyration (R_g) which gives substantial insight into the conformation of the protein. To strengthen our study, the radius of gyration is evaluated by both all-atom MD simulation as well as coarse-grained MC simulation. All-atom MD simulation was done by our collaborators and the result we got from both simulations is almost the same. Figure 4.14, Figure 4.15 and Figure 4.16 show the radius of gyration trend with temperature in the low-temperature regime in all-atom MD simulation and coarse-grain MC simulation respectively. It is seen that R_g of the protein (M-hH_V1-CTD and tD-hH_V1-CTD) decreases on increasing the temperature, which is not a typical thermal response.

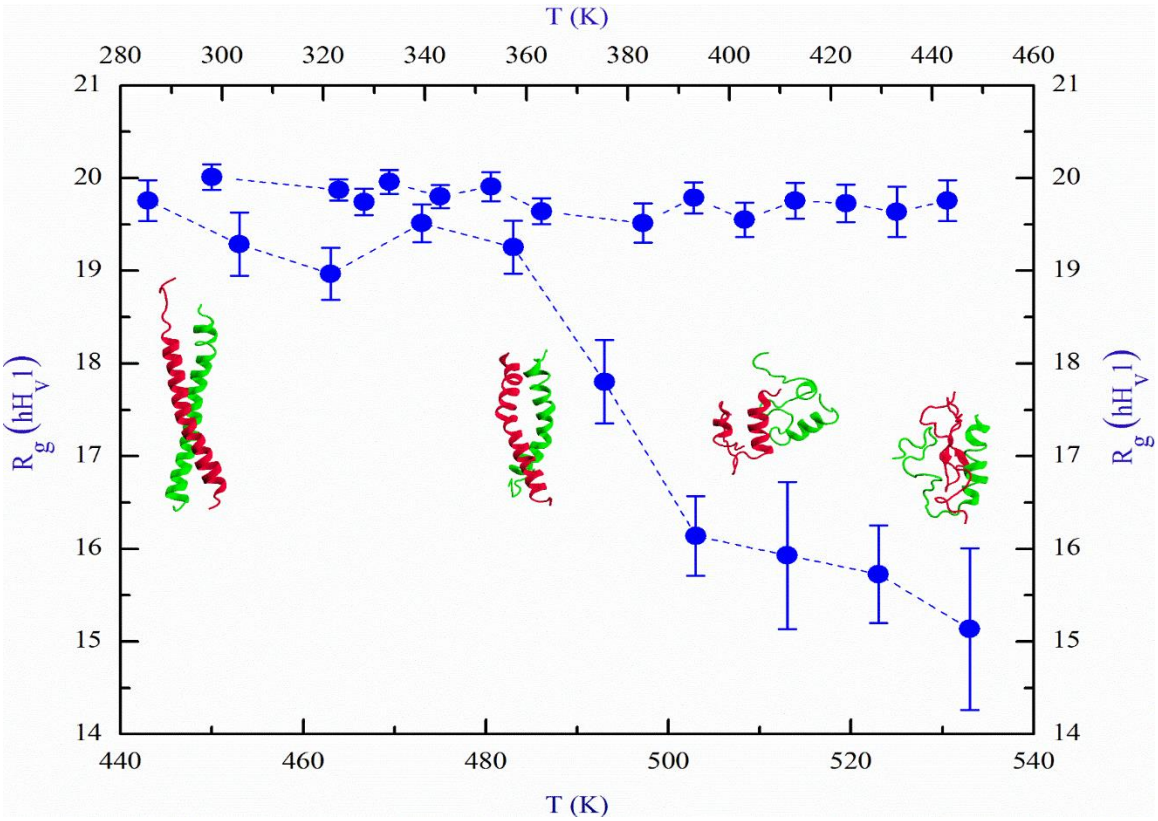


Figure 4.14 Variation of Radius of gyration (Rg) with temperature (T) obtained from all-atom MD simulation.

Note: This result is obtained by our collaborators in the low temperature regime for the protein D-hH_V1-CTD along with snapshots at particular temperatures. There is a sort of uniformity in the lower temperature (280K-440K) and after that on increasing the temperature (440K-540K) R_g drcreases. Data in the last 10ns at each independent temperature is used in estimating the average R_g.

A decay in the size of the protein on increasing the temperature is opposite what one would generally expect. Thus, the protein becomes more compact on increasing the temperature, a thermal-induced compaction. The variation of R_g with temperature resulting from the CG MC simulation data for both M-hH_V1-CTD and tD-hH_V1-CTD are presented in this section. Thermal response of the R_g of both monomer and dimer is clearly opposite in the two temperature regimes. . In the low-temperature regime (Figure 4.14, Figure 4.15 and Figure 4.16), R_g decays with the temperature while in the high

temperature regime (0, Figure 4.19 and Figure 4.20), it increases on raising the temperature and reaches a steady state.

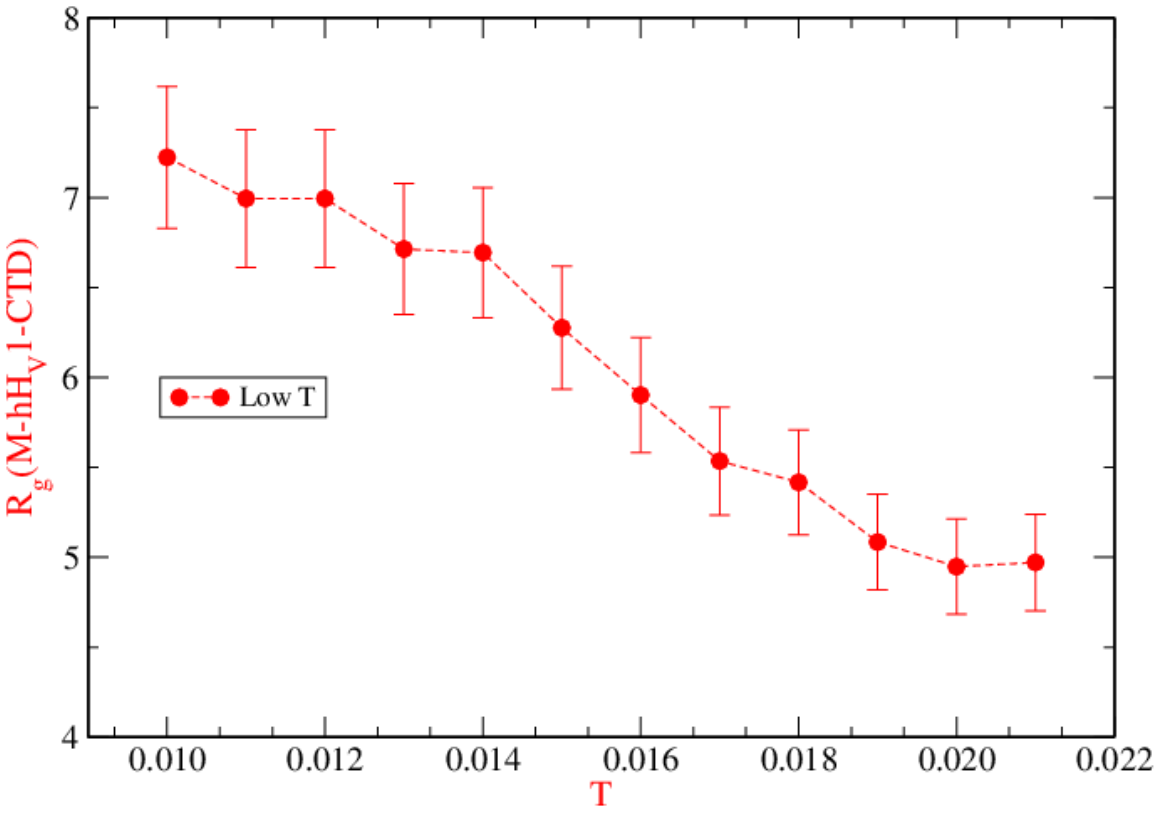


Figure 4.15 Plot of the Radius of gyration (Rg) of M-hHV1-CTD with temperature (T) in the low temperature regime along with the error bars along y-axis.

From these observations, a change in the nature of the thermal response between the two different temperature regimes was observed. The residue-residue interactions dominate over the thermal noise at low temperatures and the protein structures are almost frozen as the protein continues to perform its very slow motion. Thermal agitation stirs the self-organizing residues to adopt more compact and stable configurations with reduced entropy on raising the temperature in this regime. The decay of R_g with T continues until a characteristic value where residue-residue interactions become

comparable and the protein settles into the smallest morphology (i.e. the least entropy within constraints).

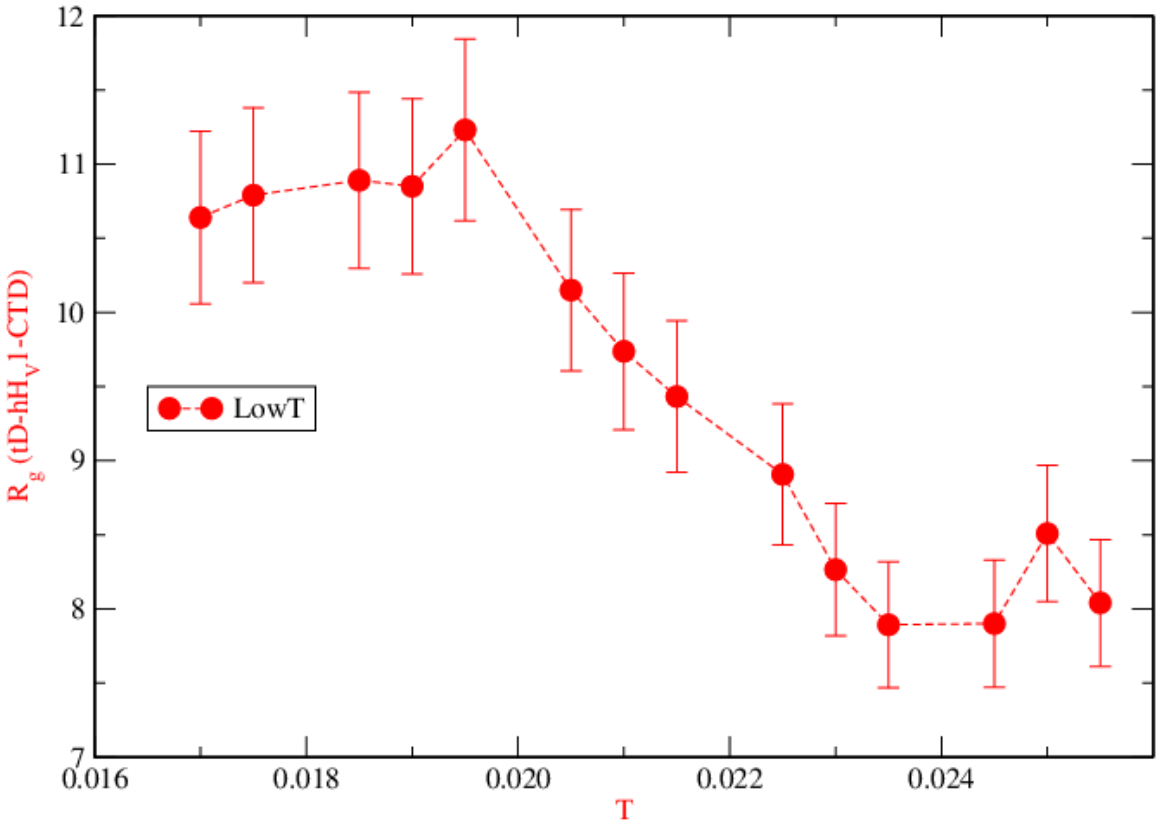


Figure 4.16 Plot of the Radius of gyration (R_g) of tD-hHV1-CTD with the temperature (T) in the low temperature regime along with the error bars along y-axis.

Thus, our finding on the thermal response of the radius of gyration of the protein (M-hH$_V$1-CTD) with coarse-grained MC simulation is consistent with the results from all-atom MD simulations at least qualitatively, as shown in Figure 4.17. From Figure 4.17, we can calibrate the temperature scale of CG MC Simulation and AA MD simulation. It can be clearly seen from the figure that the reduced temperature 0.022 in coarse-grain MC simulation is equivalent to 500K in all-atom MD simulation and the

value of 8 lattice units for R_g in coarse-grain MC simulation is nearly equal to 16 nm in all-atom MD simulation. The variation of R_g with time steps as shown in the figures above shows that the protein's structure has almost reached its steady-state equilibrium in our simulation time.

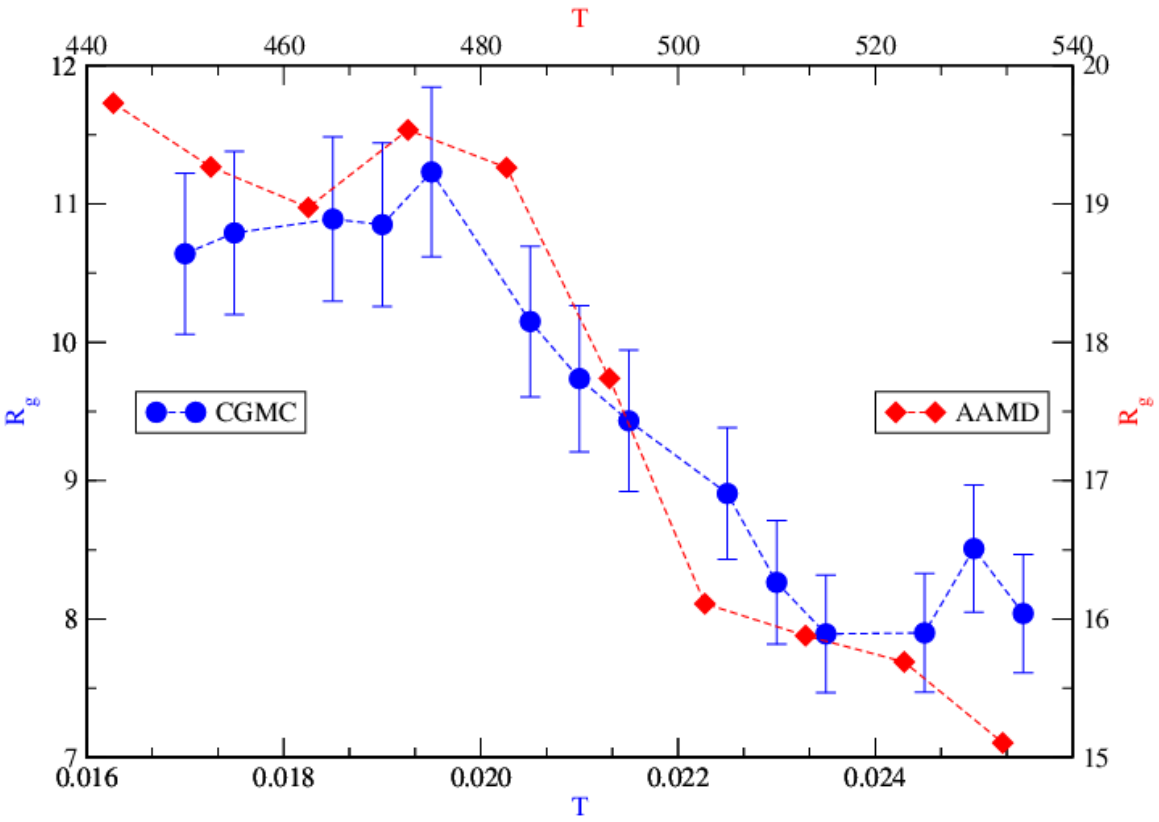


Figure 4.17 Comparative analysis of Radius of gyration (Rg) with temperature (T) in the low-temperature regime by Coarse-grained MC and All-atom MD simulation.

Note: From this we can calibrate the arbitrary value of coarse-grain model with the real value of all-atom simulation in which red line represents result of AA MD while blue line with error bars represent result of CG MC simulation in low-temperature regime.

The decrease in R_g with temperature from two different computer simulation models shows some similarity in the structural response of the protein. The coarse-grained MC approach involves an efficient and effective method to address large-scale

problems but the scales are in arbitrary units, a common drawback, while the all- atom MD approach captures atomistic detail and provides the measurements in real units. Calculation of the same quantities provides a way to calibrate the arbitrary scale at least qualitatively. Plot of the variation of R_g with temperature in high temperature regimes from all-atom MD simulation (0) is shown below.

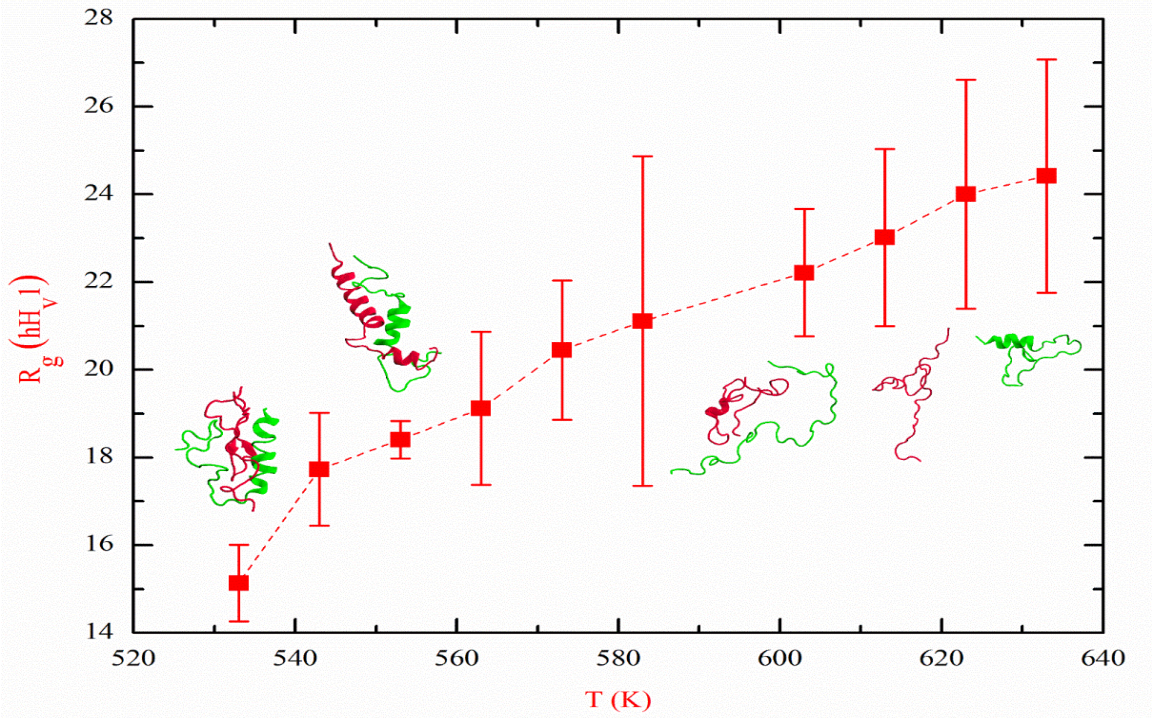


Figure 4.18 Plot of the Radius of gyration (Rg) versus the temperature (T) obtained from all-atom MD simulation in the high- temperature regime.

Note: Different snapshots at different temperatures are also shown in this figure which is obtained from all-atom MD simulation. Data in the last 10ns in each independent temperature is used in estimating the average R_g.

Then change of radius of gyration (R_g) with the temperature in the high-temperature regime by coarse-grained MC simulation of both monomer (Figure 4.19) and dimer (Figure 4.20) are shown below. Both M-hH_V1-CTD and tD-hH_V1-CTD have the

same trend of increment of R_g with the temperature in this high-temperature regime although the range of R_g and temperature are different in Figure 4.19 and Figure 4.20 which is because of different sizes of the protein.

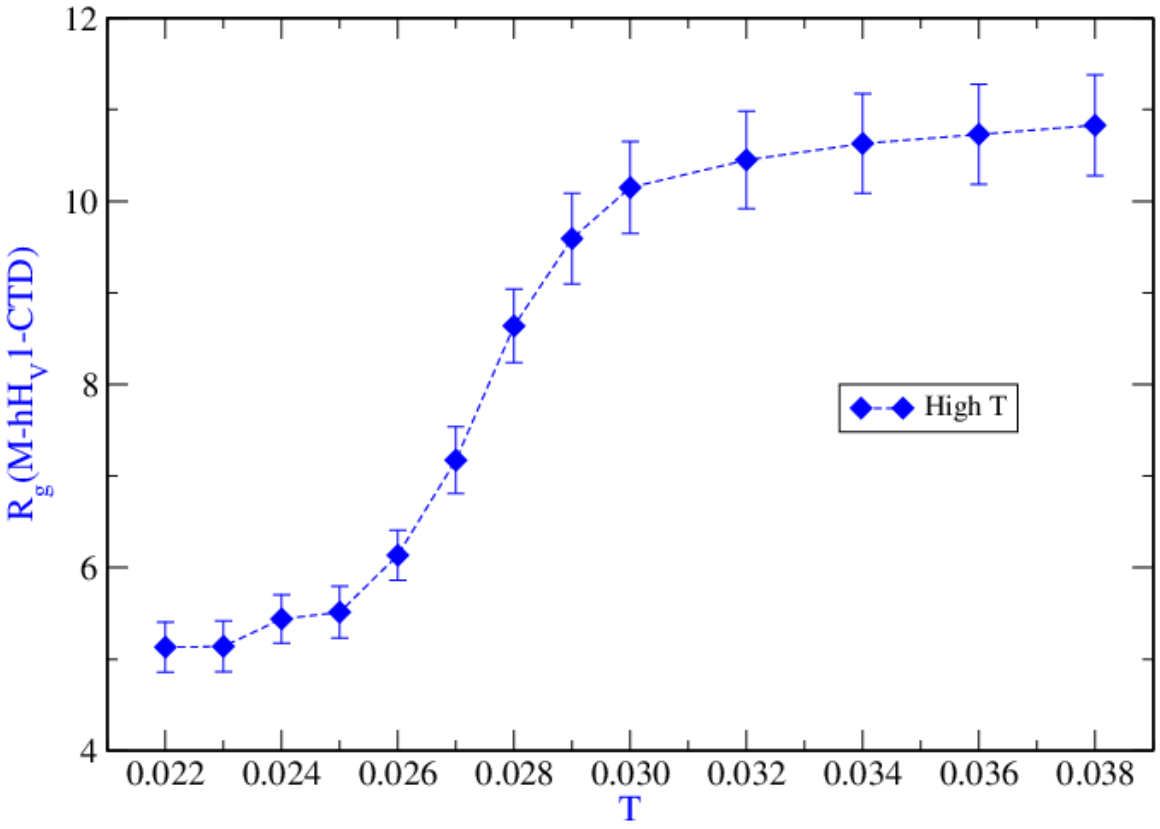


Figure 4.19 Plot of the Radius of gyration (Rg) with temperature (T) in the high-temperature regime for M-hHV1-CTD in CG MC simulations.

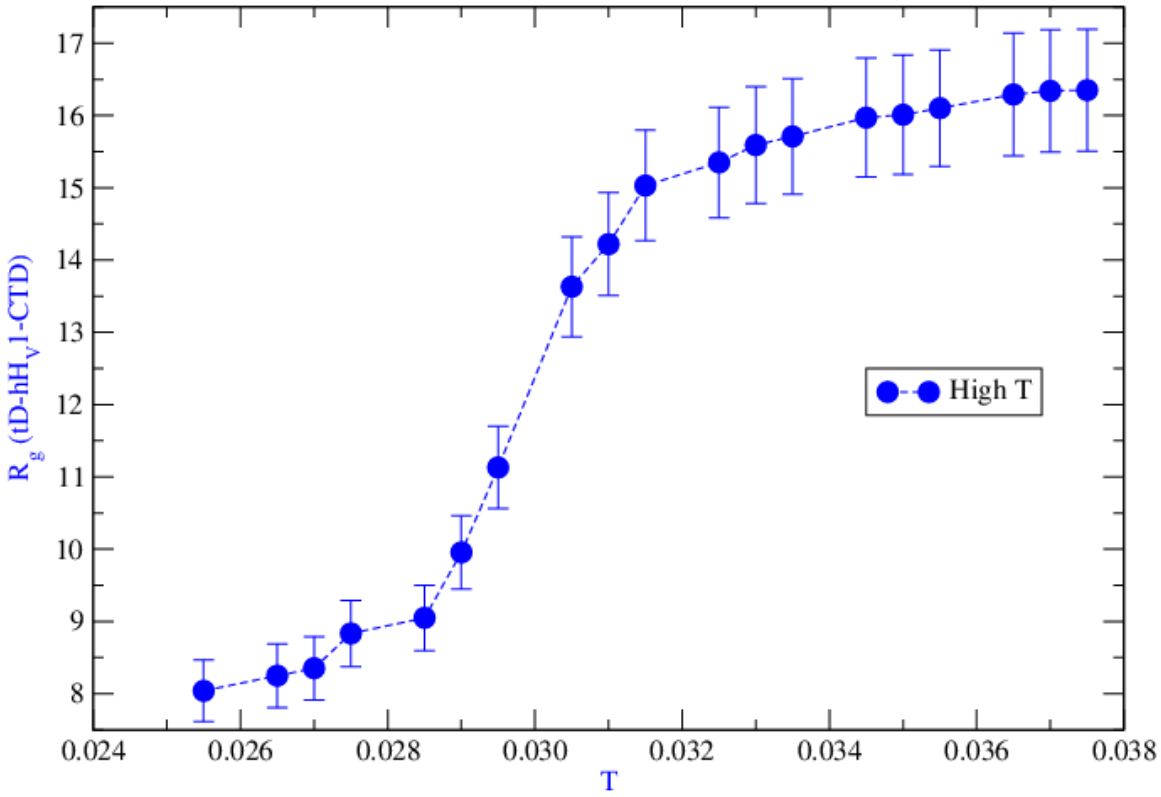


Figure 4.20 Plot of the Radius of gyration (Rg) with temperature (T) in the high-temperature regime for tD-hHV1-CTD in CG MC simulations.

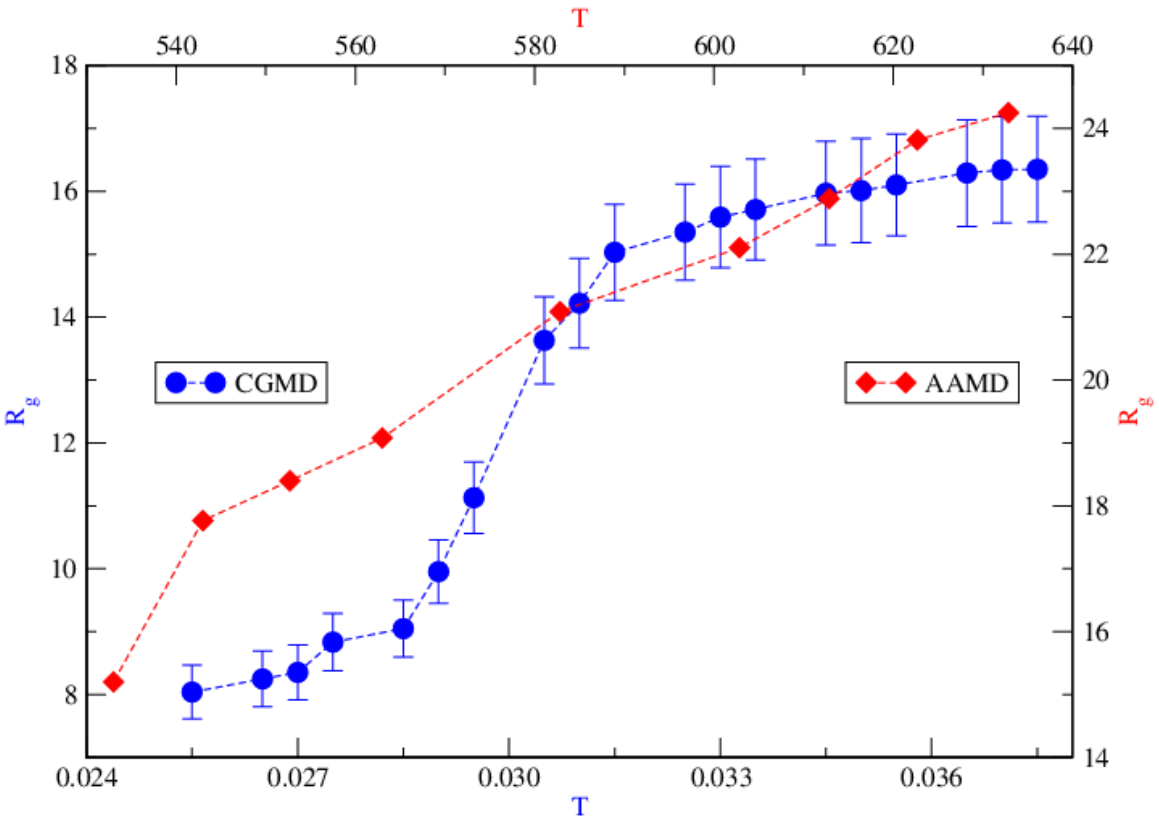


Figure 4.21 Comparative study of the dependence of radius of gyration (R_g) of the tandem dimer on temperature from two different study methods: all-atom MD and coarse-grain MC in high-temperature regime.

Note: From this we can calibrate the arbitrary value of coarse-grain model with the real value of all-atom simulation in which red line represents result of AA MD while blue line with error bars represent result of CG MC simulation in high temperature regime.

4.2.4 Structure Factor {S(q)}

The overall spatial distribution of residues in the conformational evolution of the protein may vary with the length scale and temperature. The structural variation over multiple length scales can be examined by analyzing the structure factor S(q). The variation of S(q) with the wavelength or linear length for low and high temperature regimes is presented in Figure 4.22 and Figure 4.23 for the protein tD-hH$_V$1-CTD. We can see that the variation of S(q) with the wavelength (r) at different low temperatures (Figure 4.22) is not as strong as at the high temperatures (Figure 4.23). The spread of

residues can easily be quantified by scaling the structure factor with the wavelength, comparable to R_g, of the protein, i.e. $S(q) \propto r^D$, where D is the effective dimension. The slope of S(q) versus r on a log-log scale (Figure 4.22 and Figure 4.23) are the estimates of D.

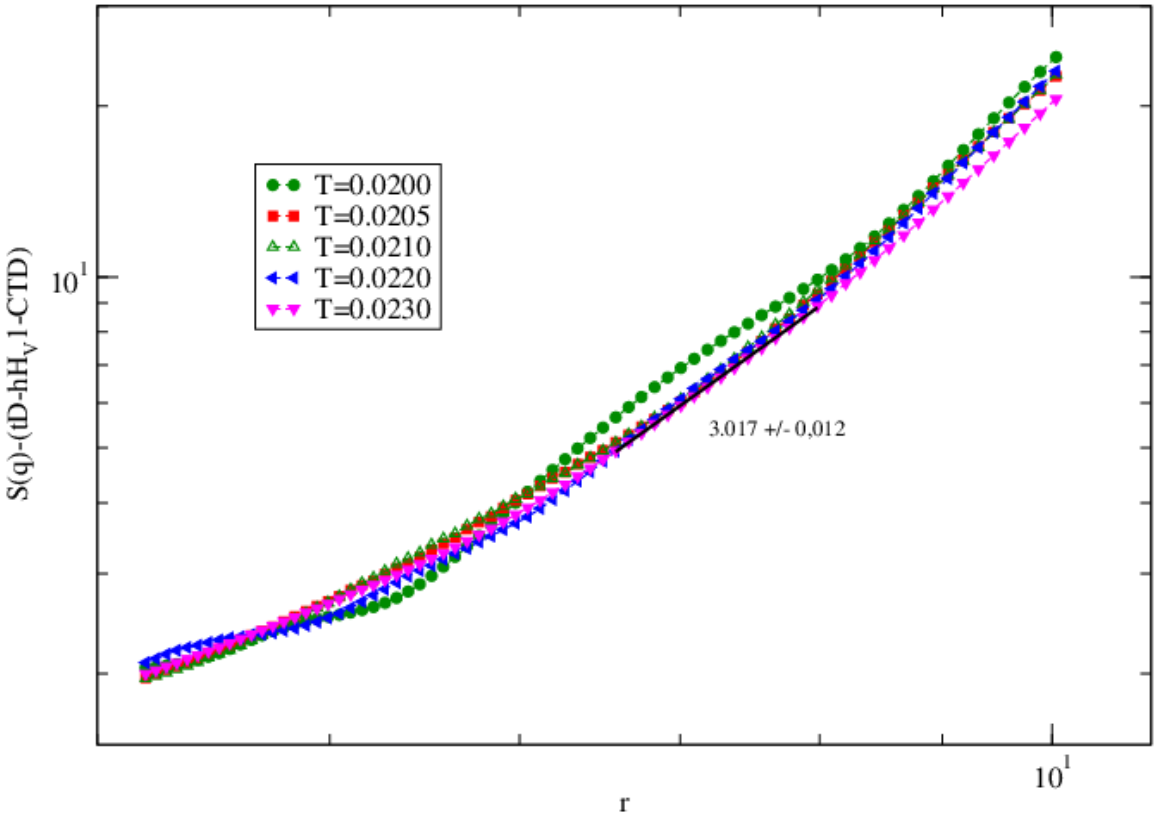


Figure 4.22 Log-log scale plot structure factor S(q) versus wave length ® at temperature $T = 0.020 - 0.023$.

Note: The slope of a set of representative data points at $T = 0.023$ over length scales comparable to its radius of gyration is an estimate of the effective dimension of the residue spread.

The R_g of the protein (tD-hH$_V$1-CTD) is in the range of 8-10 where D ~ 3 at T=0.020 indicating that the protein conforms to a globular (solid) morphology. Comparatively, there is a large variation in S(q) and R_g of the protein at higher temperatures ($T = 0.029 - 0.034$). At $T = 0.029$, the protein remains relatively

compact (D ~ 3) and becomes a random coil (D ~ 2) on raising the temperature further ($T = 0.032$) on a larger length scale. The structure of the protein appears to be linear (D ~ 1.3) at high temperature (Figure 4.23). The distribution of residues over length scale is critical in connectivity for proton transport. In general, a smaller spread (i.e. lower R_g, higher D) corresponds to larger connectivity with more pathways in a compact morphology. Thus, the diversity in structural variation in the high temperature regime suggests the proton transport along the protein conformations which show appreciable variations with the temperature.

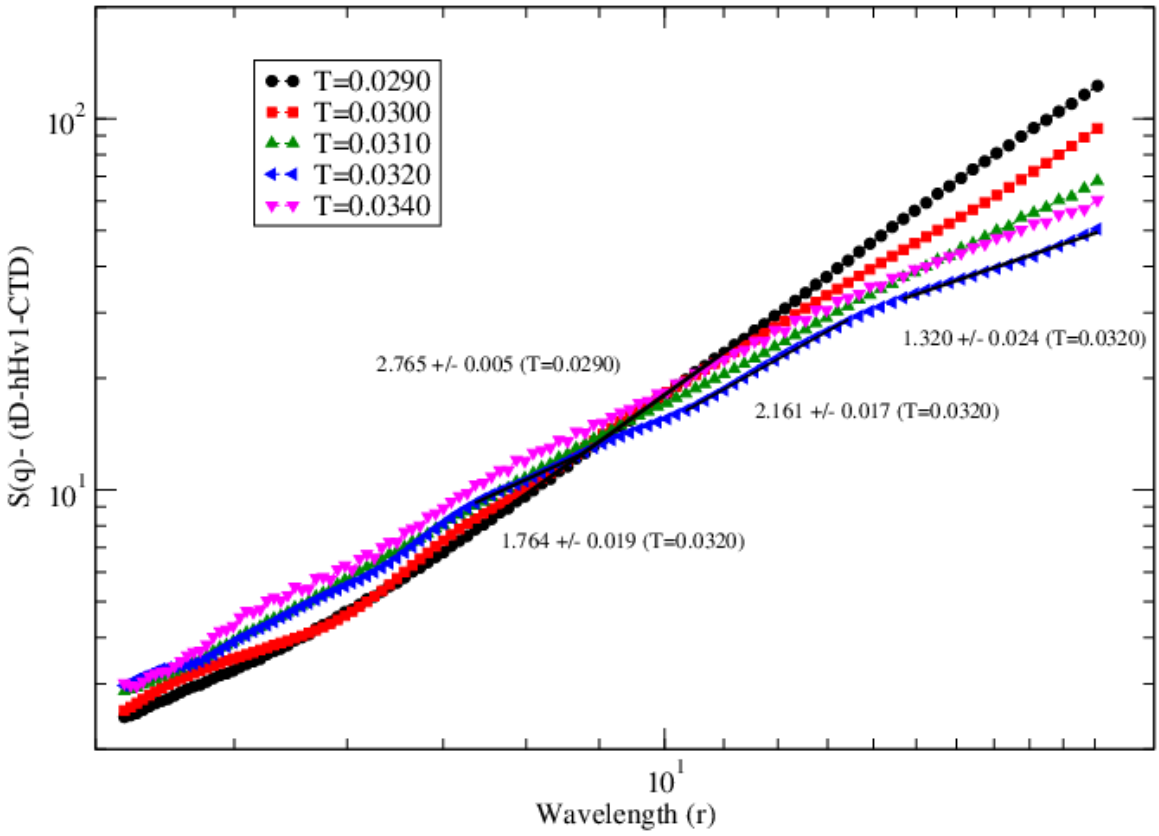


Figure 4.23 Log-log scale plot of structure factor S(q) versus wavelength r at temperatures $T = 0.029 - 0.034$.

Note: Slopes of some representative data points over relevant length scales are estimates of effective dimension of the residue spread

CHAPTER V – CONCLUSION

Conformational response of a monomer (M-hH_V1-CTD) and its tandem dimer (tD-hH_V1-CTD) for the C-terminal domain of the hH_V1 channel to temperature are studied by coarse-grained MC simulation and the main result is compared to all-atom MD simulation. All-atom MD simulations incorporate structural details starting from the atomic scale while all-residue coarse-grained MC covers the length scales spanning beyond the size of a residue and involves a knowledge-based residue-residue interaction. Because of the efficiency of implementing the CG MC approach, a range of local and global physical quantities are analyzed with a coarse-grained phenomenological interaction potential. A physical quantity such as radius of gyration can be easily calculated in both all-atom as well as all-residue approaches which provides a mean to verify results. The atomic resolution of all-atom MD simulation data involving trajectories in real space can be used for calibrating the scales (reduced units) for variables used in CG simulations as seen above. Based on data from the two approaches, we find that both monomer and dimer exhibit similar thermal responses but on different temperature scales which is because of the size of the proteins and each seems consistent with laboratory observations(Fujiwara et al., 2012; Q. Li et al., 2015).

From visual analysis of the snapshots, a general trend in variation of the overall conformational spread and structure variability (i.e. distribution of self-organizing fibrous and globular segments in the protein) with temperature is seen. We were able to identify two different temperature regimes, low and high, by visual inspection of the spread (size) of the protein. In the low-temperature regime, we find that the size of the protein decreases on increasing the temperature, an unexpected observation as adding thermal

energy by increasing the temperature generally enhances the spread. The decrease in the size of the protein implies an increase in compactness, and therefore the connected pathways, which may result in an increase in proton transport. In the high-temperature regime, the conformation of the protein spreads out with increases in temperature.

Contact maps help in synthesizing the distribution of loops (concentration and size) along the contour of the protein. Decays in loops (i.e. loss in secondary and tertiary structures) occur at high temperatures. Analysis of the contact map seems consistent with the visualization. Contact maps of tD-hH$_V$1-CTD exhibit the appearance of a mirror symmetry in structural variability around the tandem link which leads us to believe that the two monomeric units (^{1}H-^{49}I and ^{56}I-^{104}H) respond somewhat similarity. These observations are based on the segmental location of the residues. The consequences for global physical properties are quantified by analyzing the radius of gyration of the proteins, their structure factor, and mobility profiles.

Residues continue to perform their stochastic motion to organize in stable local structures as the protein chain explores its conformational phase space. We have analyzed mobility profiles and the interacting contact profiles of the residues to assess the local segmental structures and their stability. The mobility profile of the tandem dimer shows some degree of mirror symmetry in the distribution of segments with high and low mobility. For example, the localized segments with low (^{7}N–^{15}K, ^{90}K–^{98}N) and high (^{17}S–^{20}F, ^{85}F–^{88}S) mobility are symmetrically distributed around the highly mobile tandem linker in the low-temperature regime. Segmental residue contact profiles at low and high temperatures are complementary to the corresponding mobility profiles. The self-

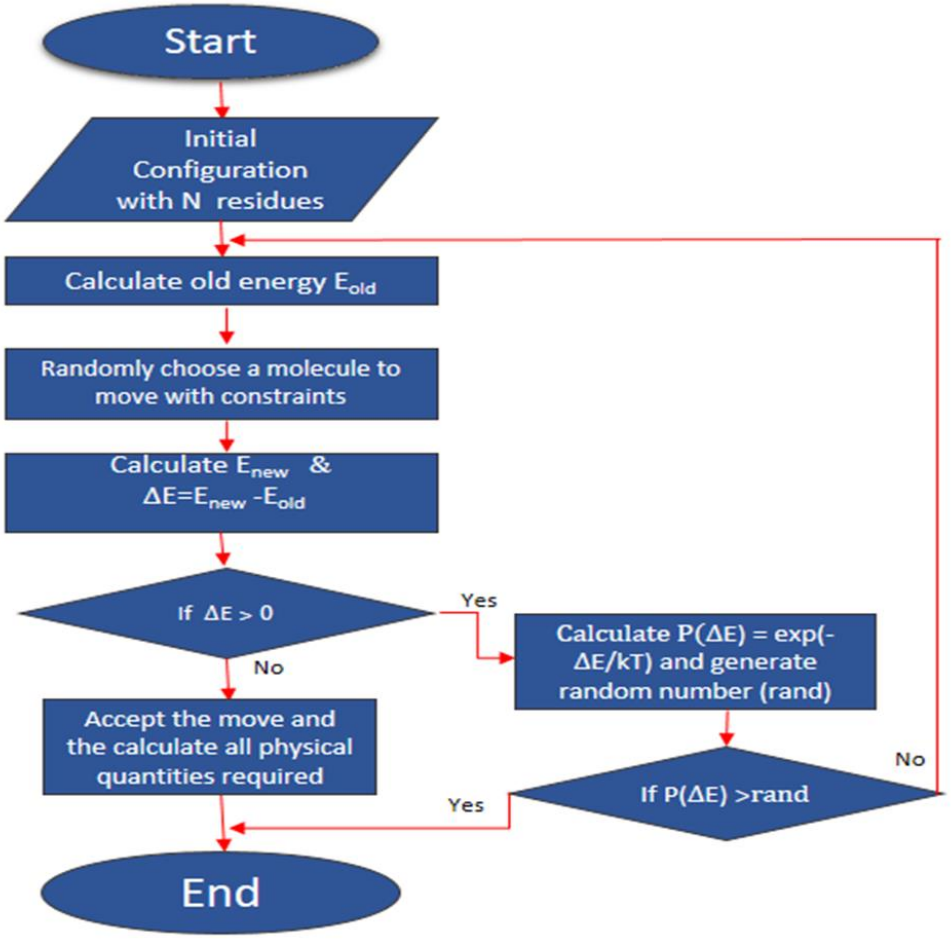
assembled segments that have a larger number of interacting residues show lower mobility.

The dependence of the radius of gyration R_g (a global physical quantity to estimate the overall size) of the protein (both M-hH$_V$1-CTD and tD-hH$_V$1-CTD) on temperature is examined in detail. From both all-atom MD and coarse-grained MC simulations, we find that the radius of gyration of the protein M-hH$_V$1-CTD decreases on increasing the temperature which is not a typical thermal response. Contrary to general expectation, the protein becomes more compact with reduced entropy on increasing the temperature in the low-temperature regime where residue-residue interaction dominates over thermal agitation – an interaction-controlled thermal-induced structural response. On further increasing the temperature, we find that the radius of gyration increases on increasing the temperature before reaching a steady-state value. Thermal agitation wins over the residue-residue interaction, the protein expands through its random-coil structures until it conforms to a linear configuration – the thermal-controlled structural response.

The spread of residues over length scales spanning over the radius of gyration can be quantified by analyzing the structure factor. Scaling of the structure factor *S(q)* of the protein with the wavelength (*r*), $S(q) \propto r^D$ provides an estimate of the effective dimension *D* for the distribution of residues. We find $D \sim 3$ at low temperatures where the protein conforms to a globular (solid) morphology. The interacting residues along the protein backbone are connected via multiple pathways (not just the peptide bonds). In the high-temperature regime, the structure of the protein undergoes various structural transformations. It transforms from a globular conformation (D ~ 3) to a random-coil (*D*

~ 2) on raising the temperature and becomes very tenuous ($D \sim 1.3$) on large scales. In general, a smaller spread (higher D) increases the connectivity with more pathways in a compact morphology. Thus, the variations in conformations may lead to corresponding variations for proton transport along the protein structure.

APPENDIX A – METROPOLIS ALGORITHM



APPENDIX B – SUPPORTING MATERIALS

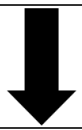
Table B.1

Symbolic Representation of Amino Acids along with Hydropathy Index

1.	I	Isoleucine (Ile, I):	H1
2.	V	Valine (Val, V):	H2
3.	L	Leucine (Leu, L):	H3
4.	F	Phenylalanine (Phe, F):	H4
5.	C	Cysteine Cys (C):	H5
6.	M	Methionine (Met, M):	H6
7.	A	Alanine (Ala, A):	H7
8.	G	Glycine (Gly, G):	H8
9.	T	Threonine (Thr, T):	P1
10.	S	Serine (Ser, S):	P2
11.	W	Tryptophan (Trp, W):	P3
12.	Y	Tyrosine (Tyr, Y):	P4
13.	P	Proline (Pro, P):	P5
14.	H	Histidine (His, H):	P6
15.	Q	Glutamine (Gln, Q):	P7
16.	N	Asparagine (Asn,N):	P8
17.	D	Aspartic Acid (Asp, D):	E1
18.	E	Glutamic Acid (Glu, E):	E2
19.	K	Lysine (Lys, K):	E3
20.	R	Arginine (Arg, R):	E4

Two-monomers-hHv1-CTD, Alanine tandem 104 bp; generated with OpenBabel 2.3.2

H,Q,R,L,L,K,N,L,R,E,I,E,Q,E,K,E,S,C,S,F,E,L,H,Q,I,K,A,A,L,Q,V,N,M,Q,K,L,R,L,L,Q,R,E,S,R,T,K,V,S,I,A,A,A,A,A,A,I,S,V,K,T,R,S,E,R,Q,L,L,R,L,K,Q,M,N,V,Q,L,A,A,K,I,Q,H,L,E,F,S,C,S,E,K,E,Q,E,I,E,R,L,N,K,L,L,R,Q,H



Two-monomers-hHv1-CTD, Alanine tandem 104 bp (in total);

14,15,20,3,3,19,16,3,20,18,1,18,15,18,19,18,10,5,10,4,18,3,14,15,1,19,7,7,3,15,2,16,6,15,19,3,20,3,3,15,20,18,10,20,9,19,2,10,1,7,7,7,7,7,7,1,10,2,19,9,20,10,18,20,15,3,3,20,3,19,15,6,16,2,15,3,7,7,19,1,15,14,3,18,4,10,5,10,18,19,18,15,18,1,18,20,3,16,19,3,3,20,15,14

REFERENCES

Adcock, S. A., & McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical Reviews*, *106*(5), 1589–615. https://doi.org/10.1021/cr040426m

Allen, M. (2004). Introduction to molecular dynamics simulation. *Computational Soft Matter: From Synthetic Polymers to …*, *23*(2), 1–28. https://doi.org/10.1016/j.cplett.2006.06.020

Allen, M. ., & Tildesley, D. . (1991). *Computer Simulation of Liquids*.

Amino acid. (2017, June 4). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Amino_acid&oldid=783745272

Amino Acids 1 | Peptide | Protein Structure. (n.d.). Retrieved June 26, 2017, from https://www.scribd.com/doc/316452518/Amino-Acids-1

Betancourt, M. R., & Thirumalai, D. (1999). Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science : A Publication of the Protein Society*, *8*(2), 361–9. https://doi.org/10.1110/ps.8.2.361

Binder, K. (2008). *Monte Carlo and Molecular Dymanics Simulations in Polymers Science*. *Vasa* (Vol. 20).

Boonamnaj, P., Subedi Paudel, S., Jetsadawisut, W., Kitjaruwankul, S., Sompornpisut, P., & Pandey, R. (2017). Conflicting thermal response of a monomer and a tandem dimer of a membrane protein segment (hHv1). *Bulletin of the American Physical Society*.

Chen, Y., Zhang, Q., & Ding, J. (2004). A coarse-grained model and associated lattice Monte Carlo simulation of the coil–helix transition of a homopolypeptide. *The Journal of Chemical Physics*, *120*(7), 3467–3474. https://doi.org/10.1063/1.1640667

Darden, T.; York, D.; Pedersen, L. (1993). No Title. *J. Chem. Phys.*, *98 (12)*, 10089–10092.

Decoursey, T. E. (2013). VOLTAGE-GATED PROTON CHANNELS: MOLECULAR BIOLOGY, PHYSIOLOGY, AND PATHOPHYSIOLOGY OF THE H V FAMILY. *Physiol Rev*, *93*, 599–652. https://doi.org/10.1152/physrev.00011.2012

DeCoursey, T. E. (2013). Voltage-Gated Proton Channels: Molecular Biology, Physiology, and Pathophysiology of the HV Family. *Physiological Reviews*, *93*(2), 599–652. https://doi.org/10.1152/physrev.00011.2012

Dill, K. A., & Maccallum, J. L. (n.d.). The Protein-Folding Problem, 50 Years On.

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (n.d.). The Protein Folding Problem. https://doi.org/10.1146/annurev.biophys.37.092707.153558

Eaton, D. C. (1985). Ionic channels of excitable membranes. Bertil Hille. Sunderland, Ma: Sinauer Associates, 1984. *Journal of Neuroscience Research*, *13*(4), 599–600. https://doi.org/10.1002/jnr.490130415

Emperador, A. (2013). Simplified models for proteins in coarse-grained, 22.

Four levels of Protein Structure. (n.d.). Retrieved June 8, 2017, from https://www.mun.ca/biology/scarr/iGen3_06-04.html

Fritsche, M., Pandey, R. B., Farmer, B. L., Heermann, D. W., & Slocik, J. (2013). Variation in Structure of a Protein (H2AX) with Knowledge-Based Interactions. *PLoS ONE*, *8*(5), e64507. https://doi.org/10.1371/journal.pone.0064507

Fujiwara, Y., Kurokawa, T., Takeshita, K., Kobayashi, M., Okochi, Y., Nakagawa, A., & Okamura, Y. (2012). The cytoplasmic coiled-coil mediates cooperative gating temperature sensitivity in the voltage-gated H+ channel Hv1. *Nature Communications*, *3*, 816. https://doi.org/10.1038/ncomms1823

Gelpi, J., Hospital, A., Goñi, R., & Orozco, M. (2015). Molecular dynamics simulations: advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, *8*, 37. https://doi.org/10.2147/AABC.S70333

Goloubinoff, P. (2014). Recent and future grand challenges in protein folding, misfolding, and degradation. *Frontiers in Molecular Biosciences*, *1*, 1. https://doi.org/10.3389/fmolb.2014.00001

Humphrey, W., Dalke, A. and Schulten, K. (1996). VMD - Visual Molecular Dynamics. Retrieved June 27, 2017, from http://www.ks.uiuc.edu/Research/vmd/allversions/cite.html

Ingólfsson, H. I., Lopez, C. A., Uusitalo, J. J., de Jong, D. H., Gopal, S. M., Periole, X., & Marrink, S. J. (2014). The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*(3), 225–248. https://doi.org/10.1002/wcms.1169

Ion channel. (2017, April 16). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Ion_channel&oldid=775702530

Jha, A. N., Vishveshwara, S., & Banavar, J. R. (2010). Amino acid interaction preferences in proteins. *Protein Science : A Publication of the Protein Society*, *19*(3), 603–16. https://doi.org/10.1002/pro.339

Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. (1983). No Title. *J. Chem. Phys.*, *79*, 926–935.

Kessel, A., & Ben-Tal, N. (2011). *Introduction to proteins*.

Kitjaruwankul, S., Khrutto, C., Sompornpisut, P., Farmer, B. L., & Pandey, R. B. (2016). Asymmetry in structural response of inner and outer transmembrane segments of CorA protein by a coarse-grain model. *Journal of Chemical Physics*, *145*(13). https://doi.org/10.1063/1.4963807

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016a). Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*, *116*(14), 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016b). Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*, *116*(14), 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163

Lässig, M., & Valleriani, A. (Eds.). (2002). *Biological evolution and statistical physics*. Berlin ; New York: Springer.

Li, Q., Shen, R., Treger, J. S., Wanderling, S. S., Milewski, W., Siwowska, K., … Perozo, E. (2015). Resting state of the human proton channel dimer in a lipid bilayer. *Proceedings of the National Academy of Sciences*, *112*(44), E5926–E5935. https://doi.org/10.1073/pnas.1515043112

Li, S. J., Zhao, Q., Zhou, Q., Unno, H., Zhai, Y., & Sun, F. (2010). The Role and Structure of the Carboxyl-terminal Domain of the Human Voltage-gated Proton Channel Hv1 *. https://doi.org/10.1074/jbc.M109.040360

Lishko, P. V., Botchkina, I. L., Fedorenko, A., & Kirichok, Y. (2010). Acid Extrusion from Human Spermatozoa Is Mediated by Flagellar Voltage-Gated Proton Channel. *Cell*, *140*(3), 327–337. https://doi.org/10.1016/j.cell.2009.12.053

MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; R, M. (1998). CHARMM22. *J. Phys. Chem. B*, *102 (18)*, 3586–616.

Membrane protein. (2017, April 25). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Membrane_protein&oldid=777200029

Miyazawa, S., & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, *18*(3), 534–552. https://doi.org/10.1021/ma00145a039

Musset, B., & Decoursey, T. (2012). Biophysical properties of the voltage-gated proton channel HV1. *Wiley Interdisciplinary Reviews: Membrane Transport and Signaling*, *1*(5), 605–620. https://doi.org/10.1002/wmts.55

Musset, B., Smith, S. M. E., Rajan, S., Cherny, V. V., Morgan, D., & DeCoursey, T. E. (2010). Oligomerization of the voltage gated proton channel. *Channels*, *4*(4), 260–265. https://doi.org/10.4161/chan.4.4.12789

Olsson, M. H.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H. (2011). PROPKA. *J. Chem. Theory Comput.*, *7 (2)*, 525–37.

Pandey, R B, Farmer, B. L., & Gerstman, B. S. (2015). Self-assembly dynamics for the transition of a globular aggregate to a fibril network of lysozyme proteins via a coarse-grained Monte Carlo simulation. *AIP ADVANCES*, *5*.

Pandey, Ras B, & Farmer, B. L. (2012). Random coil to globular thermal response of a protein (H3.1) with three knowledge-based coarse-grained potentials. *PloS One*, *7*(11), e49352. https://doi.org/10.1371/journal.pone.0049352

Pandey, Ras B., & Farmer, B. L. (2013). Conformational Response to Solvent Interaction and Temperature of a Protein (Histone h3.1) by a Multi-Grained Monte Carlo Simulation. *PLoS ONE*, *8*(10). https://doi.org/10.1371/journal.pone.0076069

Perlmutter, J. D., Drasler, W. J., Xie, W., Gao, J., Popot, J.-L., & Sachs, J. N. (2011). All-Atom and Coarse-Grained Molecular Dynamics Simulations of a Membrane Protein Stabilizing Polymer. https://doi.org/10.1021/la202103v

Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. (2005). NAMD. *J. Comput. Chem.*, *26 (16)*, 1781–1802.

Protein. (2017, May 18). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Protein&oldid=780934930

Rebolledo, S., Qiu, F., & Peter Larsson, H. (2012). Molecular structure and function of Hv1 channels. *Wiley Interdisciplinary Reviews: Membrane Transport and Signaling*, *1*(6), 763–777. https://doi.org/10.1002/wmts.49

Scott, K. A., Bond, P. J., Ivetac, A., Chetwynd, A. P., Khalid, S., & Sansom, M. S. P. (2008). Coarse-Grained MD Simulations of Membrane Protein-Bilayer Self-Assembly. *Structure*, *16*(4), 621–630. https://doi.org/10.1016/j.str.2008.01.014

Shape of proteins. (n.d.). Retrieved June 8, 2017, from http://shakes.ml/shape-of-proteins/

Sompornpisut, P.; Roux, B.; Perozo, E. (2008). PaDSAR-Pseudoatom-driven solvent accessibility refinement. *Biophys. J. 2008, 95 (11), 5349-5361.*, *95 (11)*, 5349–5361.

Vianello, R., Domene, C., & Mavri, J. (2016). The use of multiscale molecular simulations in understanding a relationship between the structure and function of biological systems of the brain: The application to monoamine oxidase enzymes. *Frontiers in Neuroscience*, *10*(JUL), 327. https://doi.org/10.3389/fnins.2016.00327

Walker, J. M. (2010). *Membrane Protein Structure Determination* (Vol. 654). https://doi.org/10.1007/978-1-60761-762-4